Spatial and Temporal Pyramids for Grammatical Expression Recognition of American Sign Language

Nicholas Michael
Computational Biomedicine
Imaging and Modeling
Rutgers University
Piscataway, NJ
nicholam@cs.rutgers.edu

Dimitris Metaxas
Computational Biomedicine
Imaging and Modeling
Rutgers University
Piscataway, NJ
dnm@cs.rutgers.edu

Carol Neidle
Linguistics Program
Boston University
Boston, MA
carol@bu.edu

ABSTRACT

Given that sign language is used as a primary means of communication by as many as two million deaf individuals in the U.S. and as augmentative communication by hearing individuals with a variety of disabilities, the development of robust, real-time sign language recognition technologies would be a major step forward in making computers equally accessible to everyone. However, most research in the field of sign language recognition has focused on the manual component of signs, despite the fact that there is critical grammatical information expressed through facial expressions and head gestures.

We propose a novel framework for robust tracking and analysis of facial expression and head gestures, with an application to sign language recognition. We then apply it to recognition with excellent accuracy ($\geq 95\%$) of two classes of grammatical expressions, namely wh-questions and negative expressions. Our method is signer-independent and builds on the popular "bag-of-words" model, utilizing spatial pyramids to model facial appearance and temporal pyramids to represent patterns of head pose changes.

Categories and Subject Descriptors

I.4.8 [**Image Processing and Computer Vision**]: Scene Analysis—*tracking*; I.5.1 [**Pattern Recognition**]: Models; I.5.2 [**Pattern Recognition**]: Design Methodology—*classifier design and evaluation, pattern analysis*

General Terms

Design, Experimentation, Performance

Keywords

Sign language recognition, face tracking, spatio-temporal pyramids, head pose estimation, expression recognition, kernel codebooks, soft quantization, pyramid match kernel

ASSETS'09, October 25-28, 2009, Pittsburgh, Pennsylvania, USA.
Copyright 2009 ACM 978-1-60558-558-1/09/10 ...$10.00.

1. INTRODUCTION

Since the seminal work of Rabiner on Hidden Markov Models (HMM) [17], their popularity has grown and many advances have been made in speech recognition technologies. For example, modern computers can now interpret voice commands in real time and they can also translate speech to text and vice versa. Through such technologies, users can accomplish many computer tasks with minimal typing, making Human Computer Interaction (HCI) an easier and more efficient experience.

On the other hand, technology for the recognition of sign language, which is widely used by the Deaf, is not nearly as well-developed, despite its many potential benefits. First of all, technology that automatically translates between signed and written or spoken language would facilitate communication between signers and non-signers by bridging the language gap. For example, users of sign language could produce signs into a camera connected to (or built into) a computer. The computer could then recognize and translate these signs to text or speech, thereby allowing a non-signer to understand the signed utterance. In a similar fashion, a non-signer could speak into a microphone connected to a computer. In real time, the computer could then perform speech recognition and translation into sign language, by synthesizing signs using realistic avatars[1]. Secondly, such technology could be used to translate sign language into computer commands, hence opening the road for the development of additional assistive technologies (in a manner analogous to existing speech recognition technologies described above) [22].

Moreover, computerized sign language recognition could facilitate the efficient archiving and retrieval of video-based sign language communication [22]. It could assist with the tedious and time-consuming task of annotating sign language video data for purposes of linguistic and computer science research. Ultimately, such research – and resulting advances in sign language recognition and generation – will have applications that could profoundly change the lives of deaf people and improve communication between deaf and hearing individuals. Non-speaking, non-deaf users of sign language, including some people with autism, aphasia, cerebral palsy, Down Syndrome, and tracheotomies, will benefit from these technologies in the same ways.

However, the task of sign language recognition is not easy. While in speech recognition systems we can model spoken

[1] Avatars are three-dimensional computer animated characters.

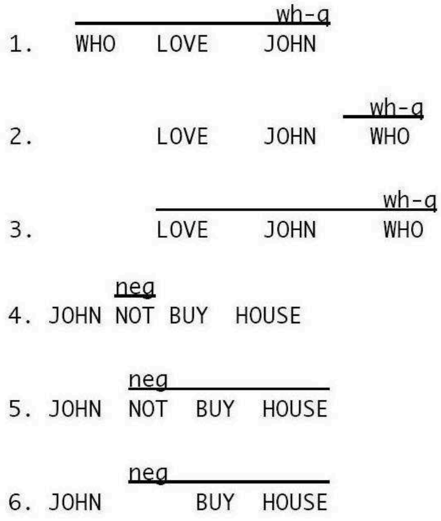

Figure 1: Several sample ASL sentences for negative and wh-question constructions, with English glosses representing the ASL signs

words as a sequence of sounds (phonemes), with sign language things get more complex. First of all, the linguistic components of a sign that must be recognized occur simultaneously rather than sequentially. For example, one or both hands may be involved in the signing and these may assume various hand shapes, orientations and types of movement in differing locations. At the same time, facial expression may also be involved in distinguishing signs, further complicating the recognition task.

