Sign Language Recognition Using the Kinect

Abraham Glasser Dr. Raja Kushalnagar
REU Accessible Multimodal Interfaces Program Site, Rochester Institute of Technology

Background and Objective

- More and more devices have voice command compatibility and this is not fully accessible by Deaf and Hard of Hearing individuals
- Sign Recognition technology is not pursued actively, many teams have worked on this but have not been continuing their work

We aim to set up a cheap and effective fundamental for sign language recognition. We hope that our model will be easy to adapt and easy to build upon in the future as technology improves. We use the Microsoft XBOX Kinect v2 as only one is needed to obtain 3D data. We analyze different approaches to using this data for Sign Language Recognition.

Methodology

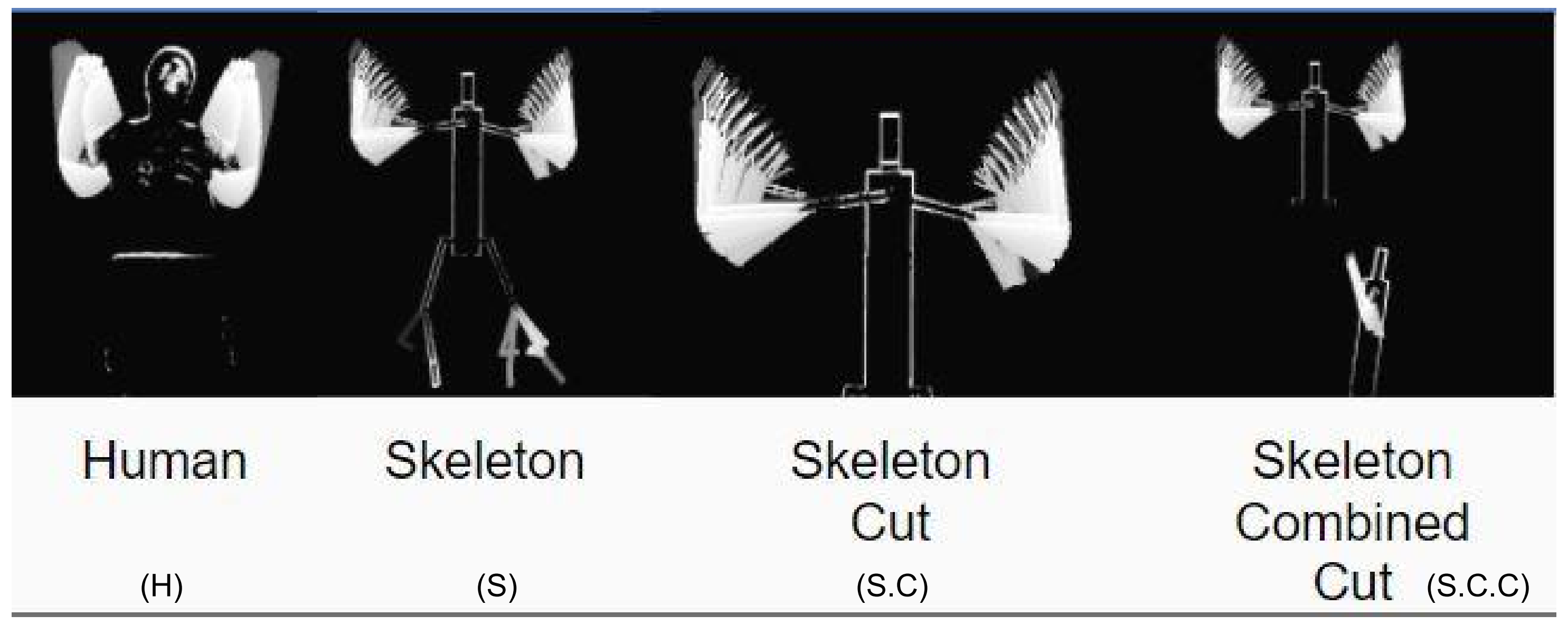
Technology Used

- Microsoft XBOX Kinect v2
- Microsoft Custom Vision AI

Procedure

- Record signers using the Kinect
- Obtain frame by frame "Human" and "Skeleton" pictures
- Create Motion History Images (MHIs)
- Use the Microsoft Custom Vision AI to run machine learning on database

Figure 1: Different Motion History Images (MHIs) signing "workout"



Evaluation

The Confusable and Separable groups had words that were estimated to be hard to tell apart and vice versa, respectively. Four different visualizations of data from the Kinect were used to make different MHIs as shown in Figure 1. Figure 3 shows an example of a "confusable" pair and a "separable" pair.

