2024 18th International Conference on Automatic Face and Gesture Recognition (FG)

A Comparative Study of Video-based Human Representations for American Sign Language Alphabet Generation

Fei Xu[1], Lipisha Chaudhary[1], Lu Dong[1],
Srirangaraj Setlur[1], Venu Govindaraju[1], Ifeoma Nwogu[1]
[1] Department of Computer Science and Engineering, University at Buffalo, Buffalo, New York, USA

***Abstract*— Sign language is a complex visual language, and automatic interpretations of sign language can facilitate communication involving deaf individuals. As one of the essential components of sign language, fingerspelling connects the natural spoken languages to the sign language and expands the scale of sign language vocabulary. In practice, it is challenging to analyze fingerspelling alphabets due to their signing speed and small motion range. The usage of synthetic data has the potential of further improving fingerspelling alphabets analysis at scale. In this paper, we evaluate how different video-based human representations perform in a framework for Alphabet Generation for American Sign Language (ASL). We tested three mainstream video-based human representations: two-stream inflated 3D ConvNet, 3D landmarks of body joints, and rotation matrices of body joints. We also evaluated the effect of different skeleton graphs and selected body joints. The generation process of ASL fingerspelling used a transformer-based Conditional Variational Autoencoder. To train the model, we collected ASL alphabet signing videos from 17 signers with dynamic alphabet signing. The generated alphabets were evaluated using automatic metrics of quality such as FID, and we also considered supervised metrics by recognizing the generated entries using Spatio-Temporal Graph Convolutional Networks. Our experiments show that using the rotation matrices of the upper body joints and the signing hand give the best results for the generation of ASL alphabet signing. Going forward, our goal is to produce articulated fingerspelling words by combining individual alphabets learned in this work.**

I. INTRODUCTION

Sign language is an advanced visual language that employs hand gestures, arm movements, facial expressions, and other body motions to convey information in conversations involving Deaf and Hard-of-hearing (D/HH) individuals. A recent report [7] showed that by 2050 nearly 2.5 billion people in the world will live with some level of hearing loss and a large fraction currently communicate via sign language. In the United States, American Sign Language (ASL) is the third most used language [18].

An automated system that can process and interpret sign languages would facilitate communication between signing-only D/HH and hearing-only individuals. Such visual-lingual systems require a gesture recognition model to process signs, identify language components such as letters and words, and generate the corresponding text content or dynamic signing in video format for communication. Although many D/HH individuals read and can communicate in writing[12], generated signing videos allow such individuals to receive information in their native language and modality, a less cognitively burdensome task[1].

Fig. 1. The top row shows frames signing the alphabet 'P'. The middle row shows the joints data as extracted with Mediapipe [4]; the bottom row shows the mesh-based SMPL-X features [22] extracted with Hand4Whole [19].

Some intrusive sign recognition systems [8], [21] request signers to wear gloves or clothes with sensors for recognition, but these have not been as well received by Deaf communities especially as sign languages comprise of both manual (involving hands) and nonmanual (involving face and head movements, lip movements, and body orientations). Vision-based systems on the other hand are much less intrusive and can more readily capture and generate all aspects and nuances of sign language.

Fingerspelling is inherently connected to its spoken language counterpart, as it was originally developed to incorporate spoken language into sign language. As a result,

[1]A spoken language and its sign counterpart do have a direct one-on-one relation. Hence, mapping them both can be a challenge with the translation requiring more cognitive processing[11].

979-8-3503-9494-8/24/$31.00 ©2024 IEEE

fingerspelling plays a crucial role in improving literacy skills, particularly for D/HH children [24]. Additionally, it serves as a primary method for signing words that lack of conventional signs. Alphabet signs, as the fundamental components of fingerspelling, involve rapid and smooth transitions between letters. Automating the translation and generation of alphabet signs can also benefit communication and education within the Deaf community.