In addition to the many possible combinations of facial expressions, hand shapes and hand movements, a recognition model should also account for variations in sign production. As with words in spoken language, a given sign is not articulated identically every time it is produced, even by the same signer. Despite the combinatorial complexity of the problem, a number of methods have shown promising results in recognizing the manual components of signs produced through movements of the hands and arms [2, 23].

Furthermore, in sign language, critical grammatical information is expressed through head gestures, such as periodic nods and shakes, and facial expressions such as raised or lowered eyebrows, eye aperture, nose wrinkles, tensing of the cheeks, and mouth expressions [1, 6, 12, 15]. These linguistically significant non-manual expressions include grammatical markings that extend over phrases to mark syntactic scope (e.g., of negation and questions). Sign language recognition cannot be successful unless these signals are also correctly detected and identified. For example, the sequence of signs JOHN BUY HOUSE could be interpreted, depending on the non-manual markings that accompany the signs, to mean any of the following: (i) "John bought the house." (ii) "John did not buy the house." (iii) "Did John buy the house?" (iv) "Did John not buy the house?" (v) "If John buys the house...". In addition, recognition of such grammatical signals can assist with the task of recognizing the manual components of signs. This is because there may be some correlations between information that is expressed manually and non-manually.

Motivated by the important role of facial expressions in sign language recognition, we present in this paper a novel framework for robust tracking and recognition of such expressions in monocular video sequences. In particular, we apply our framework to the recognition of facial expressions found in wh-questions, which involve phrases such as *who, what, when, where, why* and *how*, and head gestures of negation. Using an implementation of the robust face tracker of Kanaujia et al. [9], we accurately track the faces of American Sign Language (ASL) signers, localizing their facial components (e.g., eyes, eyebrows) and predicting their 3D head pose. Inspired by the work of Lazebnik on scene categorization [10], together with the popularity of "bag-of-words" models [11], we use spatial pyramids of features to detect lowered eyebrows and squinted eyes. We augment this information with the 3D head pose using Stacked Generalization and Majority Voting [26, 20], to recognize the presence of wh-question facial expressions in a video sequence. Additionally, we extend the idea of spatial pyramids to the temporal dimension, constructing pyramids of head pose derivatives (i.e., the change of head pose), for the recognition of head shakes that are characteristic of negative expressions. We demonstrate the effectiveness of our approach by testing on 42 videos from the Boston University American Sign Language Linguistic Research Project (ASLLRP) dataset [16] and achieving over 95% correct recognition results.

2. LINGUISTIC BACKGROUND

In ASL, there are typical facial expressions that are found with questions of different types. For wh-questions (which involve phrases such as *who, what, when, where, why* and *how*), the grammatical marking consists of lowered eyebrows and squinted eyes that occur either over either the entire wh-question or solely over a wh-phrase that has moved to a sentence-final position. The possibilities are illustrated in the example ASL sentences of Figure 1. In this figure, labeled lines indicate the signs with which the non-manual marking co-occurs. The first three examples would be translated in English as "Who loves John?". The intensity of this wh-question marking is greatest at the end of the sentence when it spreads over the entire question, as in Figure 1(3). In addition, there may be a slight, rapid side-to-side head shake over at least part of the domain of the wh-question marking.

With negation, there is a relatively slow side-to-side head shake that co-occurs with a manual sign of negation (such as NOT, NEVER), if there is one, and may extend over the scope of the negation, e.g., over the following verb phrase that is negated. These possibilities for translating an English sentence meaning "John did not buy a house" are illustrated in the bottom three examples of Figure 1(4-6). The intensity of this marking is greatest at the source of the syntactic feature being marked, as in wh-questions, but in a sentence like (5) or (6) of the same figure, this means that the intensity of the negative marking (including the amplitude of the head turns) is greatest at the left edge and diminishes as the marking continues. For further detail about distribution and intensity of non-manual grammatical markings, see [15].

3. PREVIOUS WORK

As already mentioned in the previous section, most research on computer-based sign language recognition has focused on the manual components of signs. In the work of Vogler and Metaxas [23], the manual signs are split into independent movement and hand shape channels, and an HMM framework is used to model signs as a sequence of phonemes. These independent channels allow them to handle simultaneous manual events. Bauer and Kraiss break down signs into smaller units using unsupervised clustering, achieving high recognition accuracy in isolated sign recognition experiments [2]. In [24], the authors apply techniques from speech recognition to develop a method that quickly adapts to unknown signers, in an attempt to handle interpersonal variance. Similarly, the authors of [27] present a method for sign recognition, which uses a background model to achieve accurate feature extraction and then performs feature normalization to achieve person independence. To tackle the problem of self occlusions of the hands, Martinez and Ding [7] first perform 3D hand reconstruction and then represent hand motions as 3D trajectories.