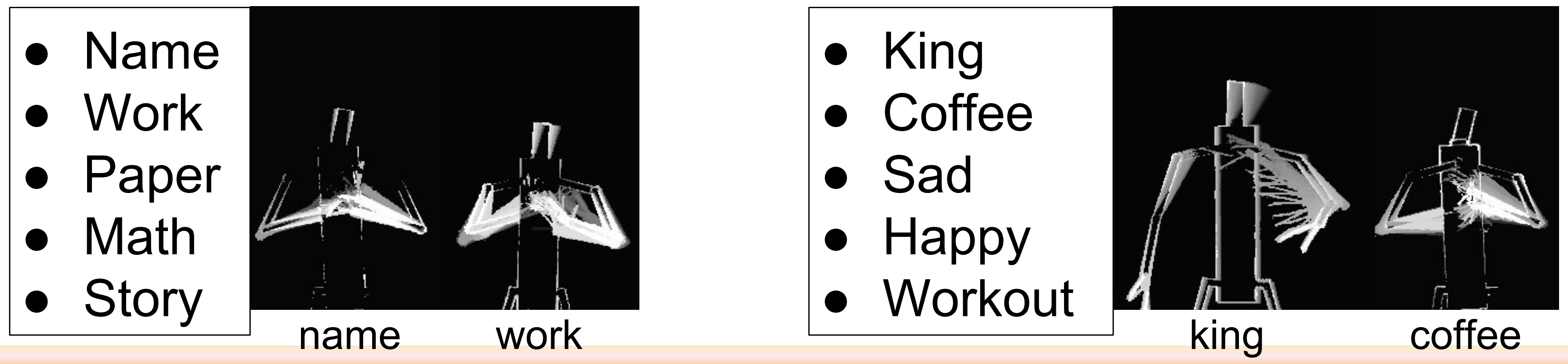
- Each group has 5 words
  - Each word was signed 5 times by 5 different people
- Each group was given to the Microsoft Custom Vision AI for analysis.
  - Images are uploaded and labelled to create training datasets. Six groups of five words were used. Each of the datasets are independent from each other

The Microsoft Custom Vision AI can be tested externally, but for reliability, we used the self-evaluation capabilities. It gives "precision (p) and recall (r) " values as a measure of its accuracy for that dataset. These values were converted into a F-score for better comparison.

$$F = 2((r*p)/(r+p))$$

The data obtained from the kinect had a little bit of noise when it was uncertain of the location of the joints. This happens when the body parts are too close to each other or if there was not a lot of motion.

