Classifying Non-Manual Markers in American Sign Language

CS 229: Machine Learning

Pamela Toman & Alex Kuefler

Task Definition:

In American Sign Language (ASL), facial expressions are grammatical. Only facial expressions distinguish whether the manual signs "HOME YOU" are a declarative sentence, a yes/no question, a negated sentence, or a command.

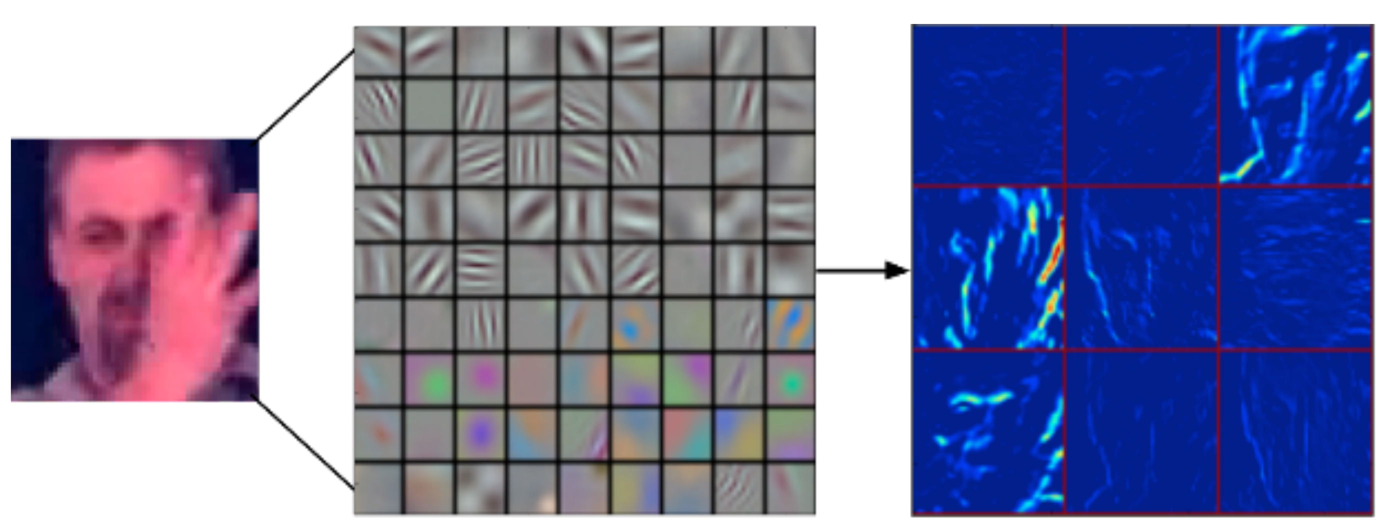
We evaluate the performance of three broad types of feature extraction, applied to a binary classification task. Given RGB video of signers' faces, our system determines whether or not a *wh*-question (who, what, where, when, why, how) is being asked.

The non-manual marker for the wh-question "when": head tilted, eyebrows lowered, eyes narrowed, lips drawn together.

Image is from ASLLRP data and is overlaid with IntraFace's facial feature, head pose, and eye gaze tracking annotations.

Data:

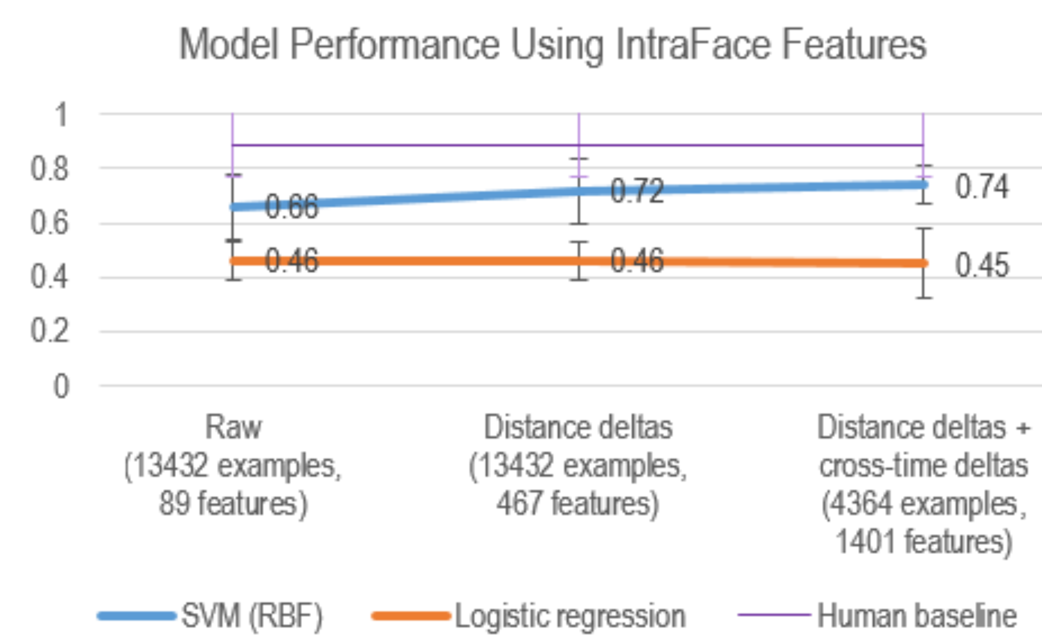
- Data from the National Center for Sign Language and Gesture Resources (NCSLGR) corpus (Neidle and Vogler, 2012): 172 utterances
- Frame-level: 7970 frames positive, 6076 negative
- Sequence-level (1.5 second): 240 positive, 123 negative
- Pre-Processing: Face bounding box recognition. Haar Cascades (Viola Jones Algorithm) extract face regions of interest from frames (Bradski, 2000), IntraFace extracts facial feature markers (Xiong & De la Torre, 2013)



Parameters fixed during pre-training represent a set of filters (middle) at each convolutional layer. Feeding image forward performs discrete convolution on input (left). Outputs from *first* conv. layer shown (right).
Image from ASLLRP data, visualizations produced using Caffe software (Jia et. al, 2014).

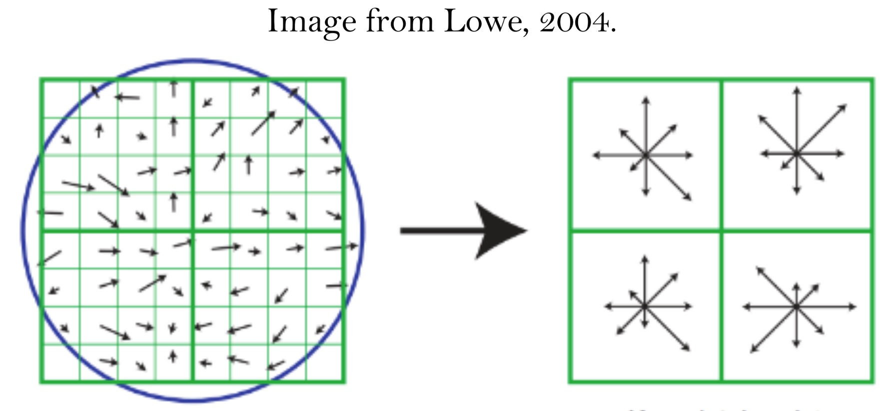
Feature Extraction:

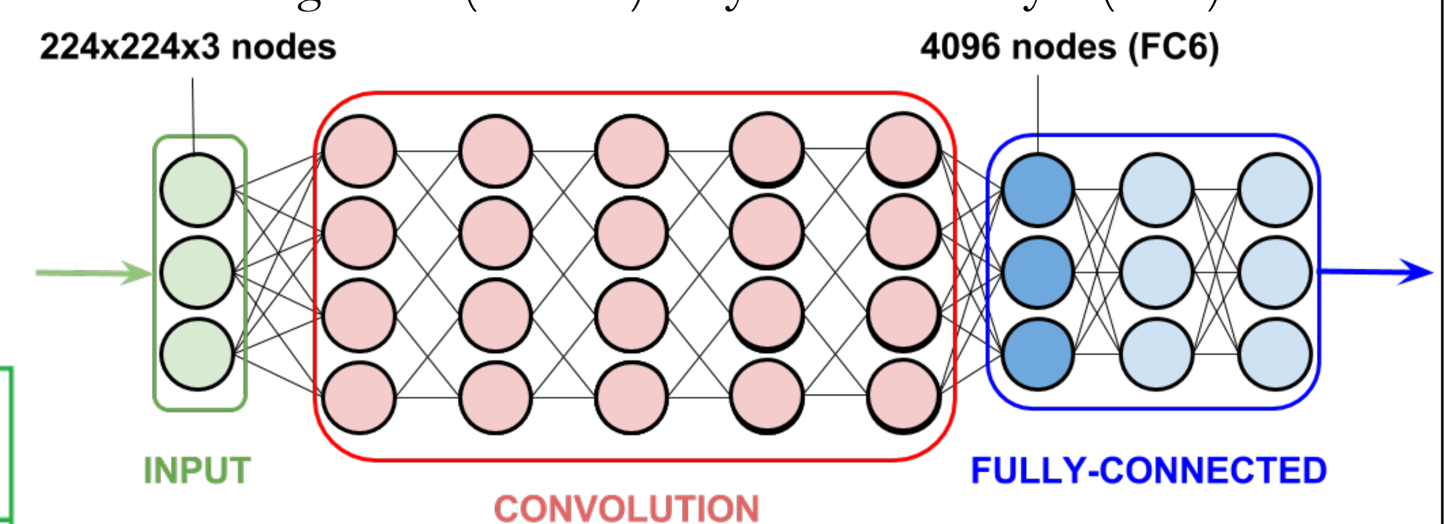
- Theory Driven Methods
  - **Human annotations: Best & Baseline**
    Frame-level prediction Jaccard similarity is 89% +/- 10% (best of 6 set ups). Features are heuristic (e.g., "body lean (forward/left)", "head mvmt: nod (onset)", "eye aperature (further squinted)").



  - **Intraface Markers: Can we perform as well without the overhead of human coding?**
    IntraFace extracts 89 human-interpretable facial features (28 facial landmarks, head pose, eye gaze and iris features). Best performance within bounds of human. Experimented with:
    - Original features alone
    - Adding facial manipulations through same-timestep landmark distance deltas
    - Adding movement features through cross-timestep distance deltas
- Data Driven Methods
  - **Convolutional Neural Network** (Pre-Trained "AlexNet")
    - Idea: Transform face images by passing them through deep net with general image knowledge. Take network outputs as mappings into feature space.
    - AlexNet: Over 60-billion parameters, trained on 1.2-million images representing 1000 classes. Features 5 convolutional layers, 3 fully connected layers (Krishevsky, Sutskever, Hinton, 2012).
    - 4096-dim. feature vectors drawn from most general (earliest) fully-connected layer (FC6).

(below) SIFT Descriptor Construction. Pixel gradients (left) are weighted by Gaussian filter. The weighted gradients of each 4x4 sub-region are summed to produce an orientation histogram describing the sub-region.
Image from Lowe, 2004.





(above) Cartoon of AlexNet architecture. Convolutional and pooling layers (red) manipulate image, before activation vectors are generated in the fully connected layers (blue).

  - **Scale-Invariant Feature Transform (SIFT)**
    - Keypoint Localization: 25 pixel extrema found by searching image at multiple scales.
    - Descriptor Generation: Gradient orientation computed at *each* pixel in region surrounding *each* keypoint. Binned to produce one 128-dim. vector for each keypoint (Lowe, 2004).
    - 15-Means Codebook: Unreliable keypoint localization in OpenCV (Bradski, 2000) created a setback. K-means was used to compress inconsistent keypoint descriptors into standard format.
- Hybrid Method: **SIFT Keypoint Localization with Intraface**
  - Eliminates need for K-means encoding, due to Intraface's precision

Results:

- 10-fold cross validation, each utterance in either test (~10k) or train (~1200)

Confusion:

- **Baseline**
  - *SVM: RBF Kernel*
  - Accuracy: 0.86 (+/- 0.11)

