Towards Dynamic Skeleton-based Handshape Subunits for Sign Language Assessment

Sandrine Tornay
Idiap Research Insitute
Martigny, Switzerland
sandrine.tornay@idiap.ch

Mathew Magimai.-Doss
Idiap Research Insitute
Martigny, Switzerland
mathew@idiap.ch

***Abstract*—Sign languages convey information through multiple channels. The handshape channel is an important manual component for conveying the message. In the literature, it is mainly modeled as a sequence of images of discrete postures even in the case of dynamic gestures, leading to blurring problems in detection. Furthermore, to model these discrete postures using deep learning frame-level labeling of the sign language videos is also required, which is time consuming and human intensive. In this paper, as opposed to modeling the handshape information through images of discrete postures, we propose dynamic modeling through skeletal information. More precisely, we develop an approach that combines HamNoSys-based prior knowledge and sign language data to derive dynamic handshape units by modeling skeletal features using hidden Markov models. We demonstrate the effectiveness of the proposed approach through sign language assessment study, sign language recognition, and handshape recognition analysis on the SMILE DSGS corpus.**

***Index Terms*—Handshape subunits, skeleton-based feature, sign language assessment, hidden Markov models.**

I. Introduction

Sign languages are visual mode of communication that involves multiple channels of information to convey meaning, namely handshape, hand movement, body posture, facial expression, mouth movement, and mouthing. While the manual information is treated as primitive components for isolated signs, the non-manuals are necessary in sentence-level signing. In both cases, the handshape channel plays an important discriminative role in the manual aspect. It is therefore the most considered feature in sign language processing [1]. Moreover, it is important to note that the manual aspect of sign languages is mainly composed of well-defined dynamic gestures, but also contains static gestures such as finger spelling hand poses. In this paper, our focus lies on effective modeling and assessment of the handshape component in sign language production.

In the literature, the handshape component is typically treated and processed as discrete subunits where the temporal relationship in-between the subunits is modeled implicitly, such as through sign-level modeling or through prior knowledge-based handshape subunit estimation itself. In the first case, image-based handshape subunits are extracted and temporal approaches such as Hidden Markov Models (HMM) [2]–[4] or Long Short-Term Memory (LSTM) are used [5]–[7] to model the sign and integrate the temporal relationship. In the second case, dynamic modeling is integrated at the subunits estimation level: for that, either 3D Convolutional Neural Networks (CNN) [8]–[10] which extracts spatio-temporal features is used or preprocessing techniques [11], [12] to select relevant posture frames are applied. In [13], the authors developed the Hand SubUNet estimator where two separate systems are used to specifically model the spatial and temporal aspects of the handshape subunits: first CNN and then LSTM and finally Connectionist Temporal Classification (CTC) is applied for sequence-to-sequence classification. The advancement of such deep learning algorithms in image processing have led to promising results; however, to develop such systems frame-level annotations of the sign language videos are needed, which in turn requires sign linguistic expertise and is expensive in terms of time and cost.

In recent years, different tools have emerged for estimating skeletal information such as, OpenPose [14] and MediaPipe [15] and have been successfully used for modeling the hand movement information for sign language assessment [16], [17] and sign language recognition [18], [19]. Moreover, skeletal information gives the possibility of isolating 3D handshape subunits from the hand movement information, leading to explicit morphologically dynamic handshape subunits, especially useful for assessing handshape component independently and potentially providing better feedback. We investigate this aspect by developing a novel hidden Markov model-based approach that explicitly models temporal relationship "within" handshape subunits based on skeletal information, akin to modeling hand movement information based on skeletal information, and validating the proposed approach through sign language assessment, sign language recognition, and handshape recognition studies.

The paper is organized as follows: Section II presents a brief background on the phonology-based sign language processing framework employed in this work. Section III the proposed handshape subunits estimation, Section IV the experimental setup, and Section V present results. Section VI finally concludes the paper.

