Towards Accessible Sign Language Assessment and Learning

NEHA TARIGOPULA, Idiap Research Institute , Switzerland and Ecole polytechnique fédérale de Lausanne, Switzerland
SANDRINE TORNAY, Idiap Research Institute, Switzerland
SKANDA MURALIDHAR, Idiap Research Institute, Switzerland
MATHEW MAGIMAI.-DOSS, Idiap Research Institute, Switzerland

Recently, a phonology-based sign language assessment approach has been proposed using sign language production acquired in 3D space using Kinect sensor. In order to scale the sign language assessment system to realistic application, there is need to reduce the dependency on Kinect, which is not accessible to wider community, and develop solutions that can potentially work with web-cameras. This paper takes a step in that direction by investigating sign language recognition and sign language assessment in 2D space either by dropping the depth coordinate in Kinect or using methods for skeleton estimation from videos. Experimental studies on Swiss German Sign Language corpus SMILE show that, while loss of depth information leads to considerable drop in sign language recognition performance, high level of sign language assessment performance can still be obtained.

CCS Concepts: • **Human-centered computing → Gestural input**; • **Applied computing → E-learning**.

Additional Key Words and Phrases: Sign language processing, sign language assessment, hidden Markov models, human skeleton estimation

ACM Reference Format:
Neha Tarigopula, Sandrine Tornay, Skanda Muralidhar, and Mathew Magimai.-Doss. 2022. Towards Accessible Sign Language Assessment and Learning. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '22), November 7–11, 2022, Bengaluru, India.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3536221.3556623

1 INTRODUCTION

Sign language (SL) is the primary mode of communication for the deaf/hard-of-hearing (DHH) community. It is a visual mode of communication that conveys information through various channels such as handshape, hand movement, facial expression, body posture etc. It is used by the DHH community to communicate with both Hearing and DHH community. The research in SL processing has mainly focused on recognition [11], translation [8, 9] and generation. Most of the recognition approaches are based on data acquisition either from sensors such as gloves and accelerometers [15, 35] or camera based systems [17–19]. Considering the sequential nature of the problem and the similarity to speech recognition, in the early research, Hidden Markov Models (HMM) played a major role in modelling the signs. Taking inspiration from speech, HMMs have been widely applied to isolated sign recognition [11, 22, 28, 33, 34]. With the advances of Deep Learning methods in the vision community with respect to body pose estimation and spatio-temporal representations, convolutional networks and recurrent neural networks have been swiftly adapted for Sign Language

*This work was funded by the SNSF through the Sinergia project SMILE-II(Scalable Multimodal Sign Language Technology for Sign language Learning and Assessment), grant agreement CRSII2_160811

Manuscript submitted to ACM

Recognition (SLR) [7, 19, 21, 31] and SL generation [25, 27]. As each SL has its own grammar, more recently there has been thrust towards development of approaches that integrate translation between spoken language representation and "gloss" representation in an end-to-end manner [8, 25].

One of the ways to bridge the communication gap between the DHH community and the Hearing community is to develop assistive technology that can help people, irrespective of whether they are hearing impaired to learn sign language, assess and improve themselves with the help of automatic systems that provide meaningful feedback. In that direction, there has been effort for more than a decade in developing interactive sign language learning platforms for both children and adults [4–6, 26, 38]. Most of the existing platforms for sign language learning/assessment test vocabulary via pre-recorded videos for later analysis. E-learning platforms such as SignAssess [10] allows to compare a user's recorded video to a reference one. In terms of real-time SL verfication, SignAll [32] and ISARA [1] applications assess if the produced sign is correct or incorrect. Validation in terms of whether the produced sign is correct or incorrect is not enough information to aid a SL learner. On the linguistics side, Willoughby *et al.* envisioned a system, My Interactive Auslan Coach [36], that provides automatic feedback on correctness of handshape and hand movement of Australian Sign Language signs. In [14], Huenerfauth *et al.* proposed a system that analyses the production of signs and gives feedback with respect to both manual and non-manual components. These two systems just depict the prototype of the feedback system and analyse the system in terms of usability through wizard-of-oz setup.

