POSTERIOR-BASED ANALYSIS OF SPATIO-TEMPORAL FEATURES FOR SIGN LANGUAGE ASSESSMENT

Neha Tarigopula Sandrine Tornay
Mathew Magimai.-Doss

Idiap-RR-11-2024

OCTOBER 2024

Received XX Month, XXXX; revised XX Month, XXXX; accepted XX Month, XXXX; Date of publication XX Month, XXXX; date of current version XX Month, XXXX.

Digital Object Identifier 10.1109/XXXX.2022.1234567

Posterior-based analysis of spatio-temporal features for Sign Language Assessment

Neha Tarigopula[1,2], Sandrine Tornay[1], Ozge Mercanoglu Sincan[3], Richard Bowden[3] and Mathew Magimai.-Doss[1]

[1]Idiap Research Institute, Martigny, Switzerland
[2]Ecole polytechnique fédérale de Lausanne (EPFL), Lausanne, Switzerland
[3]University of Surrey, UK

Corresponding author: Neha T. Author (email: neha.tarigopula@idiap.ch).

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-II), grant agreement CRSII5 193686. We thank all the collaborators in the project for their valuable work.

ABSTRACT Sign Language conveys information through multiple channels composed of manual (handshape, hand movement) and non-manual (facial expression, mouthing, body posture) components. Sign language assessment involves giving granular feedback to a learner, in terms of correctness of the manual and non-manual components, aiding the learner's progress. Existing methods rely on handcrafted skeleton-based features for hand movement within a KL-HMM framework to identify errors in manual components. However, modern deep learning models offer powerful spatio-temporal representations for videos to represent hand movement and facial expressions. Despite their success in classification tasks, these representations often struggle to attribute errors to specific sources, such as incorrect handshape, improper movement, or incorrect facial expressions. To address this limitation, we leverage and analyze the spatio-temporal representations from Inflated 3D Convolutional Networks (I3D) and integrate them into the KL-HMM framework to assess sign language videos on both manual and non-manual components. By applying masking and cropping techniques, we isolate and evaluate distinct channels of hand movement, and facial expressions using the I3D model and handshape using the CNN-based model. Our approach outperforms traditional methods based on handcrafted features, as validated through experiments on the SMILE-DSGS [1] dataset, and therefore demonstrates that it can enhance the effectiveness of sign language learning tools.

INDEX TERMS Deep Learning, Explainability, Hidden Markov Models, Sign Language Assessment, Sign Language Recognition

I. INTRODUCTION

Sign Language (SL) is a visual mode of communication, where information is conveyed through manual(handshape, hand movement) and non-manual (facial expression, body posture, mouthing) channels. Both the manual and non-manual components are crucial for effective verbal and non-verbal communication. It plays an important role in communication for the deaf and hard-of-hearing (DHH) community.

In recent years, owing to the awareness of accessibility needs of people, SL learning platforms are gaining popularity. These platforms help to bridge the communication gap between the hearing and DHH communities by developing assistive technologies that evaluate a learner's performance by providing meaningful feedback and facilitating their progress in acquiring accurate signing skills. In that direction, there has been effort for more than a decade in developing interactive sign language learning platforms for both children and adults [2]–[6].

Most existing platforms for sign language learning and assessment focus on testing vocabulary using pre-recorded videos for later analysis. E-learning platforms, such as SignAssess [7], allow users to compare their recorded videos to reference sign videos. In terms of real-time sign language verification, applications like SignAll [8] and ISARA [9] assess whether a produced sign is correct or incorrect. However, simply determining if a sign is correct or incorrect provides insufficient information to help a learner improve their production. From a linguistic perspective, Willoughby *et al.* envisioned My Interactive Auslan Coach [10], a system designed to provide automatic feedback on the correctness of handshape and hand movement for Australian Sign Language. Similarly, Huenerfauth *et al.* [11] proposed a system that analyzes sign production and offers feedback on both manual and non-manual components. However, these two systems are prototypes, with their feedback systems primarily evaluated for usability. Cory *et al.* [12] propose a distribution modeling method based on VAEs [13] and Gaussian Processes to evaluate the correctness of sign sentences, but not on a granular level.

In the context of providing automatic granular feedback, Tornay *et al.* [14], introduced a phonology-based sign language assessment system that provides feedback on two levels: (i) **Lexeme-level**, evaluating whether the produced sign matches the reference sign, and (ii) **Form-level**, assessing the correctness of each manual component. This approach models each manual channel separately, later combining them within a statistical framework using the Kullback-Leibler divergence-based Hidden Markov Model (KL-HMM) [15], [16]. The system utilizes CNN-based methods for extracting handshape features and handcrafted skeleton-based features with further processing for hand movement.

The KL-HMM system offers the advantage of modeling individual channels, such as handshape, hand movement, and facial expressions, independently. It allows for the fusion of these channels and subsequently offers a structured approach to factorize the output into distinct components. This explainability is crucial for facilitating a detailed breakdown for the assessment of individual channels. Deep learning models for spatio-temporal tasks, such as action classification [17]–[22], can be fine-tuned for sign recognition by modeling all channels—handshape, hand movement, and facial expressions—in a unified manner. These models can leverage large-scale datasets and can be trained on multilingual sign languages, making them versatile and applicable across different languages and signing variations. Although these models have shown success in recognition tasks, the inseparability of individual channels limits their ability to provide detailed, granular feedback necessary for assessment.

