Classifying Non-manual Markers in American Sign Language

Pamela Toman and Alex Kuefler // 11 December 2015

Facial expressions and other non-manual markers are the primary means by which American Sign Language (ASL) users signal sentence and clause type, negation, and a variety of adverbs. Work in the area of non-manual marker recognition has the potential to make ASL-to-English translation available in contexts where interpreters and human relay services are currently rare, such as events with non-signing family members. It also promises to generate scientific insights into signed languages, an area of linguistics of which we know relatively little. Despite the potential for work in the area of signed language recognition, there is only limited research investigating how to automatically identify non-manual markers from video streams. We contribute to the field by evaluating the performance of three feature extraction methods at the binary classification problem of identifying the presence of a *wh*-question in streaming RGB video. Unlike previous work that attempts to classify utterance types as units, we focus on feature selection for identifying the presence/absence of a non-manual marker at different times in an utterance, in order to lay a foundation for machine translation work. With theory-based feature tracking, we achieve approximately 80% frame-level accuracy and 76% accuracy on sequences. Despite limited data, our attempt to use abstract features through neural networks resulted in close to 72% frame-level accuracy and 80% accuracy on sequences. A third approach using the Scale-Invariant Feature Transform algorithm was unimpressive, with 59% accuracy.

Related Work

We have been able to identify only a few papers at the intersection of facial feature recognition, machine learning, and American Sign Language non-manual marker processing. Most recently, Benitez-Quiroz et al. (2014) learn temporal patterning characteristics in a variety of non-manual markers using detailed hand-annotated data, though they do not address the classification task nor automated extraction of features.

Figure 1. The non-manual marker for the wh-question "when": head tilted, eyebrows lowered, eyes narrowed, lips drawn together. Image is from NCSLGR data and is overlaid with IntraFace's facial feature, head pose, and eye gaze tracking annotations.

The literature contains a handful of multi-class classification approaches. In particular, Michael et al. (2010) use HMMs with the SIFT algorithm from computer vision that identifies features with interesting gradients from images, Liu et al. (2014) use a linguistically informed two-level CRF, and Vogler and Goldenstein (2008), Michael et al. (2011), and Metaxas et al. (2012) address automatic identification of non-manual expressions in continuous signing despite the occlusions of the face that are common in ASL.

We were unable to identify any benchmark datasets in the literature, and although we would like to share our own, data use restrictions complicate that endeavor. The closest are two papers that report multi-class accuracies on utterance classification: Michael et al. (2010) report 80.3% accuracy on *wh*-questions in a 4-class classification problem given 77 utterances, and Metaxas et al. (2012) report 92.5% accuracy on *wh*-questions in a 5-class classification problem given 330 utterances. Although somewhat similar to ours, these papers seek to distinguish utterance types, whereas ours explores features that help identify the particular period over which a specific marker is present in streaming video. However, the existence of this work indicates that additional attention to feature selection has the potential to be helpful to others working in automated non-manual marker recognition.

Data

We initially collected 172 utterance videos containing *wh*-questions from the National Center for Sign Language and Gesture Resources (NCSLGR) corpus (Neidle and Vogler, 2012). The

videos collected as data are compressed RGB form, and contain a total of 14,046 frames, including 7,970 positive examples, and 6,076 negative examples. They span a variety of *wh*-questions,[1] and examples come from five male speakers and three female speakers. Positive examples are frames or sequences of frames that contain a *wh*-question marker; negative examples are frames or sequences of frames from the same video that do not contain a *wh*-question marker. *Wh*-words used in rhetorical questions and in other statements are treated as negative examples.

Feature Extractors

We compare four different feature extraction methods. At the most theoretical, we use features extracted by ASL users with linguistics experience; this non-automated approach contextualizes performance on the task and offers a fully theory-driven baseline. As an automated but theory-based approach, we work with specific identifiable points on the face using the Carnegie Mellon Robotics Institute's IntraFace tool (Xiong and De la Torre, 2013). At the next higher layer of abstraction, we find clusters of data-driven high-gradient-change features through the Scale-Invariant Feature Transform (SIFT) algorithm, and represent each image as a feature vector for the presence of each cluster in each image. Finally, at the highest level of abstraction, we use a convolutional neural network (CNN) as a feature extractor. We discuss each in more detail below.

