CopyCat: Using Sign Language Recognition to Help Deaf Children Acquire Language Skills

Conference Paper · May 2021

DOI: 10.1145/3411763.3451523

Prerna Ravi
Georgia Institute of Technology
3 PUBLICATIONS **22** CITATIONS

Pranay Agrawal
Georgia Institute of Technology
1 PUBLICATION **8** CITATIONS

CopyCat: Using Sign Language Recognition to Help Deaf Children Acquire Language Skills

Dhruva Bansal
Georgia Institute of Technology
Atlanta, United States
dbansal36@gatech.edu

Prerna Ravi
Georgia Institute of Technology
Atlanta, United States
prernaravi@gatech.edu

Matthew So
Georgia Institute of Technology
Atlanta, United States
matthew.so@gatech.edu

Pranay Agrawal
Georgia Institute of Technology
Atlanta, United States
pranay.agra@gatech.edu

Ishan Chadha
Georgia Institute of Technology
Atlanta, United States
ichadha3@gatech.edu

Ganesh Murugappan
Georgia Institute of Technology
Atlanta, United States
ganeshm@gatech.edu

Colby Duke
Georgia Institute of Technology
Atlanta, United States
cduke@gatech.edu

ABSTRACT

Deaf children born to hearing parents lack continuous access to language, leading to weaker working memory compared to hearing children and deaf children born to Deaf parents. CopyCat is a game where children communicate with the computer via American Sign Language (ASL), and it has been shown to improve language skills and working memory. Previously, CopyCat depended on unscalable hardware such as custom gloves for sign verification, but modern 4K cameras and pose estimators present new opportunities. Before re-creating the CopyCat game for deaf children using off-the-shelf hardware, we evaluate whether current ASL recognition is sufficient. Using Hidden Markov Models (HMMs), user independent word accuracies were 90.6%, 90.5%, and 90.4% for AlphaPose, Kinect, and MediaPipe, respectively. Transformers, a state-of-the-art model in natural language processing, performed 17.0% worse on average. Given these results, we believe our current HMM-based recognizer can be successfully adapted to verify children's signing while playing CopyCat.

CCS CONCEPTS

• Human-centered computing → Ubiquitous and mobile computing systems and tools; **Accessibility technologies**; **• Applied computing → Computer-managed instruction**; **• Computing methodologies → Machine learning**; **Feature selection**.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
CHI '21 Extended Abstracts, May 8–13, 2021, Yokohama, Japan
© 2021 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-8095-9/21/05.
https://doi.org/10.1145/3411763.3451523

KEYWORDS

Pose Estimation, Hand Tracking, American Sign Language, Sign Language Recognition, Hidden Markov Models, Education, Deaf, Interactive Learning System

ACM Reference Format:
Dhruva Bansal, Prerna Ravi, Matthew So, Pranay Agrawal, Ishan Chadha, Ganesh Murugappan, and Colby Duke. 2021. CopyCat: Using Sign Language Recognition to Help Deaf Children Acquire Language Skills. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts (CHI '21 Extended Abstracts), May 8–13, 2021, Yokohama, Japan.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3411763.3451523

1 INTRODUCTION

Over 95% of deaf children are born to hearing parents who do not know American Sign Language (ASL) or have very low levels of proficiency [23]. Unlike hearing children of speaking parents or deaf children of signing parents, these children often lack adequate access to language at home, which is necessary for developing linguistic skills. Many of these deaf children only receive limited exposure to ASL from signing at school or existing ASL games that focus on language comprehension rather than phrase generation [1, 2, 31]. If left untreated, a lack of exposure to any language, including sign language, can lead to significantly weaker working memory [21]. These effects are especially consequential during the critical period of language development, which is between the ages of two and five [1, 24]. As a result, many pre-lingually deafened children can only repeat one or two signs in a row compared to four to six signs for children with deaf parents. This deficiency of short-term memory and other language skills can lead to Language Deprivation Syndrome (LDS), a condition with poor lifelong outcomes, including a 2-7x increase in mental health problems [3], 50% unemployment rate [25], higher rates of physical and sexual abuse [26], and a 3-30x increase in suicide rates [27].

Children who express themselves in ASL can significantly increase their short term memory [28], reducing the risk of LDS. We are developing CopyCat, an interactive and entertaining game that

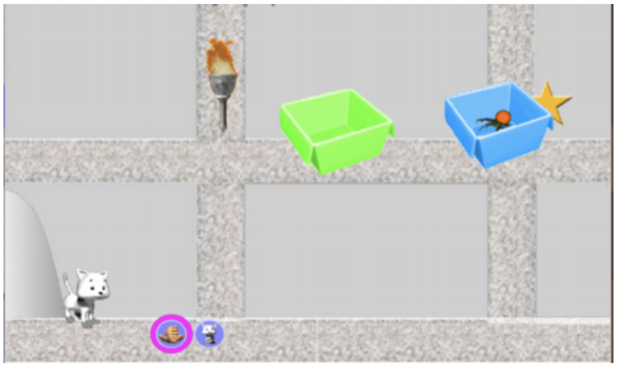


Figure 1: CopyCat game screen. Children tell the hero Iris where the monster is hiding (e.g., SPIDER IN BLUE BOX)

