Evaluating the Immediate Applicability of Pose Estimation for Sign Language Recognition

Amit Moryossef[1,2] Ioannis Tsochantaridis[2]
Joe Dinn[3] Necati Cihan Camgöz[3] Richard Bowden[3] Tao Jiang[3]
Annette Rios[4] Mathias Müller[4] Sarah Ebling[4]
amitmoryossef@gmail.com, ioannis@google.com
{j.dinn, n.camgoz, r.bowden, t.jiang}@surrey.ac.uk
{rios, mmueller, ebling}@cl.uzh.ch
[1]Bar Ilan University, [2]Google
[3]University of Surrey, [4]University of Zurich

Abstract

Sign languages are visual languages produced by the movement of the hands, face, and body. In this paper, we evaluate representations based on skeleton poses, as these are explainable, person-independent, privacy-preserving, low-dimensional representations. Basically, skeletal representations generalize over an individual's appearance and background, allowing us to focus on the recognition of motion. But how much information is lost by the skeletal representation? We perform two independent studies using two state-of-the-art pose estimation systems. We analyze the applicability of the pose estimation systems to sign language recognition by evaluating the failure cases of the recognition models. Importantly, this allows us to characterize the current limitations of skeletal pose estimation approaches in sign language recognition.

1. Introduction

Sign languages are visual languages produced by the movement of the hands, face, and body. As languages that rely on visual communication, recordings are in video form. Current state-of-the-art sign language processing systems rely on the video to model tasks such as sign language recognition (SLR) and sign language translation (SLT). However, using the raw video signal is computationally expensive and can lead to overfitting and person dependence.

In an attempt to abstract over the video information, skeleton poses have been suggested as an explainable, person-independent, privacy-preserving, and low-dimensional representation that provides the signer body pose and information on how it changes over time. Theoretically, skeletal poses contain all the relevant information required to understand signs produced in videos, except for interactions with elements in space (for example, a mug or a table).

The recording of accurate human skeleton poses is difficult and often intrusive, requiring signers to wear specialized and expensive motion capture hardware. Fortunately, advances in computer vision now allow the estimation of human skeleton poses directly from videos. However, as these estimation systems were not specifically designed with sign language in mind, we currently do not understand their suitability for use in processing sign languages both in recognition or production.

In this paper, we evaluate two pose estimation systems and demonstrate their suitability (and limitations) for SLR by conducting two independent studies on the CVPR21 ChaLearn challenge [33]. Because we perform no pretraining of the skeletal model, the final results are considerably lower than potential end-to-end approaches (§3). The results demonstrate that the skeletal representation loses considerable information. To better understand why, we evaluate our approaches (§4), categorize their failure cases (§5), and conclude by characterizing the attributes a pose estimation system should have to be applicable for SLR (§6).

2. Background

2.1. Pose Estimation

Pose estimation is the task of detecting human figures in images and videos to determine where various joints are present in an image. This area has been thoroughly researched [30, 12, 7, 20, 19] with objectives varying from the predicting of 2D/3D poses to a selection of a small specific set of landmarks or a dense mesh of a person. Vogler [38] showed that the face pose correlates with facial non-manual features.

OpenPose [7, 32, 8, 40] was the first multi-person system to jointly detect human body, hand, facial, and foot keypoints (135 keypoints in total) in 2D on single images. While this model can estimate the full pose directly from an image in a single inference, a pipeline approach is also suggested where first, the body pose is estimated and then independently the hands and face pose by acquiring higher-resolution crops around those areas. Building on the slow pipeline approach, a single-network whole-body OpenPose model has been proposed [21], which is faster and more accurate for the case of obtaining all keypoints. Additionally, with multiple recording angles, OpenPose also offers keypoint triangulation to reconstruct the pose in 3D.

DensePose [20] takes a different approach. Instead of classifying for every keypoint which pixel is most likely, similar to semantic segmentation, each pixel is classified as belonging to a body part. Then, for each pixel, knowing the body part, the system predicts where that pixel is on the body part relative to a 2D projection of a representative body model. This approach results in reconstructing the full-body mesh and allows sampling to find specific keypoints similar to OpenPose.

