Improving Keyword Search Performance in Sign Language with Hand Shape Features

Nazif Can Tamer[0000−0003−3903−7428] and Murat Saraçlar[0000−0002−7435−8510]

Department of Electrical and Electronic Engineering,
Boğaziçi University,
Istanbul, Turkey
can.tamer,murat.saraclar@boun.edu.tr

Abstract. Handshapes and human pose estimation are among the most used pretrained features in sign language recognition. In this study, we develop a handshape based keyword search (KWS) system for sign language and compare different pose based and handshape based encoders for the task of large vocabulary sign retrieval. We improved KWS performance in sign language by 3.5% mAP score for gloss search and 1.6% for cross-lingual KWS by combining pose and handshape based KWS models in a late fusion approach.

Keywords: Sign Language Recognition, Keyword Search, Handshape Recognition

1 Introduction

Sign language is the visual language of the hearing impaired. It has a different lexicon, grammar, and word ordering than spoken language; and the information in sign language is mainly carried through a mixed use of hand movements and facial expressions. To cope with this multimodal nature, recognition studies in sign language have historically divided the problem and focused on these different building blocks separately. Hand shape recognition from RGB hand patches [12][14][5] models common hand shapes, body pose based sign language recognition [9][1] mainly models places of articulation in space or along body, and other studies deal with mouthing [11] and facial expressions [2] which are all important communication channels of continuous sign language. In this paper, we introduce a hand-shape based keyword search (KWS) model for continuous sign language and compare it against KWS with pose key points [17].

Keyword search is a sub-problem of content retrieval which aims to search for a written query inside a large and unlabeled utterance. In spoken language recognition, keyword search is studied as a different problem than other retrieval problems such as keyword spotting and term discovery. Content retrieval from continuous sign language, on the other hand, is generally studied together under the umbrella term sign spotting and encompasses query-by-example search, keyword spotting, keyword search, and weakly supervised term discovery. The general approach in sign spotting requires strong supervision during training/learning.

Jantunen et al. [18] used dynamic time warping to search for citation form isolated signs in continuous sign language sentences. Yang et al. [20] used conditional random fields to search for 48 in-vocabulary signs that they learned from isolated examples. Ong et al. used hierarchical random fields to search for 48 the signs inside continuous sign language utterances. Although these approaches can obtain good retrieval performances, their search vocabulary is in the order of tens and their real-world applicability to large vocabulary retrieval systems is limited due to the amount of highly-annotated data they require during training.

Another track in sign spotting research is using weakly labeled continuous sign language in both learning time as well as testing. Most of the available sign language data are in the form of sign language interpreting and translations into the spoken language is the only form of annotation. Since there is no one-to-one relationship between signs and spoken words, several studies in the literature focus on discovering signs under the weak supervision of these translations or subtitles. Cooper and Bowden [7] used mining strategies to learn signs by matching the subtitles from TV shows. Farhadi and Forsyth [8] used HMMs to find sign boundaries assuming the sign sequence and the speech transcripts have the same word ordering. Buehler et al. [3] and Kelly et al. [10] applied multiple instance learning (MIL) based strategies to learn signs from subtitles and Kelly et al. [10] further used the isolated signs they discovered from translations to train a 30-vocabulary sign spotting framework. Being a concept adopted from speech recognition, large vocabulary keyword search methods also use weak labels in both training and test. Tamer and Saraclar [17] used graph convolution on top of skeleton joints for sign language keyword search. In this work, we introduce a hand shape based large vocabulary sign retrieval system and by combining with a pose based KWS model, we increase the recent keyword search performance in sign language by 3.5% mAP for gloss search and 1.6% mAP for cross-lingual KWS in RWTH-PHOENIX-Weather 2014T dataset.

The rest of the paper is organized as follows: In chapter 2 we briefly summarize our model and in chapters 3-6 we introduce the details. In chapter 7, we explain our experimental setting and results. Finally, in section 8 we conclude the paper.

2 Overview

The pipeline of our hand shape based KWS system is summarized in Figure 1. Our method starts with preprocessing the video to obtain hand shape feature vectors for each frame. Frame-level hand features are then fed into a 4-layer 1D temporal CNN encoder to detect movements. Keyword selection module from [17] represents the keywords in the same embedding space and focuses on relevant parts of the encoded hand shape sequence to detect the keyword.