II. Background

The research and development presented in this paper takes place in the framework of an explainable phonology-based sign language recognition [18] and assessment [16] approach developed to build assistive technology for sign language learning, where feedback on the production of different channels can be provided to the learners [17], [20]. In this framework, as illustrated in the Figure 1, *subunits* corresponding to the different channels f, such as hand movement (denoted as *hmvt*), handshape (denoted as *hshp*), are jointly modeled through hidden Markov models. This is done by: (a) estimating the posterior probability of the visual subunits vs_f: $\mathbf{z}_{t,f} = [P(vs_f^1|\mathbf{v}_t) \cdots P(vs_f^d|\mathbf{v}_t) \cdots P(vs_f^{D_f}|\mathbf{v}_t)]^{\mathrm{T}}$ for each channel $f \in \{hshp, hmvt, \dots\}$, where D_f is the feature dimension of channel f, given the sequence of visual signal frames $(\mathbf{v}_1, \dots \mathbf{v}_t, \dots \mathbf{v}_T)$ and (b) stacking the posterior probability distributions $\mathbf{z}_t = [\mathbf{z}_{t,hshp}\ \mathbf{z}_{t,hmvt}, \dots]^{\mathrm{T}}$ and using them as feature observations for a Kullback-Leibler divergence-based HMM (KL-HMM), whose states are parameterized by categorical distributions $\mathbf{y}_n = [\mathbf{y}_{n,hshp}\ \mathbf{y}_{n,hmvt}, \dots]^{\mathrm{T}}$, for $n \in \{1, \dots, N\}$ where N is the number of HMM states. The HMM parameters are estimated by optimizing a cost function based on Kullback-Leibler divergence [21], [22].

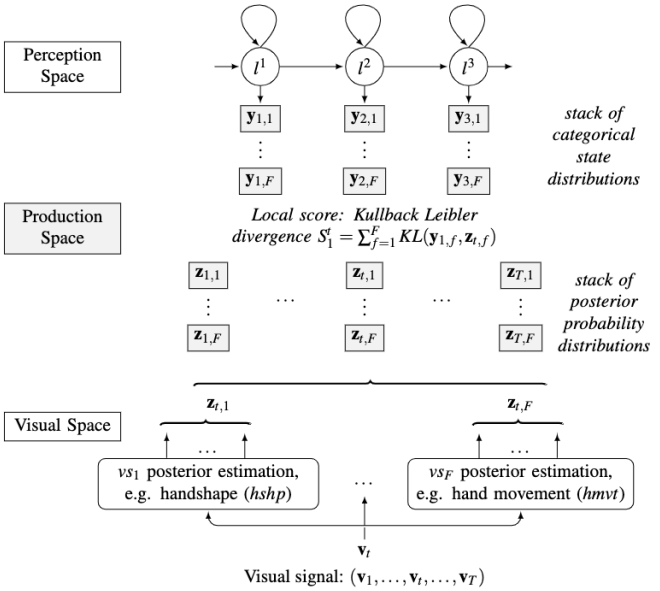

Fig. 1. Illustration of modeling production and perception phenomena in KL-HMM framework for sign language processing [18]. The visual signal is denoted by $(\mathbf{v}_1, \mathbf{v}_2, \ldots \mathbf{v}_T)$, $[\mathbf{z}_{1,1} \ldots \mathbf{z}_{t,f} \ldots \mathbf{z}_{T,F}]$ is the stack of posterior estimates of F channels obtained from the visual signal, and the emission distribution for HMM state n is parameterized by the categorical distribution $[\mathbf{y}_{n,1} \ldots \mathbf{y}_{n,f} \ldots \mathbf{y}_{n,F}]$; here it is a three HMM states example.