TN:85.6%	**FN**:14.4%
FP:13.1%	**TP**:86.9%

- **Intraface**
  - *SVM: RBF Kernel*
  - Accuracy: 0.74 (+/- 0.07)

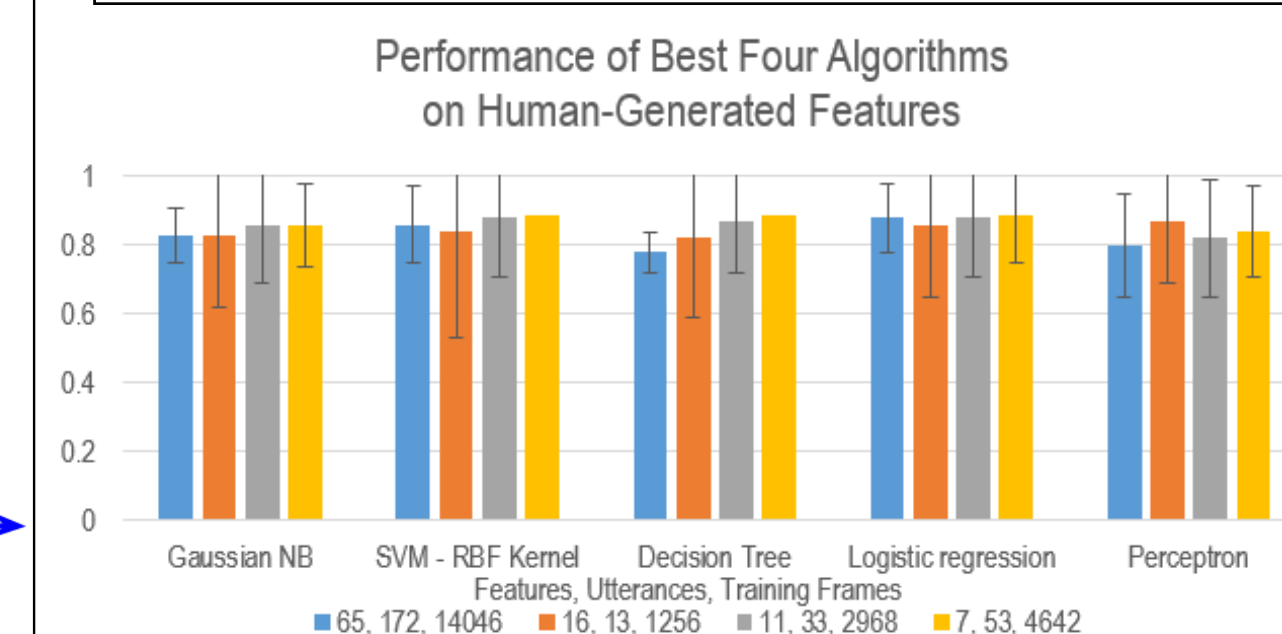
TN:67.4%	**FN**:32.6%
FP:19.1%	**TP**:80.9%

- **AlexNet**
  - *SVM: Linear Kernel*
  - Accuracy: 0.65 (+/- 0.04)

TN:60.52%	**FN**:39.47%
FP:30.22%	**TP**:69.77%

- **SIFT (Hybrid)**
  - *Logistic Regression*
  - Accuracy: 0.59 (+/- 0.04)

TN:21.46%	**FN**:78.53%
FP:14.15%	**TP**:85.84%



Next Steps:

- Whole-Sequence Classification
  - Performance may further improve at sequence level
  - Use frames to "vote" for sequence class
- Unsupervised Methods
  - Clustering: Can performance be improved by using only frames where the face is forward? Are there patterns in non-manual marker usage?

Select References:

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

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

Lowe, D. (2004). Distinctive image features from scale-invariant keypoints. In 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

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, LREC 2012, Istanbul, Turkey. Data available at http://www.bu.edu/asllrp/ and http://secrets.rutgers.edu/dai/queryPages/.

Xiong, X. & De la Torre, F. (2013). Supervised Descent Method and its Application to Face Alignment (patent pending). IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013.