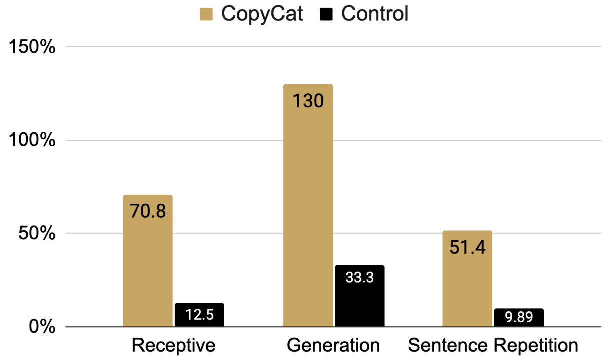


Figure 2: Six students playing CopyCat improved language skills much more quickly than six students attending standard classroom instruction

displays potential for helping deaf children with short term language memory acquisition. The game presents the child with a scene and prompts them to describe it to the mascot of the game using ASL (Figure 1). The game then uses an ASL recognizer to decide if the description is correct and, if so, allows the child to advance to the next scene. As a result, the game's effectiveness depends heavily on the accuracy with which it recognizes sentences as correct or incorrect. To test the effects of the game, previous studies purposefully tested vocabulary not in CopyCat to help differentiate general memory skill acquisition versus learning the game. They found that the game significantly increases 2 a child's score on three crucial measures of working memory—language generation, reception, and sentence repetition. We hope that adapting CopyCat to off-the-shelf hardware will provide a valuable tool for educators to help improve children's language skills.

Until now, CopyCat has relied on expensive and unscalable hardware such as custom gloves, accelerometers, and kiosks to achieve sign language recognition accuracy sufficient enough to verify the children's signing. As a result of these high costs, the low resolution of previous cameras, and lack of pose estimators, CopyCat was inaccessible to most. However, with the advent of pose estimation systems such as AlphaPose, Azure Kinect, and Google MediaPipe, along with improvements in camera resolution through devices such as the Azure Kinect and Android tablets, the goal of deploying CopyCat on off-the-shelf systems has now become viable.

1.1 Contributions

To this end, we make several contributions which bring us closer to deploying CopyCat publicly:

1. We are releasing an ASL recognition toolkit [1] incorporating Hidden Markov Model (HMM) and Transformer recognition systems. We have compared HMM performance with Transformers—current state-of-the-art technology in sign language translation—and find that HMMs outperform Transformers by over 13.1% on word accuracy.
2. We compare three pose estimators—AlphaPose, Azure Kinect, and Google MediaPipe—for ASL recognition performance (see Figure 3). We found that all pose estimators achieve similar results, suggesting CopyCat may be scalable to off-the-shelf cameras.
3. We are releasing a dataset [2] of features derived from 3914 videos collected from eight adults using the Azure Kinect. This dataset was used for evaluation throughout this paper.

2 PAST WORK

Recently, there has been significant interest in the recognition and translation of sign language using RGB cameras alone [5, 6, 8, 13, 14]. Yin et al. [30] showed state-of-the-art translation of German Sign Language. By jointly training a spatial-temporal multi-cue (STMC) network to generate glosses from videos and a transformer to generate translations from glosses, they obtained a Word Error Rate (WER) of 21.0% on the RWTH-PHOENIX-Weather-2014T (PHOENIX14T) dataset, which consists of low resolution videos of signed weather reports. Grobel et al. [12] demonstrated, at the time, state-of-the-art recognition of isolated signs using HMMs. By using multi-colored gloves to distinguish the parts of the hand, they extracted the location of the fingers, palm, and back of the hand. Using HMMs, they obtained a user-dependent accuracy rate of up to 91.1% among 262 isolated signs. However, user-independent accuracy was 47.6% and 56.2% for the two users in the dataset. Koller et al. [18] demonstrated the use of expectation maximization by integrating convolutional neural networks (CNNs) with HMMs to create a robust pose-independent hand shape classifier. This model, when trained on a 1-Million-Hands dataset, reduced error rates in a continuous sign language recognition system from 57.3% to 47.1% on the RWTH-PHOENIX-Weather dataset, demonstrating the generalizability of the model. Camgöz et al. [4] also demonstrated translation of German Sign Language. Using a novel transformer-based architecture that leverages Connectionist Temporal Classification (CTC) loss, they demonstrated translation at a WER of 24.5% also on the RWTH-PHOENIX-Weather-2014T (PHOENIX14T) dataset. Finally, Ko et al. [17] showed Sign Language Translation on Korean Sign Language by combining Openpose (for human pose estimation) with an encoder-decoder GRU model. They achieved a sentence level accuracy of 55.3% on their dataset.

