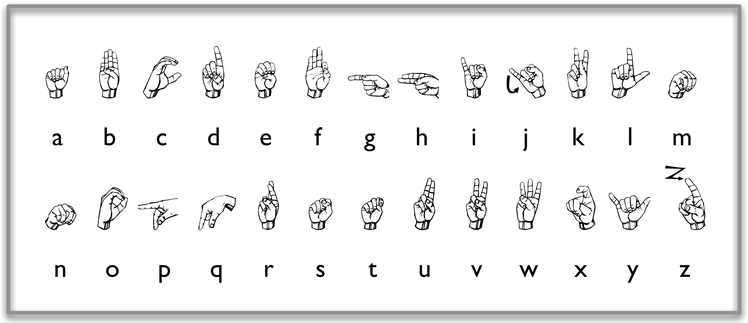
Fingerspelling Recognition through Classification of Letter-to-Letter Transitions

Susanna Ricco and Carlo Tomasi | Department of Computer Science, Duke University

Problem Statement

In order to communicate uncommon words in American Sign Language, signers use *fingerspelling*, spelling desired words with gestures corresponding to single letters in the English alphabet. Most of these gestures are completely defined by the positions of the fingers of the signer's dominant hand at a single moment.

Existing fingerspelling recognition systems rely on an initial temporal segmentation step to identify isolated candidate frames on which to attempt recognition. Sadly, locating the correct static images within a fingerspelling sequence performed by a proficient signer is non-trivial. Our goal is to develop an approach to fingerspelling recognition that does not require explicit temporal segmentation, giving it the possibility of scaling to conversational speed, native fingerspelling.



The 26 gestures forming the ASL Manual Alphabet.

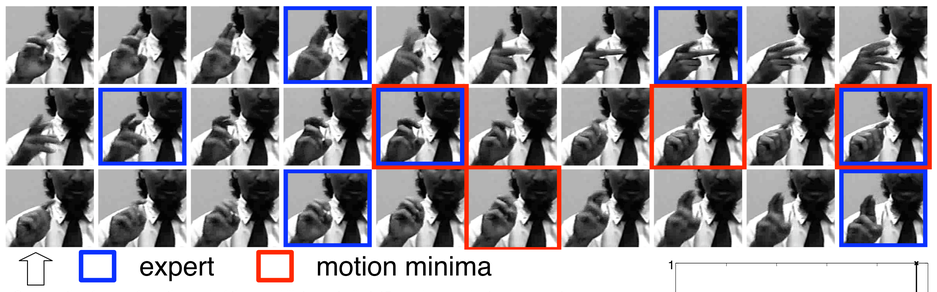
Our Approach

We abandon the misleading notion that fingerspelling consists of a separable sequence of static hand poses. Instead, we consider a *transition between a pair of letters* to be the basic elements of a fingerspelling sequence. Each letter-to-letter transition is a dynamic gesture.

Benefits

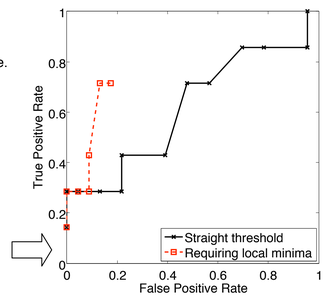
- Eliminates need for explicit temporal segmentation
- Incorporates information from traditionally discarded frames to reduce ambiguities
- Eliminates need for special case for J and Z
- Unifies fingerspelling recognition with word-level sign recognition

The Case For Transitions



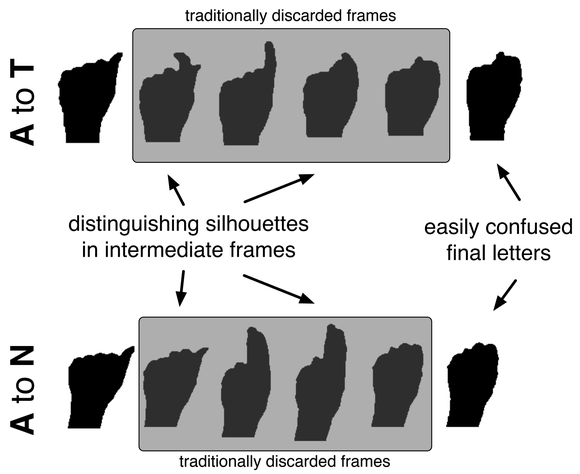
Consecutive frames from the end of the word "interpreter" at 45 words per minute. An expert segmentation only matches the extracted motion minima for two of the seven letters in the sequence.

