Automatic Recognition of Fingerspelled Words in British Sign Language

Stephan Liwicki & Mark Everingham

UNIVERSITY OF LEEDS

Contribution

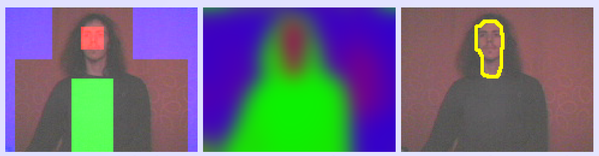
We propose a method for recognition of continuous British Sign Language (BSL) fingerspelling. The method uses only *appearance* of the hands, not motion, and gives highly accurate recognition for a lexicon of 100 words. By using appearance alone we require only *letter* examples for training, not words, improving scalability over previous approaches, which rely on co-articulation between pairs of letters. The method comprises four stages:

- Hand segmentation using graphcut with automatically bootstrapped spatial color models.
- Hand shape description using state-of-the-art HOG-based descriptor.
- Discriminative hand shape classification.
- Hidden Markov Model (HMM) representation of words built from text alone, without training examples of words.

Hand Segmentation

The hands are segmented to provide a normalized coordinate frame. An MRF formulation is used, solved using maxflow.

Color models for skin/body/background, a spatial prior, and a mask for the face region are automatically bootstrapped from a face detection.



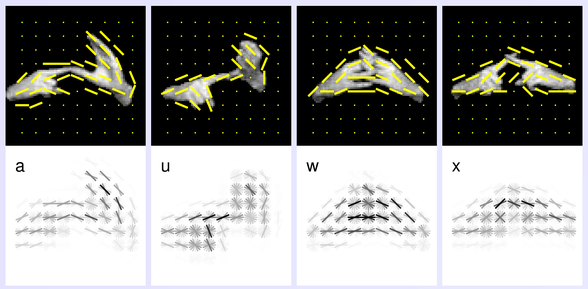
Automatic labeling | Learnt spatial prior | Learnt face mask

The spatial prior, learnt by smoothing the initial automatic segmentation, enables correct segmentation of the hands when (i) the background is skin-colored; (ii) the clothing color is close to shading on the hand.

Spatial prior | No prior

Hand Shape Descriptor

Most previous work has used simple descriptors of the hand silhouette alone e.g. shape moments. Here we use a variation of the Histogram of Oriented Gradients (HOG) descriptor. This gives some invariance to hand pose and shading, while capturing both *silhouette* and *internal* features.

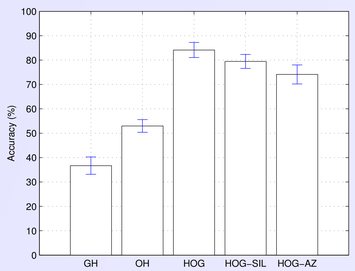


Letter Recognition

Multi-class logistic regression is used to train classifiers for each letter and "non-sign". This gives a posterior estimate $P(letter|descriptor)$ needed in the word recognition stage. Comparable letter-level accuracy is obtained with linear/kernel SVM.

We compared simple silhouette-based features, orientation histograms and HOG with/without internal features.

Training data was either (i) alphabet A-Z; (ii) pangrams e.g. "The quick brown fox jumps over the lazy dog."



Method	Accuracy
GH	36.7 ± 3.5
OH	53.0 ± 2.6
HOG	$\mathbf{84.1 \pm 3.1}$
HOG-SIL	79.5 ± 2.9
HOG-AZ	74.1 ± 3.9

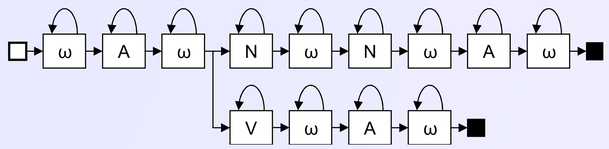
- HOG performs significantly better than shape moments/orientation histogram: **84.1%**.
- Capturing internal features improves accuracy.
- Training using pangrams gives significantly better results, by representing variation in the signs.
- Previous work [1] achieved 75.4%/57.9% for single letter/word training, using motion features.