There are three visual representations that are widely used in general video-based action recognition tasks: two-stream inflated 3D ConvNet (I3D), 3D landmarks of body joints, and rotation matrices of body joints. In this paper, we present a comparative study of these video-based human representations for dynamic ASL alphabet generation. We also tested the effects of using selected body joints and different skeleton graphs and present our findings. To accomplish this, we collected signing videos from 17 native ASL signers to create a dataset with dynamic alphabet signings. Following the work from Petrovich et al. [23], we trained a transformer-based Conditional Variational Auto-Encoder (CVAE) to generate the appropriate ASL fingerspellings when conditioned on fixed input embeddings representing the alphabets. The generated alphabets were evaluated using automatic metrics of quality such as FID, and we also considered supervised metrics by recognizing the generated entries using Spatio-Temporal Graph Convolutional Networks. Our experiments show that using the rotation matrices of the upper body joints and the signing hand give the best results for the generation of ASL alphabet signing.

II. Related Work

In this work, we are concerned with how different video-based human representations work for ASL alphabet generation. Some of these representations rely on methods for Human Pose Estimation (HPE) which we briefly describe in Section II-A. We then provide a brief description for each of the three widely used video-based human representations in Section II-B. Then, we finally provide a brief summary of recent methods for synthetic data generation for action recognition II-C.

A. Human Pose Estimation

Most HPE methods have two stages: human detection and keypoints estimation. Depending on which stage performs first, HPE methods can be categorized top-down or bottom-up approaches. Previous empirical studies [26] suggest that top-down HPE can be more accurate in detecting human poses. This is likely due to their ability estimate the humans as complete objects first. This effectively reduces the ambiguity of merging independent joints into full skeletons, and it also helps the joint feature extraction concentrate on more precise regions.

It is challenging to analyze fingerspellings due to their attributes. Specifically, some fingerspelling characters share some level of similarity from the same view, and depth features can provide important information to differentiate such characters. However, 3D human pose estimation (HPE) from 2D images presents additional challenges due to the unknown complexity of the depth dimension. In this work we considered two top-down HPE frameworks to extract data per video frame: we use Mediapipe [4] to extract body joints' 3D landmarks, and Hand4Whole [19] is used to extract rotation matrices of body joints.

B. Human Representations for Action Recognition

Human action in videos can be represented by a sequence of human pose changes. This recognition task become essential due to its derivability to several downstream applications such as video retrieval, surveillance systems, and human behavior analysis. It is also a challenging task in computer vision because of its environmental and temporal dynamics. There are three visual representations that are widely used in general video-based action recognition tasks: two-stream inflated 3D ConvNet (I3D), 3D landmarks of body joints, and rotation matrices of body joints. In this section, we will briefly describe these three representations.

3D convolutional models can directly work on a sequence of raw images. Previous works [13], [3] used 3D CNN models to extract dynamics from RGB, gradient and optical flow channels for action recognition. Inflating from pretrained 2D convolutional networks, I3D convolution modules [6] are employed to extract 3D convolutional representations that encompass both spatial and temporal information for video analysis. The procedure to recover the 3D convolutional representations back to a sequence of 2D images is non-trivial, however, due to its wide usage in various video-based analysis, we test it as one representation of interest.

Skeleton-based body joints, specially 3D Landmarks, are widely used in action recognition due to the view-invariant and the interpretation of human body motion to be a continuous evolution of rigid segments from an articulated system. To include temporal information for skeleton change, previous works [5], [14], [20] projected body joints dynamics into 2D motion images for action recognition. Yan et al. [27] and Shi et al. [25] construct representations that contain joint dynamics on spatial and temporal skeleton-based graphs. Recent works [22], [19] take advantage of geometric-constrained human models to increase the accuracy of the extracted joint data.

Rotation Matrices, on the other hand, are a different way to represent 3D human poses. Unlike the landmark based representations, the rotation matrices focus the relative orientations between connected landmarks.

C. Synthetic Data Generation for Action Recognition

While several deep-learning based approaches have been presented in the literature for generating novel static images of humans, there have been significantly fewer works in dynamic human motion generation[2]. Li et al [16] used generative adversarial networks (GAN) to synthesize human

[2]Generation here refers to *synthesizing* human motion purely from a random sample, unlike many works that predict the next step given a previous one.

movements when prompted by short text phrases representing actions. Ahn et al [1] and Ahuja and Morency [2] also present conditioning generative models for synthesizing human motion when presented with textual descriptions. Similarly, Petrovich et al. [23], presented a generative model to produce human motions when conditioned on one of 40 action words.