The goal of this work is to address these limitations by proposing an approach that combines the merits of both deep learning approaches and statistical methods to enable fine-grained assessment of sign language videos. We propose leveraging the spatio-temporal representations extracted from I3D model [17] trained on MeineDGS [23], combined with masking and cropping techniques to isolate hand movement and facial expressions, and integrating these representations into a KL-HMM framework for evaluation. For handshape analysis, we employ a CNN-based model like proposed in [14], [24].

The contributions of this paper are as follows:

1) Integration and analysis of deep learning-based spatio-temporal representations with the statistical framework of KL-HMM for sign language assessment.
2) Isolating individual components in manual and non-manual channels by using masking and cropping of videos
3) Leveraging the I3D model that uses temporal context for facial action unit detection and extending the assessment system to evaluate facial expressions.

II. BACKGROUND

This work takes place in the context of sign language learning as illustrated in Figure 1, where a learner's sign production is evaluated and feedback is given based on the correctness of the production at different levels. The rest of this section describes the framework for the phonology-based assessment system for Sign Language as proposed in [14]. The method takes inspiration from spoken language, as both spoken and sign language have a production phenomenon and a perception phenomenon. The production involves the generation of the signal, in speech, it is movement of articulators such as vocal folds, tongue, jaw, lips etc., that produce a 1D acoustic signal. Whereas, in SL a 2D visual signal is generated with varied hand movements, facial expressions and body postures. The perception phenomenon involves interpreting the signal in terms of linguistic units of words or phrases. In [15], [16] a KL-HMM based approach was used to model articulatory features (AF) as **posterior** representations for speech recognition, which was extended for SL [25].

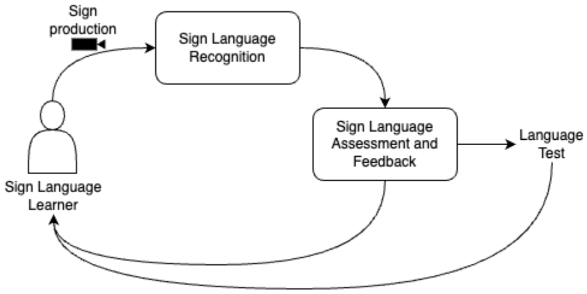

Sign production
Sign Language Recognition
Sign Language Assessment and Feedback
Language Test
Sign Language Learner

FIGURE 1. **Illustration of the assessment framework for sign language learning.**

The framework consists of two phases (i) Training Phase: To build reference KL-HMM models for the signs (ii) Assessment Phase: Validation of the produced sign against the reference sign.

A. Training Phase

In the training phase, subunit posteriors corresponding to different channels of handshape (*hshp*), hand movement (*hmvt*), facial expressions (*fexp*) etc., are modelled through KL-HMMs as depicted in Figure 2. More precisely, given the visual signal $(\mathbf{v}_1, \dots \mathbf{v}_t, \dots \mathbf{v}_T)$, the posterior probability of subunits corresponding to each of the channels are estimated as $\mathbf{z}_{t,f}$ where $f \in \{hshp, hmvt, \dots\}$. The posteriors corresponding to different channels are stacked $\mathbf{z}_t = [z_{t,1}, \dots z_{t,f}, \dots z_{t,F}]$ and used a feature observations to train a HMM, whose states are parameterized by categorical distributions $\mathbf{y}_i = [y_{i,1}, \dots y_{i,f}, \dots y_{i,F}]^{\mathrm{T}}$, for $i \in \{1, \dots N\}$ where N 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) [15], [16].

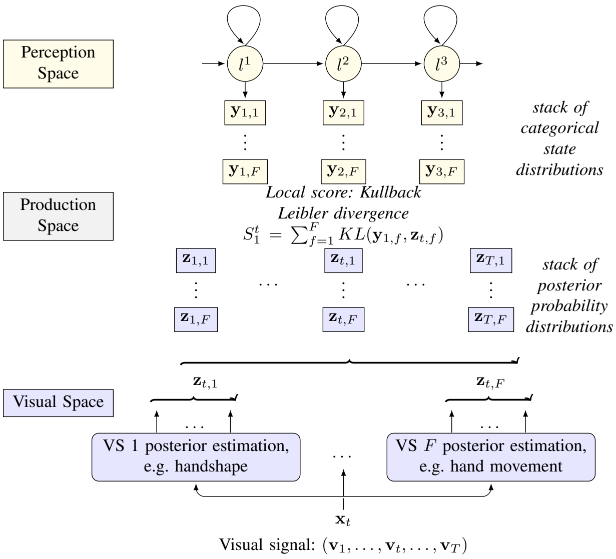

Perception Space
l^1
l^2
l^3
$\mathbf{y}_{1,1}$
$\mathbf{y}_{2,1}$
$\mathbf{y}_{3,1}$
$\mathbf{y}_{1,F}$
$\mathbf{y}_{2,F}$
$\mathbf{y}_{3,F}$
stack of categorical state distributions
Production Space
Local score: Kullback Leibler divergence
$S_1^t = \sum_{f=1}^{F} KL(\mathbf{y}_{1,f}, \mathbf{z}_{t,f})$
$\mathbf{z}_{1,1}$
$\mathbf{z}_{t,1}$
$\mathbf{z}_{T,1}$
$\mathbf{z}_{1,F}$
$\mathbf{z}_{t,F}$
$\mathbf{z}_{T,F}$
stack of posterior probability distributions
Visual Space
$\mathbf{z}_{t,1}$
$\mathbf{z}_{t,F}$
VS 1 posterior estimation, e.g. handshape
VS F posterior estimation, e.g. hand movement
$\mathbf{x}_t$
Visual signal: $(\mathbf{v}_1, \dots, \mathbf{v}_t, \dots, \mathbf{v}_T)$