Only recently have researchers begun to address the importance of facial expressions for sign recognition systems. Von Agris et. al. [25] provide an extensive review of recent developments in visual sign recognition, together with a system that uses both the manual and the non-manual components of signs. However, their system poses the restriction that the signer must be wearing a glove with colored markers, in order to enable robust hand tracking and hand posture reconstruction. Additionally, in their system, the tracked facial features are not used to recognize facial expressions which have grammatical meaning. Vogler and Goldenstein present a 3D deformable model for face tracking, which emphasizes outlier rejection and occlusion handling [21, 22] at the expense of slower run time. They use their system to demonstrate the potential of face tracking for the analysis of facial expressions encountered in sign language, but they do not use it for any actual recognition of facial expressions. In our work, we use the more robust Active Shape Model (ASM) face tracker [5, 9] to do *real time tracking* with better handling of partial occlusions and head rotations. Additionally, we demonstrate the effectiveness of our method in recognizing facial expressions by extending ideas originally proposed for scene categorization [10]. This allows us to move away from the complexity of training Hidden Markov Models, which have been the dominant tools in this domain.

4. METHOD

Our framework for facial tracking and facial expression recognition consists of the following steps for each video sequence that is processed:

1. Face tracking and pose estimation
   (a) Feed video sequence into ASM tracker to localize and track signer's face
   (b) ASM tracker outputs (x,y) positions of 79 facial landmarks and the 3D head pose for each frame
2. Feature Extraction for each tracked frame utilizing ASM tracker's output
   (a) Compute bounding box of eyes and eyebrows and extract dense SIFT feature descriptors from it [13]
   (b) *Soft* quantize the SIFT descriptors and the head pose using separate feature codebooks [8]
   (c) Build pyramid representation of frames and video sequences
      i. Build spatial pyramids of computed SIFT descriptors for *each frame*
      ii. Build temporal pyramid of head pose derivatives (change in pose) for the *entire sequence*
3. Recognize video sequences containing Negative expressions using the temporal pyramid representation of pose derivatives and a Support Vector Machine (SVM) [3] with pyramid matching kernel [10]
4. Recognize video sequences containing Wh-Question expressions
   (a) Use a Stacked Support Vector Machine [20, 26], which combines the score obtained from classifying the spatial pyramid representation of SIFT descriptors and the score obtained from classifying the pose angle, to classify *each frame* in the video sequence
   (b) Apply majority voting [20] on the results of the previous step, to classify the *entire sequence* based on the classification of *each frame* within the sequence (if the majority of the frames are classified as depicting a Wh-Question expression, the entire sequence is also classified as such)

The following subsections explain the components of our system in more detail.

4.1 Face Tracking

Face tracking is a challenging problem because the tracker needs to generalize well to unseen faces and must handle illumination changes. It should also cope with partial occlusions and pose changes, such as head rotations, which cause drastic changes in the shape of the face, causing it to lie on a non-linear manifold. This basically means that as the head rotates by a certain amount, the shape of the different parts of the face, as viewed from a two dimensional perspective, does not change uniformly and by an equal amount in all places. This effect is more severe during head rotations which approach profile poses.

Kanaujia et al. [9] tackle the problem with an Active Shape Model (ASM) [5], which is a statistical model of facial shape variation. In the ASM framework, a facial shape $\vec{S}$ is represented by N landmarks, each of which is characterized by its (x, y) image coordinates, so that $\vec{S} = \{x_1, y_1, \ldots, x_N, y_N\}$. Through the application of Principal Component Analysis (PCA) on an aligned training set of facial shapes, a subspace is learned which captures the major modes of shape variation, by projecting shapes along the eigenvectors of the shape covariance matrix with the highest eigenvalues. Essentially, this allows us to learn a model of the permissible ways in which faces of different people differ, so that we can then apply this model to images of unseen faces and still be able to localize and track them. In this way, an aligned shape $\vec{X} = \Phi(\vec{S})$, where Φ is the linear transformation that aligns a shape $\vec{S}$ to the mean shape $\vec{\bar{X}}$ of the subspace, can be approximated as:

$$\vec{X} \approx \vec{\bar{X}} + \vec{P}\vec{b} \ , \qquad (1)$$

where $\vec{P}$ is the eigenvector matrix and $\vec{b}$ is a column vector of shape parameters (encoding).