[1] https://github.com/Accessible-Technology-in-Sign/ASLRT
[2] https://github.com/Accessible-Technology-in-Sign/copycat-data-chi

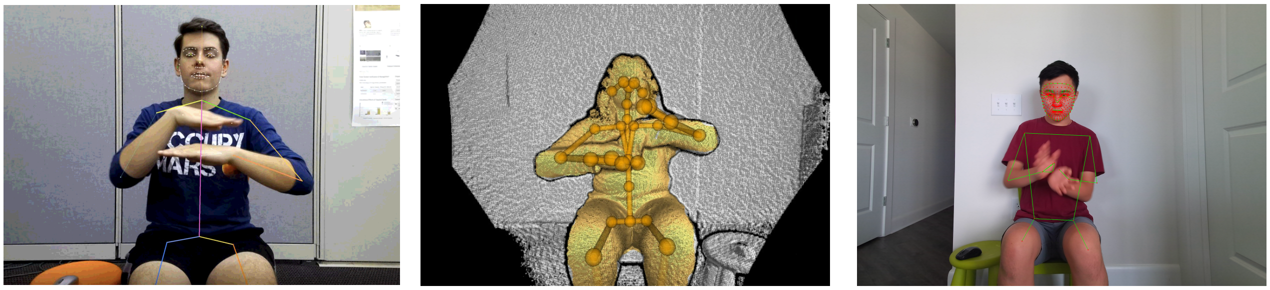


Figure 3: Pose estimation with AlphaPose (left, sign "above"), Azure Kinect (middle, sign "in"), and MediaPipe (right, sign "alligator") showing difficult signs

Weaver et al. [28] introduced the first iteration of CopyCat. This iteration of the game entailed a quest by the main character to collect items to remedy a problem. Children instruct the main character via sign language, which was verified by an automatic computer sign language recognizer. To quantify the impact of the game, they conducted a study at a local school of the deaf and reported significant improvements in language reception, generation, and sentence repetition abilities—critical measures of working memory. The study was composed of 12 participants, aged between 6 and 11. They asked participants at both the beginning and end of the study to configure plastic toys based on signed instruction (reception), express an event depicted in stop-motion animation (generation), and repeat a signed phrase (sentence repetition). The students who played CopyCat improved their scores on reception, generation, and sentence repetition measures by 70.8%, 130.0%, and 51.4% as compared to 12.5%, 33.3%, and 9.9%, respectively, for learners who followed the normal classroom curriculum. (Figure 2). Our iteration of the game improves upon the story line and design elements while also retaining interactions and methodology.

Brashear et al. [2] initiated efforts towards building a sign language recognition for CopyCat. They released two datasets—vision and accelerometer—and demonstrated how HMMs can be used to recognize American Sign Language. Their vision-based dataset consisted of videos from a camera mounted on a cap, giving them a direct view of the signing. On the other hand, their accelerometer-based dataset consisted of time series data from two accelerometers embedded into wearable gloves. Their models achieved 52.4% and 65.9% user independent sentence accuracy on the vision and accelerometer dataset respectively. Unfortunately, the system's dependence on accelerometer data from expensive custom gloves (Figure 4) made it infeasible for widespread deployment.

To move towards cheaper and ubiquitous equipment, Zafrulla et al. collected a dataset with over one thousand ASL phrases collected using Kinect and demonstrated recognition using HMMs [31]. They also compared recognition on data collected while standing versus sitting. Building upon the work done by Brashear et al., they used HMMs along with features including the shape, velocity, and acceleration of the user's hands. Their models achieved 36.2% and 36.3% user independent sentence accuracy on data collected while standing and sitting respectively.

3 METHODOLOGY

In this iteration of the CopyCat game, we have revamped game design elements, story line, calibration system, data collection, feature extraction, and sign language recognition. In this section, we detail these changes and also elaborate upon our machine learning pipeline for our 3 different human pose recognition models.

3.1 CopyCat Improvements

Calibration of the distance between the signer and the camera is integral to CopyCat as it can greatly reduce variability between users. The original calibration system standardizes the distance between the signer and the camera by using the front view of the chair to guide the positioning of the CopyCat chair (Figure 5a). We improved upon this calibration system by adding a side view of the chair and re-designing the user experience (Figure 5b).

Addition of side view: We performed two user studies to quantify the impact of adding a side view. The first user study compared the ease of use and efficiency of calibration using only a front view with using both views. Nine adults participated in our study, five of whom have no experience with CopyCat's calibration system. On average, the two-view calibration system decreased the time required to calibrate by 63% and decreased the NASA-TLX score by 27%, showing that the two-view system is faster and easier to use. The second user study determined the accuracy of two-view chair calibration and found that using both views created a highly accurate calibration with a low average difference from desired to actual position of 2.4/16 inches. Since using both views created a faster, easier-to-use, and accurate calibration system, we decided to integrate both views into the CopyCat calibration system.