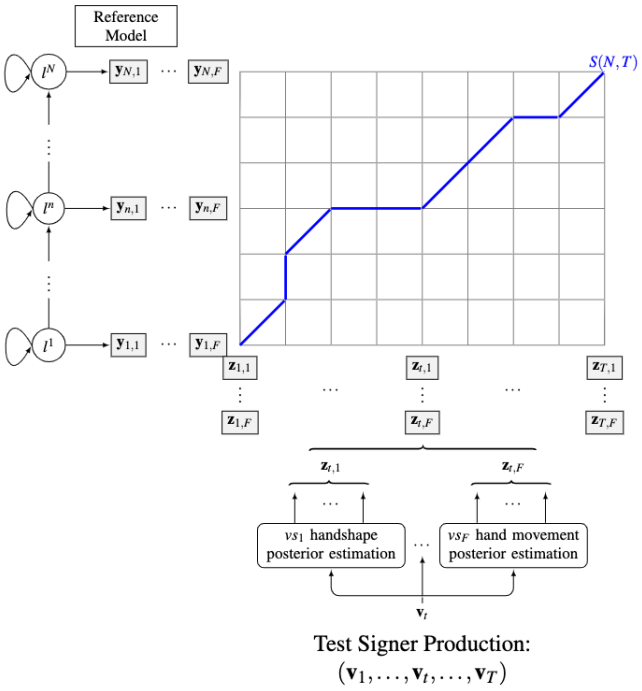

Fig. 2. Illustration of the assessment framework [16]. $[\mathbf{z}_{1,1} \ldots \mathbf{z}_{t,f} \ldots \mathbf{z}_{T,F}]$ is the stack of posterior estimates of F visual subunits obtained from the test signer production. Each state l_n of the reference KL-HMM model is parameterized by the categorical distribution $[\mathbf{y}_{1,1} \ldots \mathbf{y}_{n,f} \ldots \mathbf{y}_{N,F}]$. The DTW score is given by $S(N,T)$.

After training sign-level HMMs, isolated sign language recognition is carried out by decoding the most likely sign at the output. Sign language assessment, as illustrated in Figure 2, is carried out by matching the expected reference sign production with the stacked posterior feature sequence estimated from the visual signal of the test sign production using Dynamic Time Warping (DTW) with local score based on symmetric KL divergence. A threshold is applied to the resulting global score $S(N,T)$ after path-length normalization to carry out sign-level assessment (i.e., whether the produced sign is targeting the correct sign or not). Whilst, form-level assessment (i.e., whether the produced hand movement and handshape are correct or not) is carried out by factoring out each channel's score from the global score and applying a threshold to the resulting channel-wise score. For more details, the reader is referred to [16].

When compared to the previous works [16], [18], the focus of this paper is on modeling the handshape subunits using skeletal information.

III. Proposed Approach

In this section, we present the proposed dynamic handshape subunits modeling based on skeleton information. The proposed method is inspired by the skeleton-based hand movement subunits modeling proposed in [16], [18]. As illustrated in Figure 3, the proposed method consists of three steps:

1) **Skeletal feature extraction:** This is done by extracting 21-dimensional 3D skeletal joints of each hand using tools such as, MediaPipe [15] or OpenPose [14], and aligning the hand skeletons of each frame at the wrist-based coordinate center to remove the hand movement of the produced sign. The dominant and non-dominant hands space[1] are then unified by mirroring the joints of the non-dominant hand skeleton. It is worth mentioning that the hand dominance is preserved at the model level. Finally, since hand joints are highly related (hand-finger structure), a 60-dimensional decorrelated feature vector (20 joints $\times$ 3D) per frame is extracted by applying the Karhunen-Loeve transformation (KLT) without dimensionality reduction. The KLT matrix is estimated on the training set data.
2) **HMM-based handshape subunits inference:** This is done by grouping the dominant and the non-dominant handshapes of each sign in the vocabulary into handshape classes by using the HamNoSys annotation [23]. More precisely, by extracting the handshape symbols from the sign-based HamNoSys annotation which describes how the sign should be produced and grouping the HamNoSys annotations into a set of unique handshapes. To this set of handshape classes, we added a *waiting class* representing the non-dominant hand of the one-handed sign. Each of the handshape classes is then modeled by left-to-right HMM/GMMs (Gaussian Mixture Models) using the skeleton-based features as the feature observation and handshape subunits are inferred through cross-validation. More precisely, by training HMM/GMMs with a fixed number of states and mixtures per state for all the handshape classes and selecting the setup that yields the best handshape classification on the development set. Similar to the case of hand movement subunit extraction and modeling [16], [18], to better segment the beginning and end of the handshape movement, a three-state HMM common to all the handshape classes is added at the beginning and end of the HMM. The resulting HMM states serve as the handshape subunits.
3) **Estimation of handshape subunits posterior probabilities**

[1]In sign languages, the signer usually has a dominant hand (left or right) for producing one-handed signs.

$\mathbf{z}_{t,hshp}$**:** This is done by aligning the handshape HMMs on the training data and training a neural network based classifier that takes the skeleton-based features as input and classifies the HMM states of all handshape classes at the output. The neural network is trained with cross entropy error criterion.