Human Baseline

Using linguistics theory-based features painstakingly encoded by humans (see Neidle and Vogler, 2012), we identify a baseline goal for performance that helps contextualize machine performance. Sample binary frame-level features used at this highly theoretical phase include "eye aperture => (lowered lid)", "head pos: tilt fr/bk => (slightly back)", and "POS => (Noun)".

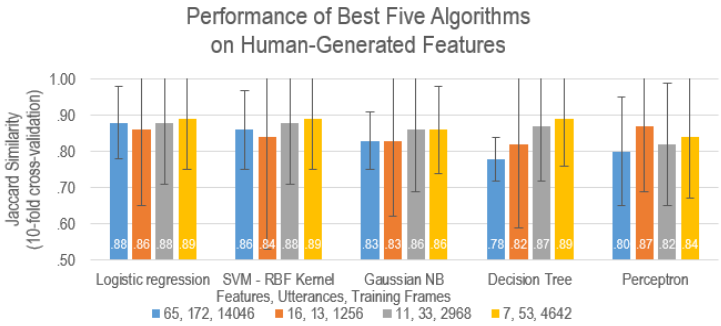

Figure 2. In general, the more training examples, the better the performance – even when the number of features decreases. These results suggest that this domain will benefit from additional good data.

Because distinct utterances were not annotated with a consistent set of features, we experimented with preprocessing to balance number of features and number of examples. As per Figure 2, we found the best performance when including the most samples, either by recoding missing target variables as non-presence of the target feature or by limiting to a very small set of features and using the few thousand associated examples. Using the smaller number of examples that shared a larger set of features did not perform as well, suggesting that more training data – even when slightly incorrect – is very valuable for this question domain.

In the absence of a benchmark dataset, we treat the best human annotation performance of 0.89 +/- 0.14 as a contextualizing performance goal, despite recognition that human annotations and machine annotations are distinct, and that there may be correlational biases or missed information in the human annotations that are not reflected in the machine learned annotations.

Tracked Facial Markers

Xiong and De la Torre (2013) develop an algorithm for supervised nonlinear optimization in the realm of computer vision and apply it to facial feature detection. We utilize their executable IntraFace to extract 28 facial landmarks, head pose, eye gaze, and iris features.

[1] The configuration of non-manual gestures combined in the *wh*-question marker in ASL is used for multiple question words (e.g., who, what, where, when, why, how, what-for, how-many). Linguistically, the *wh*-question marker indicates an interrogative pro-form.

We augment the tracked features in three ways. First, we extract the Euclidean distance deltas between each of the facial landmarks to capture rotation invariant aspects of the facial layout and directly measure theoretically important variables like eye/mouth aperture and eyebrow height. Second, for each feature we add time-sequence deltas from the previous and next frames, and we remove the neighboring frames from the dataset to not contaminate the evaluation. Finally, after creating all the raw features we normalize each feature within each video utterance. Our results suggest that appropriate processing to make cross-video features more consistent and incorporate new features improves performance substantially (see Figure 3).

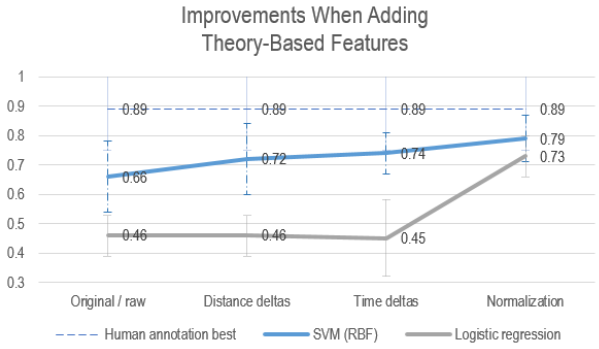

Figure 3. By augmenting the tracking features, we were able to achieve +0.13 (SVM) and +0.27 (logistic regression) improvements in performance (with accuracy measured as Jaccard similarity), while simultaneously shrinking variance.

