Automatic Recognition of American Sign Language from Signing Videos

Elahe Vahdani[1], Longlong Jing[1], Yingli Tian[1,2] ,Matt Huenerfauth[3],
[1]Ph.D. Program in Computer Science, The Graduate Center, CUNY
[2]Department of Electrical Engineering, City College, CUNY
[3]The Rochester Institute of Technology

Funded by National Science Foundation

Background

American Sign Language (ASL) has become the 4th most studied language at U.S. colleges. ASL is a natural language conveyed through movements and poses of the hands, body, head, and face.

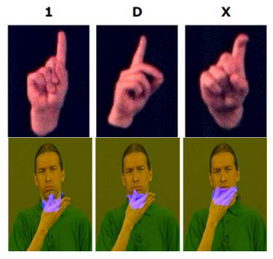
Example images of lexical facial expressions along with hand gestures for signs: **NEVER**, **WHO**, and **WHAT**.

Examples of ASL grammatical elements from multimodalities including facial expressions, head movements, and hand gestures: WH-Question (**WHQ**), Yes-No-Question (**YNQ**), Negative (**NEG**), Conditional (**COND**), **Topic**, and Rhetorical Question (**RHQ**).

Challenges

1) **Visual Complexity**: Slight difference in one hand's phonemes makes another sign.

2) **Occlusion**: Hand/hand or Hand/face occlusion. The occluded pixel are shown in blue.



Continuous Sign Language Recognition Challenges:

3) No temporal boundaries of signs are provided.

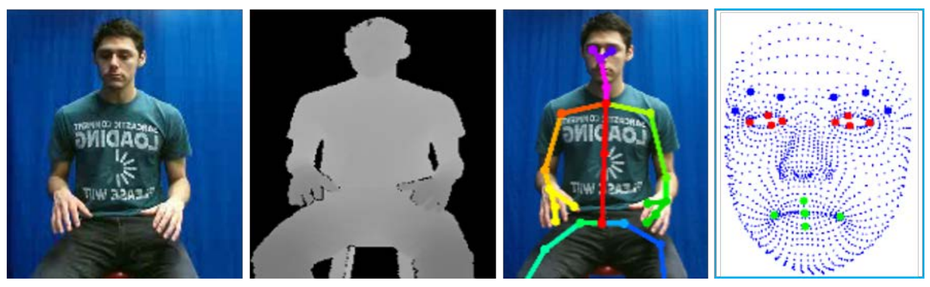
4) Transition movements between two consecutive signs are subtle, diverse and hard to detect.

Our Contributions

- **Isolated Sign Language:** proposed a framework by using 3DCNNS for ASL recognition in RGB-D videos by fusing multi-channels including RGB, depth, motion, and skeleton joints. Also, created a new ASL dataset, **ASL-100-RGBD,** including multiple modalities (facial movements, hand gestures, and body pose) and multiple channels (RGB, depth, skeleton joints, and HDface).

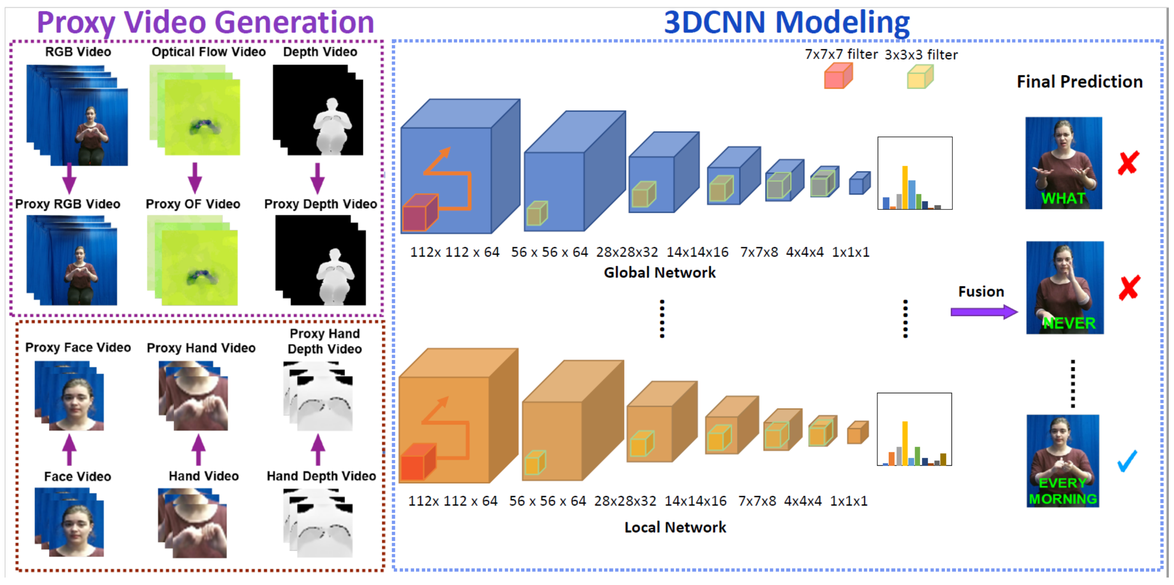
- **Continuous Sign Language:** proposed the first framework for automatic detection of ASL grammatical mistakes in continuous signing videos, based on a 3D multimodal network to recognize grammatically important manual signs and nonmanual signals from continuous signing videos. In addition, collected a continuous sign language dataset named as **ASL-HW-RGBD**, consisting of 1,026 continuous videos of signing several sentences by fluent and student ASL signers. Our dataset covers different aspects of ASL grammar.