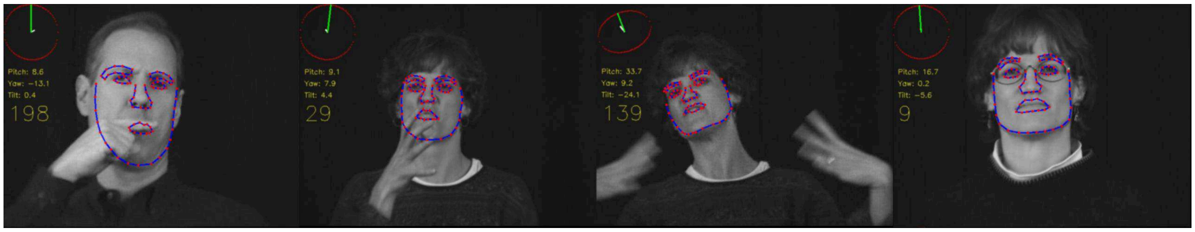

Figure 2: Sample frames (best viewed in color) showing accurate tracking under challenging scenarios (partial occlusions, fast movements and glasses). Here, red dots represent tracked landmarks. The predicted head pose is shown in the top left corner of each frame as a 3D vector

The authors of [9] additionally propose a piecewise approximation to the non-linear shape manifold using overlapping linear subspaces, where each subspace corresponds to a different group of poses. Basically this means learning separate ASM models for each major pose (e.g., looking frontally, left, right, up, down, etc.) and dynamically switching subspaces as the pose of the tracked face changes through a head rotation. Their system is made to run in real time by incorporating a Sum of Squared Intensity Differences (SSID) point tracker [18], which tracks image patches across successive frames assuming small displacements, and only fitting the ASM model periodically to correct any tracking error. Moreover, using a Bayesian Mixture of Experts they are able to estimate the 3D pose of the head from the tracked landmarks (refer to [9] for a more thorough treatment). Figure 2 illustrates the abilities of the ASM tracker on a few challenging input frames exhibiting partial occlusions, as well as rapid head movements and rotations.

The middle columns of Figure 3 show sample frames with the 79 tracked landmarks, along with the predicted 3D head poses for each one. 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 to compute their bounding box in each frame (we will refer to the region inside this bounding box as the *eye region*). We then apply the computer vision algorithm of the Scale Invariant Feature Transform (SIFT) by David Lowe [13], to extract dense discriminative features from within this bounding box to characterize the local texture and appearance of this eye region, and learn to recognize lowered eyebrows and squinted eyes. The computed SIFT features are invariant to scale and rotation changes, meaning that they can still be detected in challenging the ASL video sequences in which the face moves closer or further to the camera and changes pose.

4.2 Codebook Construction

The codebook approach, inspired by the word-document representation used in text retrieval, was first applied to images in the work of Leung and Malik [11]. This approach allows classification of images by representing them as a bag of features, for example SIFT features [13], which are in turn represented as discrete prototypes [8]. Typically researchers use unsupervised clustering to obtain a codebook, V, of prototypes, w, from a random subset of the training data and label each feature by its best representing prototype. Then they count how many times each prototype occurs in an image and stack these frequencies in a vector, which becomes the new representation of the image and can later be used for classification purposes. This codebook representation is essentially a histogram of prototypes.

However, quantizing features in this manner creates problems. For example, if some feature is too distant from all available prototypes, forcing such a hard assignment could mean that the resulting encoding is implausible. Moreover, if a feature is very close to more than one prototype, it becomes ambiguous as to which one would represent it the best. The authors of [8] overcome these problems of *codeword plausibility* and *codeword ambiguity* by employing ideas from kernel density estimation. They propose a soft assignment of prototypes, resulting in a Kernel Codebook (KCB) representation of an image for each prototype, w:

$$KCB(w) = \frac{1}{w}\sum_{i=1}^{n} K_\sigma(D(w, r_i)) \ , \qquad (2)$$

where n is the number of features in the image, r_i is the i^{th} feature, $D(w, r_i)$ is the distance of prototype w from the i^{th} feature, and σ is the smoothing parameter of kernel K. In our work, we adopt this method of soft quantization, setting K to be a Gaussian kernel and using Euclidean distance as our distance metric, $D(w, r_i)$.

4.3 Pyramid Representation

After we extract and softly quantize [8] the discriminative SIFT features of each frame, we utilize the work on pyramid representation of Lazebnik et al. [10], which enables us to model the spatial relationships among features and also provides the means for measuring feature similarity between frames, using a pyramid match kernel.

Denote the set of quantized features extracted from two frames as X and Y. To build a pyramid with L levels, for each level $l = 0, 1, ..., L$, we divide the frame into an imaginary grid of $2^{2\times l}$ cells, along both the x and y dimensions, so that the cells in level l are bigger than the cells in level $l+1$ above it. We histogram the quantized features that fall in each cell (for each feature we know its position within the frame it came from), yielding separate histograms for each cell for each of the L levels. These histograms represent the feature distribution of a particular cell, in terms of the relative frequency of occurrence of each feature prototype within that cell. Because cells at different levels have different sizes, their histograms are computed over image subregions of different sizes, yielding an image representation of different levels of resolution. The topmost layer, having the smallest sized cells, forms the most detailed representation of the feature distribution within an eye region, while the bottommost layer the least detailed. Collectively, the histograms at each level form the pyramid representation of