MediaPipe Holistic [19] attempts to solve the 3D pose estimation problem directly by taking a similar approach to OpenPose, having a pipeline system to estimate the body and then the face and hands. It uses a dense mesh model for the face pose containing 468 points, but resorts to skeletal joints for the body and hands. Unlike OpenPose, the poses are estimated using regression rather than classification and are estimated in 3D rather than 2D.

2.2. Sign Language Recognition

Sign language recognition (SLR) is the task of recognizing a sign or a sequence of signs from a video. This task has been attempted both with computer vision models, assuming the input is the raw video, and with poses, assuming the video has been processed with a pose estimation tool.

2.2.1 Video to Sign

Camgöz et al. [4] formulate this problem as one of translation. They encode each video frame using AlexNet [24], initialized using weights that were trained on ImageNet [16]. Then they apply a GRU encoder-decoder architecture with Luong Attention [25] to generate the signs. In a follow-up work [6], they use a transformer encoder [37] to replace the GRU and use Connectionist Temporal Classification (CTC) [18] to decode the signs. They show a slight improvement with this approach over the previous one.

Adaloglou et al. [1] perform a comparative experimental assessment of computer vision-based methods for the SLR task. They implement various approaches from previous research [3, 15, 36] and test them on multiple datasets [22, 4, 39, 36] either for isolated sign recognition or continuous sign recognition. They conclude that 3D convolutional models outperform models using only recurrent networks because they better capture temporal information and that convolutional models are more scalable given the restricted receptive field, which results from their "sliding window" technique.

2.2.2 Pose to Sign

Upper body poses have been widely used as a feature for computational sign language research [14], due to their signer-invariant representation capabilities. They have been included into recognition [17], translation [5], or detection [28] frameworks, either in raw coordinate form or as linguistically meaningful symbols extracted from joint coordinates [13].

Before the deep learning era, most sign language systems utilized specialized sensors, such as Kinect [43, 10], to estimate signers pose in real-time [31]. There have also been attempts to train models on sign language data [29, 11, 26] which extract low-resolution skeletons, i.e., few joints. However, these approaches suffered from noisy estimations and had deficient hand joint resolution.

As with any subfield of computer vision, human pose estimation also improved with the introduction of deep learning-based approaches. Open source, general-purpose human pose estimation models, such as convolutional pose machines [41] and their predecessor OpenPose [7], became widely used in sign language research. Ko et al. [23] utilized a transformer-based translation based purely on skeletal information. Albanie et al. [2] proposed using pose estimates to recognize co-articulated signs. They further used the pose estimates to train knowledge distillation networks and learn meaningful representations for downstream tasks.

3. Experiments

To evaluate whether pose estimation models are applicable for SLR, we participated in the CVPR21 ChaLearn challenge for person-independent isolated SLR on the Ankara University Turkish Sign Language (AUTSL) [34] dataset. Even though the dataset includes Kinect pose estimations, Kinect poses have not been made available for the challenge. We processed the dataset using two pose estimation tools: 1. OpenPose Single-Network Whole-Body Pose Estimation [21]; and 2. MediaPipe Holistic [19]; and made the data available via an open-source sign language datasets repository [27].

We approach the recognition task with two independent experiments performed by different teams unaware of the other team's work throughout the validation stage. In the validation stage, each team focussed on one pose estimation approach, and in the test stage, both teams got access to both pose estimation outputs. We eventually submitted three systems: 1. based on *OpenPose* poses; 2. based on *Holistic* poses; 3. based on both *OpenPose* and *Holistic* poses combined (concatenated).

