Computer-based recognition of facial expressions in ASL: From face tracking to linguistic interpretation

Nicholas Michael*, **Carol Neidle**†, **Dimitris Metaxas***

*Computational Biomedicine Imaging & Modelling Center, Rutgers University
{nicholam, dnm}@cs.rutgers.edu
†Linguistics Program, Boston University
carol@bu.edu

Abstract

Most research in the field of sign language recognition has focused on the manual component of signing, despite the fact that there is critical grammatical information expressed through facial expressions and head gestures. We, therefore, propose a novel framework for robust tracking and analysis of nonmanual behaviors, with an application to sign language recognition. Our method uses computer vision techniques to track facial expressions and head movements from video, in order to recognize such linguistically significant expressions. The methods described here have relied crucially on the use of a linguistically annotated video corpus that is being developed, as the annotated video examples have served for training and testing our models. We apply our framework to *continuous recognition* of three classes of grammatical expressions, namely wh-questions, negative expressions, and topics. Our method is signer-independent, utilizing spatial pyramids and Hidden Markov Models (HMMs) to model the temporal variations of facial shape and appearance.

1. Introduction

Nowadays, speech recognition technologies have become standard components of modern operating systems, allowing average users to interact with computers verbally. Unfortunately, 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 (Vogler and Goldenstein, 2008b; Michael et al., 2009; Neidle et al., 2009). First of all, technology that automatically translates between signed and written or spoken language would facilitate communication between signers and non-signers, thus bridging the language gap. Secondly, such technology could be used to translate sign language into computer commands, favoring the development of additional assistive technologies. Moreover, it could facilitate the efficient archiving and retrieval of video-based sign language communication and could assist with the tedious and time-consuming task of annotating sign language video data for purposes of linguistic and computer science research.

However, sign language recognition poses many challenges. First, many of 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 different locations. At the same time, facial expression may also be involved in distinguishing signs, further complicating the recognition task. Secondly, there is variation in production of a given sign, even by a single signer. Additional variation is introduced by the *co-articulation* problem, meaning that the articulation of a sign is influenced by preceding and following signs. This can result in departures from the expected hand shape, location, and/or orientation found at the edge of a sign, and there may also be movement transitions between signs (sometimes referred to as "movement epenthesis"). Nevertheless, many methods (Vogler and Metaxas, 1998; Bauer and Kraiss, 2002; Vogler and Metaxas, 2004) have shown promising results in recognizing manual components of signs.

Furthermore, in sign language, critical grammatical information is expressed through head gestures (e.g., periodic nods and shakes) and facial expressions (e.g., raised or lowered eyebrows, eye aperture, nose wrinkles, tensing of the cheeks, and mouth expressions (Baker-Shenk, 1983; Coulter, 1979; Liddell, 1980; Neidle et al., 2000)). These linguistically significant nonmanual expressions include grammatical markings that extend over phrases to mark syntactic scope (e.g., of negation and questions). For example, in *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 the entire wh-question or solely over a wh-phrase that has moved to a sentence-final position. 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. The eyes may squint or close. Lastly, *topics* are characterized by raised eyebrows, wide eyes, head tilted back, and an optional nod.

Sign language recognition cannot be successful unless these nonmanual signals are also correctly detected and identified. For example, the sequence of signs JOHN BUY HOUSE could be interpreted, depending on the accompanying nonmanual markings, 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...".

Motivated by the importance of facial expressions and head gestures, we present a novel framework for robustly tracking and recognizing such nonmanual markings associated with *wh-questions*, *negative* sentences and *topics*. Our method extends prior work (Michael et al., 2009; Neidle et al., 2009), in which the signer's head is tracked and appear-

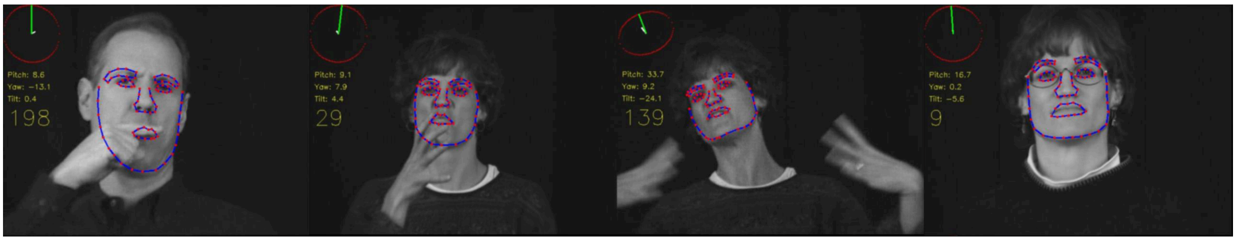

Figure 1: Sample frames showing the accuracy of tracking under challenging scenarios (partial occlusions, fast movements, and glasses), using our face tracker (Kanaujia et al., 2006). Here, red dots represent tracked landmarks. The predicted head pose is shown in the top left corner of each frame as a 3D vector