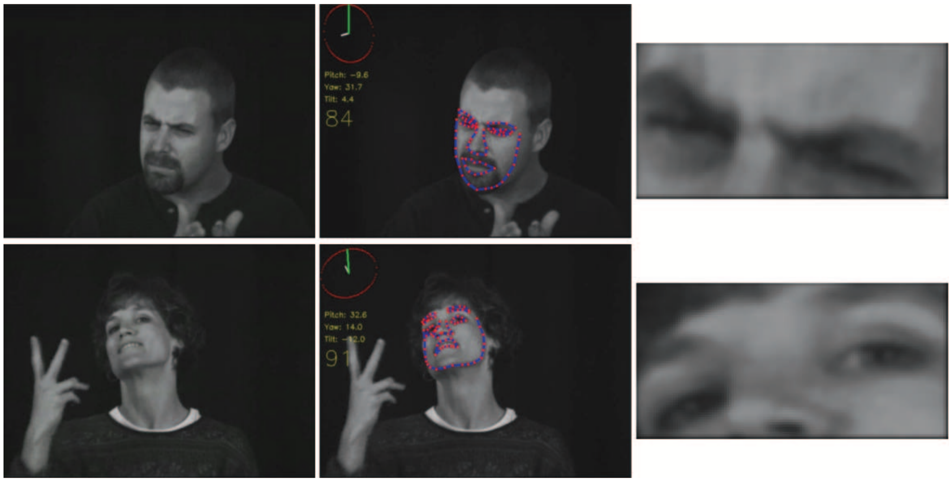

Figure 3: First column shows the input frame, second column shows the tracked face with the estimated 3D pose, and third column shows the extracted eye and eyebrow region. The top signer is producing a Wh-Question, while the bottom signer is producing a negative expression (best viewed in color)

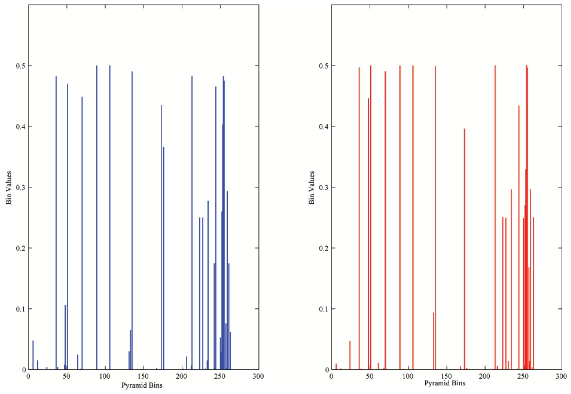

Figure 4: Spatial pyramids of SIFT descriptors (50-word codebook, $\sigma = 0.2$). Pyramid levels are increasing with increasing bin index. Left plot is for a Wh-Question. Right plot is for a Negative expression

the feature distribution within an image, which is effectively a concatenated vector of the bin values of all the histograms in the pyramid.

Figure 4 shows two spatial pyramid representations extracted from video sequences containing different facial expressions. The pyramid on the left corresponds to a frame in which the signer was producing a wh-question expression, while the pyramid on the right comes from a video of a negative expression. Examining the two plots, the difference in the pyramids is evident, especially in the levels of finer resolution (finer resolution bins are on the left). The input frames, together with the tracked faces and the extracted eye regions, that generated these spatial pyramids are shown in Figure 3.

In order to measure the distance between the feature sets X and Y, and eventually measure the dissimilarity in appearance between any pair of frames,we just need to compare their pyramid representations, essentially meaning comparing the bins of these histograms to see how much they match. Similar to [10], we measure histogram similarity at each level l, using the histogram intersection function presented in the work of Swain and Ballard [19] and defined as:

$$I(H_X^l, H_Y^l) = \sum_{j=1}^{C} \min\left(H_X^l(j), H_Y^l(j)\right) \ , \qquad (3)$$

where H_X^l and H_Y^l are the histogram representations of the two frames at level l, C is the number of cells at level l, while $H_X^l(j)$ and $H_Y^l(j)$ are the respective histograms of frames X and Y in the j^{th} cell of level l.

Since higher levels are of a finer resolution, it is intuitive to weigh the similarity match of cells in these levels with a higher weight than that used for the lower levels of coarser resolution. Moreover, if a match is found at a level l, it will also be found in the coarser level $l-1$, so when comparing feature sets, we should only consider the new matches found at each levels. This leads to the following match kernel for spatial pyramids having L levels:

$$K^L(X,Y) = \frac{1}{2^L} I^0 + \sum_{l=1}^{L} \frac{1}{2^{L-l+1}} I^l \ , \qquad (4)$$

where I^0 is the intersection score at level 0 and I^l is the intersection score at level l [10].

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 and softly quantize the yaw derivatives using a codebook that we compute from a random subset of the training set. Then we construct a temporal pyramid for each video, by dividing a sequence of frames into cells, in a similar fashion as done for spatial pyramids and using the same match kernel. In this way, we form a representation which allows us to detect the head shake of a signer. This is because we expect to see a distinct uniform pattern of yaw angle derivatives resulting from a head shake during a negative expression, which is distinct from the pattern of yaw derivatives resulting from other ASL expressions. This difference in yaw angle derivative patterns is illustrated in Figure 5.