- **System Design:** designed an educational system to provide ASL students immediate feedback whether their signing is fluent.

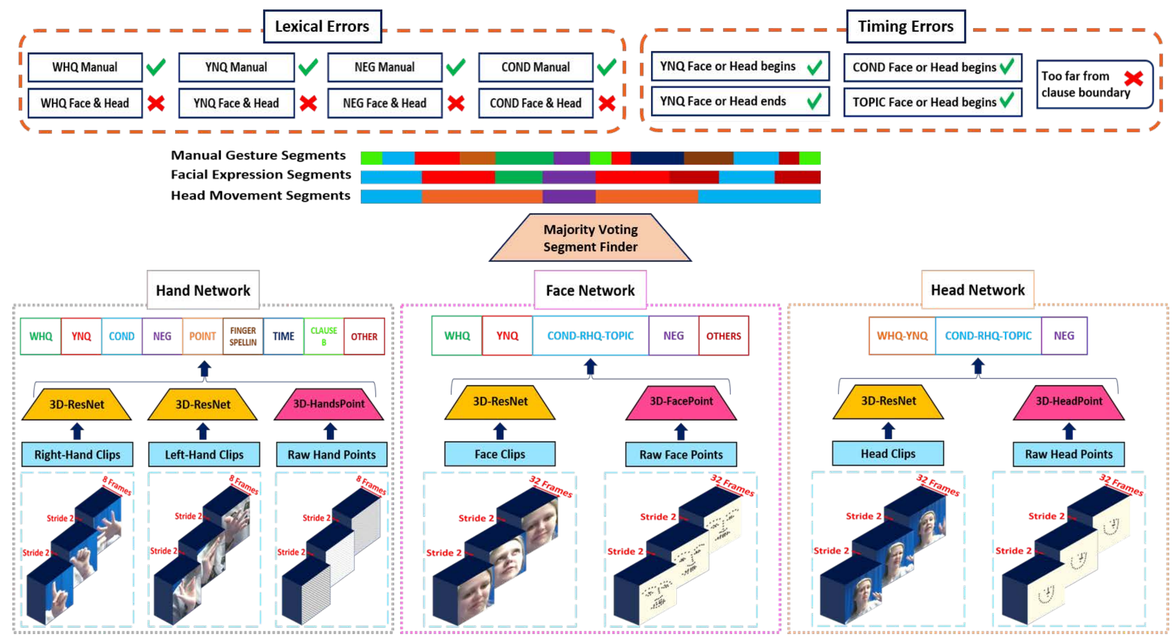


RGB **Depth** **Skeleton** **HD-Face**

Our Models



The pipeline of multi-channel multi-modal 3DCNN framework for ASL Isolated Sign Recognition. Proxy videos are generated to represent the overall temporal dynamics, and are fed into the multi-stream 3DCNN component. The predictions of these networks are weighted to obtain the final results.



The pipeline of the proposed multimodal 3DCNN framework for ASL grammar recognition. Raw coordinates and RGB clips of face and hands are fed to the networks to extract the pose and spatiotemporal information. Temporal boundaries of manual gestures and nonmanual signals are estimated using temporal sliding windows. The final predictions are determined via Majority Voting algorithm and segments of all classes are extracted. Temporal correspondences of modalities are compared to detect lexical and timing errors.

Experimental Results

Channels	Fusions			
RGB	√	√		√
Depth	√	√	√	√
RGBflow	√	√	√	
RGB of Hands	√	√	√	√
RGB of Face		√	√	√
Performance	91.19%	92.48%	92.48%	92.88%

Performance of our framework for recognizing Isolated Signs on ASL-100-RGBD dataset.

Error Type	Ground Truth	Recognized	TP Rate
Error-YNQ-Lexical	12	7	58.3%
Error-NEG-Lexical	36	14	38.8%
Error-WHQ-Lexical	40	23	57.5%
Error-COND-Lexical	24	18	75.0%
Error-TOPIC-Beginning	23	19	82.6%
Total	135	81	60.0%

ASL nonmanual grammatical error recognition results on ASL-HW-RGBD dataset.

The example RGB images and their corresponding attention maps from the fifth convolution layer of the 3DResNet-34 on ASL-100-RGBD test set. This confirms the attention is focused on hands and face.

Error Type	Manual Sign	Nonmanual Signal
WHQ-Lexical	WHQ word	Not (WHQ or RHQ)
YNQ-Lexical	YNQ word	Not YNQ
NEG-Lexical	Negative word	Not Negative
COND-Lexical	Conditional word	Not Conditional
YNQ-Beginning	Clause Boundary	YNQ begins
YNQ-End	Clause Boundary	YNQ ends
COND-Beginning	Clause Boundary	Conditional begins
TOPIC-Beginning	Clause Boundary	Topic begins

ASL grammar errors recognized by our system on ASL-HW-RGBD dataset.

Summary

We collected two ASL datasets for Isolated Signs as well as Continuous Sign Language Recognition. We developed models for recognition of ASL words and finding the grammatical mistakes in signing videos. Our system provides feedback for ASL students whether their signing is fluent or not.

Funded by National Science Foundation under award numbers 1400802, 1400810, and 1462280.