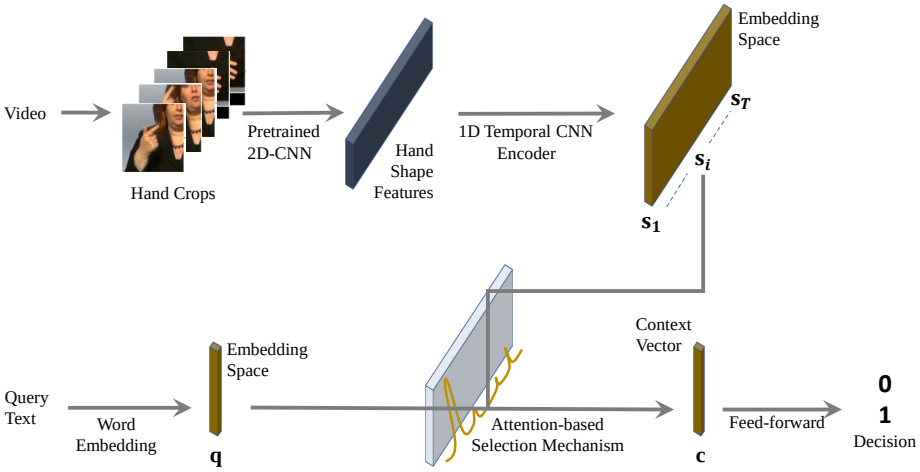

Fig. 1. Sign language KWS with Hand Shape Features. After the hand crops are obtained with the help of openpose, the hand shape features are extracted from frame. Word embeddings, 1D Temporal CNN Encoder, selection mechanism and the final feed-forward layer are trained end-to-end to represent video and text in the same embedding space.

3 Pre-processing

For each frame, the right hand wrist spatial locations are extracted with the help of OpenPose [6] pose estimation toolkit and square hand crops centered around the wrist joint are obtained. By feeding hand crops into one of the two pre-trained 2D CNN options, we represent hand crop with a vectoral feature. The frames are omitted if the OpenPose cannot estimate the location of the wrist joint. For the two 2D CNN options, the resulting one-dimensional hand shape features are 1024-dimensional for DeepHand and 2048-dimensional for MultiTask respectively.

3.1 Deephand Feature Extraction

We used the pre-trained CNN from DeepHand [12] to extract hand shape features. The model takes hand crops and classifies them into 60 pre-defined hand shape classes or a junk class. Their training data consists of two isolated sign language corpora (Danish and New Zealand SL), and continuous Phoenix-2014 Weather dataset. Since the third dataset they used in training is almost identical to our experiment data and the amount of supervision they used in training is more than that of our keyword search models, we believe the pre-trained DeepHand model can be viewed as the topline for hand shape encoders in this dataset. In our implementation, we used 1024 dimensional features from the second-last layer of DeepHand CNN.

3.2 Multitask Feature Extraction

Multitask features are introduced as a tokenization layer for sign language translation [14]. The network is trained for hand shape recognition in two datasets: the first one is the Danish and New Zealand SL corpora from DeepHand [12] excluding RWTH-PHOENIX-Weather 2014, and the second one is a framewise labeled and smaller Turkish SL dataset [16]. The network shares parameters at the start, and the final layers are different for matching different hand shape classification tasks. While the first one is 60 hand shapes and a junk class, the target for the smaller dataset also includes specific classes for hands showing certain body parts, thus, incorporating background information to some extent. Since the domain data is not used in training of Multitask features, it can be thought of as a real-world scenario for RGB hand shape based KWS. In feature extraction, we used 2048 dimensional vectors from the shared part of the multitask network.

4 1D Temporal CNN Encoder

At the end of pre-processing step, we obtain each frame represented with raw hand shape features. Since duration of a sign is greater than a single frame, however, we cannot learn keyword embeddings with these raw hand shape features and a further sequential modeling step is necessary. To model a 1-second-long temporal sliding window, we used 4-layer 1D convolutional network with kernel size 7 and same padding. We used leaky ReLU as the activation between layers. The first layer has 1024 channels for DeepHand and 2048 for MultiTask features. Then the channel sizes at the end of each layer are 512 for layer 1, 256 for layer 2, 128 for layer 3 and 256 for the last layer, respectively. With the help of same padding during convolution, we kept the encoded sequence length same with raw hand features. Each time step at the encoded sequence has access to 25 time steps of raw hand shape features resulting in a temporal range of 1 second in 25-fps RWTH-PHOENIX-Weather 2014T dataset.

5 Keyword Search Module

The keyword search module follows from [17] and consists of word embeddings, attention-based selection mechanism, and the final feed-forward layer.