Our work is taking place in the context of development of SL production assessment system for SL learning. In that regard, automatically assessing whether a produced sign is correct or incorrect by itself would not be sufficient, as sign language conveys information in parallel through multiple channels, i.e., handshape, hand movement, facial expression etc. Therefore, providing channel-wise feedback would be beneficial. In that direction, in a recent work [29], a "phonology" based SL production assessment approach was proposed, where both lexeme-level assessment (assessing whether a sign production is acceptable or not) and form-level assessment (assessing channel-wise errors) can be carried out in a seamless manner. In [29], the framework was validated with skeleton data acquired through Kinect 3D camera. Dependency on the Kinect sensor makes the framework inaccessible to a wider community, so, it is desirable to have a system that can operate on RGB images, such as captured through a web-camera. This paper investigates that with the following research questions: (i) RQ1 - what is the impact of loss of depth information on SL production assessment? (b) RQ2 - can the impact of loss of depth information on SL production assessment be handled through 3D joint estimation techniques from RGB videos? and (c) RQ3 - SL assessment framework inherently involves sign language recognition components. So, a question that arises is: with loss of depth information, what is the trade-off between SL production assessment performance and SLR performance? We investigate these research questions through a study on Swiss German Sign Language database SMILE DSGS [12].

The paper is organized as follows. Section 2 provides a background on the SL assessment approach. Section 3 presents the experimental setup and Section 4 presents the results and analysis.

2 BACKGROUND

In this section, we provide a brief overview on the phonology-based sign language assessment framework that was developed in [29]. The framework consists of two phases: training phase and assessment phase.

In the training phase, "subunits" corresponding to the different channels f such as, hand movement (denoted as *hmvt*), handshape (denoted as *hshp*) are jointly modeled through HMMs [31]. More precisely, as illustrated in Figure 1 , this is done by estimating the posterior probability of the subunits for each channel $f \in \{hshp, hmvt, \dots\}$ given the

visual signal $(\mathbf{v}_1, \dots \mathbf{v}_t, \dots \mathbf{v}_T)$. Mathematically, this is represented as

$$\mathbf{z}_{t,f} = [P(vs_f^1|\mathbf{v}_t) \cdots P(vs_f^d|\mathbf{v}_t) \cdots P(vs_f^{D_f}|\mathbf{v}_t)]^{\mathrm{T}} \qquad f \in \{hshp, hmvt, \dots\},$$

where vs_f^d denotes visual sub-unit d corresponding to channel f. The posterior probability of the subunits corresponding to the different channels are stacked and used as feature observation $\mathbf{z}_t = [\mathbf{z}_{t,hshp}\ \mathbf{z}_{t,hmvt}, \dots]^{\mathrm{T}}$ for an HMM, whose states are parameterized by categorical distributions $\mathbf{y}_i = [\mathbf{y}_{i,hshp}\ \mathbf{y}_{i,hmvt}, \dots]^{\mathrm{T}}$, for $i \in \{1, \dots, I\}$ where I is the number of HMM states. The parameters of the HMM are estimated by optimizing a cost based on Kullback-Leibler (KL) divergence. This HMM is referred to as Kullback Leibler divergence based HMM (KL-HMM) [2, 3].

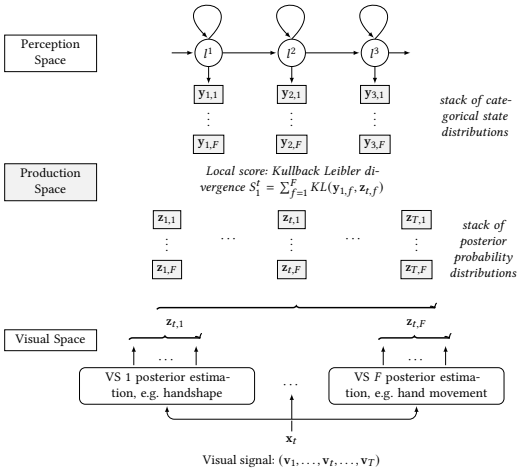


Fig. 1. Illustration of modeling production and perception phenomena in KL-HMM framework for sign language processing [31]. The visual signal is denoted by $(v_1, v_2 \dots v_T)$, $[z_{1,1} \dots z_{t,f} \dots z_{T,F}]$ is the stack of posterior estimates of F channels obtained from the visual signal, and the emission distribution for HMM state i is parameterized by the categorical distribution $[y_{i,1} \dots y_{i,f} \dots y_{i,F}]$.