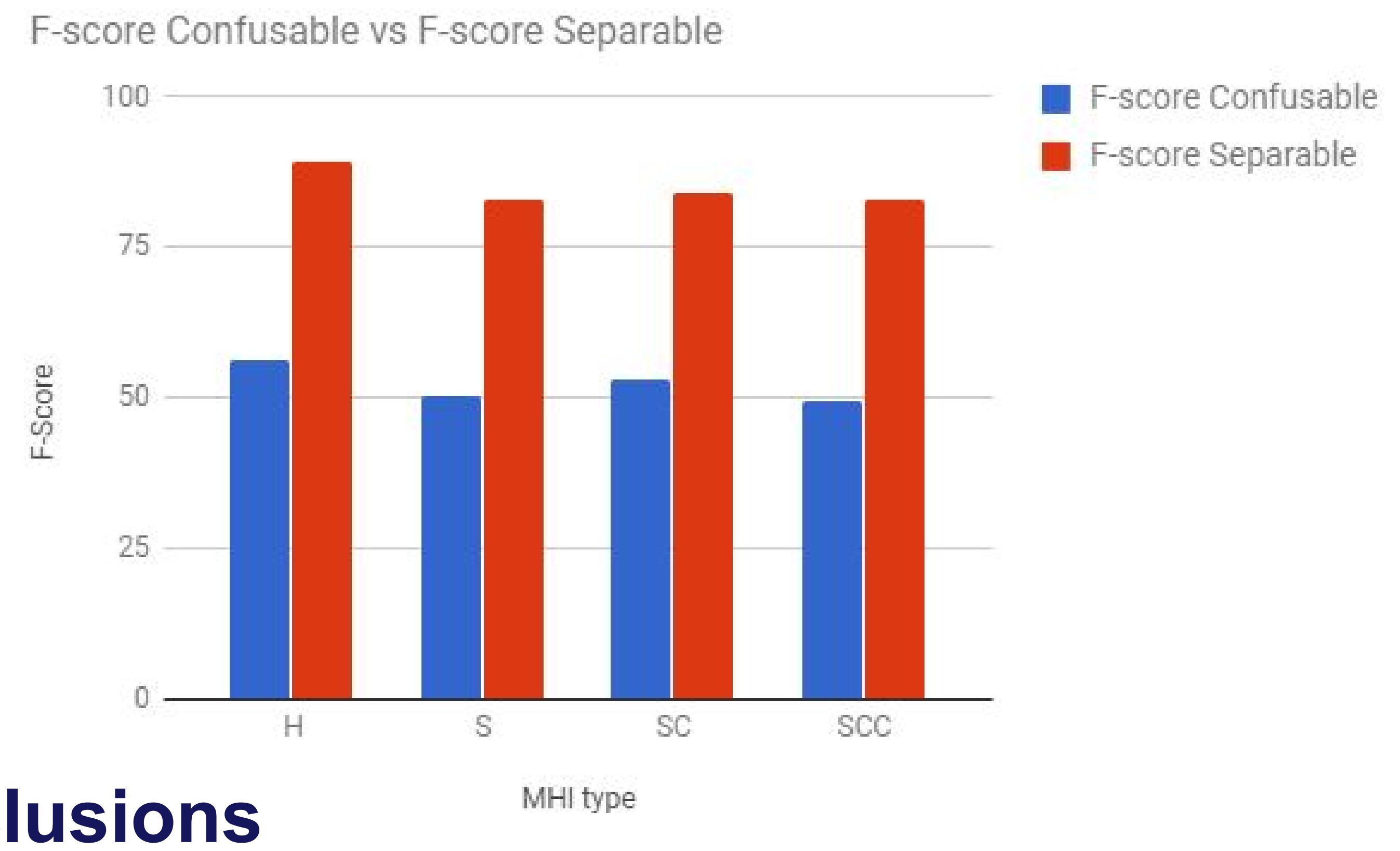
Figure 3: Example of 1 out of 3 groups of Confusable (left) vs Separable (right) signs



- Name
- Work
- Paper
- Math
- Story

- King
- Coffee
- Sad
- Happy
- Workout

Figure 2: Negative Precision and Recall comparison for the Confusable and Separable groups



Conclusions

- Separable groups had a significantly better result than the Confusable groups.
  - Shows that the Kinect currently performs better for signs that have a larger range of motion, different locations, and are distinguishable.
- Human and Skeleton groups performed similarly.
- Our data collection was not in optimal settings and we had some noisy data which caused the 3D data to perform worse than expected. However, it was only behind the Human data by a few percent.

MHI is good for "separable" signs and that more needs to be done to use it for "confusable" signs. Possibilities include better Kinect setup and more hand information from the kinect.

References

1. Vogler, Christian, and Dimitris Metaxas. *Adapting Hidden Markov Models for ASL Recognition by Using Three-Dimensional Computer Vision Methods.*
2. Starner, Thad, and Alex Pentrand. Visual Recognition of American Sign Language Using Hidden Markov Models.
3. https://www.microsoft.com/en-us/research/wp-content/uploads/2012/08/Xilin_Chen.pdf
4. https://www.customvision.ai/
5. http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=6612186

Funding

This work has been generously supported by an NSF REU Site Grant (#1460894) awarded to Dr. Raja Kushalnagar, PI

Contact (Abraham Glasser; atg2036@rit.edu,
Dr. Raja Kushalnagar: raja.kushalnagar@gallaudet.edu)