Re-designing user experience: We re-designed the user interface by leveraging depth features from Kinect Depth Camera SDK's K4AViewer Tool, as depicted in Figure 5b. Specifically, we used the x, y, and z coordinates provided by the Azure Kinect SDK to display an outline of the chair (Figure 5c). Using these coordinates, we incorporated the outline of the chair into the calibration system to determine when the chair is correctly calibrated (Figure 5b) [10, 15]. These outlines turn green if the values of each of the depth coordinates of points A, B, and C (Figure 5c) are located within a set range, signaling the success of the task to the user.

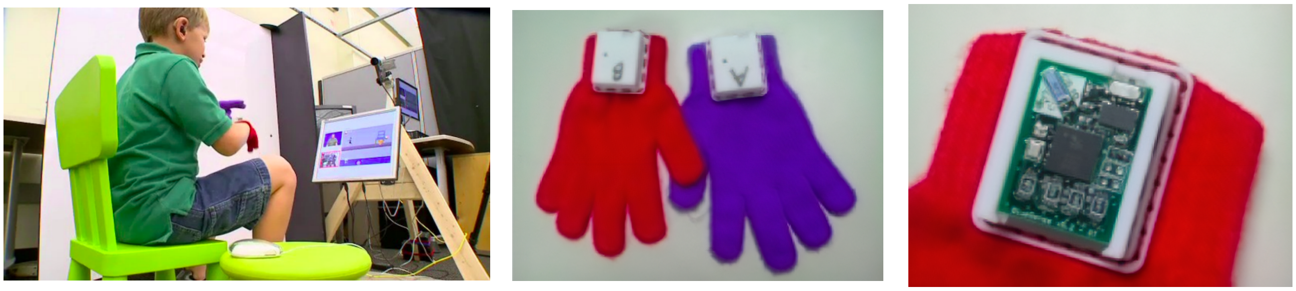


Figure 4: Originally, CopyCat used specially-designed kiosk (left) and custom sensor gloves (center) with embedded accelerometers (right) to achieve sufficient recognition accuracy for gameplay.

3.2 Data Collection

The 4K RGBD dataset used in this study was collected using the Azure Kinect from 8 adults at Georgia Tech with little experience with ASL. Adult novice signers were chosen for multiple reasons. Firstly, the target population of CopyCat—deaf children born to hearing parents—are disfluent and slow, much like novice signers. Additionally, adult novice signers are widely available since anyone familiar with the phrases can record. In comparison, we have a limited pool of deaf children since they are only available through local institutions. We plan on using them to test the effectiveness of the final game. In total, 3914 videos were recorded. Each user recorded up to ten sets of a standardized set of 58 phrases. Each phrase consisted of three to five words each and followed the structure (A)NP(A)N, where A represents adjectives, N represents nouns, and P represents prepositions. All recording sessions were supervised by a second member with full familiarity with the phrase set. The files generated were processed using our pose estimation systems to determine the coordinates of various key points on the user's hands, upper body, and face.