In the assessment phase, a test sign production is matched against an expected reference sign production. As illustrated in Figure 2, this is done by first matching the KL-HMM corresponding to the reference sign, i.e. the sign that was expected to be produced, with the stacked posterior feature sequence estimated from the visual signal of test sign production using dynamic time warping with local score based on symmetric KL-divergence. Lexeme-level assessment is carried out by applying a threshold on the path length normalized global score $S(N, T)$. Form-level assessment, i.e. assessment of the different channels, is carried by factoring out the score of each channel from the global score and applying a threshold on the resulting channel-wise score. For further details, the reader is referred to [29]. This assessment framework as such can integrate all the channels in sign language. In this work, we limit ourselves to handshape and hand movement channels, as reliably estimating other channels such as, facial expression, mouthing information is still an open research.

3 EXPERIMENTAL SETUP

To address the research questions posed in Section 1, we used SMILE DSGS database, as it is the only linguistically annotated SL dataset available for sign language assessment system development. In the remainder of the section,

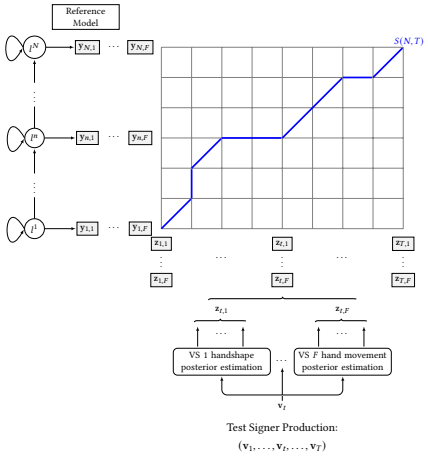


Fig. 2. Illustration of the assessment framework [29]. $[z_{1,1} \ldots z_{t,f} \ldots z_{T,F}]$ is the stack of posterior estimates of F visual sub-units obtained from the test signer production. Each state l_n of the reference KL-HMM model is parameterized by the categorical distribution $[y_{1,1} \ldots y_{n,f} \ldots y_{N,F}]$. The DTW score is given by $S(N,T)$

we first present the SMILE DSGS database. We then present the human skeleton estimation methods that are used to answer the research questions. We then present the SL assessment and SL recognition systems developed for evaluation.

3.1 SMILE DSGS Database

The SMILE DSGS [1] dataset[12] was created for the development of an assessment system for lexical signs of Swiss German Sign Language. It is the *only* dataset that is available in the context of linguistically annotated sign language dataset that aids the development of production level SL assessment. The database has 100 lexical signs of DSGS vocabulary performed three times by 11 adult L1 signers and 19 adult L2 signers. The videos were collected with the Microsoft Kinect v2 sensor, the dataset includes both RGB and depth data obtained by the Kinect v2 sensor and the gloss(meaning label associated to the sign in related spoken language) annotations. Figure 3 shows a snippet of Swiss-German signs from the SMILE dataset.

Fig. 3. Video snippets from SMILE dataset. The first row corresponds to the sign VON, second row corresponds to the sign BLAU and the third row corresponds to the sign EINVERSTANDEN

In addition to the gloss, a category annotation developed by linguistic experts is also provided for each sign. The category of sign production marked 1 through 6, evaluates the acceptance of the sign produced according to whether it

[1]DSGS stands for DeutschSchweizerische GebärdenSprache

is the same lexeme(word or words) as target sign, has the same meaning as target sign and whether it has same form (shape, movement) as target sign. (1) *Category 1* - Same lexeme as target sign: same meaning, same form, (2) *Category 2* - Same lexeme as target sign: same meaning, slightly different form, (3) *Category 3* - Same lexeme as target sign: same meaning, different form, (4) *Category 4* - Same lexeme as target sign: slightly different meaning, slightly different form, (5) *Category 5* - Different lexeme than target sign: same meaning, different form and (6) *Category 6* - Different lexeme than target sign: different meaning, different form. Category 1 and 2 are the linguistically-valid acceptable sign productions and are used to build the components of the assessment framework. The acceptable sign productions of category 1 and 2 were partitioned into 1125 training samples from 15 signers, 581 test samples from 8 signers and 509 development samples from 7 signers. There are 412 samples corresponding to category 3 and 4, and 183 samples corresponding to category 5 and 6. *We used exactly the same dataset protocol as in [29] for sign language assessment and sign language recognition studies*. For the details, the reader is referred to [29].