FIGURE 2. Illustration of modeling production and perception phenomena in KL-HMM framework for sign language processing [25]. 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 $[y_{i,1} \dots y_{i,f} \dots y_{i,F}]$.

B. Assessment Phase

In the assessment phase, the produced sign is matched with the expected sign production in a Dynamic Time Warping framework (DTW). As illustrated in Figure 3, the method matches the stack of posterior sequences from the produced sign video $Z_T = [z_{1,1}, \dots z_{t,f}, \dots z_{T,F}]$ with the sequence of KL-HMM states of the expected sign characterized by categorical distributions $Y_N = [y_{1,1}, \dots y_{n,f} \dots y_{N,F}]$, where n is the state. The local score given by $S(n,t)$ is based on symmetrical KL-divergence. A threshold applied on the path length normalized global score $S(N,T)$ is used for lexeme-level assessment i.e., to assess whether the produced sign matches the reference sign. Form-level assessment that corresponds to assessing the sign at the level of different production channels is done by factoring out the score for each channel from the global score and applying a threshold on channel-wise scores.

The match is obtained by dynamic programming with the recursion following

$$S(n,t) = l(y_n, z_t) + min[S(n, t-1) + c_x, \\ S(n-1, t-1) + c_x]$$

where $c_x = -log(0.5)$ is the transition cost and $l(y_n, z_t)$ is the local score given by:

$$l(y_n, z_t) = \sum_{f=1}^{F} SKL(y_{n,f}, z_{t,f})$$

$$SKL(y_{n,f}, z_{t,f}) = \frac{1}{2} \sum_{d=1}^{D_f} y_{n,f}^d log(\frac{y_{n,f}^d}{z_{t,f}^d}) + z_{t,f}^d log(\frac{z_{t,f}^d}{y_{n,f}^d})$$

where D_f corresponds to the dimension of the f^{th} channel.

Based on the best matching path (t_n^b, t_n^e) for each state n, the score for lexeme-level assessment is calculated as:

$$S_{lex} = \frac{1}{N} \sum_{n=1}^{N} \frac{\sum_{t_n^b}^{t_n^e} l(y_n, z_t)}{t_n^e - t_n^b + 1}$$

The form-level assessment scores for each channel can be factored from this as;

$$S_{form}^f = \frac{1}{N} \sum_{n=1}^{N} \frac{\sum_{t_n^b}^{t_n^e} SKL(y_{n,f}, z_{t,f})}{t_n^e - t_n^b + 1}$$

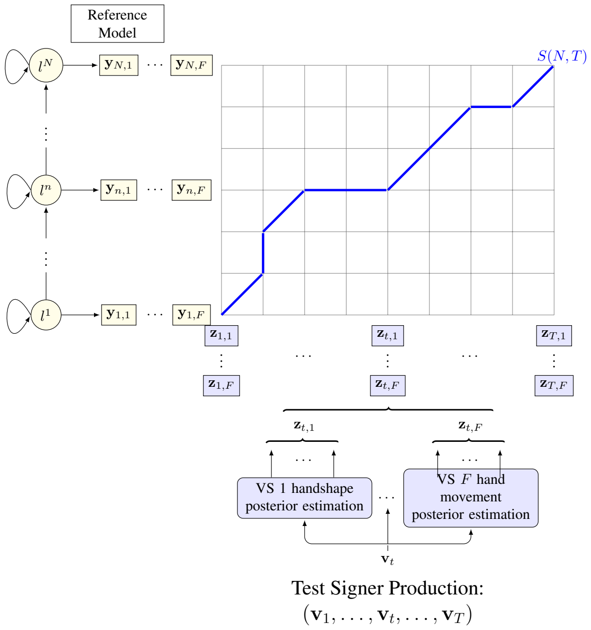


FIGURE 3. Illustration of the assessment framework [14]. $[z_{1,1} \dots z_{t,f} \dots z_{T,F}]$ is the stack of posterior estimates of F visual sub-units obtained from the test signer production. Each state l_n of the reference KL-HMM model is parameterized by the categorical distribution $[y_{1,1} \dots y_{n,f} \dots y_{N,F}]$. The DTW score is given by $S(N,T)$

III. PROPOSED METHODS

In this section, we describe the methods used to extract posterior features for different channels of handshape, hand movement, and facial expressions, for developing the KL-HMM recognition models and assessment systems. The proposed framework is illustrated in Figure 4.