3.1. Team 1

Team 1 worked with OpenPose [21] pose estimation output and used the SLR transformer architecture from Camgöz et al. [6]. The model takes as input a series of feature vectors, in this case, human upper body skeletal coordinates extracted from the video frames. These are each projected to a lower dimension hidden state vector. The size of the hidden state remains constant throughout the subsequent operations. A sinusoidal positional encoding is added to provide temporal information. This is then passed to a subnetwork consisting of a multiheaded self-attention layer, followed by a feedforward layer. After each of these layers, the output is added to the input and normalized. This subnetwork can be repeated any number of times. Finally, the output is fed to a linear layer and softmax to give probabilities for each class (Figure 1).

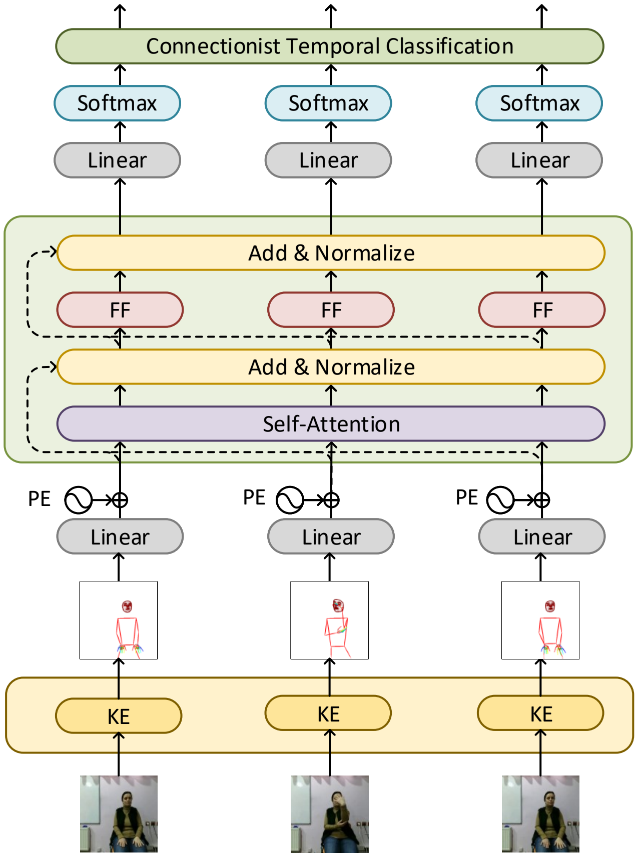

Figure 1. Diagram of *Team 1*'s model with one subnetwork (in green). (KE: Keypoint extraction, PE: Positional encoding, FF: feed forward)

The model is trained using CTC loss. This is designed to allow the output to be invariant to alignment; however, this is not a significant concern when there should only be one output symbol. The final prediction is obtained via CTC beam search decoding, collapsing multiple same class outputs into one. As the model is trained to predict a single class per video, it does not predict different classes within a sequence.

The number of layers, heads, hidden size, and dropout rate affect the model complexity. There is, therefore, a tradeoff between sufficient complexity to model the data and overfitting.

Additionally, as a baseline, the pose estimation keypoints were replaced with the output of three off-the-shelf image-based frame feature extractors, giving us small dense representations for each frame. Three extractors were used: 1. EfficientNet-B7 [35]; 2. I3D trained on Kinetics [9]; and 3. I3D trained on BSL1K [2].

3.2. Team 2

Team 2 worked with the MediaPipe Holistic [19] pose estimation system output. From the 543 landmarks, the face mesh was removed which consists of 468 landmarks and the