3.2 Human skeleton estimation

To investigate the research questions, we used only the RGB videos from the SMILE dataset. To investigate RQ1, we used 2D joint pose estimation. More precisely, we used OpenPose [39] and MaskRCNN [13] based key-point estimation, referred to as "OpenPose2D" and "MaskRCNN" in our experiments. MaskRCNN extends faster-RCNN [24] to first obtain bounding boxes of the detected objects/persons and then adding a branch of instance segmentation over it. It treats the problem of key-point estimation as instance segmentation by viewing each key-point as a one-hot encoded binary mask, which helps to identify instance based poses. To investigate RQ2, we applied methods to estimate 3D skeleton from RGB videos without any depth information, namely, VideoPose3D [23] and Vibe3D [16], referred to as "VideoPose3D" and "Vibe3D" in our experiments. VideoPose3D operates by first predicting 2D key-points from images using MaskRCNN [13] and then the 3D key-points of the video are estimated through temporal convolutions over the 2D key-points. Owing to lack of 3D annotated motion data, Vibe leverages adversarial learning to discriminate real human motions from the ones generated by temporal shape and pose regressors to predict reliable 3D SMPL [20] human models from in-the-wild videos. We use the pretrained models available for all the methods.

3.3 Posterior feature vector estimation

To extract handshape posterior feature $\mathbf{z}_{t,hshp}$, as done in [29], we used the pre-trained handshape classifier neural network SubUNets [7]. It uses a ResNeXt-101 [37] based model for handshape classification. The first classifier was trained on the One-Million-Hands dataset [18] that uses only the top 30 commonly occurring handshapes out of 60. In addition to this, a second classifier was trained to classify the 30 handshapes and a transition shape. The input to neural network is hand patches obtained by applying Openpose [39] 2D pose estimation method to localize the wrist and cropping hand patches from the image using the wrist coordinates. The output is handshape class-conditional posterior probability vector $\mathbf{z}_{t,hshp}$.

To estimate hand movement posterior features $\mathbf{z}_{t,hmvt}$, we used the same procedure as in [30] and [29]. More precisely, this involves two steps, (i) **hand movement subunit inference**: a sequence of hand movement feature vectors based on skeletal information are extracted for each sign. The feature vector consists of data corresponding to the coordinates of left and right hands relative to head, hip and shoulder and their velocities. To normalize the variation in between signers, the neck joints of all the signers are aligned with respect to a randomly chosen signer and scaled by the shoulder width. Given the sequence of features, the hand movement subunits are inferred by training left-to-right HMMs with different number of states for each of the signs and selecting the number of states that yields

best performance on a development set. (ii) **hand movement subunit posterior probability ($\mathbf{z}_{t,hmvt}$) estimator training**: this is done by obtaining an alignment of the features with the HMM states and training of a multilayer perceptron (MLP) to classify the HMM states with a cost function based on cross entropy. The architecture of the MLP is determined in a cross validation manner.
It is worth mentioning that loss of depth information affects hand movement subunits posterior estimation not handshape.

