A Framework for the Recognition of Non-Manual Markers in Segmented Sequences of American Sign Language

Nicholas Michael[1]
http://www.cs.rutgers.edu/~nicholam
Peng Yang[1]
http://www.cs.rutgers.edu/~peyang
Qingshan Liu[1]
http://www.cs.rutgers.edu/~qsliu
Dimitris Metaxas[1]
http://www.cs.rutgers.edu/~dnm
Carol Neidle[2]
http://www.bu.edu/asllrp/carol.html

[1] CBIM Center
Rutgers University,
Piscataway, NJ 08854, USA
http://cbim.rutgers.edu

[2] Linguistics Program
Boston University,
Boston, MA 02215, USA

Most research in the field of sign language recognition has primarily focused on the manual component of signing. However, 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 [2, 5]). 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 non-manual signals are also correctly detected. For example, depending on the accompanying non-manual markings, the sequence of signs JOHN BUY HOUSE could be interpreted 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...".

Therefore, we propose a novel framework for robust tracking and analysis of non-manual behaviours, with an application to sign language recognition. The novelty of our method is threefold. First, we propose a *dynamic feature representation* (see Figure 1). Instead of using only the features available in the current frame (e.g., head pose), we additionally aggregate and encode the feature values in neighbouring frames to better encode the dynamics of expressions and gestures (e.g., head shakes). Second, we use Multiple Instance Learning [3] to handle *feature misalignment* resulting from drifting of the face tracker and partial occlusions. Third, we utilize a *discriminative* Hidden Markov Support Vector Machine (HMSVM) [1] to learn finer temporal dependencies between the features of interest. We apply our signer-independent framework to *segmented recognition* of five classes of grammatical constructions conveyed through facial expressions and head gestures: wh-questions, negation, conditional/when clauses, yes/no questions and topics. By this, we mean that for each video sequence we only attempt to classify segments of frames containing some non-manual marker of one of these five classes.

More specifically, our method extends prior work in [4], where a face tracker first locates facial landmarks and then appearance and head pose features are fed to a Support Vector Machine (SVM) for classification. Once we track the facial landmarks in each frame, we focus on an extended rectangular region of interest (ROI), which includes the eyes, eyebrows and nose, so as to capture a wider range of upper face expressions, e.g., nose wrinkling and cheek tensing. We divide this ROI into a set of smaller patches (henceforth referred to as parts), which correspond roughly to areas of the face relevant for these specific grammatical expressions, e.g., inner and outer eyebrows. We extract from each part a histogram of Local Binary Patterns (LBP) [6]. These are effective for texture classification [8], faster to compute and more robust to illumination variations than SIFT, which is used in [4]. We then handle feature misalignment, arising from tracking inaccuracies and partial facial occlusions, by computing a Multiple Instance Feature (MIF) [3] for each part.

In addition to the head pose and texture features per frame, we explicitly calculate eyebrow height as the average Euclidean distance between the subjects' inner eyebrows and their nose tip. The 2D positions of the eyebrows and nose tip are computed from the tracked landmarks and filtered by a Kalman filter with linear dynamics and Gaussian noise. The final feature descriptor is augmented with a "summary" of the features of future and past frames sampled at regular intervals in the neighbourhood of the current frame, which we call *Oracle Features* (see Figure 1). This representation aims to encode the dynamic nature of facial expressions and head gestures encountered in non-manual grammatical markers. We demonstrate the validity of our method with experimental results based on a linguistically annotated corpus of ASL, as produced by native signers and collected at Boston University. This data is searchable via: `http://secrets.rutgers.edu/dai/queryPages`. In particular, we show that our method outperforms generative Hidden Markov Models (HMMs) [7], which can over-fit the training data in the absence of sufficient training examples, by utilizing a margin-maximizing, discriminative HMSVM [1] for recognition of. We also show that performance is degraded when either MIF or Oracle Features are not used.

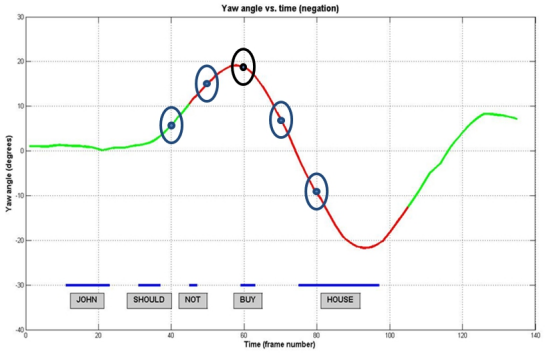

Figure 1: Plot of head yaw angle over time for a sequence containing negation (red segment marks when the non-manual marker occurs), also illustrating computation of yaw oracle features for frame 60

[1] Y. Altun, I. Tsochantaridis, and T. Hofmann. Hidden markov support vector machines. In *International Conference on Machine Learning*, pages 3–10, 2003.

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

[3] Z. Lin, G. Hua, and L.S. Davis. Multiple instance feature for robust part-based object detection. *IEEE Conf. on Computer Vision and Pattern Recognition*, 0:405–412, 2009.

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

[5] 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.

[6] T. Ojala and M. Pietikäinen. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *PAMI*, 2002.

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

[8] G. Zhao and M. Pietikäinen. Dynamic texture recognition using volume local binary patterns. *European Conference on Computer Vision*, 2006.