We also experimented with trimming frames in which the facial orientation was an extreme in at least one of pitch/roll/yaw to limit all training and testing examples to a frontal orientation. However, we found that trimming as preprocessing offered only mild improvement (on the order of 0.01 +/- 0.10 for SVMs). This finding suggests that given well-preprocessed data and a model that accounts of interaction effects, *wh*-questions may be identifiable even with extremes of head pose, and that any trimming is better used to eke out late-stage improvements rather than to be a primary source of concern.

Flattened SIFT Descriptors from IntraFace Keypoints

The second feature extractor relies on Scale-Invariant Feature Transform (SIFT). Given points of interest around the image, SIFT generates a set of orientation histograms, where each bin contains the number of gradients facing a certain direction within a sub-region of a 5x5 pixel area surrounding the keypoint (Lowe, 2004). The intent of this work was to identify informative features of the face through gradients and textures. Initial work learned the points of interest automatically and then clustered them, creating a feature vector for each frame based on the clusters observed; performance of this approach was low (39.62% frame-level accuracy on true positives, 96.81% on true negatives). Leveraging the high precision of the IntraFace facial landmarks, we then generated 28 keypoints around signers' faces, flattening the SIFT descriptor generated at each point into distinct feature vectors for each frame. Performance on SIFT given IntraFace-selected points of interest was also low (85.84% on true positives, 21.46% on true negatives), leading to us focus more of our efforts on the tracked facial features and neural network feature extractors.

Deep Convolutional Feature Extraction

By tuning connection weights across hidden layers, convolutional neural network (CNN) models essentially learn a set of filters that automate the process of feature extraction. Since CNNs can involve tens of millions of parameters, which require extensive time and large datasets that were unavailable for this problem domain, we used a pre-trained CNN with the same architecture and learned parameters as AlexNet (Jia et al., 2014). AlexNet consists of over 60-billion parameters

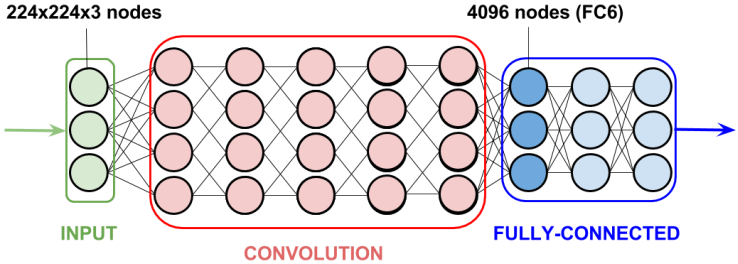

Figure 4. Cartoon of AlexNet architecture. Convolutional and pooling layers (red) apply a series of filters to 224x224 RGB image input, before activation vectors are generated in the fully-connected layers (blue).

spread across eight layers, trained for a 1000-class image recognition task on the ImageNET dataset (Krishevsky, Sutskever & Hinton, 2012; Russakovsky et al., 2015). We pre-processed the data using OpenCV's implementation of the Viola-Jones Algorithm (Bradski, 2000), which extracted from each frame a 57x57 pixel region around the subject's face. We fed the resized ROIs discovered during pre-processing into the input layer of the network and used as features the vector of activations across one of the three fully-connected (FC) layers that lie downstream of the network's learned filters. This procedure is described by Donahue et al (2014), who reported best results on an image recognition task using the 4096-dimensional outputs from AlexNet's first fully-connected layer (FC6). In testing, we also observed peak performance using FC6-features, which occur earliest after convolution and thus capture the most domain-generalizable information of the three FC layers.

Evaluation

We carried out two experiments to evaluate the approaches to feature extraction. The first is a frame-level binary classification task. The second is a time-series-sequence level binary classification task. We standardized models across features as an SVM with RBF kernel and performed 10-fold cross-validation in each experiment. Due to the highly correlated appearance of adjacent frames in a single utterance, we ensured that each utterance appeared in only a single fold, and we avoided testing on frames and sequences that come from the same utterances as those in the test set (see Figure 6 for distinctions between utterances, frames, and sequences). Our best models achieve approximately 80% accuracy on the frame-level task and 90% accuracy on the sequence-level task. The rest of this section provides details about each evaluation.