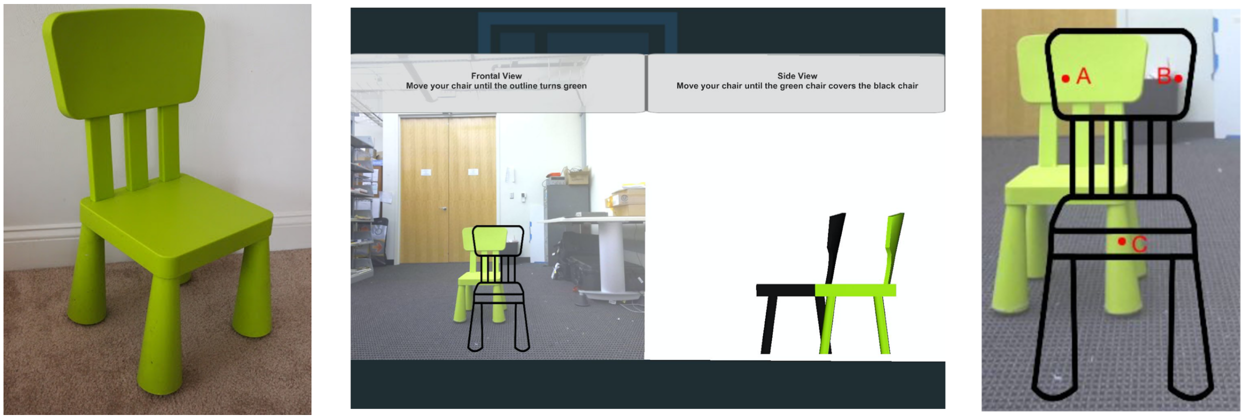
3.3 Feature Extraction

The v1.3.0 Azure Kinect SDK leverages 4K RGB videos and depth data to extract the absolute 3D location of over 32 joints [7], resulting in a total of 96 features. Unlike Kinect, both the Halpe Full-Body 136 AlphaPose model and MediaPipe v0.8.2 extract the corresponding 2D points using only the 4K RGB videos [9, 19, 20, 29]. When visualizing the features with Kinect, we noticed that hand landmarks with signs "in" and "above" were difficult to distinguish between. Since both signs are performed with hands close together near the chest level, the pose estimator was inconsistent when it came to extracting hands that were occluded (see Figure 3). In some videos, Kinect did not extract hand features for the occluded hand, and other times would extract the wrong location of the occluded hand. Furthermore, the Kinect pose estimator does not respond rapidly to fast movements, with delays between the video and extracted features present. In the case of AlphaPose, we noticed that when extracting features for signs like "alligator" and "above" where one or both hands are on the horizontal plane, not all hand features (finger joints for example) get identified (see Figure 3). For signs like "in" where one hand is directly behind the other with respect to the camera, AlphaPose does not extract hand features for the occluded hand for a few frames. This combined with the previous problem leads to confusion in differentiating "above" from "in". Overall, MediaPipe provided the worst feature location estimates for signs where the hands must be close together due to occlusion of one of the hands. Some examples of these signs include "alligator", "above", "in", and "below", such as the one shown in Figure 3. The absolute locations from each respective computer vision framework were then used to generate four additional types of features. Firstly, delta features were calculated using the change in location between the current and previous frame. Next, features with respect to the position of the signer's nose were calculated using the absolute distance between the feature and the nose. Additionally, Z-score normalized features were calculated by subtracting the mean location from the current location and then dividing by its standard deviation. Finally, min-max normalized features were calculated by linearly mapping the 5% quartile to -100 and the 95% quartile to +100. Finally, since aggregating these features for training and testing resulted in an excessively long list, we performed feature selection to select the most predictive features to be used as training and testing data to feed into the HMMs and Transformers.

3.4 American Sign Language Recognition

Hidden Markov Models: We hypothesize that HMMs are best suited for ASL recognition in this setting due to their great performance on time series and pattern recognition problems along with low training data requirements when compared to deep learning models such as LSTMs and Transformers. HMMs for each word are trained by first performing a flat initialization and then performing several iterations of the Baum-Welch re-estimation (shown in algorithm 1). Every 20-25 iterations, the number of mixtures in each HMM model is also increased. Finally, Viterbi Decoding (shown in algorithm 2) is used to find the sequence of HMM models which yields the highest observation likelihood. The number of states used is varied between 14 and 20 depending on the pose estimator used. A left-to-right HMM topology with no skip transitions was found to produce the best results based on extensive experimentation.

Transformers: The Transformer model used by Comgez et al. [4] for German Sign Language Translation was adapted to American Sign Language Recognition to compare results with HMMs and establish a baseline. We used PyTorch to implement and Google Colab to train end-to-end transformers which use frameworks above to extract keypoints and then recognize the resulting sequence data.



Frontal View
Move your chair until the outline turns green

Side View
Move your chair until the green chair covers the black chair

(a)

(b)

(c)

Figure 5: CopyCat calibrates a commonly available IKEA chair (left) using a graphical interface (center) to make sure the user is seated in an appropriate position. Points A, B, and C (right) are used determine the location of the chair when calibrating.

Algorithm 1 Baum-Welch Re-estimation

```
$X = Features, Y = Label, T = TotalTimesteps$
Initialize HMM Parameters $\theta = (A, B, \pi)$ randomly
$\alpha(X_0) = P[Y_0, X_0] = P[Y_0|X_0]P[X_0]$
$\beta(X_T) = 1$
while $i \leq iterations$ do
    for $k = 0 \rightarrow T$ do
        $\alpha(X_k) = \sum_{X_{k-1}} \alpha(X_{k-1})P(X_k|X_{k-1})P(Y_k|X_k)$
    end for
    for $k = N \rightarrow 0$ do
        $\beta(X_k) = \sum_{X_{k+1}} \beta(X_{k+1})P(X_{k+1}|X_k)P(Y_{k+1}|X_{k+1})$
    end for
    $\eta(X_k) = \frac{\alpha(X_k)\beta(X_k)}{\sum_{X_k} \alpha(X_k)\beta(X_k)}$
    $\epsilon(X_k, X_{k+1}) = \frac{\alpha(X_k)\beta(X_{k+1})P[X_{k+1}|X_k]P[Y_{k+1}|X_{k+1}]}{\sum_{X_k} \alpha(X_k)\beta(X_{k+1})P[X_{k+1}|X_k]P[Y_{k+1}|X_{k+1}]}$
    $\pi_0^* = \eta(X_0)$
    $A_{ij}^* = \frac{\sum_k \epsilon(X_k=j, X_{k-1}=i)}{\sum_k \eta(X_{k-1}=i)}$
    $B_{ij}^* = \frac{\sum_k \eta(X_k=i)1_{Y_{k=j}}}{\sum_k \eta(X_k=i)}$
end while
```

Rather than using video frames directly, we use these keypoints as the embeddings for the Transformer. Since keypoints can represent all relevant information in a smaller feature space than frames, they allow Transformers to generalize better to our dataset compared to using frames directly. However, we retain positional embeddings since they add temporal information to the dataset. The model consists of two encoder and decoder layers, four multi-head attention layers, and a 2048 dimensional feed forward layer. We initialize all layers using Xavier initialization and train the model using the Adam optimizer [16].