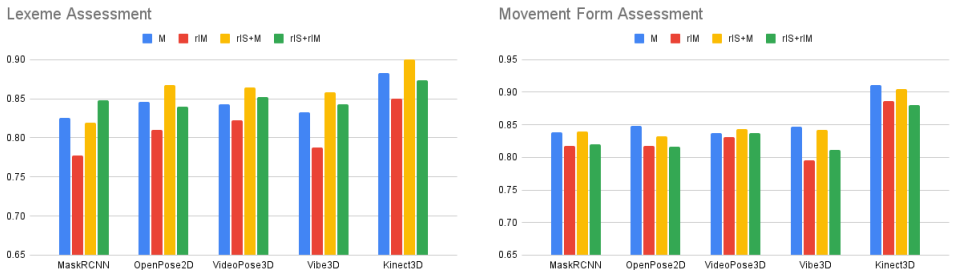
3.4 KL-HMM systems

As done in [29], we trained different whole sign KL-HMM models for each sign, namely, (i) **M**: only the hand movement sub-units obtained from both right and left hands(combined) are modelled. (ii) **rlM**: the hand movement sub-units are obtained separately for left and right hands, and the two set of hand movement posteriors obtained from two MLPs are stacked and modelled using KL-HMM and, (iii) **M+rlS** and (iv) **rlM+rlS** are systems that use the concatenation of hand movement and hand-shape probability posteriors in accordance with the systems described above. The number of states is determined by training KL-HMMs with 3 to 30 states on acceptable sign production data, and selecting the model that yielded best recognition accuracy on the development set. For SL assessment studies, similar to [29], the thresholds for lexeme assessment , hand movement form assessment and handshape form assessment are determined from the development set consisting of data from Category 1 and 2. This is done by creating a set of correct sign scores by matching the same sign instances and by creating a set of incorrect sign scores by matching instances from different signs. The thresholds are then set as the ones that yielded the best F_1 score for lexeme and form assessment on the development data. Besides SL assessment studies, to address RQ3, we carried out SL recognition studies.

4 RESULTS AND ANALYSIS

In this section, we present and analyse the experiment results that were performed to address the research questions presented earlier. We show the assessment results that address RQ1 and RQ2 with respect to various 2D and 3D skeleton estimation methods used in Figure 4. Since the handshape modelling is the same across the different methods, we only focus on the lexeme and hand movement form assessment in our results. We compare the F1 scores of lexeme and hand movement form assessment of the skeleton estimation methods used against the results obtained by training the framework using data from Kinect sensor (referred to as Kinect3D) as in [29]. We obtain similar results for Kinect3D as given in [29]. We can see that Kinect3D performs the best in all cases, as it has additional true depth information. Nevertheless, the drop in F1 scores from moving from Kinect skeleton to other skeleton estimation methods is not very high. The best assessment performance in 2D space is obtained by MaskRCNN in 3D space by VideoPose3D.

To analyse the statistical significance of the two best performing methods, we perform a McNemar's test on the lexeme and movement assessment results. We observed that the lexeme assessment is not significantly different, but the hand movement assessment is significantly different at 95%, owing to the differences in 2D and 3D movement features. To further analyse the movement form assessment results of MaskRCNN and VideoPose3D, we split the signs in the test-set based on the direction of hand movement in the signs. We identified 5 scenarios: (i) x,y and z, (ii) x, y (iii) z (iv) in place wrist rotation (v) static sign. For example, in Figure 3, the first sign has movement focused only in z direction, the second sign in xyz direction and the third sign in xy direction. For each of the scenarios, we calculate the percentage of correctly identified assessments. In Table 1 we present the results only for the best reference system(rlM+rlS), similar trends are observed across all the reference systems, we only present results on rlM+rlS in this paper, due to lack of space. From the table, we can observe that 3D estimation for videos is beneficial compared to 2D estimation in terms of movement assessment.



(a) Lexeme Assessment (b) Movement Form Assessment

Fig. 4. SL assessment F1 scores

Table 1. Percentage of correctly identified movement assessment in signs across different directions of movement

	Movement Direction				
	xy	**xyz**	**z**	**rotation**	**static**
MaskRCNN	66.99	51.94	61.32	64.13	50
VideoPose3D	**78.47**	**74.42**	**73.58**	**81.52**	**59.09**

Table 2. SL recognition accuracy

	MaskRCNN	**OpenPose**	**VideoPose3D**	**Vibe3D**	**Kinect3D**
Recognition Acc	**69.46**	65.38	65.92	62.59	75.6

To address RQ3, we used the developed KL-HMM reference models to perform isolated sign language recognition studies. Table 2 presents the SL recognition accuracy of the rlM+rlS reference system. As in assessment, Kinect3D outperforms all the other methods because of true depth information. It is worthwhile to note that MaskRCNN results in the better recognition accuracy as opposed to VideoPose3D which gave better assessment performance, this is true across different reference systems, even though we report the accuracy only on rlM+rlS. This observation is inline with the one in [5], where the trend in performance on acceptability ratings of SL is not correlated with the SL recognition performance. Overall, the drop in recognition accuracy is more than the drop in assessment when moving away from Kinect sensor based data.

5 CONCLUSION

In this paper, we investigated the impact of carrying out automatic sign language assessment in 2D space as opposed to 3D space, to make the assessment system accessible for all. Our investigations on SMILE corpus show that moving from 3D space to 2D space leads to an acceptable drop in sign language assessment performance. We also observe that the model that gives the best assessment performance does not automatically translate to best recognition performance. This indicates that, although the trained KL-HMM models loose discrimination across signs, they still are able to capture information related to acceptable sign production. Overall, the experimental studies provide promising direction to move on from Kinect capturing systems to more readily available capturing systems like web-cameras. Our ongoing work is focusing in that direction. More precisely, remote sign language assessment system development using web-cameras.