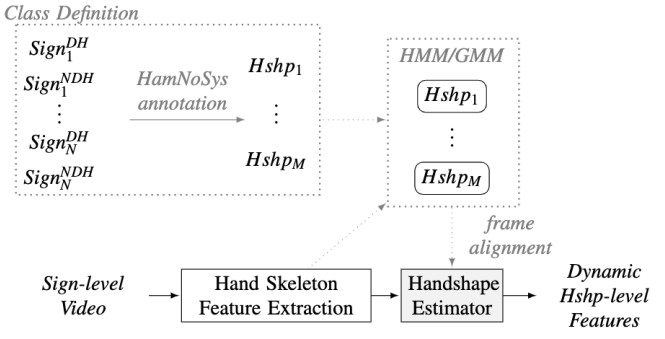

Fig. 3. Illustration of the proposed approach. *Hshp* stands for handshape, *DH* for dominant hand and *NDH* for non-dominant hand.

IV. Experimental Setup

This section presents the experimental setup to validate the proposed skeleton-based handshape subunit extraction approach.

A. SMILE DSGS database

The SMILE Swiss German Sign Language database [24], referred to as SMILE DSGS database, is composed of 100 Swiss German Sign Language (DSGS[2]) signs produced three times by 28 adult signers. The data collection was done using the Microsoft Kinect v2 sensor. In our experimental setup, we used the second pass out of the three which was manually annotated through 6 categories that evaluate the acceptability of a sign production according to linguistic criteria (lexeme/sign, meaning and form) (see 'Category of sign produced' in [24]). Categories 1 and 2, linguistically annotated as acceptable signs, were used to build the different components of the proposed systems and were partitioned in a signer-independent manner into 1125 training set samples from 13 signers, 509 development set samples from 7 signers, and 581 test set samples from 8 signers.

We used a DSGS HamNoSys dictionary to get the sign-based HamNoSys annotations. Extracting handshape symbols for each sign and grouping them resulted in 28 unique handshape classes (Step 2 of the proposed approach).

B. Handshape subunit posterior $\mathbf{z}_{t,hshp}$ estimation

Baseline (Hand SubUNet): We compare the proposed approach against the image-based handshape subunit estimation approach that was employed in [16]. More precisely, we used the off-the-shelf handshape subunits neural network, originally proposed by Camgöz et al. in [13] and trained on the One-Million-Hands dataset [25]. This neural network estimates posterior probabilities for 61 handshape subunits (including a transitional subunit). For a fair comparison, we used the reduced channel of 31 handshape classes, which contains all the handshapes produced in the SMILE DSGS dataset. For more details, the reader is referred to [16].

Proposed: In the case of the proposed skeleton-based approach, in Step 1 we used the MediaPipe [15] estimator to extract the handshape skeletal joints. Step 2 resulted in 141 skeleton-based handshape subunits (5 HMM states $\times$ 28 handshape classes + one transition state). In Step 3, a multilayer perceptron (MLP) classifier with softmax nonlinearity was trained with different numbers of hidden layers (0, 1, 2, 3) and hidden units (600, 800, 1000) using Quicknet software [26]. The MLP that yielded the best frame-level accuracy on the cross-validation data was selected for $\mathbf{z}_{t,hshp}$ estimation. The resulting architecture was an input layer of dimension 540 (60 feature dimension $\times$ (1 + 4 frames preceding + 4 frames following context)) and a softmax output layer of dimension 141.

[2]*Deutschschweizerische Gebärdensprache*

C. Hand movement subunits posterior $\mathbf{z}_{t,hmvt}$ estimation

For the posterior estimation of the hand movement subunits, we implemented the method presented in [16] using the same skeletal joint estimator as for the proposed handshape estimator, i.e., the MediaPipe [15] estimator. Two separate estimators were developed for the dominant and non-dominant hand. For the sake of completeness, we also present experimental studies with Kinect 3D based hand movement subunit posterior estimation, as done in the previous works [16], [18].