The above human motion synthesis methods focus on gross body movements whereas, our ASL alphabet motion generation focuses on very fine motor movements that can be easily missed or confused with one another. To address this, we trained a human motion generative model for focused fine motor activities and we tested it out on various video-based human representations. Though there are various generative models for video generation, the analysis of such models is out of the scope of this work.

III. The Generation Pipeline

We followed the CVAE [23] as the main generation pipeline because it learns the sequence-level conditional embedding for action generation, and it can generate sequences with control of sequence length. Here we give a brief description of the CVAE model [23] that we used for generating ASL alphabet signings.

A. The CVAE Model

This CVAE model uses 26 one-hot embeddings that represent the 26 English alphabets as the condition c in both the encoder and the decoder to generate human motions. The variational lower bound of this CVAE model is:

$$\begin{aligned}\tilde{L}_{CVAE} &= \log q(X|c) - D_{KL}[q(z|X,c)||p(z|X,c)] \\ &= E(\log q(X|(\Sigma,\mu))) - D_{KL}[q(z|X,(\Sigma,\mu))||p(z|b)] \quad (1)\end{aligned}$$

where X is a sequence of motion representations, and $q(\cdot)$ and $p(\cdot)$ represent the encoder and decoder separately. Given the input X, the alphabet-related tokens will be prepended to X before passing it through a positional embedding layer. The transformer-based encoder will then generate conditional distribution parameters μ and Σ to sample sequence-level motion latent embedding z. The decoder uses an extra learnable bias b to shift the embedding z based on the motion information, and the duration r is given to the positional embedding layer to control the length of the input to the decoder. Fig. 2 illustrates the architectural flow of the model.

We consider the reconstruction loss of motion representations and KL divergence loss of distributions between generated and real data. To evaluate the generation quality, we use an STGCN [27] classifier that uses spatial-temporal skeleton graphs for action recognition task. The spatial connections between joints are defined naturally based on the body joints, and the same joints from consecutive frames are connected along the temporal dimension. We use both generated sign accuracy and the Fréchet inception distance (FID) metric to compare the distributions between generated data and real data to evaluate the generated ASL alphabet signings.

B. Comparisons of video-based human representations

We compared three video-based human representations for ASL alphabet generation: I3D features (III-B.1), 3D landmarks of body joints(III-B.2), and rotation matrices of body joints (III-B.3).

In Sec.III-C, we discuss the modifications we made to the skeleton graphs used as inputs to train the STGCN classifiers for different representations.

1) I3D Features: To understand the effects of using basic raw RGB video frames, we adapt the concept of using 3D convolutional features. We make use of I3D (Inflated 3D Networks) [6] and consider a sliding window technique to obtain overlapping segments. For this work, based on the observation of the rapid transitions of alphabet signings, we have chosen a sliding window size of 8 and a stride size of 2. The overlapping of the frames guarantees a broader view of the information in an iterative fashion ensuring near about maximum details to be encoded in the resultant segments.

We make use of a pre-trained I3D model[15] as a feature extractor to obtain a 1024 embedding for each video segment. To enable the direct usage of this representation with STGCN for evaluation later, we took inspiration from ViT [10] and divided each frame into 4×4 patches before feature extraction. The corresponding graph structure will be introduced in Sec. III-C.

2) Body Joint 3D landmarks: In our analysis of generating alphabet data, we used Mediapipe to obtain both body and hands 3D landmarks. Blazepose [4] was integrated to provide 33 3D body joints. For hand joints, a single-shot detector [17] was utilized to identify the palm region first, followed by applying a hand model to localize 21 joints per hand, including one wrist point and four keypoints per finger. All joints data is normalized by the average of the two distances between each side of shoulder to hip (Fig. 3 (a)). Note that for ASL alphabet generation, we only include joints of the signing hand. We then merged the body and hand joints using the wrist joint coordinates.