A. Handshape

We use the pre-trained frame-wise handshape classifier based on SubUNets [26] to extract handshape posteriors. This classifier utilizes a CNN-LSTM-based model trained on the One-Million Hands dataset [27] for handshape classification. The model is trained on the 30 most commonly occurring handshapes out of the 60 in the dataset. Additionally, another classifier was trained to include these 30 handshapes along with a transitional shape. Consequently, we extract 61-dimensional vectors for each hand.

B. Hand Movement

For modeling the hand movement, we leverage the I3D [17] model for action recognition, which was pre-trained on MeineDGS (Deutsch GebärdenSprache - German SL) dataset [23] for sign spotting, to recognize isolated signs in specific frame window. We expect the model to capture sign language-specific movements, rather than merely functioning as an action recognition model. We leverage cross-lingual knowledge by utilizing a model trained on one sign language to enhance its applicability across different sign languages. To isolate the handshape information from hand movement features, we mask the hand region in the frame before extracting the movement features.

Unlike the handshape case, where we use the model output (after softmax) as our posterior representation, for hand movement, we use the 1024-dimensional representation from the penultimate layer of I3D and then later convert them into posteriors for DSGS (DeutschSchweizerische GebärdenSprache - Swiss German SL) for assessment, as the final layer representations are more tailored to DGS.

We analyze two methods for converting the feature representations into posteriors for integration with the KL-HMM framework: (i) Language dependent subunit extraction and (ii) Conversion using softmax. In the first approach, we generate posteriors by classifying subunit-like movement structures for each frame. Given a sequence of I3D feature representations for each sign, we train left-to-right HMMs with varying numbers of states for each sign, selecting the optimal number based on development set performance for DSGS sign recognition. We then align the features with the HMM states and train a multilayer perceptron (MLP) to classify these states that serve as movement subunits, using a cross-entropy-based cost function. The MLP output serves as our movement posterior representation. The process is illustrated in Figure 5. In the second approach, we apply a softmax function to the feature representations to transform them into posterior-like outputs.

C. Facial Expression Analysis

The Facial Action Coding System (FACS) [28], [29] is a taxonomy of facial action units (FAUs) used to encode facial expressions based on the activation of specific muscles or muscle groups (e.g., cheek raise, cheek puff, brow raise). Facial expressions are typically dynamic, involving onset, peak, and offset phases. Some transitions can be subtle, requiring analysis of a sequence of frames to capture them effectively, rather than a single frame. [30]–[32] apply temporal modeling using LSTM-based models for FAU detection. Due to the high cost of labeling action units, video datasets for FAU detection are relatively scarce. One notable large-scale dataset with frame-level annotations is Aff-Wild2 [33]. In this work, we leverage the Aff-Wild2 dataset and employ the I3D [17] model to capture effective spatio-temporal representations for FAU detection.

IV. EXPERIMENTAL SETUP

In this section, we outline the setup for (i) analyzing I3D-based posterior features for hand movement, (ii) developing a recognition system, and (iii) creating an assessment system.

A. Analysis of I3D-based posteriors

As mentioned in Section III, we evaluate two methods to convert the I3D representations into posteriors. In this section, we analyze the separability of features obtained by the two methods and also compare them with the handcrafted skeleton-based representations from [25]. We conducted the study by plotting the distribution of positive and negative distances between sign instances. Positive distances represent distances obtained by matching instances within the same sign class, whereas negative distances correspond to the distance obtained by matching instances of different sign classes. We use dynamic time warping (DTW) with a cost function based on Symmetric KL divergence, cosine similarity, and Bhattacharya distances for the analysis. The degree of overlap between the positive and negative distance distributions provides insight into the separability of the features.

B. Datasets

1) Sign Language Assessment

The **SMILE-DSGS** dataset [1] was created in the context of developing an assessment system for the lexicon of Swiss German Sign Language. It is the only database that has *linguistically annotated* SL data to aid production-level SL assessment. The dataset is composed of 100 isolated signs from DSGS. The data was acquired from 11 adult L1 signers and 19 adult L2 learners performing the signs of a DSGS vocabulary production test. The videos were collected with a Microsoft Kinect v2 sensor, the dataset includes both RGB and depth data obtained by the sensor and the gloss(meaning label associated with the sign in related spoken language) annotations. The linguistic annotations evaluate the acceptability of the signs through six categories, based on linguistic