D. Sign-level Reference Systems

The sign-level reference systems used in sign language recognition and assessment tasks were implemented using the KL-HMM framework described in Section II. All the KL-HMM systems were trained using 3 to 30 KL-HMM states and the system that yielded the best recognition accuracy on the development set was chosen as the reference. Three different systems were implemented depending on which subunits were stacked and modeled to train the KL-HMM, namely, the **rlS** system which refers only to the handshape subunits, the **rlM** system for the hand movement subunits and the **rlS+rlM** system for both subunits.

We have conducted two different studies to validate the proposed approach, namely, (a) sign language recognition study. In this case, the performance is measured in terms of recognition accuracy and (b) sign language assessement study, where we carry out sign-level assessment and handshape form-level assessment. For both types of assessment, the performance is measured in terms of F_1 score. For setting up the thresholds for the sign-level assessment and the form-level assessment, we followed the same procedure as in [16]. Briefly, on the development data, a set of correct sign scores is obtained by matching the same sign instances and a set of incorrect match scores is obtained by matching instances of different signs, and a threshold that yields the lowest F_1 score on the development set is used.

To evaluate the sign language recognition study, we used the test set composed by categories 1 and 2 according to 'Category of sign produced' in [24]. To evaluate the sign-level assessment, we separated the correct/incorrect test data as the following: categories 1 to 4 were set as correct target signs and categories 5 and 6 as incorrect target signs. Since the incorrect set was unbalanced, we created additional data by matching each sample of the categories 1 and 2 with a randomly selected incorrect reference. To evaluate the handshape form-level assessment, we followed the same setup with only categories 1 and 2 samples as correct test set.

V. Results

In this section, we first present results of the sign language assessment study. We then corroborate the findings of the sign language assessment study through sign language recognition and handshape recognition analysis.

A. Sign language assessment study

We have conducted sign language assessment studies using System **rlS** and combined System **rlS+rlM**, where System **rlM** is based on MediaPipe. Table I presents the F_1 scores of (i) the sign-level assessment which verifies whether the produced sign is targeting the correct reference sign, and (ii) the handshape form-level assessment which verifies whether the produced handshapes are correct. It can

TABLE I
F_1 SCORES OF THE SIGN-LEVEL ASSESSMENT (SIGN) AND THE HANDSHAPE FORM-LEVEL ASSESSMENT (HSHP) USING EITHER THE PROPOSED APPROACH OR THE HAND SUBUNET ESTIMATOR FOR THE HANDSHAPE SUBUNITS ESTIMATION. **RLM** IS BASED ON THE MEDIAPIPE FEATURES

	Sign-level		**Hshp-level**	
	rlS	**rlS+rlM**	**rlS**	**rlS+rlM**
Proposed	0.82	0.88	0.83	0.72
Hand SubUNet	0.76	0.84	0.69	0.64

be observed that the proposed skeleton-based handshape subunit approach yields better systems for both sign-level and handshape form-level assessment. Moreover, the comparison between System **rlS** and **rlS+rlM** shows that the hand movement slightly helps in *sign-level* assessment, while it introduces confusion in *hshp-level* assessment.

B. Sign language recognition analysis

Table II presents the sign language recognition (SLR) accuracy for the baseline Hand SubUNet approach and for the proposed handshape subunit modeling approach. For handshape alone modeling SLR case

TABLE II
SIGN LANGUAGE RECOGNITION ACCURACIES USING EITHER THE PROPOSED APPROACH OR THE HAND SUBUNET ESTIMATOR FOR THE HANDSHAPE SUBUNITS ESTIMATION

	KL-HMM References				
		rlM using MediaPipe		*rlM using Kinect*	
	rlS	**rlM**	**rlS+rlM**	**rlM**	**rlS+rlM**
Proposed	68.0	35.6	80.7	57.1	87.4
Hand SubUNet	30.5		64.0		74.2

(System **rlS**), we can observe that the proposed approach yields significantly better SLR system than Hand SubUNet approach. For hand movement alone modeling case (System **rlM**), modeling skeletal information using Kinect yields better system than using MediaPipe. However, with the proposed approach modeling handshape skeletal information using MediaPipe, System **rlS** yields better system than System **rlM** based on MediaPipe or Kinect, while this is not the case with System **rlS** based on Hand SubUNet. Finally, as observed in previous SLR studies [16], [18], the system combining both handshape and hand movement information, i.e. System **rlS+rlM**, consistently yields better system than modeling either handshape or hand movement information alone. The improvements scale for both MediaPipe and Kinect skeletal-based hand movement modeling cases.