ance features, in the form of spatial pyramids (Lazebnik et al., 2006) of SIFT descriptors (Lowe, 2004), are extracted from the eye and eyebrow region (which we will refer to, henceforth, as the eye region). First, we extract additional shape features in the form of spatial pyramids of histograms of oriented gradients (PHOG) (Bosch et al., 2007). Second, we use spectral clustering (Ng et al., 2002), to reduce the dimensions of the augmented appearance and shape feature vectors. Third, by utilizing Hidden Markov Models (HMMs) (Rabiner, 1989), our method can perform *continuous* recognition in unsegmented video sequences.

2. Previous Work

As already mentioned, most research on computer-based sign language recognition has focused on the manual components of signing. A thorough review of early such efforts is presented in Pavlovic et al. (1997). Only recently have researchers begun to address the importance of facial expressions for sign recognition systems (Ong and Ranganath, 2005). An extensive review of recent developments in visual sign recognition, together with a system that captures both manual and nonmanual signs is provided by von Agris et al. (2008). However, their system requires the signer to be wearing a glove with colored markers to enable robust hand tracking and hand posture reconstruction. Additionally, in their system, the tracked facial features (lip outline, head pose, eye gaze, etc.) are not used to recognize facial expressions that have grammatical meaning. Vogler and Goldenstein (2008a; 2008b) present a 3D deformable model for face tracking, which emphasizes outlier rejection and occlusion handling 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. Lastly, the authors of (Michael et al., 2009; Neidle et al., 2009) use a method based on spatial pyramids (Lazebnik et al., 2006) to do *isolated* recognition of *wh-questions* and *negative* sentences. In this paper, we extend that work, so that we are now able to recognize in a *continuous* fashion *wh-questions* and *negative* sentences, as well as *topics* (i.e., no segmentation of test sequences is needed).

3. Face Tracking

Face tracking is a challenging problem because the tracker needs to generalize well to previously unseen faces and to varying illumination. 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. Kanaujia et al. (2006) tackle these problems with an Active Shape Model (Cootes et al., 1995), which is a statistical model of facial shape variation, where shapes are represented by a set of facial landmarks. Through the application of Principal Component Analysis (PCA) on an aligned training set of facial shapes, they learn a model of the permissible ways in which different people's faces differ, which is then used for face tracking.

Moreover, using a Bayesian Mixture of Experts model they are able to estimate the 3D pose of the head from the tracked landmarks. This model uses linear regressors and a multiclass classifier to map landmark configurations to predictions of head pose. Figure 1 shows the output of the ASM tracker on a few challenging input frames exhibiting rapid head movements and rotations, and partial occlusions.

Following ideas in (Michael et al., 2009; Neidle et al., 2009), the first step of our recognition framework involves tracking the signer's head using the above described framework (Kanaujia et al., 2006), localizing the facial components (e.g., eyes, eyebrows) and predicting the 3D head pose (i.e., pitch, yaw, tilt). We then extract from the eye region the features described in the next section.

4. Feature Representation and Recognition

In order to train machine learning algorithms for recognition of facial expressions, we first need a discriminative feature representation. Therefore, we extract dense SIFT descriptors over a regular grid from the eye region of each tracked frame; these are invariant to linear transformations such as scaling and rotation (Lowe, 2004). We cluster the SIFT descriptors of a random subset of the training frames, to obtain a codebook of prototypes and then encode all other descriptors by the index of their nearest prototype.

Next, we divide each frame into imaginary grids of cells and count the relative frequency of occurrence of each encoded feature in each cell. This collection of histograms becomes the spatial pyramid SIFT representation (PSIFT). In order to measure the dissimilarity in appearance between any pair of frames, we just need to compare their PSIFT representations, essentially comparing the bins of these histograms to see how much they match, using a weighted Spatial Pyramid Match Kernel (SPMK) with the histogram intersection function (Swain and Ballard, 1991; Grauman and Darrell, 2005; Lazebnik et al., 2006).