4 CURRENT RESULTS

In Table 1a, we compare user adaptive results from HMMs on all three pose estimation frameworks with results from Transformers.

Algorithm 2 Viterbi Decoding

```
create path matrix $viterbi[N, T]$
for $s = 1 \rightarrow N$ do
    $viterbi[s, 1] = \pi_s * b_s({}_1)$
    $backp[s, 1] = 0$
end for
for $= 2 \rightarrow T$ do
    for $s = 1 \rightarrow N$ do
        $viterbi[s, t] = \max viterbi[s', t-1]a_{s',s}b_s({}_t)$
        $backp[s, t] = \arg\max viterbi[s', t-1]a_{s',s}b_s(o_t)$
    end for
end for
$bestpathprop = \max viterbi[s, T]$
$bestpathpointer = \arg\max viterbi[s, T]$
$bestpath$ = path starting at $bestpathpointer$, follows backp[]
    to states back in time
```

Models	HMMs	Transformers
AlphaPose	98.3 (95.8)	89.1 (82.4)
Kinect	98.1 (94.9)	94.8 (91.4)
MediaPipe	**98.8 (96.8)**	97.4 (95.3)

(a) User adaptive

Models	HMMs	Transformers
AlphaPose	**98.5 (96.1)**	84.8 (74.8)
Kinect	97.2 (91.6)	81.5 (68.9)
MediaPipe	98.1 (94.5)	91.1 (84.1)

(b) User dependent

Table 1: User adaptive and user dependent word (sentence) percent accuracy

User adaptive refers to models initialized with a user independent model and updated with data from a target user [11]. To emulate

Participant	P1	P2	P3	P4	P5	P6	P7	P8	Average
AlphaPose	83.3 (62.9)	96.8 (88.9)	85.4 (61.9)	91.8 (77.8)	93.8 (77.0)	91.3 (69.4)	84.2 (43.4)	98.8 (85.5)	**90.6 (71.9)**
Kinect	86.5 (64.4)	94.8 (82.0)	81.3 (49.7)	92.1 (77.0)	91.2 (71.4)	92.6 (77.3)	92.4 (71.4)	93.0 (79.4)	90.5 (71.6)
MediaPipe	79.6 (67.9)	94.7 (82.4)	80.6 (48.3)	98.0 (93.0)	93.6 (78.0)	94.2 (78.3)	73.9 (29.8)	98.9 (96.6)	90.4 (71.7)

Table 2: User independent word (sentence) percent accuracy using HMMs

Participant	P1	P2	P3	P4	P5	P6	P7	P8	Average
AlphaPose	65.8 (47.0)	69.3 (50.0)	44.7 (28.7)	57.9 (44.0)	79.8 (67.5)	73.7 (56.0)	61.7 (41.0)	84.6 (74.5)	67.2 (51.1)
Kinect	57.0 (38.4)	88.3 (78.8)	56.6 (40.8)	77.3 (65.2)	80.0 (70.3)	79.8 (65.5)	80.7 (71.1)	85.4 (73.3)	75.7 (62.9)
MediaPipe	71.4 (60.5)	87.8 (79.4)	51.8 (32.8)	85.5 (78.8)	85.0 (75.9)	84.7 (73.9)	65.0 (44.0)	88.9 (81.8)	**77.5 (65.9)**