C. Handshape recognition analysis

As a second analysis, we conducted a frame-level handshape recognition study. However, since the SMILE DSGS dataset only contains sign-level handshape transcriptions (referred to as true label(s)) and no frame level annotations, we evaluated the handshape recognition accuracy (RA) in the following manner: we supposed that each sign production should contain the true label(s) and the transition label. So, at the frame-level if a true label or a transition label was predicted then we considered it to be correctly predicted. In the case of the Hand SubUNet approach, we made a correspondence between the 31 output handshape classes and the HamNoSys symbols of the 28 handshape classes present in the SMILE DSGS dataset using a matching Table [3]. The first column of Table III gives the resulting handshape RA using for the proposed handshape subunits estimator and the Hand SubUNet estimator. The second and third columns of Table III further provides the percentage split of each case separately.

TABLE III
HANDSHAPE RECOGNITION ACCURACIES OF THE TRUE LABEL(S) AND/OR TRANSITION LABEL OF THE PROPOSED APPROACH AND THE HAND SUBUNET SYSTEM

	true & trans. labels	**trans. label**	**true label**
Proposed	92.1	58.3	33.8
Hand SubUNet	62.3	40.6	21.7

We can observe that both approaches follow a similar proportion of true and transition labels, where a little over a third of the predicted labels are true label(s). This similar proportion of splitting indicates that the proposed approach significantly improves the handshape RA for both transition and true label(s) handshape detection. Furthermore, this suggests that each video, i.e. each production of isolated sign, is roughly divided into three equal parts: beginning transition (such as going up), main part, ending transition (such as going down). Further analysis is needed to validate this assumption. In this direction, as an insight, we manually labeled three videos: (i) the longest video of the test set and 50% was labeled as true label(s), (ii) the shortest, 45% and (iii) one video of average length, 33%.

VI. CONCLUSION

This paper develops an approach for modeling dynamically handshape information using skeletal information for the development of sign language assessment systems. This approach combines both prior knowledge and data to derive and model handshape subunits. More precisely, HamNoSys prior is used to group signs into handshape classes, and subunits that model the handshape classes are derived in a data-driven manner using HMMs. Sign language assessment study, as well as sign language recognition and handshape recognition analysis on the SMILE DSGS corpus, show that the proposed approach yields considerably better systems than the image-based handshape classification approach Hand SubUNet, where the handshape subunits are based on prior knowledge alone. As evident from the present and previous studies [16], [18], Hand SubUNet can be trained on sign language-independent data and used for sign language assessment and recognition on other sign languages. Our future work will investigate whether such sign language independence is exhibited by the handshape subunits obtained through the proposed skeleton information-based approach.

ACKNOWLEDGMENT

This work was funded by the SNSF through the Sinergia project SMILE-II (Scalable Multimodal Sign Language Technology for Sign Language Learning and Assessment - Phase 2), grant agreement CRSII5_193686. We thank all the collaborators in the project for their collaboration.

[3]https://www.sign-lang.uni-hamburg.de/dgs-korpus/files/inhalt_pdf/HamNoSys_Handshapes.pdf

REFERENCES

[1] S. Subburaj and S. Murugavalli, "Survey on sign language recognition in context of vision-based and deep learning," *Measurement: Sensors*, vol. 23, pp. 100385, 2022.

[2] M. Z. Mahmoud and I. S. Samir, "Sign language recognition using a combination of new vision based features," *Pattern Recognition Letters*, vol. 32, no. 4, pp. 572–577, 2011.

[3] H. Cooper, E.J. Ong, N. Pugeault, and R. Bowden, "Sign language recognition using sub-units," *Journal of Machine Learning Research 13*, 2012.

[4] O. Koller, O. Zargaran, H. Ney, and R. Bowden, "Deep sign: hybrid CNN-HMM for continuous sign language recognition," in *Proc. of the BMVC*, 2016.

[5] A. A. Hosain, P. S. Santhalingam, P. Pathak, H. Rangwala, and J. Košecká, "Finehand: Learning hand shapes for american sign language recognition," in *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, 2020, pp. 700–707.