3) Body Joint Rotation Matrices: Considering the subtle and rapid finger motions for ASL alphabet signing, we used Hand4Whole [19] to get the body joints' rotation matrices, as the framework specifically considers metacarpophalangeal (MCP) joints to improve wrist point estimation for the hand data. The extracted data of body and hand joints are compatible with the SMPL-X [22] human model, which has 22 joints for the body (Fig. 3 (b)) and 15 joints on each hand.

C. Skeleton Graph for training STGCN

As described in Sec.III-A, the STGCN model uses a spatial-temporal skeleton graph during training. Since ASL alphabets are gestured using a single hand, we focus on the signing hand and upper body motion. We tested two types of selected body joints (Sec. V-B) using pose data. Each one uses a different skeleton graph that is modified from the original one accordingly.

As shown in Fig. 3, Blazepose [4] and SMPL-X [22] use different skeleton graphs. To test the effect of using different

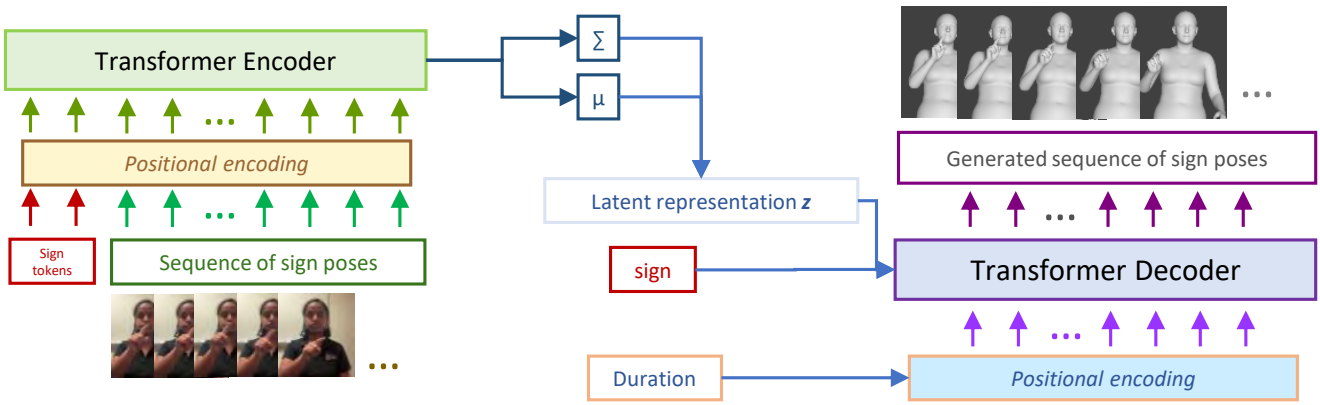


Fig. 2. The CVAE pipleline [23] used for FS signing generation. Input to the encoder includes two sign-related tokens, which are used to produce alphabet-conditional distribution parameters μ and Σ to get latent representation z to generate sequence of motions with decoder.

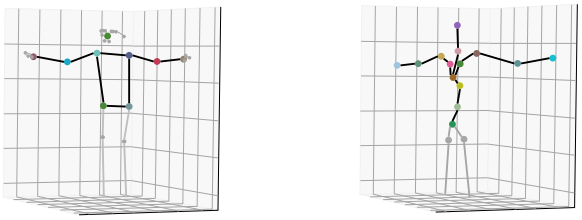
skeleton graphs on generating ASL alphabet signing, we transformed the body joint's 3D landmarks from one skeleton graph to another. Specifically, the shoulder and hip joints at Blazepose skeleton (Fig. 3 (a)) are used to create the joints along the spine in SMPL-X skeleton (Fig. 3 (b)) with estimated geometric proportions.

To utilize STGCN as the recognition classifier and run comparisons among all three representations with uniform conditions, we created a grid-like 4×4 graph for I3D features, where each positional patch of the video segment is treated as one node on the spatial-temporal graph. Every two patch-level video segments that represent as positional neighbors are connected on the spatial skeleton graph. We also tried other sizes for splitting patches, but they did not make a significant difference with respect to recognition accuracy. Therefore, we do not include such results in Sec.V.