Bosch et al. (2007) also build spatial pyramids. Instead of SIFT descriptors, their idea is to quantize the gradient

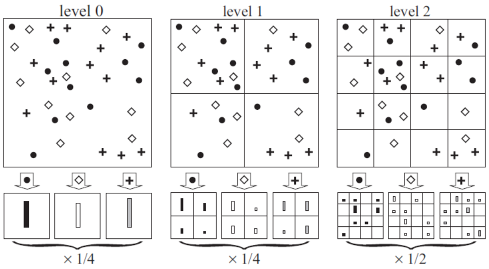

Figure 2: Toy illustration of spatial pyramid construction (Lazebnik et al., 2006), where, for simplicity, we assume there are only 3 codewords (circle, diamond, cross)

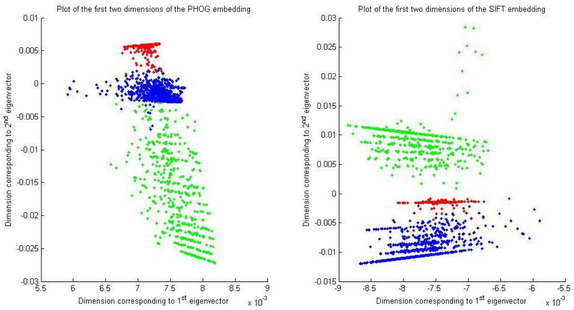

Figure 3: Spectral feature embedding of each frame (red: negative, green: topics, blue: wh-questions)

orientations of pixels into uniform bins, with each pixel's vote being proportional to the magnitude of its gradient, forming what they call a PHOG descriptor. We compute PHOG features in the same way, but for measuring PHOG similarity we use the weighted SPMK. By combining appearance (PSIFT) and shape (PHOG) features we obtain a more discriminative representation of eye regions.

5. Recognition Models

Although combining appearance and shape features improves the discriminative power of our representation, it increases the dimensionality of our input. As such, it increases the amount of training data that we need in order to learn accurate recognition models, and this also causes an increase in complexity, thus slowing down computations.

Spectral clustering (Ng et al., 2002) is a popular method of dimensionality reduction. The feature vector of each training example is represented as a node in a graph that is connected with a weighted edge to its nearest neighbors in the training set (weights reflect degree of similarity). The algorithm then applies an eigenvalue decomposition on the matrix representing this graph, reducing the feature vector dimensionality in a way that preserves the neighborhood structure. We use SPMK as the similarity measure and reduce the dimension of PSIFT and PHOG features separately. Figure 3 shows the resulting embedding of the training set, where we see that the classes are well separated.

The final feature descriptors per frame are the combined SIFT and HOG features of reduced dimensionality together with the 3D head pose and its first order derivatives. These are used to train HMM models (Rabiner, 1989). An HMM is a probabilistic model popular for time series data, consisting of a set of hidden states. At each time step, it transitions state based on a transition probability and it emits an observation. For our recognition task, we divide frames from each training sequence into four sets, one for each class of expressions we want to recognize. We train a separate HMM for each class, using sequences segmented by class.

	None	Negative	Topic	Wh-Q
Training	10144	997	1604	1208
Testing	9359	1053	1248	1182

Table 1: Dataset composition (number of frames per class)

	Predicted Class			
	None	Negative	Topic	Wh-Q
True None	92.8%	2.9%	2.2%	2.1%
True Negative	7.7%	80.3%	5.8%	6.2%
True Topic	9.2%	4.5%	81.2%	5.1%
True Wh-Q	8.3%	5.3%	4.5%	81.9%

Table 2: Confusion matrix of HMM continuous recognition

6. Use of the Annotated Video Corpus

The machine learning fundamental to our approach has been carried out using a linguistically annotated corpus of ASL (as produced by native signers) created at Boston University. This publicly available corpus, including 15 short narratives plus hundreds of additional elicited utterances, includes multiple synchronized views of the signing (generally 2 stereoscopic front views plus a side view and a close-up of the face), which have been linguistically annotated using SignStream™ (Neidle, 2002; Neidle et al., 2001) software, which enables identification of the start and end points of the manual and nonmanual components of the signing. Annotation conventions are documented (Neidle, 2002/2007), and the annotations are available in XML.