[6] A. A. Hosain, P. S. Santhalingam, P. Pathak, J. Košecké, and H. Rangwala, "Body pose and deep hand-shape feature based american sign language recognition," in *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, 2020, pp. 207–215.

[7] S. Aly and W. Aly, "Deeparslr: A novel signer-independent deep learning framework for isolated arabic sign language gestures recognition," *IEEE Access*, vol. 8, pp. 83199–83212, 2020.

[8] L. Pigou, Sander Dieleman, Pieter-Jan Kindermans, and Benjamin Schrauwen, "Sign language recognition using convolutional neural networks," in *Computer Vision - ECCV 2014 Workshops*, Lourdes Agapito, Michael M. Bronstein, and Carsten Rother, Eds., Cham, 2015, pp. 572–578, Springer International Publishing.

[9] P. Molchanov, X. Yang, S. Gupta, K. Kim, S. Tyree, and J. Kautz, "Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural networks," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4207–4215.

[10] Y. Liao, P. Xiong, W. Min, W. Min, and J. Lu, "Dynamic sign language recognition based on video sequence with blstm-3d residual networks," *IEEE Access*, vol. 7, pp. 38044–38054, 2019.

[11] C. Sun, T. Zhang, B.-K. Bao, and C. Xu, "Latent support vector machine for sign language recognition with kinect," in *2013 IEEE International Conference on Image Processing*, 2013, pp. 4190–4194.

[12] M. Al-Hammadi, G. Muhammad, W. Abdul, M. Alsulaiman, M. A. Bencherif, T. S. Alrayes, H. Mathkour, and M. A. Mekhtiche, "Deep learning-based approach for sign language gesture recognition with efficient hand gesture representation," *IEEE Access*, vol. 8, pp. 192527–192542, 2020.

[13] N. C. Camgöz, S. Hadfield, O. Koller, and R. Bowden, "Subunets: End-to-end hand shape and continuous sign language recognition," in *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[14] 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.

[15] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. G. Yong, J. Lee, W.-T. Chang, W. Hua, M. Georg, and M. Grundmann, "Mediapipe: A framework for building perception pipelines," 2019.

[16] S. Tornay, N. C. Camgoz, R. Bowden, and M. Magimai.-Doss, "A phonology-based approach for isolated sign production assessment in sign language," in *ICMI '20 Companion*, Oct. 2020.

[17] S. Tornay, *Explainable Phonology-based Approach for Sign Language Recognition and Assessment*, Ph.D. thesis, Ecole polytechnique fédérale de Lausanne (EPFL), Switzerland, 2021.

[18] S. Tornay, M. Razavi, N. C. Camgoz, R. Bowden, and M. Magimai.-Doss, "HMM-based approaches to model multichannel information in sign language inspired from articulatory features-based speech processing," in *Proc. in the IEEE ICASSP*, 2019.

[19] N. Tarigopula, S. Tornay, S. Muralidhar, and M. Magimai Doss, "Towards accessible sign language assessment and learning," in *Proceedings of the 2022 International Conference on Multimodal Interaction*, New York, NY, USA, 2022, ICMI '22, p. 626–631, Association for Computing Machinery.

[20] "web SMILE demonstrator," https://lab.idiap.ch/sap/smile/en/how.

[21] G. Aradilla, J. Vepa, and H. Bourlard, "An acoustic model based on Kullback-Leibler divergence for posterior features," in *Proc. of the IEEE Proc. of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2007.

[22] G. Aradilla, H. Bourlard, and M. Magimai.-Doss, "Using KL-based acoustic models in a large vocabulary recognition task," in *Proc. of Interspeech*, 2008.

[23] T. Hanke, "HamNoSys - Representing sign language data in language resources and language processing contexts," *Workshop proceedings : Representation and processing of sign languages*, pp. 1–6., 2004.

[24] 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, "SMILE Swiss German Sign Language dataset," in *Proc. of the LREC*, 2018.

[25] O. Koller, H. Ney, and R. Bowden, "Deep hand: How to train a CNN on 1 million hand images when your data is continuous and weakly labelled," in *Proc. of the IEEE CVPR*, June 2016.

[26] D. Johnson et al., "ICSI Quicknet Software Package," 2004.