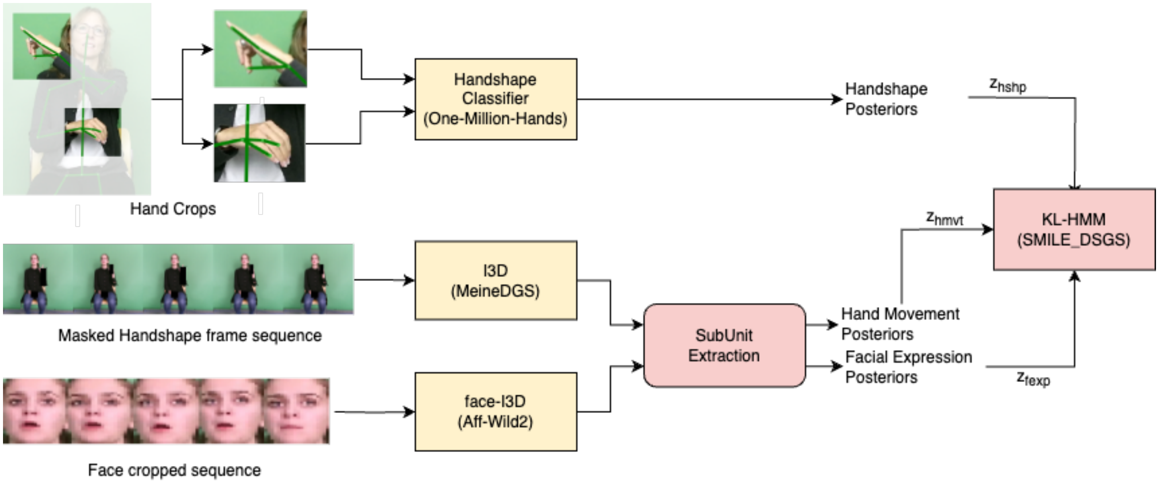


FIGURE 4. Proposed framework for the development of KL-HMM-based systems for sign language assessment. Framewise hand crops are used to extract handshape posteriors, hand masked sequences of 16 frames are used to extract hand movement posteriors, and a sequence of 16 face crops are used to extract facial expression posteriors. A stack of the posteriors are used to train the KL-HMM models.

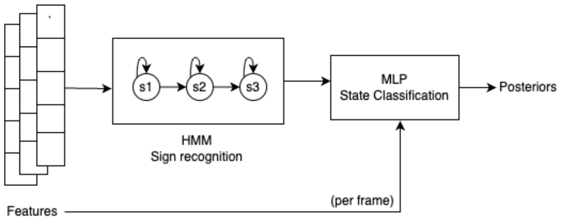


FIGURE 5. Subunit-based posterior extraction

criteria (lexeme, meaning, and form). The category evaluates the acceptance of the produced sign according to whether it is the same lexeme (word), has the same meaning, and has the same form as the 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

The linguistically acceptable productions of Category 1 and 2 are used to build the KL-HMM reference models for assessment. The data was 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 categories 3 and 4, and 183 samples corresponding to categories 5 and 6.

Aff-wild2 [33]–[43] is a large, in-the-wild dataset for classifying basic expressions and action units. It was introduced within the Affective Behavior Analysis in the Wild (ABAW) competition. It consists of 564 videos with about 2.6 million frames. It has a huge diversity in terms of age, ethnicity, gender, nationalities, and environment. The data is annotated on a per-frame basis for the seven basic expressions (i.e., happiness, surprise, anger, disgust, fear, sadness, and the neutral state) and twelve action units. Since facial expressions usually have onset, peak, and offset stages, temporal modeling is beneficial for classification. We use the Aff-wild2 videos to train our I3D-based model for representing facial expression features.

C. Posterior feature extraction

We employ the methods described in Section III to extract posterior representations for handshape, hand movement, and facial expressions.

(i) Handshape: We use the SubUnets [26] model mentioned in Section III to extract the handshape posteriors. Openpose [44] 2D pose estimation method was used to localize the wrist and these coordinates were used to obtain a hand patch, that serves as the input to the model. The output of the SubUnets classifier is used as the handshape posterior probability vector $\mathbf{z}_{t,hshp}$.

(ii) Hand movement: We use the I3D-model pre-trained on MeineDGS data to extract spatio-temporal representations for DSGS data. The model takes 16 frames of size 224x224 as input, with necessary padding if signs last shorter than 16 frames. The model was trained to optimize cross-entropy loss using SGD optimizer [45] with a momentum of 0.9, batch size of 4, and an initial learning rate of 0.01 with decay. Feature representations for each frame are extracted using a sliding window approach, with the representation assigned to the central frame within the window. We then use the subunit extraction method described in Section III to obtain the hand movement posterior vector $\mathbf{z}_{t,hmvt}$.

(iii) Facial expression: We train an I3D model on AffWild2 dataset for facial action unit classification. It is a multi-label classification problem, as more than one FAU can be activated in a frame. We use the MTCNN model [46] for face detection, cropping faces from individual frames in the Aff-Wild2 dataset. The input to the model consists of 16 frames of cropped faces, each resized to 224x224. We extract and resize the face crops and apply slight augmentations. The model is trained with a batch size of 16, optimizing binary cross-entropy loss using an SGD optimizer with a momentum of 0.9 and a learning rate of 0.01. We employ the subunit extraction method to obtain posterior representations for facial expressions. $\mathbf{z}_{t,fexp}$

D. KL-HMM reference systems

As proposed in [14], we trained different configurations of KL-HMM models for each sign, as follows:

1) **M**: Models only the hand movement subunits obtained from both the dominant and non-dominant hands (combined)
2) **M+S**: Models the stack of hand movement subunits and handshape subunits.
3) **M+S+F**: Models the stack of all three subunits of handshape, hand movement, and facial expressions.

We train the KL-HMMs with a varying number of hidden states, ranging from 3 to 30, using only the **acceptable** sign production data belonging to Category 1 and 2. This variation in the number of states allows us to capture different levels of granularity in the sign's temporal structure. The optimal number of states is determined by selecting the model that achieves the best recognition accuracy on the development set. Once we have the KL-HMM reference models for all the signs, we use them to match with the test signer productions to obtain the lexeme and form-level scores.