REFERENCES

[1] ISARA application. 2016. *ISARA app.* Retrieved May 2022 from https://isara.app/features

[2] G. Aradilla, H. Bourlard, and M. Magimai.-Doss. 2008. Using KL-Based Acoustic Models in a Large Vocabulary Recognition Task . In *Proceedings of Interspeech*. 928–931.

[3] G. Aradilla, J. Vepa, and H. Bourlard. 2007. An Acoustic Model Based on Kullback-Leibler Divergence for Posterior Features. In *ICASSP*. 657–660.

[4] O. Aran et al. 2009. SignTutor: An Interactive System for Sign Language Tutoring. *IEEE MultiMedia* 16, 1 (2009), 81–93.

[5] J. Arendsen et al. 2008. Acceptability ratings by humans and automatic gesture recognition for variations in sign productions. In *Proc. of IEEE International Conference on Automatic Face and Gesture Recognition*. 1–6.

[6] H. Brashear et al. 2006. American Sign Language Recognition in Game Development for Deaf Children. In *Proc. of the International ACM SIGACCESS Conference on Computers and Accessibility*. 79–86.

[7] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, and Richard Bowden. 2017. SubUNets: End-to-End Hand Shape and Continuous Sign Language Recognition. In *2017 IEEE International Conference on Computer Vision (ICCV)*. 3075–3084. https://doi.org/10.1109/ICCV.2017.332

[8] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018. Neural Sign Language Translation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7784–7793. https://doi.org/10.1109/CVPR.2018.00812

[9] Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020. Sign Language Transformers: Joint End-to-end Sign Language Recognition and Translation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[10] J. Christopher. 2012. SignAssess – Online Sign Language Training Assignments via the Browser, Desktop and Mobile. In *Computers Helping People with Special Needs*, Klaus Miesenberger, Arthur Karshmer, Petr Penaz, and Wolfgang Zagler (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 253–260.

[11] H. Cooper, B. Holt, and R. Bowden. 2011. Sign Language Recognition. In *Visual Analysis of Humans, 2011.* https://doi.org/10.1007/978-0-85729-997-0_27

[12] S. Ebling, N. C. Camgöz, P. Boyes Braem, K. Tissi, S. Sidler-Miserez, S. Stoll, S. Hadfield, T. Haug, R. Bowden, S. Tornay, M. Razavi, and M. Magimai-Doss. 2018. SMILE Swiss German sign language dataset. In *Proc. of the Language Resources and Evaluation Conference*.

[13] K. He, G. Gkioxari, P. Dollár, and R. Girshick. 2017. Mask R-CNN. In *Proc. of ICCV*. 2980–2988.

[14] Matt Huenerfauth, Elaine Gale, Brian Penly, Sree Pillutla, Mackenzie Willard, and Dhananjai Hariharan. 2017. Evaluation of Language Feedback Methods for Student Videos of American Sign Language. *ACM Trans. Access. Comput.* 10, 1, Article 2 (apr 2017), 30 pages. https://doi.org/10.1145/3046788

[15] Mohammed Kadous. 1996. Machine Recognition of Auslan Signs Using PowerGloves: Towards Large-Lexicon Recognition of Sign Language. In *Procs. of Wkshp : Integration of Gesture in Language and Speech*.

[16] M. Kocabas, N. Athanasiou, and M. J. Black. 2020. VIBE: Video Inference for Human Body Pose and Shape Estimation. In *Proc. of CVPR*.

[17] Oscar Koller, Jens Forster, and Hermann Ney. 2015. Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *Computer Vision and Image Understanding* 141 (Dec. 2015), 108–125.

[18] O. Koller, H. Ney, and R. Bowden. 2016. Deep Hand: How to Train a CNN on 1 Million Hand Images When Your Data Is Continuous and Weakly Labelled. In *Proc. of CVPR*.

[19] O Koller, O Zargaran, H Ney, and R Bowden. 2016. Deep Sign: Hybrid CNN-HMM for Continuous Sign Language Recognition. In *Proceedings of the British Machine Vision Conference 2016*.