Frame-level Task

Our first evaluation of our three proposed feature extraction methods considers the setting in which each of the 14,046 frames in the 172 video corpus are treated as individual examples. Each frame was assigned a binary label corresponding to whether or not it occurred between the onset and offset of a *wh*-question. Positive examples account for approximately 58% of the dataset. The error rates and confusion matrices of each SVM extractor are given below:

Tracked Facial Markers
Accuracy: 79% +/- 8%

	Actual +	Actual -
Pre. +	80.82%	19.18%
Pre. -	23.05%	76.96%

SIFT Descriptors
Accuracy: 58% +/- 12%

	Actual +	Actual -
Pre. +	57.73%	42.27%
Pre. -	28.89%	71.11%

CNN (FC6 Only)
Accuracy: 72% +/- 8%

	Actual +	Actual -
Pre. +	75.57%	24.45%
Pre. -	32.93%	67.07%

Figure 5. Error rates and confusion matrices for the frame-level task. Accuracies calculated as Jaccard similarity.

We find that tracked facial markers perform best for the frame-level evaluation, followed by the CNN-as-feature-extractor. The SIFT descriptors are not competitive, so we focused our efforts on improving the other two.

Sequence-level Task

Our second evaluation of our three proposed feature extraction methods divides each

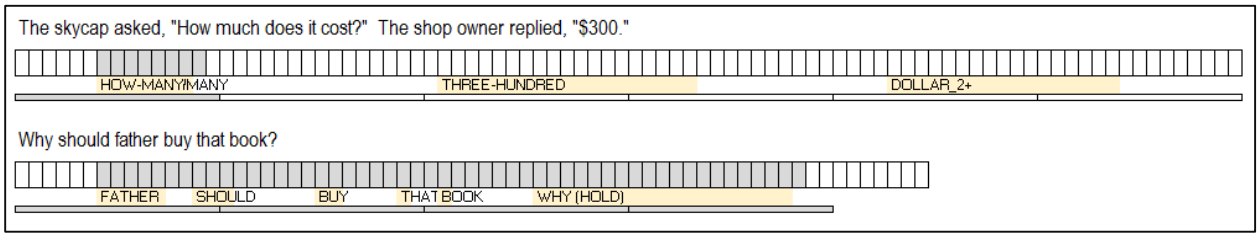

Figure 6. Two utterances from the dataset, with details regarding frame and sequence evaluation. *First line:* English translation. *Second line:* Frames with extent of *wh*-question non-manual marker shaded. *Third line:* Glosses and span of manual signs. *Fourth line:* Sequences with positive examples shaded.

utterance into 15-frame sequences; positive examples contain a *wh*-question in at least a third of constituent frames. The sequence setting balances the desire for extracting detailed information about timelines of feature presence with the desire for larger units that simplify the challenge identifying a non-manual marker from a single still frame. To balance the amount of positive and negative data, we collected 29 additional samples from the NCSLGR database that do not include *wh*-questions. This evaluation set consisted of 19,094 frames divided into 1,276 sequences across 201 utterances. Approximately 42% of these sequences were positive.

In this task, frames in a sequence vote as to whether they are part of a *wh*-sequence, and a hyper-parameter-learned threshold distinguishes predicted positive and negative examples. In order to fit the threshold hyper-parameter, we held out a single test set consisting of 22 utterances, and used labeled 10-fold cross validation to split the remaining data into training and validation sets. The threshold value maximizes the mean classification accuracy across the 10 validation sets.

Tracked Facial Markers
Accuracy: 76%
(Validation: 63% +/ 12%)

	Actual +	Actual -
Pre. +	86.00%	14.00%
Pre. -	50.00%	50.00%

CNN (FC6 Only)
Accuracy: 80%
(Validation set: 63% +/- 8%)

	Actual +	Actual -
Pre. +	84.61%	15.39%
Pre. -	22.23%	77.77%

Figure 7. Error rates and confusion matrices for the sequence-level task. Accuracies are calculated as Jaccard similarity on the single held-out dataset. The 10-fold hyperparameter setting results appear in parentheses.

We find that performance improves on the sequence-level task compared to the frame-level task, and again that tracked facial features outperform the neural network approach, though both perform well.