Table 3: User independent word (sentence) percent accuracy using Transformers

this, we combine all data and perform stratified 10-fold Cross Validation (CV) where test sets are not part of training sets. Word and sentence accuracies were averaged across all folds. On user adaptive tests, HMMs outperformed Transformers by 9.2%, 3.3%, and 1.4% word accuracy for AlphaPose, Kinect, and MediaPipe respectively. In Table 1b, we compare user dependent results from HMMs on all three pose estimation frameworks with results from Transformers. User dependent refers to models which are trained and tested on data from the same user [22]. To emulate this, we performed 10-fold CV for each user where test sets were not part of training sets. Word and sentence accuracies were averaged across all ten folds and eight users. HMMs again outperformed Transformers by 13.7%, 15.7%, and 7.0% word accuracy for AlphaPose, Kinect, and MediaPipe respectively. While we expect user dependent rates to be better than user adaptive rates, the opposite is true here. This is often an indicator of a small dataset, which is expected at the current stage of the project. However, we note that Transformers are especially affected, suggesting that they require more training data than HMMs to achieve parity. Table 2 shows user independent results for HMMs while Table 3 shows user independent results for Transformers. HMMs were again found to outperform Transformers by 23.4%, 14.8%, and 12.9% word accuracy for AlphaPose, Kinect, and MediaPipe respectively. Since HMMs can leverage the grammatical structure of the dataset and perform well on small datasets, these results are not surprising. However, models trained on AlphaPose or MediaPipe data, which only generate features using RGB videos, produced results similar to models trained on Kinect data, which uses additional depth data to generate features. Despite producing a more descriptive feature set, MediaPipe's processing speed - about 80 frames per second - was significantly greater than that of AlphaPose, which processed data at around 5.5 frames per second. Since most off-the-shelf cameras lack depth information, the quick processing speed of a 2D pose estimator like MediaPipe suggests that CopyCat could potentially perform well in real-time using common video recording devices.

5 FUTURE WORK

Zafrulla et al. [31] demonstrated how verification could also be used to confirm whether signed phrases are correct. Switching from recognition to verification improved their sentence accuracy from 67.0% to 82.0% on the CopyCat children dataset. As noted above, this dataset consisted of data from customized gloves. This verification system was used by Weaver et al. [28] at the deaf residential school to achieve good educational results. On the CopyCat adult dataset, Zafrulla et al. achieved 51.0% sentence accuracy. Given that AlphaPose and HMMs achieve better user independent sentence accuracy (71.9%), future work will focus on integrating ASL verification and collecting data from deaf children.

Our current dataset contains data from only eight adults. Additional data from a wider variety of signers will help models generalize and reach higher user-independent sentence accuracies. From the explanation provided in the methodology above, training on more adult novice subjects, due to their disfluencies and the limited deaf child population, is a critical step towards transitioning to the target population: deaf children.

ACKNOWLEDGMENTS

We would like to thank Prof. Thad Starner for his mentorship and resources, without which this project would not have been possible. We would also like to thank everyone who participated in our user study and recorded the dataset.

REFERENCES

[1] Helene Brashear, Valerie Henderson, Kwang-Hyun Park, Harley Hamilton, Seungyon Lee, and Thad Starner. 2006. American sign language recognition in game development for deaf children. In *Proceedings of the 8th international ACM SIGACCESS conference on Computers and accessibility*. 79–86. https://doi.org/10.1145/1168987.1169002

[2] Helene Brashear, Thad Starner, Paul Lukowicz, and Holger Junker. 2006. Using multiple sensors for mobile sign language recognition. In *Proceedings of the 8th international ACM SIGACCESS conference on Computers and accessibility*. 79–86. https://doi.org/10.1109/iswc.2003.1241392

[3] P Margaret Brown and Andrew Cornes. 2015. Mental health of deaf and hard-of-hearing adolescents: what the students say. *Journal of deaf studies and deaf education* 20, 1 (2015), 75–81. https://doi.org/10.1093/deafed/enu031

[4] Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020. Sign Language Transformers: Joint End-to-End Sign Language Recognition and Translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.1109/cvpr42600.2020.01004

[5] Helen Cooper and Richard Bowden. 2010. Sign language recognition using linguistically derived sub-units. In *Proceedings of 4th workshop on the representation and processing of sign languages: corpora and sign language technologies*. 57–61.

[6] Helen Cooper, Eng-Jon Ong, Nicolas Pugeault, and Richard Bowden. 2012. Sign language recognition using sub-units. *The Journal of Machine Learning Research* 13, 1 (2012), 2205–2231. https://doi.org/10.1007/978-3-319-57021-1_3

[7] David Coulter, Yijie Wang, and Phil Meadows. [n.d.]. Azure Kinect body tracking joints. https://docs.microsoft.com/en-us/azure/kinect-dk/body-joints. Accessed: 2020-01-07.

[8] M. Ebrahim Al-Ahdal and M. T. Nooritawati. 2012. Review in Sign Language Recognition Systems. In *2012 IEEE Symposium on Computers Informatics (ISCI)*. 52–57. https://doi.org/10.1109/ISCI.2012.6222666

[9] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. 2017. RMPE: Regional Multi-person Pose Estimation. In *ICCV*. https://doi.org/10.1109/iccv.2017.256

[10] Barak Freedman, Alexander Shpunt, Meir Machline, and Yoel Arieli. 2013. Depth mapping using projected patterns. US Patent 8,493,496.