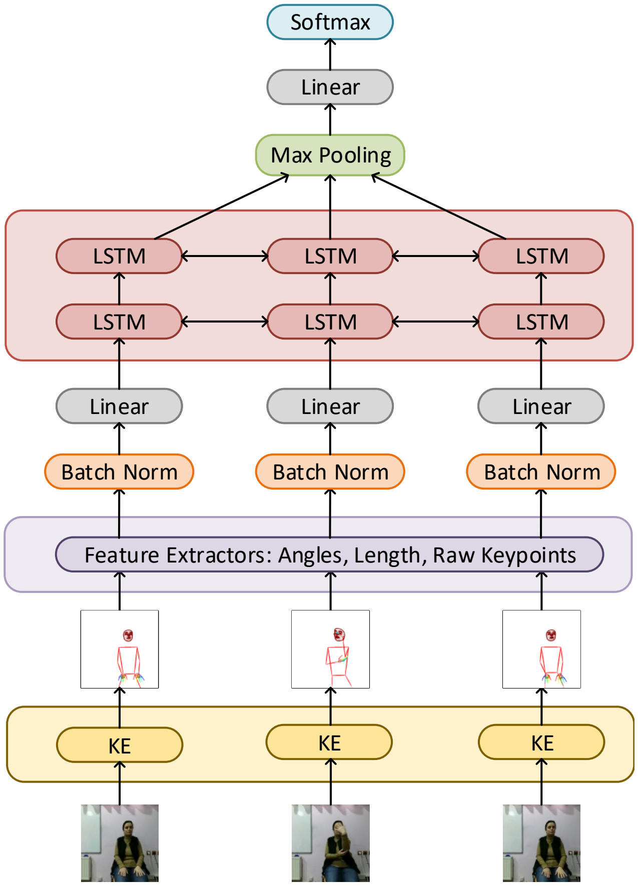

Figure 2. Diagram of *Team 2*'s model. (KE: Keypoint extraction)

remaining 75 landmarks were used for the body and hands.

A standard sequence classification architecture was used. The model takes as input a series of feature vectors, constructed from a flat vector representation of the pose concatenated with the 2D angle and length of every limb, using the *pose-format*[1] library. These representations are subjected to a 20% dropout, normalized using 1D batch normalization, and are projected to a lower dimension hidden state vector (512 dimensions). This is then passed to a two-layer BiLSTM with hidden dimension 256, followed by a max-pooling operation to obtain a single representation vector per video. Finally, the output is fed to a linear layer and softmax to give probabilities for each class (Figure 2).

The model is trained using cross-entropy loss with the Adam optimizer (with default parameters) and a batch size of 512 on a single GPU. *No* data augmentation or frame dropout is applied at training time, except for horizontal frame flip to account for left-handed signers in the dataset.

4. Results

Table 1 shows our teams' results on the validation set. We note that both teams' approaches using pose estimation performed similarly, with validation accuracy ranging between 80% and 85%. It rules out trivial errors and implementation issues that, despite working independently, and with two separate pose estimation tools, both teams achieve similar evaluation scores. Furthermore, from a comparison between the pose estimation based systems (80-85%) and the pretrained image feature extractors (38-68%), we can see that pose estimation features do indeed generalize better to the nature of the challenge, including unseen signers and backgrounds.

	Team 1	Team 2
EfficientNet-B7	38.80%	—
I3D (Kinetics)	47.46%	—
I3D (BSL1K)	68.65%	—
OpenPose	83.25%	79.99%
Holistic	85.63%	82.14%
OpenPose+Holistic	84.16%	82.89%

Table 1. Results evaluated on the validation set with various frame-level features.

We submitted *Team 2*'s test set predictions to the official challenge evaluation. On the test set, both *OpenPose* and *Holistic* performed **equally well** despite making different predictions, each with 78.35% test set accuracy. However, our combined system, which was trained using both pose estimations, achieves 81.93% test set accuracy.

5. Analysis

The interpretability of skeletal poses allows us to assess them qualitatively using visualisation. We manually review our model's failure cases and categorize them into two main categories: hands interaction and hand-face interaction.

Hands Interaction When there exists an interaction between both hands, or one hand occludes the other from the camera's view, we often fail to estimate the pose of one of the hands (Figure 3) or estimate it incorrectly such that the interaction is not clearly shown (Figure 4).