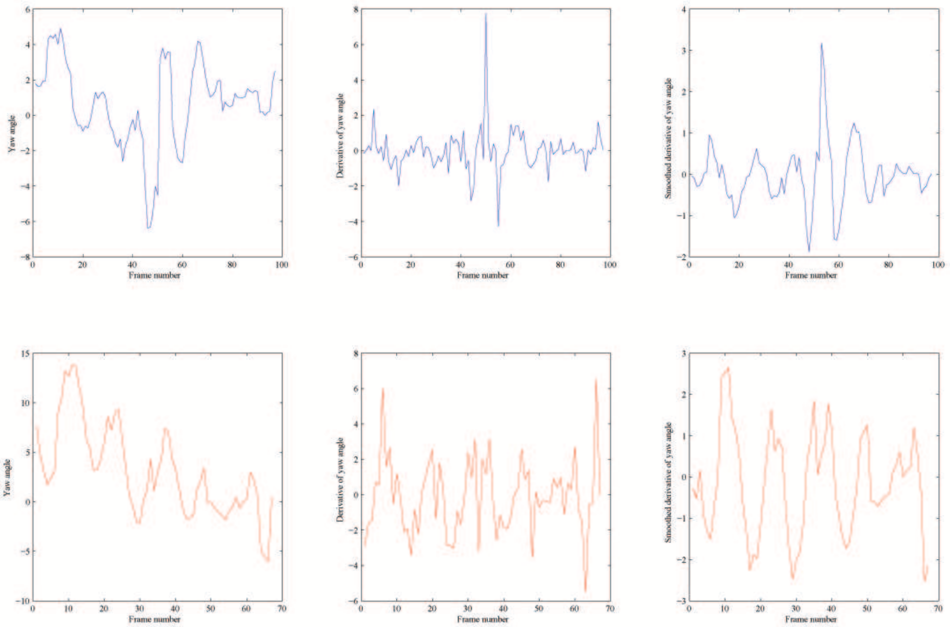

Figure 5: Sample plots of yaw angles, yaw derivatives and smoothed derivatives for two video sequences of different class. Top row plots are from a Wh-Question. Bottom plots are from a Negative construction

Table 1: Dataset Composition

	Training	Testing	Total
Wh-question	25	11	36
Non-Wh-question	25	11	36
Total	50	22	72
Negative Expression	22	10	32
Non-Negative Expression	22	10	32
Total	44	20	64

5. EXPERIMENTAL RESULTS

The Boston University American Sign Language Linguistic Research Project (ASLLRP) dataset used for the research reported here consists of 15 spontaneous short narratives plus over 400 additional elicited utterances collected from several native signers of ASL [16]. Synchronized video cameras captured the signing from multiple viewpoints (two stereoscopic front views plus a side view and a close-up of the face). The data were annotated using SignStream®, software[2] developed by our group specifically for linguistic annotation of visual language data [14]. The annotations include identification of start and end frames of individual signs as well as labeling of facial expressions and head movements that have grammatical significance.

In our experiments we used the close up view of the face only from isolated utterances. We selected a total of 36 video sequences showing wh-questions and 32 sequences showing negative expressions. These formed our set of positive examples for each of the two classes. An equal number of negative examples were collected by randomly selecting video sequences from different classes. We then randomly split our two datasets of wh-questions and negative expressions into a training and validation set, and into a test set, ensuring that both sets contained data from different signers. The duration of the video sequences ranged from 1.6 seconds to 6 seconds. The training sets contained about 70% of the total data, while the remaining data formed the testing set. Table 1 shows the dataset composition in more detail.

Table 2: Performance metrics

	Precision	Recall	Accuracy
Stacked Wh-question	*91.7%*	*100%*	*95.5%*
SIFT Wh-question	90.9%	90.9%	90.9%
Pose Wh-question	63.6%	63.6%	63.6%
Negative Expression	*90.9%*	*100.0%*	*95.0%*

Table 3: Confusion matrix for recognition of negative expressions

	Predicted as as Negative	Predicted as Non-Negative
True Negative	10	0
True Non-Negative	1	9

We used the ASM face tracker of [9] to track the signer's face in each sequence and extract their eye region, as well as predict their 3D head pose. Figure 3 shows sample results of tracking, pose prediction and localization of the eye region. The pose angle predictions were smoothed with a one-sided Gaussian filter with $\sigma = 2$ and a length of 7 frames, so that the pose in a given frame was a weighted combination of the pose predictions in that frame and of those in the 6 frames before it, in order to filter out noise. Pose angle derivatives were computed, as the difference in pose angle between two successive frames, and then a random subset was used to construct a codebook of 75 codewords using soft assignment