5.1 Word Embeddings

A query in the form of text is first converted into an index in the vocabulary, and for all unique queries, a simple linear word embedding $\mathbf{q}$ is learned to match encoded sequence frame $\mathbf{s_i}$ in a mutual embedding space.

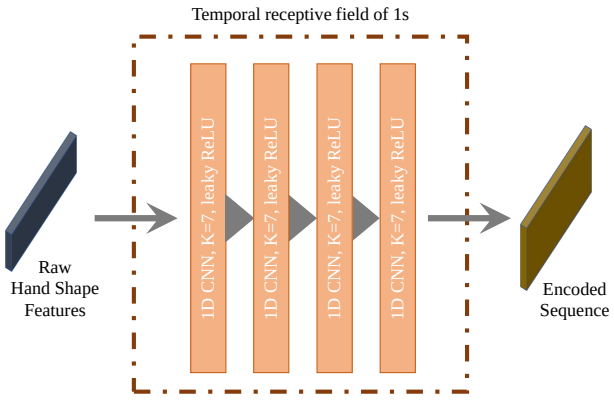

Fig. 2. 1D Temporal CNN Encoder

5.2 Attention-based Selection Mechanism

Attention is a widely known concept from Neural Machine Translation, where it helps the decoder to focus on relevant parts of the source sentence when predicting the next word in the target sequence [13]. In a similar fashion, we use attention to focus on the most relevant part of the encoded sequence $\mathbf{s_{1:T}}$ to the query $\mathbf{q}$. The relevance $score(\mathbf{q}, \mathbf{s}_i)$ between the ith element of the encoded sequence $\mathbf{s_i}$ and the query $\mathbf{q}$ is measured by a cosine-similarity-based function with a learnable parameter β:

$$score(\mathbf{q}, \mathbf{s}_i) = \beta \left[\frac{\mathbf{q} \cdot \mathbf{s}_i}{\|\mathbf{q}\| \cdot \|\mathbf{s}_i\|} \right]^2 \tag{1}$$

The context vector $\mathbf{c}$ is the weighted average of relevance scores after $softmax$ function is applied:

$$\mathbf{c} = \sum_i \left[\frac{\exp\left(score(\mathbf{q}, \mathbf{s}_i)\right)}{\sum_{i'} \exp(score(\mathbf{q}, \mathbf{s}_{i'}))} \right] \cdot \mathbf{s}_i \tag{2}$$

Once the context vector $\mathbf{c}$ is obtained, it is then fed into a one-layer feed-forward network with sigmoid activation to decide whether the query $\mathbf{q}$ is found inside the weakly-labeled sequence $\mathbf{s}_{1:T}$.

Although it only has β parameter and the weights of the final feed-forward layer as the learnable parameters, the selection mechanism is the most important layer of the network since it makes the weakly supervised learning of keyword embeddings possible. All the keywords in the vocabulary are searched in tandem in the same sequence. The keywords that appear in the transcription sequence are labeled positive whilst keywords that are not apparent in the transcription are trained to match negative labels.

6 Combining Hand Shape and Pose Based KWS Models

Hand shape and pose keypoints are among the most common pretrained features used in sign language recognition studies. We applied a late fusion approach to test the effectiveness of combining keyword search models trained with hand shapes and pose keypoints.

6.1 Pose based KWS

The model in [17] is employed for pose based keyword search. OpenPose [6] pose estimation features are extracted from each frame and fed into the Spatial-Temporal Graph Convolution Encoder [19] in 4 different layouts:

Pose1 (UB+RH+LH; x, y, conf): Upper body and right-left hand keypoints with confidence scores alongside (x, y) spatial locations,

Pose2 (UB+RH+LH; x, y): Upper body and right-left hand keypoints with (x, y) spatial locations omitting the OpenPose confidence scores,

Pose3 (UB+RH; x, y, conf): Upper body and right hand keypoints with confidence scores alongside (x, y) spatial locations, and

Pose4 (UB; x, y, conf): Only upper body keypoints with (x, y) spatial locations and OpenPose keypoint prediction confidences.

6.2 Fusion Strategy

In a setting where we search for a single keyword in a single sign language utterance, let $l \in 0, 1$ represent the binary label, h the prediction of the hand shape based KWS model, and k the prediction of the pose keypoints based KWS model respectively. The fused prediction p is found as

$$\log p = (1 - \gamma) \cdot \log h + \gamma \cdot \log k \tag{3}$$

The blending ratio γ is the number maximizing the mean average precision (mAP) score in train and development sets and fusion results are reported using this γ-value in the test set.