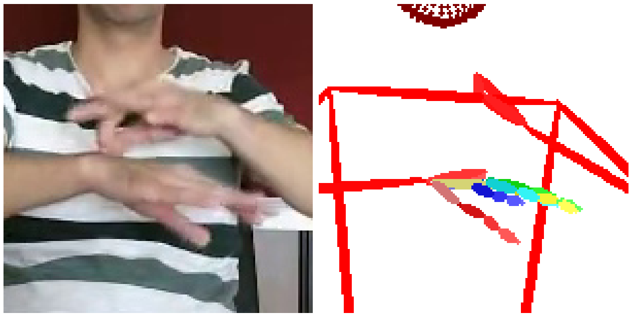

Figure 3. Example of hands interaction, where the pose estimation fails for one of the hands (Holistic).

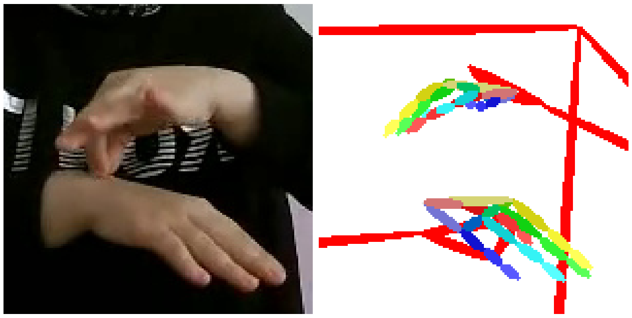

Figure 4. Example of hands interaction, where the pose estimation does not reflect the existing interaction (Holistic).

Hand-Face Interaction When there exists an interaction between a hand and the face, or one hand overlaps with the face from the camera's angle, we often fail to estimate the pose of the interacting hand (Figure 5).

These cases of missed interactions between the different body parts often lose the essence of the sign, where the interaction and the hand shape are the main distinguishing features for those signs, and thus hinder the model's ability to extract meaningful information from the pose that is relevant to the sign.

Presence or absence of hand pose We describe a number of failure cases of Holistic pose estimation above. Many of

[1] https://github.com/AmitMY/pose-format

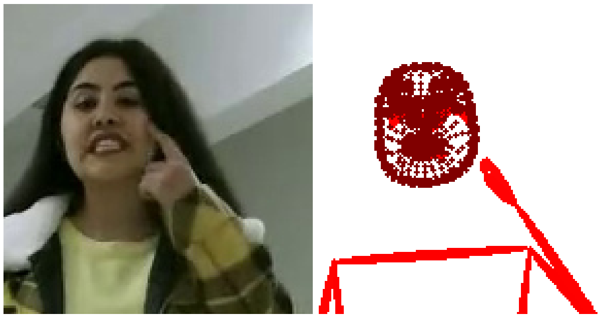

Figure 5. Example of hand-face interaction, where the pose estimation fails for the interacting hand (Holistic).

them mean that keypoints for the hands are not available at all, since Holistic can omit them if it fails to detect the hand. As a complementary quantitative analysis, we correlate prediction outcomes with the average number of frames where hand pose was present (Figure 6).

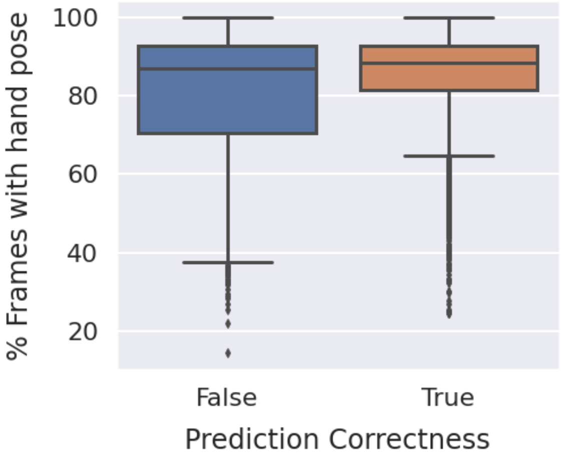

Figure 6. Distribution of percent of frames containing the Holistic pose estimation of the dominant hand in each validation sample, grouped by whether the final prediction of our model was correct.