IV. Dataset

For this work, we collected video recordings of continuous ASL alphabet signing from 17 native or fluent signers. 10 of these signers are collected from YouTube public channels that give instructional ASL videos to teach alphabet signing. We selected videos that do not have interruptions (e.g. camera shakes, video edition effect, etc.) during signing to avoid unexpected visual challenges in video processing. Such challenges will make it much harder to learn the conditional embedding of alphabet signings, especially when the alphabet signs themselves typically involve rapid and subtle motions. Similarly, the other 7 signers from local communities recorded themselves with a fixed camera and provided signing videos with sufficient environmental brightness and high contrast between foreground and background. The average frame numbers for training and testing clips are 43.3 and 89.2 before further preprocessing.

All videos are first converted to the same frame rate (30 FPS), and segmented and annotated using the video annotation tool [9]. Each video clip includes both the transition and the clear gesture of a single alphabet signing, and a human bounding box is annotated per segment to cover the upper body with both hands and the face of the signer in each frame. Various representations are then extracted from the annotated region for analysis. Since the signer's handiness won't affect the signing appearance, for a left-handed signer, the video frame will be horizontally flipped and treated as right-handed before processing. Both the raw videos and their extracted features will be publicly available.



(a) Mediapipe body topology (b) Body topology of SMPL-X

Fig. 3. Body joints topology for different visual features. Note that the grey-out joints are not considered in this work

V. Experiment

We tested several components in the generation of video-level ASL alphabet data. For each experimental configuration, we utilized the transformer-based CVAE [23] (Sec. III-A). A STGCN [27] classifier was trained for each configuration to assess recognition accuracy on the generated testing set. We also include FID metrics for evaluating generated FS signing. In the first three experiments, all CVAE and STGCN models are trained only with real training data. The STGCN classifier trained for the fourth experiment uses the generated training data from the CVAE with the best performance among previous experimental conditions. In the Tab. I, FID_{train} and FID_{test} refer to the FID metrics of the generated training and testing data separately. Acc of the first three experiments refers to the recognition performance of the classifier that trained using real training data and tested on generated testing data, while in the fourth experiment (Rot_{GEN} in I) the Acc is a classifier that trained using gener-

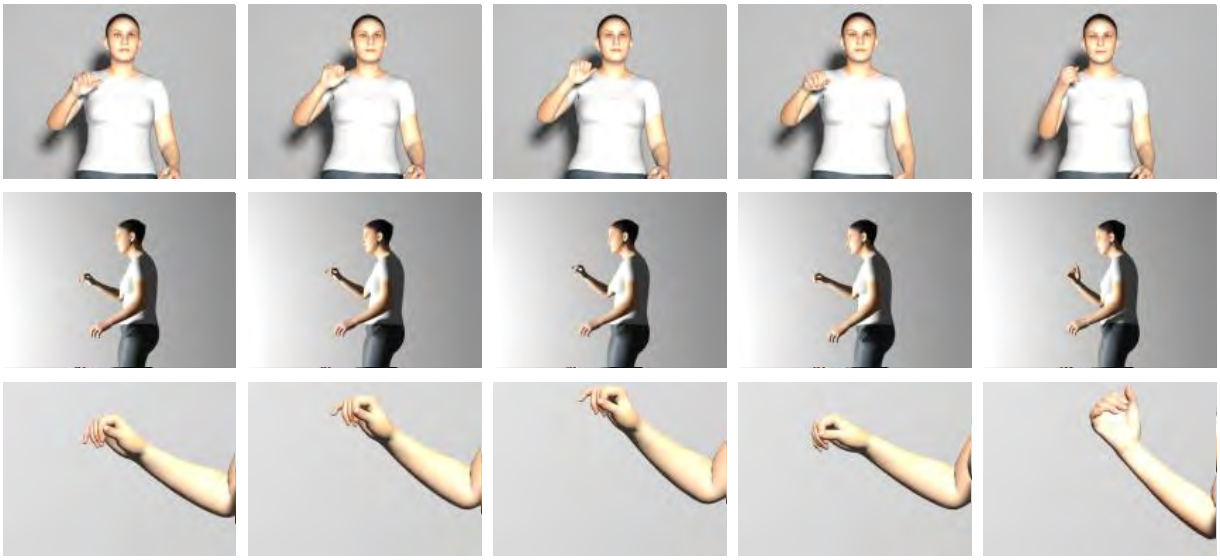


Fig. 4. Generated avatar sequence of "J" using both SMPL-X model [22] and body joint rotation matrices data. The top row shows the front view and the bottom two rows show the right-side view to indicate more details on the signing hand fingers.

