Concept Embedding through Canonical Forms: A Case Study on Zero-Shot ASL Recognition

A. Kamzin, V. N. S. A. Amperayani ,P. Sukhapalli, A. Banerjee and S. K.S. Gupta, IMPACT Lab CIDSE, Arizona State University, Tempe, AZ, USA
Email: akamzin, vamperay, psukhapa, abanerj3, sandeep.gupta@asu.edu
IMPACT Lab: impact.asu.edu

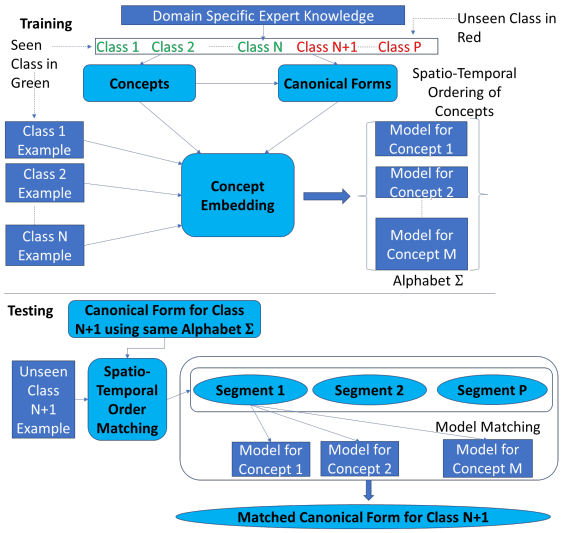
Motivation

- Canonical forms of classes are used to recognize unseen gestures
- Gesture understanding requires a language model
- Gesture-based searching and mining
- Automated Transcription of gestures
- To Recognize unseen gestures without access during training

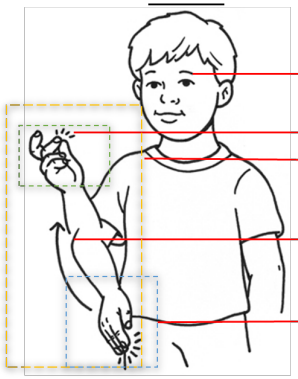
Solution Approach

- A gesture parser splits a gesture video into concepts following a grammar
- Utilize transfer learning models for each concept

Concept Embedding for zero-shot recognition



ASL



- Canonical form of gesture representation

$$
\begin{aligned}
Hand &\rightarrow \Sigma_H \quad \text{Handshape Alphabet}\\
Mov &\rightarrow \Sigma_M \quad \text{Movement Alphabet}\\
Loc &\rightarrow \Sigma_L \quad \text{Location Alphabet}\\
GE &\rightarrow GE_{Left}GE_{Right}\\
GE_X &\rightarrow Hand|\epsilon, \text{ where } X \in \{Right, Left\}\\
GE_X &\rightarrow Hand\ Loc\\
GE &\rightarrow Hand\ Loc\ Mov\ Hand\ Loc
\end{aligned}
$$

Temporal Sequencing

Evaluation Dataset

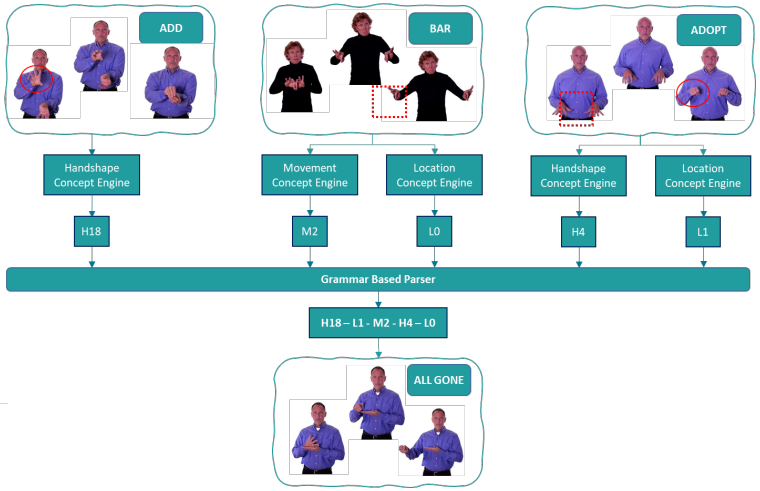
IMPACT Lab dataset:

- Using Learn2Sign mobile application
- 23 gestures from 130 learners with 3 repetitions

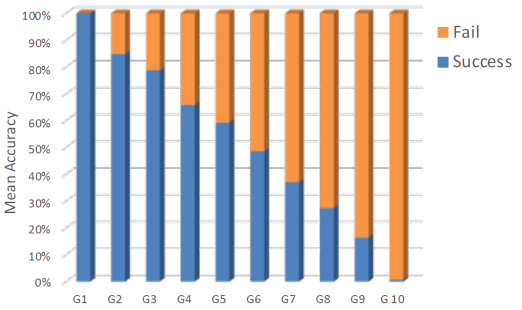
ASLTEXT dataset:

- Subset of ASL Lexicon Video Dataset from Boston University
- 250 unique gestures. 1598 videos out of which we utilize 1200 videos of 190 gestures not in the IMPACT dataset.

Application of Concept Embedding



Results



- Overall normalized accuracy of 66% out of 1200 videos
- Closest state of the art using 3D-CNN reports 51.4%
- While utilizing part of ASLTEXT as training set

Defined canonical form representation of gestures to Zero-Shot Learning and Developed an ensemble system that recognize novel unseen gestures