[20] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. 2015. SMPL: A Skinned Multi-Person Linear Model. *ACM Trans. Graph.* 34, 6, Article 248 (oct 2015), 16 pages. https://doi.org/10.1145/2816795.2818013

[21] Katerina Papadimitriou and Gerasimos Potamianos. 2020. Multimodal Sign Language Recognition via Temporal Deformable Convolutional Sequence Learning. In *INTERSPEECH*.

[22] VN Pashaloudi and KG Margaritis. 2002. Hidden Markov model for sign language recognition: A review. In *Proc. 2nd Hellenic Conf. AI, SETN-2002*. 11–12.

[23] D. Pavllo, C. Feichtenhofer, D. Grangier, and M. Auli. 2019. 3D human pose estimation in video with temporal convolutions and semi-supervised training. In *Proc. of CVPR*.

[24] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1* (Montreal, Canada) *(NIPS'15)*. MIT Press, Cambridge, MA, USA, 91–99.

[25] Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2021. Continuous 3D Multi-Channel Sign Language Production via Progressive Transformers and Mixture Density Networks. *Int. J. Comput. Vision* 129, 7 (jul 2021), 2113–2135.

[26] G. Spaai et al. 2005. Elo: An electronic learning environment for practising sign vocabulary by young deaf children. In *Proc. of International Congress for Education of the Deaf*.

[27] Stephanie Stoll, Necati Cihan Camgöz, Simon Hadfield, and Richard Bowden. 2020. Text2Sign: Towards Sign Language Production Using Neural Machine Translation and Generative Adversarial Networks. *Int. J. Comput. Vis.* 128, 4 (2020), 891–908.

[28] Sandrine Tornay, Oya Aran, and Mathew Magimai.-Doss. 2020. An HMM Approach with Inherent Model Selection for Sign Language and Gesture Recognition. In *Proc. of the International Conference on Language Resources and Evaluation LREC 2020*.

[29] S. Tornay, N. C. Camgoz, R. Bowden, and M. Magimai.-Doss. 2020. A Phonology-based Approach for Isolated Sign Production Assessment in Sign Language. In *Companion Publication of the 2020 International Conference on Multimodal Interaction (ICMI '20 Companion)*.

[30] S. Tornay and M. Magimai.-Doss. 2019. Subunits Inference and Lexicon Development Based on Pairwise Comparison of Utterances and Signs. *Information* 10 (2019). https://doi.org/10.3390/info10100298

[31] S. Tornay, M. Razavi, N. C. Camgoz, R. Bowden, and M. Magimai.-Doss. 2019. HMM-based Approaches to Model Multichannel Information in Sign Language inspired from Articulatory Features-based Speech Processing. In *Proc. in the IEEE ICASSP*.

[32] SignAll Technologies Inc. (USA). 2021. *A communication bridge between deaf and hearing - SIGNALL.* Retrieved May 2022 from https://www.signall.us

[33] C. Vogler and D. Metaxas. 1998. ASL recognition based on a coupling between HMMs and 3D motion analysis. *Procs. of ICCV*, 363–369.

[34] C. Vogler and D. Metaxas. 1999. Parallel hidden Markov models for American sign language recognition. In *Proc. of the Seventh IEEE International Conference on Computer Vision (ICCV)*, Vol. 1. 116–122 vol.1. https://doi.org/10.1109/ICCV.1999.791206

[35] M.B. Waldron and Soowon Kim. 1995. Isolated ASL sign recognition system for deaf persons. *IEEE Transactions on Rehabilitation Engineering* 3, 3 (1995), 261–271. https://doi.org/10.1109/86.413199

[36] Louisa Willoughby, Stephanie Linder, Kirsten Ellis, and Julie Fisher. 2015. Errors and Feedback in the Beginner Auslan Classroom. *Sign Language Studies* 15 (2015), 322 – 347.

[37] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He. 2016. Aggregated Residual Transformations for Deep Neural Networks. *arXiv preprint arXiv:1611.05431* (2016).

[38] Z. Zafrulla et al. 2011. CopyCat: An American Sign Language game for deaf children. In *IEEE International Conference on Automatic Face and Gesture Recognition*.

[39] C. Zhe, S. Tomas, W. Shih-En, and S. Yaser. 2017. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. In *Proc. of CVPR*.