In order for pattern recognition algorithms to correctly identify a class of interest, they must be trained with both positive examples and negative examples. These are easily obtainable from the annotated corpus. From this corpus we selected a training set of 77 video clips of isolated utterances (negative: 17, topic: 40, wh: 20). Our testing set contained 70 such clips (negative: 15, topic: 38, wh: 17). The exact composition of these sets, in terms of numbers of frames per class, is shown in Table 1. Both sets contained three different signers. Using the methods described in previous sections, we tracked the signer's head, extracting pose, PHOG and PSIFT features, the dimensionality of which was then reduced using spectral clustering. We then trained class-specific HMMs, optimized to recognize frame sequences of their class. To evaluate their performance at continuous recognition, we used a sliding window approach. We fed subsequences of all *unsegmented* test sequences to each HMM, classifying each frame as negative, topic, wh, or none, based on which HMM output had the highest probability of having generated each subsequence.

Recognition accuracy is summarized in the confusion matrix of Table 2.

7. Discussion

We presented a novel framework for robust *real time* face tracking and facial expression analysis from a single uncalibrated camera. Our feature representation comprises spatial pyramids of SIFT and HOG features, and head pose features, which are reduced in dimensionality using a spectral decomposition. We demonstrated that our framework is successful at continuous recognition of wh-questions, negative expressions, and topics in unsegmented video data.

Feature fusion will be crucial in helping to recognize classes of nonmanual markings that are only subtly different. Therefore, as part of our future research we will be looking at combining facial features and looking at intensity and temporal patterning of nonmanual gestures (in relation, as well, to manual signing).

8. Acknowledgements

We gratefully acknowledge substantial contributions to this project by Benjamin Bahan, Lana Cook, Quinn Duffy, Robert G. Lee, Dawn MacLaughlin, Joan Nash, Michael Schlang, and Norma Bowers Tourangeau at Boston University. Signstream$^{\text{TM}}$ programming has been carried out by David Greenfield and Iryna Zhuravlova. The research reported here has been partially funded by grants from the National Science Foundation: #CNS-04279883, #EIA-9809340, and #IIS-9912573.

9. References

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

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

A. Bosch, A. Zisserman, and X. Munoz. 2007. Representing shape with a spatial pyramid kernel. In *CIVR*, pages 401–408.

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

G. R. Coulter. 1979. American Sign Language Typology. Unpublished PhD Dissertation.

K. Grauman and T. Darrell. 2005. The pyramid match kernel: Discriminative classification with sets of image features. In *ICCV*, pages 1458–1465, October.

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

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

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

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

N. Michael, D. N. Metaxas, and C. Neidle. 2009. Spatial and temporal pyramids for grammatical expression recognition of American Sign Language. In *ASSETS*, pages 75–82, October.

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

C. Neidle, S. Sclaroff, and V. Athitsos. 2001. Signstream$^{\text{TM}}$: A tool for linguistic and computer vision research on visual-gestural language data. *Behavior Research Methods, Instruments, and Computers*, 33(3):311–320.

C. Neidle, N. Michael, J. Nash, and D. Metaxas. 2009. 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.

C. Neidle. 2002. Signstream$^{\text{TM}}$: A database tool for research on visual-gestural language. *Journal of Sign Language and Lignuistics*, 4(1/2):203–214.

C. Neidle. 2002/2007. SignStream annotation: Conventions used for the American Sign Language Linguistic Research Project. Technical report, American Sign Language Linguistic Research Project Nos. 11 and 13 (Addendum), Boston University. Also available at http://www.bu.edu/asllrp/reports.html.

A. Y. Ng, M. I. Jordan, and Y. Weiss. 2002. On spectral clustering: Analysis and an algorithm. *NIPS*, 14:849–856.

S. C. W. Ong and S. Ranganath. 2005. Automatic sign language analysis: a survey and the future beyond lexical meaning. *IEEE TPAMI*, 27(6):873–891, June.

V. I. Pavlovic, R. Sharma, and T. S. Huang. 1997. Visual interpretation of hand gestures for human-computer interaction: A review. *IEEE TPAMI*, 19:677–695.

L. R. Rabiner. 1989. A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

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

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

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

C. Vogler and D. Metaxas. 1998. ASL recognition based on a coupling between HMMs and 3D motion analysis. In *ICCV*, pages 363–369.

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

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