We find that on average, for all correct predictions the percentage of frames that do contain hand keypoints (85.13%) is significantly higher[2] than for all incorrect predictions (79.78%). This is in line with our qualitative analysis.

6. Conclusions

Although many teams outperformed our models that use only off-the-shelf skeletal representations, with the best submission reaching 98.4% test set accuracy, it is unclear how well such approaches will generalise to other datasets. Our initial questions related to how good skeletal representations are for recognition, given their natural ability to generalise. However, performance in the ChaLearn challenge suggests that despite their benefits, considerable information is lost in the skeletal representation that must be represented in the image domain. A qualitative analysis of our models' failure cases shows that pose estimation tools suffer from shortcomings when body parts interact. We conclude that pose estimation tools are not immediately applicable for the use in sign language recognition – the current representations are not sufficiently expressive, and that further improvements with regard to interacting body parts is crucial for their applicability.

[2] We tested for a significant difference of the mean values with a Wilcoxon rank-sum test [42], $p < 0.0001$.

References

[1] Nikolas Adaloglou, Theocharis Chatzis, Ilias Papastratis, Andreas Stergioulas, Georgios Th Papadopoulos, Vassia Zacharopoulou, George J Xydopoulos, Klimnis Atzakas, Dimitris Papazachariou, and Petros Daras. A comprehensive study on sign language recognition methods. *arXiv preprint arXiv:2007.12530*, 2020. 2

[2] Samuel Albanie, Gül Varol, Liliane Momeni, Triantafyllos Afouras, Joon Son Chung, Neil Fox, and Andrew Zisserman. BSL-1K: Scaling up co-articulated sign language recognition using mouthing cues. In *ECCV*, 2020. 2, 3

[3] Necati Cihan Camgöz, Simon Hadfield, Oscar Koller, and Richard Bowden. Subunets: End-to-end hand shape and continuous sign language recognition. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3075–3084. IEEE, 2017. 2

[4] Necati Cihan Camgöz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. Neural sign language translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7784–7793, 2018. 2

[5] Necati Cihan Camgöz, Oscar Koller, Simon Hadfield, and Richard Bowden. Multi-channel transformers for multi-articulatory sign language translation. In *European Conference on Computer Vision*, pages 301–319, 2020. 2

[6] Necati Cihan Camgöz, Oscar Koller, Simon Hadfield, and Richard Bowden. 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*, pages 10023–10033, 2020. 2, 3

[7] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 1, 2

[8] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017. 2

[9] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition. *A new model and the kinetics dataset. CoRR, abs/1705.07750*, 2(3):1, 2017. 3

[10] Xiujuan Chai, Guang Li, Yushun Lin, Zhihao Xu, Yili Tang, Xilin Chen, and Ming Zhou. Sign Language Recognition and Translation with Kinect. In *Proceedings of the International Conference on Automatic Face and Gesture Recognition (FG)*, 2013. 2

[11] James Charles, Tomas Pfister, Mark Everingham, and Andrew Zisserman. Automatic and Efficient Human Pose Estimation for Sign Language Videos. *International Journal of Computer Vision (IJCV)*, 110(1), 2014. 2

[12] Yu Chen, Chunhua Shen, Xiu-Shen Wei, Lingqiao Liu, and Jian Yang. Adversarial posenet: A structure-aware convolutional network for human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1212–1221, 2017. 1

[13] Helen Cooper and Richard Bowden. Sign Language Recognition using Linguistically Derived Sub-Units. In *Proceedings of the 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*, 2010. 2

[14] Helen Cooper, Brian Holt, and Richard Bowden. Sign Language Recognition. In *Visual Analysis of Humans*. Springer, 2011. 2

[15] Runpeng Cui, Hu Liu, and Changshui Zhang. A deep neural framework for continuous sign language recognition by iterative training. *IEEE Transactions on Multimedia*, 21(7):1880–1891, 2019. 2

[16] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2

[17] Biyi Fang, Jillian Co, and Mi Zhang. DeepASL: Enabling Ubiquitous and Non-Intrusive Word and Sentence-Level Sign Language Translation. In *Proceedings of the ACM Conference on Embedded Networked Sensor Systems (SenSys)*, 2017. 2