[2] http://www.bu.edu/asllrp/SignStream/

Table 4: Confusion matrix for recognition of wh-expressions using pose information only

	Predicted as Wh-Exp	Predicted as Non-Wh-Exp
True Wh-Exp	7	4
True Non-Wh-Exp	4	7

Table 5: Confusion matrix for recognition of wh-expressions using spatial pyramids only

	Predicted as Wh-Exp	Predicted as Non-Wh-Exp
True Wh-Exp	10	1
True Non-Wh-Exp	1	10

Table 6: Confusion matrix for recognition of wh-expressions using spatial pyramids and pose information

	Predicted as Wh-Exp	Predicted as Non-Wh-Exp
True Wh-Exp	11	0
True Non-Wh-Exp	1	10

[8] with a Gaussian kernel and $\sigma = 0.1$ (larger size codebooks did not achieve better recognition). Temporal pyramids with three levels (i.e. $L = 2$) were then constructed for each video sequence and a Support Vector Machine (SVM) with the pyramid match kernel discussed in Section 4.3 was trained via cross-validation, and used to classify the test set sequences into negative and non-negative expressions. An SVM is a popular machine learning algorithm that can be trained to discriminate two classes of objects, using some characteristic features of these classes (in this case the temporal pyramid representation), when presented with example instances from each class.

The SVM classifier achieved a precision accuracy of 90.9% and a recall rate of 100%, with an overall recognition accuracy of 95%. Using more levels in the temporal pyramid hurt the performance. Here, by recognition *accuracy* we refer to the percentage of instances in the training set that were correctly classified (i.e. $\frac{\text{tp + tn}}{\text{N}}$). *Precision* is the ratio $\frac{\text{tp}}{\text{tp + fp}}$ and *recall* is the ratio $\frac{\text{tp}}{\text{tp + fn}}$, where N stands for the total number of test set instances, tp stands for true positive, fp for false positive and fn stands for false negative. The detailed classification results are shown in Table 2, while Table 3 shows the confusion matrix, where we see that our proposed method of temporal pyramids yields only one false positive.

Similarly, from the localized eye regions we have extracted dense SIFT features [13], which we also quantized using soft assignment [8] with a Gaussian kernel and $\sigma = 0.2$. 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 4. We experimented with different codebook sizes but we found that a codebook of 100 words for the spatial pyramids performed adequately: for recognizing the wh-questions a single SVM classifier (which we call the “SIFT Wh-question” recognizer) only achieved a recognition accuracy of 90.9%.

However, as seen in Table 5 this classifier had room for improvement. We trained a second SVM (which we call the “Pose Wh-question” recognizer) on the pitch angle of the signer’s head in each frame, which achieved an accuracy of 63.6%, revealing an unknown weak correlation between head 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”) [20, 26] to combine the predictions of the “SIFT Wh-question” recognizer with those of the “Pose Wh-question” recognizer. Stacking [20, 26] is a machine learning method for training a classifier, which learns to smartly combine the individual predictions of multiple base classifiers in order to improve classification accuracy, by utilizing the specific expert knowledge learned during training by each of the base classifiers. The stacked SVM took as input features the prediction scores output by each of the other two SVM classifiers, and classified frames using an RBF kernel, with its σ parameter chosen by cross validation, into frames depicting wh-questions and frames depicting non-wh-questions. Majority voting was used to decide the class label prediction of each sequence, based on the predicted labels of their constituent frames. Recognition results are summarized in Table 2 and the confusion matrices of the base SVM classifiers and of the stacked SVM classifier are shown in Tables 4, 5 and 6, respectively. By looking at these tables, one can see that by combining both the appearance features and the head pose information, helped improve the recognition accuracy of the overall method, by correctly classifying the one false negative instance of the “SIFT Wh-question” recognizer. It is likely that, as research continues, for differentiation of non-manual markings that differ very subtly from one another, it is going to be crucial to combine multiple evidence (e.g., use head positions and movements, appearance features around the nose, etc.).

6. FUTURE WORK

Currently our system uses spatial pyramids and pose angles to determine whether there is a wh-question facial expression in each frame, and then uses majority voting to do the final recognition for the video sequence. In other words, predictions are made for each frame in isolation from its neighboring frames. However, successive frames in a video sequence are dependent. Thus, modeling the temporal dependency between features extracted from successive frames can make the system more robust to noisy measurements, enabling it to make more accurate predictions. To address this we began experimenting with spatio-temporal pyramids, which also take into account the time dimension and have already been successfully used to match objects in video shots [4]. Similarly, in order to improve our negative expression recognizer we are also investigating methods for temporal alignment and normalization of our features, which could help improve the discrimination power of our temporal pyramid representation.

As mentioned at the end of Section 5, combining multiple evidence will be crucial in helping recognize classes of non-manual markings that are only subtly different, so as part of our future research we will be looking at combining appearance information obtained from multiple regions of interest on the face, for example, the nose and even the mouth. Ultimately, for automatic translation of sign language to written or spoken language, automatic recognition of sentence boundaries will also be an interesting and important avenue to pursue.