7 Experiments

7.1 Dataset

RWTH-PHOENIX-Weather 2014T [4] dataset is used for our experiments. This dataset is originally introduced for translation task and includes weather forecasts in German, their sign language interpreting in video format, and gloss sequence corresponding to the signs in the interpreting. The video footage is in 25 fps and in low resolution with heavy amount of blur. There is 9.2 hours of training, 37 minutes of development and 43 minutes of test partitions in the dataset.

For both the gloss and the cross-lingual keyword search, we used the original dataset labels in the following fashion: We formed our vocabulary from the training set. Each video is weakly labeled with 0 or 1 for every keyword in our vocabulary by looking at whether the keyword is in the label sequence or not. We dropped glosses starting with "__" since they contain on/off tokens and ambiguous signs, and we did not utilize lemmatization for German keywords. At the end, we have 1085 glosses in our gloss vocabulary and 2887 German keywords in cross lingual vocabulary. Since 392 of the glosses and 942 of the German keywords are shared between train and test sets, we report our results on this shared vocabulary. Out-of-vocabulary keyword search is not supported in this implementation.

7.2 Evaluation Metrics

For each keyword, we sorted utterances that give highest prediction scores and used 4 different information retrieval metrics to measure the quality of the keyword search performance. The first three are based on precision values at different ranks and the last one, nDGC, measures the ranking quality.

Mean Average Precision (mAP) is the mean of average precision scores so that all the keywords are equally important no matter how frequent they are in the test set. Average precision (AP) for a keyword q is defined as:

$$\text{AP} = \frac{1}{|N|} \sum_{n=1}^{|N|} \text{Precision@}n(q) \tag{4}$$

Precision at 10 (p@10) is the mean of precision scores at first ten retrieved utterances. It is a common metric in information retrieval for historical reasons, however, if the keyword is seen only once in the test set and that utterance is retrieved correctly, we still get p@10 score of 10% for this keyword.

Precision at N (p@N) is the mean of precision scores at first N_{test} retrieved utterances where N_{test}, the number of positive utterances in the test set, is different for each keyword.

Normalized Discounted Cumulative Gain (nDCG) is a measure of ranking quality normalized with the ideal possible ranking. It weights the first retrieved utterances more and the gain gets smaller once we move into higher ranks.

7.3 Effect of Different Encoder Structures on KWS Performance

Keyword search results with various handshape and pose based models are compared in Tables 1 and 2. From the comparison of Pose1 and Pose2 models in both

tables, we can see that using confidence scores of OpenPose keypoint estimations increase the retrieval performance in every metric. Both hand shape based gloss search models in Table 1 perform better than Pose4, pose based gloss search with only upper body; but addition of right hand in other pose based models increase the retrieval performance drastically. DeepHand features, which are trained on the domain data, perform universally better than Multitask features in both settings.

The results with the fusion of handshape based and pose based KWS model are also summarized in Tables 1 and 2. We applied a late fusion approach described in Section 6.2 with γ values learned from development set. We see that using fusion of handshape based features and Pose1, we can surpass the recent KWS performance in both gloss and cross-lingual KWS. When we compare the fusion models, combining Pose1 with DeepHand is better than combination with the Multitask based one in many of the metrics. However, the Multitask features are trained with only out-of-domain data, and the difference between using Multitask features instead of DeepHand is minimal. Thus, we opted for combining Multitask with Pose1 as our go-to structure.

Table 1. Gloss search results (in %, the higher the better) with different encoder structures. Both Multitask and DeepHand features are extracted from right hand only. UB: upper body, RH: right hand, LH: left hand, and conf refers to the use of OpenPose confidence scores alongside (x, y) spatial locations [17]. In fusion, $\gamma > 0.5$ denotes increasing reliance on the pose model.

Gloss Search Models	**mAP**	**p@10**	**p@N**	**nDCG**
Pose1 (UB + RH + LH; x, y, conf)	29.24	26.25	25.84	47.52
Pose2 (UB + RH + LH; x, y)	28.05	24.97	24.38	47.02
Pose3 (UB + RH; x, y, conf)	29.21	26.15	25.94	47.68
Pose4 (UB; x, y, conf)	22.80	21.45	19.95	43.15
Multitask	23.54	23.03	20.71	42.89
Multitask + Pose1, γ=0.54	32.22	27.98	27.66	50.08
DeepHand	24.93	23.65	22.27	43.86
DeepHand + Pose1, γ=0.58	32.78	27.88	28.67	50.02