[18] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376, 2006. 2

[19] Ivan Grishchenko and Valentin Bazarevsky. Mediapipe holistic. https://ai.googleblog.com/2020/12/mediapipe-holistic-simultaneous-face.html, 2020. 1, 2, 3

[20] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7297–7306, 2018. 1, 2

[21] Gines Hidalgo, Yaadhav Raaj, Haroon Idrees, Donglai Xiang, Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Single-network whole-body pose estimation. In *ICCV*, 2019. 2, 3

[22] Jie Huang, Wengang Zhou, Qilin Zhang, Houqiang Li, and Weiping Li. Video-based sign language recognition without temporal segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. 2

[23] Sang-Ki Ko, Chang Jo Kim, Hyedong Jung, and Choongsang Cho. Neural Sign Language Translation based on Human Keypoint Estimation. *Applied Sciences*, 9(13), 2019. 2

[24] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 2

[25] Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal, Sept. 2015. Association for Computational Linguistics. 2

[26] Marcos Luzardo, Matti Karppa, Jorma Laaksonen, and Tommi Jantunen. Head Pose Estimation for Sign Language Video. *Image Analysis*, 2013. 2

[27] Amit Moryossef. Sign language datasets. https://github.com/sign-language-processing/datasets, 2021. 2

[28] Amit Moryossef, Ioannis Tsochantaridis, Roee Yosef Aharoni, Sarah Ebling, and Srini Narayanan. Real-time sign-language detection using human pose estimation. 2020. 2

[29] Tomas Pfister, James Charles, Mark Everingham, and Andrew Zisserman. Automatic and Efficient Long Term Arm and Hand Tracking for Continuous Sign Language TV Broadcasts. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2012. 2

[30] Leonid Pishchulin, Arjun Jain, Mykhaylo Andriluka, Thorsten Thorm ä hlen, and Bernt Schiele. Articulated people detection and pose estimation: Reshaping the future. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3178–3185. IEEE, 2012. 1

[31] J. Shotton, Andrew Fitzgibbon, M. Cook, Toby Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time Human Pose Recognition in Parts from Single Depth Images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011. 2

[32] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *CVPR*, 2017. 2

[33] Ozge Mercanoglu Sincan, Julio C. S. Jacques Junior, Sergio Escalera, and Hacer Yalim Keles. Chalearn LAP large scale signer independent isolated sign language recognition challenge: Design, results and future research. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2021. 1

[34] Ozge Mercanoglu Sincan and Hacer Yalim Keles. Autsl: A large scale multi-modal turkish sign language dataset and baseline methods. *IEEE Access*, 8:181340–181355, 2020. 2

[35] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019. 3

[36] Hamid Vaezi Joze and Oscar Koller. Ms-asl: A large-scale data set and benchmark for understanding american sign language. In *The British Machine Vision Conference (BMVC)*, September 2019. 2

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 2

[38] Christian Vogler and Siome Goldenstein. Analysis of facial expressions in american sign language. In *Proc, of the 3rd Int. Conf. on Universal Access in Human-Computer Interaction, Springer*, 2005. 1

[39] Ulrich Von Agris and Karl-Friedrich Kraiss. Towards a video corpus for signer-independent continuous sign language recognition. *Gesture in Human-Computer Interaction and Simulation, Lisbon, Portugal, May*, 11, 2007. 2

[40] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *CVPR*, 2016. 2

[41] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional Pose Machines. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[42] Frank Wilcoxon. Individual comparisons by ranking methods. In *Breakthroughs in statistics*, pages 196–202. Springer, 1992. 5

[43] Zahoor Zafrulla, Helene Brashear, Thad Starner, Harley Hamilton, and Peter Presti. American Sign Language Recognition with the Kinect. In *Proceedings of the ACM International Conference on Multimodal Interfaces (ICMI)*, 2011. 2