For sign language assessment, following the approach in [14], the thresholds for lexeme assessment, hand movement form assessment, facial expression form assessment, and handshape form assessment are calculated using the development set, which contains data from Categories 1 and 2. This is done by creating a set of positive sign scores by matching the same sign instances and by creating a set of negative sign scores by matching instances from different signs. We select the threshold that produces the highest F1 score for both lexeme and form assessment on the development set.

We report the recognition accuracy of the KL-HMM systems and the F_1 scores for lexeme and form-level assessment on the test set.

V. RESULTS

In this section, we present the results of the analysis of I3D features, KL-HMM recognition systems for DSGS sign recognition, and the F1 scores for assessment.

A. Analysis of I3D features

The histogram plots for the feature separability analysis are shown in Figure 6. The overlap between the positive and negative distance distributions provides insight into the separability of the features. Lower overlap indicates better separability, as the feature representations for the same sign class are closer than those of different sign classes. Conversely, higher overlap suggests poor separability, making it more difficult for the model to distinguish between different signs accurately. From Figure 6, we observe that the sub-unit-based I3D posteriors exhibit the least overlap, making them the preferred choice for building the assessment system.

B. Recognition

Table 1 presents the recognition performance of various KL-HMM systems trained using skeleton-based posteriors and I3D sub-unit-based posteriors as hand movement features. The KL-HMM system configuration is denoted by M for movement, S for handshape, and F for facial expression. The results indicate that I3D-based features outperform skeleton-based methods in the DSGS sign recognition task. For these results, movement features are isolated from handshape information by applying hand masking in the input frames.

TABLE 1. KL-HMM recognition accuracy for different model configurations

	M	M+S	M+S+F
Skeleton	55.77%	74.18%	-
I3D	**66.09%**	**75.81%**	**75.34%**

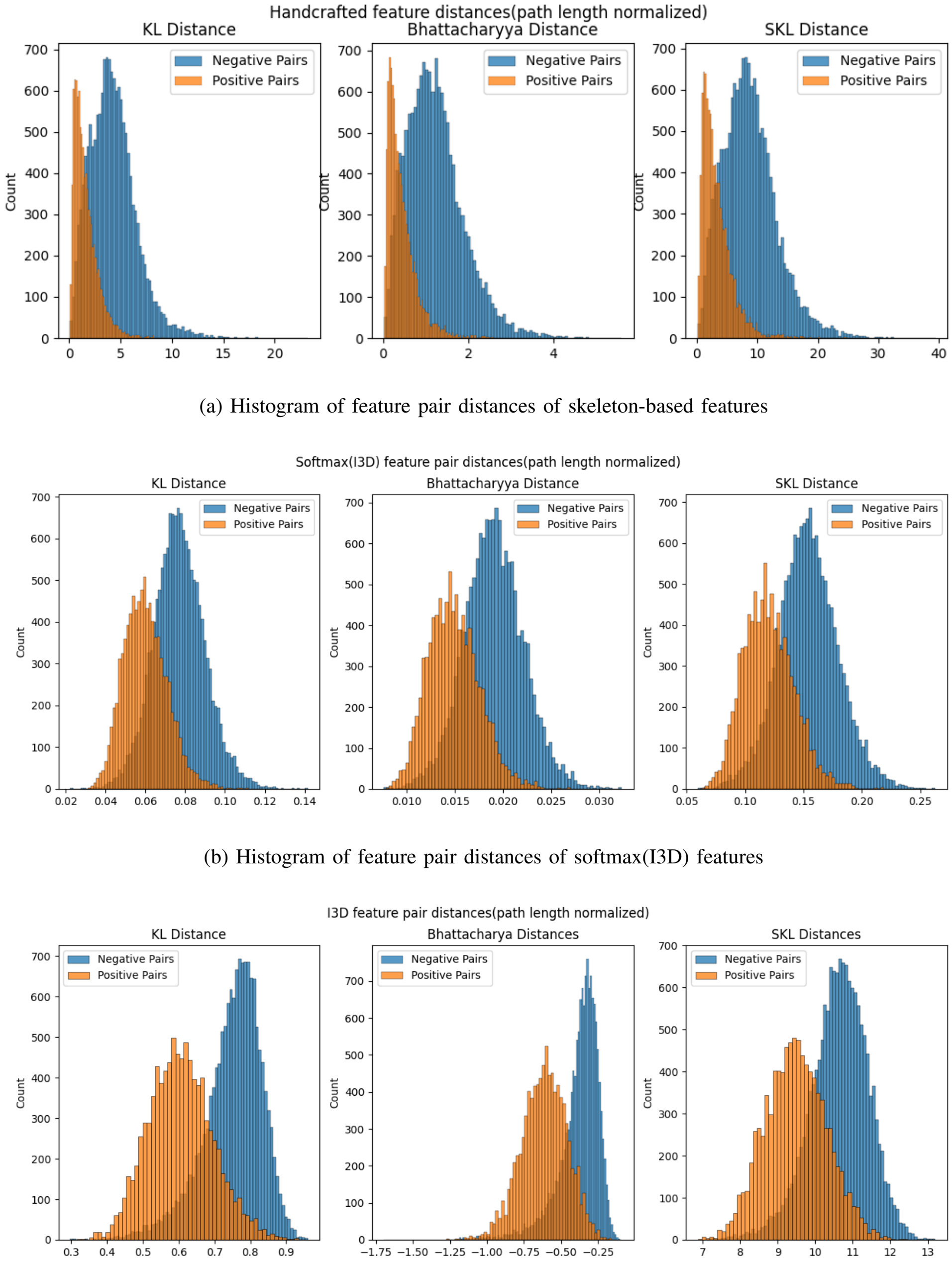
To further examine the impact of hand masking on hand movement feature extraction, we conducted additional experiments without applying hand masking during hand movement feature extraction. We provide the recognition results without hand masking in Table 2. In the unmasked case, handshape is integrated into the hand movement features, leading to better performance in sign classification than in the masked case. We also observe that incorporating facial expression posteriors does not significantly impact performance. We hypothesize that this is because isolated signs are less influenced by facial expressions compared to continuous signing (like in sentences).