7. CONCLUSIONS

We presented 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 non-manual components of signs encountered in isolated utterances of ASL video, by successfully recognizing both wh-questions and negative expressions with excellent accuracy. Lastly, we discovered a correlation between a signer's head pose, in particular their head's pitch angle, and the appearance of their eyes and eyebrows, and used it to improve classifier performance, in a stacked SVM framework. This finding reinforces the belief that combination of multiple evidence will be crucial in distinguishing the subtle differences between certain classes of non-manual markings. Future steps include extending our work to perform accurate recognition in longer and more challenging video sequences, as well as modeling temporal dependencies between successive frames to improve recognition accuracy.

8. ACKNOWLEDGMENTS

The research reported here has been partially supported by funding from the National Science Foundation (grants number #CNS-04279883, #CNS-0428231).

9. REFERENCES

[1] C. Baker-Shenk. A Micro-analysis of the Nonmanual Components of Questions in American Sign Language. Unpublished PhD Dissertation, 1983.

[2] B. Bauer and K.-F. Kraiss. Video-based sign recognition using self-organizing subunits. In *ICPR*, volume 2, pages 434–437, 2002.

[3] C. J. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2:121–167, 1998.

[4] J. Choi, W. J. Jeon, and S.-C. Lee. Spatio-temporal pyramid matching for sports videos. In *MIR '08: Proceeding of the 1st ACM international conference on Multimedia information retrieval*, pages 291–297, New York, NY, USA, 2008. ACM.

[5] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models – their training and application. In *Comp. Vis. Image Underst.*, pages 38–59, 1995.

[6] G. R. Coulter. American Sign Language Typology. Unpublished PhD Dissertation, 1979.

[7] L. Ding and A. M. Martinez. Three-dimensional shape and motion reconstruction for the analysis of American Sign Language. In *CVPRW '06: Proceedings of the 2006 Conference on Computer Vision and Pattern Recognition Workshop*, page 146, Washington, DC, USA, 2006. IEEE Computer Society.

[8] J. C. Gemert, J.-M. Geusebroek, C. J. Veenman, and A. W. Smeulders. Kernel codebooks for scene categorization. In *ECCV '08: Proceedings of the 10th European Conference on Computer Vision*, pages 696–709, Berlin, Heidelberg, 2008. Springer-Verlag.

[9] A. Kanaujia, Y. Huang, and D. Metaxas. Tracking facial features using mixture of point distribution models. In *ICVGIP*, 2006.

[10] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, pages 2169–2178, 2006.

[11] T. Leung and J. Malik. Representing and recognizing the visual appearance of materials using three-dimensional textons. *IJCV*, 43:29–44, 2001.

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

[13] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, 2004.

[14] C. Neidle. Signstream^{TM}: A Database Tool for Research on Visual-Gestural Language, 2000. Boston MA: American Sign Language Linguistic Research Project No. 10, Boston University.

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

[16] C. Neidle, N. Michael, J. Nash, and D. Metaxas. A method for recognition of grammatically significant head movements and facial expressions, developed through use of a linguistically annotated video corpus. *Proc. of 21st ESSLLI Workshop on Formal Approaches to Sign Languages*, July 2009.

[17] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[18] J. Shi and C. Tomasi. Good features to track. In *CVPR*, pages 593–600, 1994.

[19] M. J. Swain and D. H. Ballard. Color indexing. *IJCV*, 7:11–32, 1991.

[20] C.-F. Tsai and C. Hung. Automatically annotating images with keywords: A review of image annotation systems. *Recent Patents on Computer Science*, 1:55–68, 2008.

[21] C. Vogler and S. Goldenstein. Facial movement analysis in ASL. *Univers. Access Inf. Soc.*, 6(4):363–374, 2008.

[22] C. Vogler and S. Goldenstein. Toward computational understanding of sign language. *Technology and Disability*, 20(2):109–119, 2008.

[23] C. Vogler and D. Metaxas. *Handshapes and movements: Multiple-channel ASL recognition*, pages 247–258. LNAI. Springer, Berlin, 2004.

[24] U. von Agris, D. Schneider, J. Zieren, and K.-F. Kraiss. Rapid signer adaptation for isolated sign language recognition. In *V4HCI*, 2006.

[25] U. von Agris, J. Zieren, U. Canzler, B. Bauer, and K.-F. Kraiss. Recent developments in visual sign language recognition. *Univers. Access Inf. Soc.*, 6(4):323–362, 2008.

[26] D. H. Wolpert. Stacked generalization. *Neural Networks*, 5:241–259, 1992.

[27] J. Zieren and K.-F. Kraiss. Robust person-independent visual sign language recognition. In *Proceedings of the 2nd Iberian Conference on Pattern Recognition and Image Analysis*, volume LNCS, 2005.