Systems identify frames to recognize by assuming that letters occur at motion minima. In reality, fluent signers blend letters together through an entire word. As a result, existing techniques miss frames containing signed letters and select frames that correspond to changes in the direction of motion instead of actual letters.



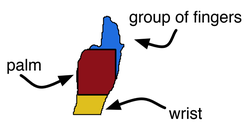
Identification of letter frames based on the amount of motion in the sequence compared to expert segmentation. Only approximately 75% of the frames can be correctly identified by looking for local minima of motion, even with the most generous threshold possible.

Fluent signers do not pause at each letter.

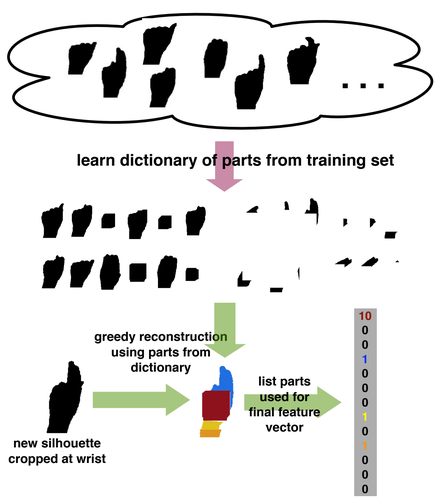


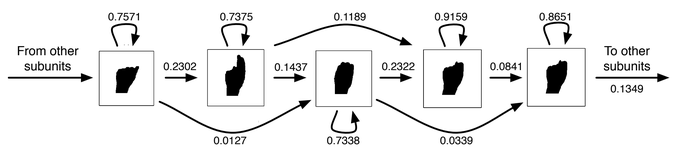
Ignoring frames during transitional motions discards useful information.

Modeling Transitions

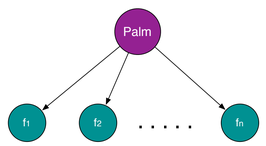


Hand silhouettes decompose into parts.



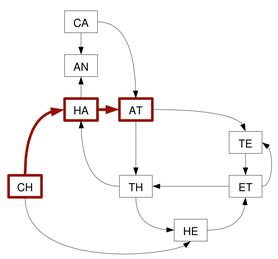


Each letter-to-letter transition is modeled with a five-state Hidden Markov Model.



$$P(palm, f_1, \ldots, f_n) = P(palm) \prod_{i=1}^{n} P(f_i|palm)$$

Each state defines distribution over possible observations.



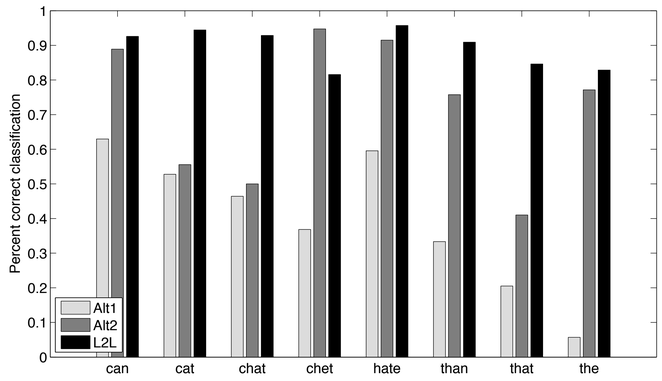
Individual letter-pair HMMs chain together to form final HMM. Fingerspelled word recognized using Viterbi algorithm.

(see Rabiner, 1989)

Test Dataset

- Over 17 minutes of video (> 30,000 frames)
- Divided equally into separate training and testing sets
- Single signer signs instances of eight words:
  can, cat, chat, chet, hate, than, that, the
- Vocabulary includes n and t, letters easily confused in isolation
- No artificial pauses at each letter
- Controlled background and lighting

Recognition Results

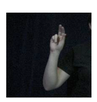


Alt1 - static letters, no grammar
Alt2 - static letters, with grammar
L2L - proposed algorithm (letter pairs and implicit grammar)

Recognizing letter-to-letter transitions improves performance, particularly when individual letters are ambiguous. No explicit temporal segmentation is required.

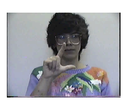
Remaining Challenges

Current Status



- 25 wpm
- Restricted vocabulary
- Reliable skin segmentation possible because of controlled background

Goal System



- 45 wpm
- Full alphabet
- No restrictions on background, robust to errors in skin detection