Discussion and Conclusions

The best performance is from the processing highly theory-based facial feature tracking, rather than with the highly abstracted CNNs or hybrid SIFT method. We suspect this is driven in large part by the limited amount of data available, such that data-driven methods are at a disadvantage. We also suspect the hybrid SIFT model might have performed better given substantially more pre-processing, such that the meaningful patterns were more apparent to it.

We found with the facial feature approach that appropriate pre-processing improved performance substantially. As such we would support a rigorous evaluation of preprocessing options used in the literature – which range in complexity from the appropriate-for- streaming-translations methods used here to more complex suggested methods like 3-D deformable face models – with regard to effectiveness, complexity, and runtime.

We believe the use neural networks in this application area is novel. Given the 72% / 80% performance on limited examples of a generically trained network, we believe that there is promise in the further application of neural networks to this problem domain. Until large annotated ASL databases are developed, we hope that work with pre-trained neural nets may drive interest in bringing Big Data tools to bear on this problem domain, thereby allowing learning CNNs to become a viable option for frame-level classification and recurrent neural networks to become a viable option for labeling at the level of sequences.

Based on our experimental results, sequence classification through threshold-voting appears to be a good alternative that performs on par with HMMs (Michael et al., 2011). However, we should maintain some skepticism, as the test set we evaluated on was relatively small (22 utterances, 1917 frames). Nevertheless, our results indicate that frames are indeed a useful unit of analysis for this question, and that voting shows early promise as a means for leveraging single-frame information to make judgments about series and for identifying onset and offset of non-manual markers in ASL, as is needed to make progress toward machine translation.

References

Benitez-Quiroz, C.F., Gökgöz, K., Wilbur, R. B., & Martinez, A. M. (2014). Discriminant Features and Temporal Structure of Nonmanuals in American Sign Language. PLoS ONE 9(2): e86268.

Bradski, G. (2000). The OpenCV Library. Dr. Dobb's Journal of Software Tools.

Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., & Darrell, T. (2014). DeCafe: A deep convolutional activation feature for generic visual recognition. ICML, 2014.

Krishevsky, A., Sutskever, I., & Hinton, G. E (2012). ImageNet Classification with Deep Convolutional Neural Networks. Advances in Neural Information Processing Systems 25.

Liu, J., Liu, B., Zhang, S., Yang, F., Yang, P., Metaxas, D. N., & Neidle, C. (2014). Non-manual grammatical marker recognition based on multi-scale, spatio-temporal analysis of head pose and facial expressions. Image and Vision Computing 32(10), 671-681.

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

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Gaudarrama, S., & Darrelll, T. (2014). Convolutional architecture for fast feature embedding. In arXiv preprint arXiv:1408.5093.

Metaxas, D., Liu, B., Yang, F., Yang, P., Michael, N., & Neidle, C. (2012). Recognition of Nonmanual Markers in American Sign Language (ASL) Using Non-Parametric Adaptive 2D-3D Face Tracking. LREC 8.

Michael, N., Neidle, C. & Metaxas, D. (2010). Computer-based recognition of facial expressions in ASL: from face tracking to linguistic interpretation. Proceedings of the 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies.

Michael, N., Yang, P., Liu, Q., Metaxas, D., & Neidle, C. (2011). A Framework for the Recognition of Non-Manual Markers in Segmented Sequences of American Sign Language. LREC 8.

Neidle, C. & Vogler, C. (2012). A New Web Interface to Facilitate Access to Corpora: Development of the ASLLRP Data Access Interface. Proceedings of the 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon. Data available at http://www.bu.edu/asllrp/.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., & Fei-Fei, L. (2015). ImageNet large scale visual recognition challenge. IJCV, 2015.

Pedregosa, F., Varoquauz, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R. Dubourg, V., Vanderplas, J., Passos, A., Cornapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830.

Vogler, C. & Goldenstein, S. (2008). Toward computational understanding of sign language. Technology and Disability 20, 109-119.

Xiong, X. & De la Torre, F. (2013). Supervised Descent Method and its Application to Face Alignment. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013. Executable available at http://www.humansensing.cs.cmu.edu/intraface.