Word Recognition

To recognize continuous fingerspelling, each word in a known lexicon is modeled as an HMM. This enables (i) suppression of repeated output for a single sign; (ii) disambiguating visually ambiguous signs.

The HMMs can be built *without* examples of words, since there is a 1:1 correspondence between letter signs and states, allowing scalability to large lexicons. The state is *unobserved* – the letter classifier provides an estimate of $P(state_t|descriptor_t)$.

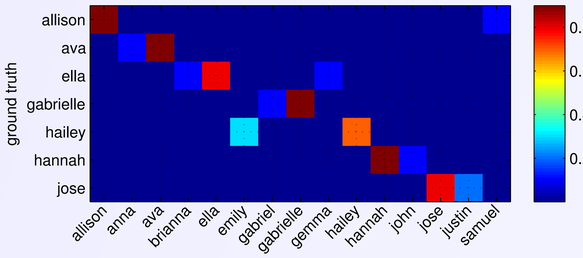
Inference is efficient by building a tree structure of HMMs and applying the max product algorithm.



Part of tree of HMMs for two words: "Anna" and "Ava"

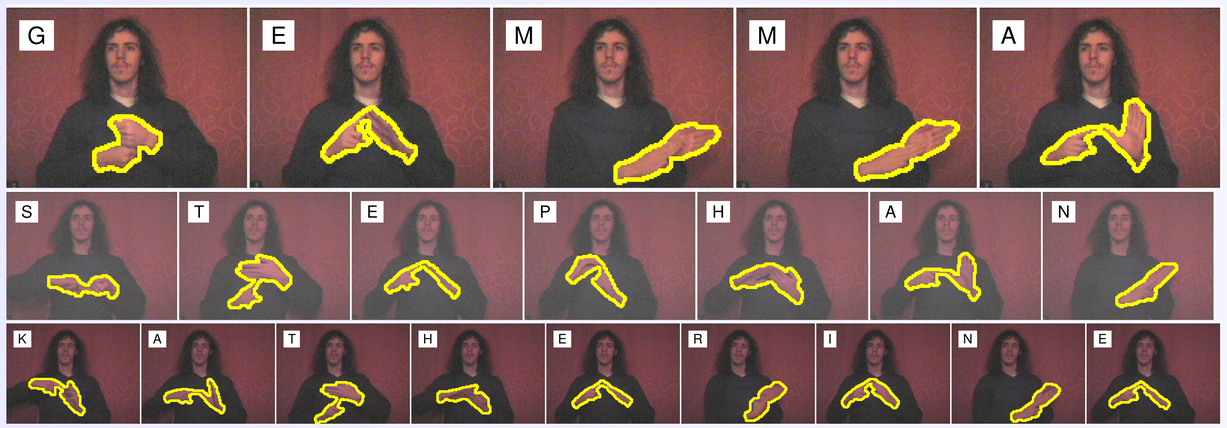
Experiments used 1,000 videos of the 100 most common male/female forenames in USA 2000–2007.

- Overall accuracy is **98.9%** – only 11 out of 1,000 videos (of 7 words) are misclassified.
- Previous work [1] achieved 88.6% for 20 words, and required videos of *words* for training.
- Estimated accuracy is 87.4% for 1,000 words, 81.0% for 2,000 words.



Confusion matrix for the 7 words with accuracy <100%.

Example Results



Future Work

- Further experiments on fluent and cross-signer recognition are needed.
- Improved recognition should exploit clues from multiple frames e.g. "early" formation of the hand shape, without relying on co-articulation which limits scalability.
- Our CVPR paper [2] exploits a robust hand tracker and HOG descriptors, giving good results on fluent fingerspelling in TV footage.

References

[1] P. Goh, E.-J. Holden. Dynamic fingerspelling recognition using geometric and motion features. In *ICIP 2006*

[2] P. Buehler, M. Everingham, A. Zisserman. Learning sign language by watching TV (using weakly aligned subtitles) In *CVPR 2009*

Funding

Mark Everingham is supported by an RCUK Academic Research Fellowship.