[11] Enrique Garcia-Ceja, Michael Riegler, Anders K. Kvernberg, and Jim Torresen. 2020. User-adaptive models for activity and emotion recognition using deep transfer learning and data augmentation. *User Modeling and User-Adapted Interaction* 30, 3 (July 2020), 365–393. https://doi.org/10.1007/s11257-019-09248-1

[12] K. Grobel and M. Assan. 1997. Isolated sign language recognition using hidden Markov models. In *1997 IEEE International Conference on Systems, Man, and Cybernetics. Computational Cybernetics and Simulation*, Vol. 1. 162–167 vol.1. https://doi.org/10.1109/ICSMC.1997.625742

[13] Jie Huang, Wengang Zhou, Qilin Zhang, Houqiang Li, and Weiping Li. 2018. Video-based sign language recognition without temporal segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[14] Hamid Reza Vaezi Joze and Oscar Koller. 2018. Ms-asl: A large-scale data set and benchmark for understanding american sign language. *arXiv preprint arXiv:1812.01053* (2018).

[15] Kourosh Khoshelham and Sander Oude Elberink. 2012. Accuracy and resolution of kinect depth data for indoor mapping applications. *Sensors* 12, 2 (2012), 1437–1454. https://doi.org/10.3390/s120201437

[16] Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. http://arxiv.org/abs/1412.6980 cite arxiv:1412.6980Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015.

[17] Sang-Ki Ko, Chang Jo Kim, Hyedong Jung, and Choongsang Cho. 2019. Neural sign language translation based on human keypoint estimation. *Applied Sciences* 9, 13 (2019), 2683. https://doi.org/10.3390/app9132683

[18] Oscar Koller, Hermann Ney, and Richard Bowden. 2016. Deep Hand: How to Train a CNN on 1 Million Hand Images When Your Data Is Continuous and Weakly Labelled. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.1109/cvpr.2016.412

[19] Yong-Lu Li, Liang Xu, Xinpeng Liu, Xijie Huang, Yue Xu, Shiyi Wang, Hao-Shu Fang, Ze Ma, Mingyang Chen, and Cewu Lu. 2020. PaStaNet: Toward Human Activity Knowledge Engine. In *CVPR*.

[20] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. 2019. MediaPipe: A Framework for Building Perception Pipelines. *CoRR* abs/1906.08172 (2019). arXiv:1906.08172 http://arxiv.org/abs/1906.08172

[21] Rachel I Mayberry and Ellen B Eichen. 1991. The long-lasting advantage of learning sign language in childhood: Another look at the critical period for language acquisition. *Journal of memory and language* 30, 4 (1991), 486–512. https://doi.org/10.1016/0749-596x(91)90018-f

[22] S. Mazilu, M. Hardegger, Z. Zhu, D. Roggen, G. Tröster, M. Plotnik, and J. M. Hausdorff. 2012. Online detection of freezing of gait with smartphones and machine learning techniques. In *2012 6th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth) and Workshops*. 123–130. https://doi.org/10.4108/icst.pervasivehealth.2012.248680

[23] Ross E Mitchell and MICHAELA KARCHMER. 2004. Chasing the mythical ten percent: Parental hearing status of deaf and hard of hearing students in the United States. *Sign language studies* 4, 2 (2004), 138–163. https://doi.org/10.1353/sls.2004.0005

[24] Elissa L Newport. 1990. Maturational constraints on language learning. *Cognitive science* 14, 1 (1990), 11–28. https://doi.org/10.1207/s15516709cog1401_2

[25] Robin E Perkins-Dock Ph D, Terrilyn R Battle MS, Jaleassia M Edgerton MS, and Jaqueline N McNeill MS. 2015. A survey of barriers to employment for individuals who are deaf. *JADARA* 49, 2 (2015), 3.

[26] Patricia M Sullivan and John F Knutson. 2000. Maltreatment and disabilities: A population-based epidemiological study. *Child abuse & neglect* 24, 10 (2000), 1257–1273. https://doi.org/10.1016/s0145-2134(00)00190-3

[27] Oliver Turner, Kirsten Windfuhr, and Navneet Kapur. 2007. Suicide in deaf populations: a literature review. *Annals of General Psychiatry* 6, 1 (2007), 26. https://doi.org/10.1186/1744-859X-6-26

[28] Kimberly A Weaver, Harley Hamilton, Zahoor Zafrulla, Helene Brashear, Thad Starner, Peter Presti, and Amy Bruckman. 2010. Improving the language ability of deaf signing children through an interactive American Sign Language-based video game. (2010).

[29] Yuliang Xiu, Jiefeng Li, Haoyu Wang, Yinghong Fang, and Cewu Lu. 2018. Pose Flow: Efficient Online Pose Tracking. In *BMVC*.

[30] Kayo Yin. 2020. Sign Language Translation with Transformers. *ArXiv* abs/2004.00588 (2020).

[31] Zahoor Zafrulla, Helene Brashear, Pei Yin, Peter Presti, Thad Starner, and Harley Hamilton. 2010. American sign language phrase verification in an educational game for deaf children. In *2010 20th International Conference on Pattern Recognition*. IEEE, 3846–3849. https://doi.org/10.1109/icpr.2010.937