TABLE I
RECOGNITION ACCURACIES OF DIFFERENT HUMAN REPRESENTATIONS.

Features	FID_{train}	FID_{test}	Acc
I3D	225.4	325.6	0.064
3Dlandmarks	125.6	118.4	0.159
Rotation matrix	19.2	**95.0**	**0.979**
Rotation matrix$_{UP}$	**12.1**	112.4	0.949
3D landmarks$_{C}$	120.8	199.6	0.226
Rotation matrix$_{GEN}$	-	-	**0.959***

ated training data and tested on real testing data. The Rot_{UP} (Sec. V-B) refers to the case that uses rotation matrices of partial upper body joints. $3Dlandmark_C$ (Sec. V-C) is the configuration that converted the joints' 3D landmarks to a new version that matches a different skeleton graph. Rot_{GEN} (Sec. V-D) is one that trained the STGCN classifier using generated rotation matrices data, and tested on real testing data.

A. Comparison of Video-based Human Representations

We test the effectiveness of different human representations in generating alphabet signing data including I3D feature, 3D landmarks joint data ($3Dlandmarks$), and rotation matrices joint data (Rot). The first three rows ($I3D$, $3Dlandmarks$, and Rot) of Table I display FS recognition performance of generated testing data using different representations. Compared with the other two representations, joint rotation matrix performs better to represent the rapid and subtle dynamics in ASL alphabet signings for video-based recognition task. According to the FID_{test} metrics, this might be related to the fact that using body joint rotation matrix helps to produce FS signings while keeping the better quality of generated data.

B. Comparison of different selected body joints

To test if learning from essential body joints alone would help the generation of ASL alphabet, we tested two types of selected joints using their rotation matrices. The first type (Rot in Table I) includes joints of upper body and the signing hand. The second one (Rot_{UP}) uses all joints of the signing hand but only includes partial body joints from the pelvis joint to the signing arm wrist joint. Results show that concentrating on signing-related joints alone helps ASL FS generation during training, but removing non-signing arm information hurts the performance of generating testing sets.

C. Comparison of Different Skeleton Graphs

Here we first convert the landmark data from Blazepose (Fig. 3 (a)) to SMPL-X skeleton graph (Fig. 3 (b)) and get a new version of 3D landmark data ($3Dlandmarks_C$ in Table I) that includes new joints from the pelvis to the collar joint. Results show that though using a different skeleton graph that has more body joints shows some improvement in recognition task on generated data (accuracy increases from 15.9% to 22.6%), but it downgrades the quality of generated FS for testing set based on the FID_{test}.

D. Generated Data Quality

From all representation configurations of these three experiments, we choose the best configuration (Rot), which uses rotation matrices of the entire upper body joints and signing hand. To verify the quality of the generated ASL alphabet signing data, we ran a reverse experiment. We took the CVAE model that was trained with the best configuration to create generated training data. This generated training set

is then used to train an STGCN classifier. We perform this classifier on real testing data to show the quality of data generation. The results of row Rot_{GEN} show the classifier that trained with generated training joint pose data gives a similar level of recognition performance on real testing data.

This indicates the generated data is diverse, covering the original distribution of the FS dataset.

VI. Conclusion

In this paper, we test several configurations of video-based human representations. We identified the best of all configurations to learn conditional embedding for generating ASL alphabet data. Based on the results obtained we concluded, that the use of rotation matrices data from the signer's entire upper body and signing hand works better compared to the other considered conditions and representations. Our goal is to generate natural and smooth articulated fingerspelled words by combining individual alphabets learned in this work. Since such rotation matrices data can also be utilized to control signing videos with human-like avatars (see Fig. 4), we aim to generate FS videos that are more acceptable among hard-of-hearing community using avatar models. This work can be furthermore extended to create more smooth and fluid co-articulated signs - figerspell and otherwise - by following signing rules used by fluent signers while fingerspelling.