TABLE 2. KL-HMM recognition accuracy for masked and unmasked hands for hand movement posterior extraction

	M	M+S	M+S+F
I3D-unmasked	**88.77%**	**89.65%**	**89.44%**
I3D-masked	66.09%	75.81%	75.34%

C. Assessment

Table 3 presents the F1 scores for assessment on lexeme level, form (handshape, hand movement, and facial expression) level. The I3D-based features perform better than



(a) Histogram of feature pair distances of skeleton-based features

(b) Histogram of feature pair distances of softmax(I3D) features

(c) Histogram of feature pair distances of I3D features

FIGURE 6. Histogram of Positive and Negative DTW Distances(with different cost functions) for Feature Separability Analysis: The positive distances represent distances between instances of the same sign class, while the negative distances correspond to distances between different sign classes

skeleton-based features for movement assessment. Since the best path for calculating the assessment score is obtained based on all the channels, it also leads to a better assessment of handshape in some cases.

VI. CONCLUSION

In this paper, we presented a method to integrate deep learning-based feature representations into the statistical framework of KL-HMM for sign language assessment. Our experiments demonstrated the effectiveness of using I3D-based models for hand movement feature extraction; however, this approach is flexible and can be adapted to other action recognition models fine-tuned on large-scale sign language datasets. Furthermore, language dependence can be incorporated through the subunit extraction method outlined in our study. Interestingly, our findings indicate that facial expressions do not significantly contribute to isolated sign recognition, suggesting a limited role in this context. However, the influence of facial expressions in continuous signing remains an open question and a potential direction

TABLE 3. F1 scores for lexeme and form assessment. hshp corresponds to handshape assessment, hmvt corresponds to hand movement assessment and fexp corresponds to facial expression assessment.

	Model Conf	hshp	hmvt	fexp	lexeme
Skeleton	M	-	0.9003	-	0.8771
	M+S	0.7960	0.9049	-	0.8993
I3D	M	-	0.9222	-	0.9123
	M+S	0.8053	0.9090	-	0.9234
	M+S+F	0.8041	0.9201	0.8612	0.9192

for future research. Overall, our work lays the groundwork for developing more adaptable and comprehensive sign language assessment systems, contributing to the advancement of sign language learning tools.

REFERENCES

[1] 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 Language Resources and Evaluation Conference*, 2018.

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

[3] H. Brashear et al., "American sign language recognition in game development for deaf children," in *Proc. of the International ACM SIGACCESS Conference on Computers and Accessibility*, 2006, pp. 79–86.

[4] J. Arendsen et al., "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*, 2008, pp. 1–6.

[5] O. Aran et al., "Signtutor: An interactive system for sign language tutoring," *IEEE MultiMedia*, vol. 16, no. 1, pp. 81–93, 2009.

[6] Z. Zafrulla et al., "CopyCat: An American sign language game for deaf children," in *IEEE International Conference on Automatic Face and Gesture Recognition*, 2011.

[7] J. Christopher, "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., Berlin, Heidelberg, 2012, pp. 253–260, Springer Berlin Heidelberg.

[8] SignAll Technologies Inc. (USA), "A communication bridge between deaf and hearing - signall," 2021.

[9] ISARA application, "Isara app," 2016.

[10] Louisa Willoughby, Stephanie Linder, Kirsten Ellis, and Julie Fisher, "Errors and feedback in the beginner auslan classroom," *Sign Language Studies*, vol. 15, pp. 322 – 347, 2015.

[11] Matt Huenerfauth, Elaine Gale, Brian Penly, Sree Pillutla, Mackenzie Willard, and Dhananjai Hariharan, "Evaluation of language feedback methods for student videos of american sign language," *ACM Trans. Access. Comput.*, vol. 10, no. 1, apr 2017.

[12] Oliver Cory, Ozge Mercanoglu Sincan, Matthew Vowels, Alessia Battisti, Franz Holzknecht, Katja Tissi, Sandra Sidler-Miserez, Tobias Haug, Sarah Ebling, and Richard Bowden, "Modelling the distribution of human motion for sign language assessment," 2024.

[13] Diederik P. Kingma and Max Welling, "An introduction to variational autoencoders," *CoRR*, vol. abs/1906.02691, 2019.

[14] S. Tornay, N. C. Camgoz, R. Bowden, and M. Magimai.-Doss, "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)*, Oct. 2020.

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

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

[17] João Carreira and Andrew Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4724–4733.

[18] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lucic, and C. Schmid, "Vivit: A video vision transformer," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Los Alamitos, CA, USA, oct 2021, pp. 6816–6826, IEEE Computer Society.

[19] Karen Simonyan and Andrew Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1*, Cambridge, MA, USA, 2014, NIPS'14, p. 568–576, MIT Press.

[20] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu, "3d convolutional neural networks for human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 221–231, 2013.

[21] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He, "Slowfast networks for video recognition," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 6201–6210.

[22] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool, "Temporal segment networks: Towards good practices for deep action recognition," 10 2016, vol. 9912.

[23] Reiner Konrad, Thomas Hanke, Gabriele Langer, Dolly Blanck, Julian Bleicken, Ilona Hofmann, Olga Jeziorski, Lutz König, Susanne König, Rie Nishio, Anja Regen, Uta Salden, Sven Wagner, Satu Worseck, Oliver Böse, Elena Jahn, and Marc Schulder, "Meine dgs – annotiert. öffentliches korpus der deutschen gebärdensprache, 3. release / my dgs – annotated. public corpus of german sign language, 3rd release," 2020.

[24] 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.

[25] 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.

[26] Necati Cihan Camgoz, 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)*, 2017, pp. 3075–3084.

[27] O. Koller, O. Zargaran, H. Ney, and R. Bowden, "Deep sign: hybrid CNN-HMM for continuous sign language recognition," in *Proc. of the British Machine Vision Conference (BMVC)*, 2016.

[28] Leon Rothkrantz, Dragos Datcu, and Pascal Wiggers, "Facs-coding of facial expressions," in *Proceedings of the International Conference on Computer Systems and Technologies and Workshop for PhD Students in Computing*, New York, NY, USA, 2009, CompSysTech '09, Association for Computing Machinery.

[29] Paul Ekman and Wallace V. Friesen, *Facial Action Coding System: A Technique for the Measurement of Facial Movement*, Consulting Psychologists Press, 1978.

[30] Elena Pyumina, Denis Dresvyanskiy, and Alexey Karpov, "In search of a robust facial expressions recognition model: A large-scale visual cross-corpus study," *Neurocomputing*, vol. 514, 10 2022.

[31] Ciprian Corneanu, Meysam Madadi, and Sergio Escalera, "Deep structure inference network for facial action unit recognition," in *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part XII*, Berlin, Heidelberg, 2018, p. 309–324, Springer-Verlag.

[32] Mani Kumar Tellamekala, Ömer Sümer, Björn W. Schuller, Elisabeth André, Timo Giesbrecht, and Michel Valstar, "Are 3d face shapes expressive enough for recognising continuous emotions and action unit intensities?," 2023.

[33] Dimitrios Kollias, Panagiotis Tzirakis, Alice Baird, Alan Cowen, and Stefanos Zafeiriou, "Abaw: Valence-arousal estimation, expression recognition, action unit detection & emotional reaction intensity estimation challenges," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 5888–5897.

[34] Stefanos Zafeiriou, Dimitrios Kollias, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, and Irene Kotsia, "Aff-wild: Valence and arousal 'in-the-wild'challenge," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*. IEEE, 2017, pp. 1980–1987.

[35] Dimitrios Kollias, Panagiotis Tzirakis, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, Björn Schuller, Irene Kotsia, and Stefanos Zafeiriou, "Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond," *International Journal of Computer Vision*, pp. 1–23, 2019.

[36] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou, "Face behavior a la carte: Expressions, affect and action units in a single network," *arXiv preprint arXiv:1910.11111*, 2019.

[37] Dimitrios Kollias and Stefanos Zafeiriou, "Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface," *arXiv preprint arXiv:1910.04855*, 2019.

[38] Dimitrios Kollias and Stefanos Zafeiriou, "Affect analysis in-the-wild: Valence-arousal, expressions, action units and a unified framework," *arXiv preprint arXiv:2103.15792*, 2021.

[39] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou, "Distribution matching for heterogeneous multi-task learning: a large-scale face study," *arXiv preprint arXiv:2105.03790*, 2021.

[40] D Kollias, A Schulc, E Hajiyev, and S Zafeiriou, "Analysing affective behavior in the first abaw 2020 competition," in *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)(FG)*, 2020, pp. 794–800.

[41] Dimitrios Kollias and Stefanos Zafeiriou, "Analysing affective behavior in the second abaw2 competition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3652–3660.

[42] Dimitrios Kollias, "Abaw: Valence-arousal estimation, expression recognition, action unit detection & multi-task learning challenges," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2328–2336.

[43] Dimitrios Kollias, "Abaw: learning from synthetic data & multi-task learning challenges," in *European Conference on Computer Vision*. Springer, 2023, pp. 157–172.

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

[45] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton, "On the importance of initialization and momentum in deep learning," in *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*. 2013, ICML'13, p. III–1139–III–1147, JMLR.org.

[46] Jia Xiang and Gengming Zhu, "Joint face detection and facial expression recognition with mtcnn," in *2017 4th International Conference on Information Science and Control Engineering (ICISCE)*, 2017, pp. 424–427.