References

[1] H. Ahn, T. Ha, Y. Choi, H. Yoo, and S. Oh. Text2action: Generative adversarial synthesis from language to action. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5915–5920. IEEE, 2018.

[2] C. Ahuja and L.-P. Morency. Language2pose: Natural language grounded pose forecasting. In *2019 International Conference on 3D Vision (3DV)*, pages 719–728. IEEE, 2019.

[3] J. Arunnehru, G. Chamundeeswari, and S. P. Bharathi. Human action recognition using 3d convolutional neural networks with 3d motion cuboids in surveillance videos. *Procedia computer science*, 133:471–477, 2018.

[4] V. Bazarevsky, I. Grishchenko, K. Raveendran, T. Zhu, F. Zhang, and M. Grundmann. Blazepose: On-device real-time body pose tracking. *arXiv preprint arXiv:2006.10204*, 2020.

[5] C. Caetano, J. Sena, F. Brémond, J. A. Dos Santos, and W. R. Schwartz. Skelemotion: A new representation of skeleton joint sequences based on motion information for 3d action recognition. In *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–8, 2019.

[6] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[7] C. A. Chadha S, Kamenov K. The world report on hearing. *Bull World Health Organ.*, 2021.

[8] H. Cooper, B. Holt, and R. Bowden. Sign language recognition. In *Visual Analysis of Humans: Looking at People*, pages 539–562. Springer, 2011.

[9] K. Davila and R. Zanibbi. Whiteboard video summarization via spatio-temporal conflict minimization. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, pages 355–362. IEEE, 2017.

[10] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[11] M. Gärdenfors. *Writing in deaf and hard-of-hearing children: A bimodal bilingual perspective on their written products and writing processes*. PhD thesis, Department of Linguistics, Stockholm University, 2023.

[12] M. Huenerfauth and V. Hanson. Sign language in the interface: access for deaf signers. *Universal Access Handbook. NJ: Erlbaum*, 38:14, 2009.

[13] S. Ji, W. Xu, M. Yang, and K. Yu. 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):221–231, 2012.

[14] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid. A new representation of skeleton sequences for 3d action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3288–3297, 2017.

[15] D. Li, C. Xu, X. Yu, K. Zhang, B. Swift, H. Suominen, and H. Li. Tspnet: Hierarchical feature learning via temporal semantic pyramid for sign language translation. *Advances in Neural Information Processing Systems*, 33:12034–12045, 2020.

[16] X. Lin and M. R. Amer. Human motion modeling using dvgans, 2018.

[17] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 21–37. Springer, 2016.

[18] R. E. Mitchell and T. A. Young. How many people use sign language? a national health survey-based estimate. *Journal of Deaf Studies and Deaf Education*, 28(1):1–6, 2023.

[19] G. Moon, H. Choi, and K. M. Lee. Accurate 3d hand pose estimation for whole-body 3d human mesh estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2308–2317, 2022.

[20] Q. Nie, J. Wang, X. Wang, and Y. Liu. View-invariant human action recognition based on a 3d bio-constrained skeleton model. *IEEE Transactions on Image Processing*, 28(8):3959–3972, 2019.

[21] C. Oz and M. C. Leu. American sign language word recognition with a sensory glove using artificial neural networks. *Engineering Applications of Artificial Intelligence*, 24(7):1204–1213, 2011.

[22] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. A. Osman, D. Tzionas, and M. J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019.

[23] M. Petrovich, M. J. Black, and G. Varol. Action-conditioned 3d human motion synthesis with transformer vae. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10985–10995, 2021.

[24] J. A. Scott, S. G. Hansen, and A. R. Lederberg. Fingerspelling and print. *American Annals of the Deaf*, 164(4):429–449, 2019.

[25] L. Shi, Y. Zhang, J. Cheng, and H. Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12026–12035, 2019.

[26] F. Xu, K. Davila, S. Setlur, and V. Govindaraju. Skeleton-based methods for speaker action classification on lecture videos. In *International Conference on Pattern Recognition*, pages 250–264. Springer, 2021.

[27] S. Yan, Y. Xiong, and D. Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.