7.4 Gloss-Specific Comparison of Hand Shape and Pose Based Encoders

In this section, we show that some glosses can be retrieved more easily with handshape features whilst pose based KWS models are better for others. We qualitatively compare the model performances by looking 6 isolated sign samples. Since there is no ground truth labels in RWTH-PHOENIX-Weather 2014T, we use citation form isolated signs taken from SignDict [15] German sign language dictionary for illustration. When selecting these 6 signs, we simply sorted all

Table 2. Cross-lingual search results (in %, the higher the better) with different encoder structures. Both Multitask and DeepHand features are extracted from right hand only. UB: upper body, RH: right hand, LH: left hand, and conf refers to the use of OpenPose confidence scores alongside (x, y) spatial locations [17]. In fusion, $\gamma > 0.5$ denotes increasing reliance on the pose model.

Cross-Lingual KWS Models	**mAP**	**p@10**	**p@N**	**nDCG**
Pose1 (UB + RH + LH; x, y, conf)	13.14	10.57	10.39	32.54
Pose2 (UB + RH + LH; x, y)	12.61	10.31	10.16	31.79
Pose3 (UB + RH; x, y, conf)	12.92	10.59	10.06	32.52
Pose4 (UB; x, y, conf)	10.79	8.77	8.73	29.99
Multitask	10.44	9.05	8.75	29.30
Multitask + Pose1, γ=0.68	14.34	11.52	11.27	33.66
DeepHand	11.11	9.62	9.14	29.85
DeepHand + Pose1, γ=0.60	14.75	11.43	11.63	33.97

Fig. 3. Hand-picked definitive single frames for the signs in Table 3. Frames are taken from isolated videos in SignDict dictionary [15]. Hand shapes are the most important feature in defining the sign for the first three whilst places of articulation along body are more definitive for the rest.

Table 3. Gloss-specific AP scores for different models. Both MultiTask and DeepHand features are extracted from right hand only. UB: upper body, RH: right hand, LH: left hand. All the pose based KWS models shown in this table are are with (x, y) spatial locations and OpenPose confidence scores.

Gloss Search AP (%)	WENIG	BESSER	ELF	APRIL	GLEICH	NAH
Pose1 (UB+RH+LH)	7.47	17.22	19.24	85.24	76.39	50.81
Pose3 (UB+RH)	4.40	3.44	15.53	49.17	48.98	12.31
Pose4 (UB)	2.62	30.20	17.09	50.83	31.68	5.60
Multitask	55.06	100.00	74.34	8.12	1.48	2.40
DeepHand	62.94	81.25	60.51	3.52	13.83	3.32
Multitask + Pose1	43.10	75.00	36.12	67.19	61.48	45.65

gloss queries according to the difference between Multitask and Pose1 models and picked the top 3 that also have a dictionary entry in SignDict for both extremes.

In Table 3, we can see that for the signs WENIG, BESSER, and ELF, both Multitask and DeepHand handshape based KWS models perform better than pose based ones. From the dictionary entries for these signs in Figure 3, we see that all three of these signs are single-handed and formed of simple hand shapes. The signs APRIL, GLEICH and NAH are the among the signs where Pose based models perform significantly better than handshape based ones. When we do some qualitative analysis, we can see that places of articulation are more important in defining these signs. In Figure 3, the sign for APRIL includes the thumb touching the nose and for GLEICH and NAH, we see hands interacting with each other. In Table 3, it can be seen that both Multitask and DeepHand handshape based encoders performed poorly compared to Pose1 model that includes upper body and both hands in the graph layout. Lastly, by observing the average performance in all these signs, we conclude that our Multitask + Pose1 fusion model performs reasonably better than relying on either hand shape or pose based models individually.

7.5 Analysis of the Fusion Model

The performance of Multitask + Pose1 model on different gloss vocabulary subsets are shown in Figure 4. When using weak labels during training, a single utterance is usually not enough to learn which temporal region is relevant for the sign. Thus, we also report our results in smaller vocabulary subsets. For 168 glosses with number of training samples $N_{train} \geqslant 50$, the mAP score is over 55%. For 115 glosses with $N_{train} \geqslant 100$, more than 7 out of 10 first retrieved utterances are correct. The results in Figure 4 follows a linear fashion other than the sharp increase in precision@10 scores. It is needed to have at least 10 positive utterances in the test set and this is true for most signs with $N_{train} \geqslant 100$.

8 Conclusion

In this paper, we introduce handshape based keyword search (KWS) models with Multitask[14] and DeepHand[12] pretrained features. We compared the performance of pose and handshape based KWS models in RWTH-PHOENIX-Weather 2014T dataset [4]. We improved the keyword search performance in sign language by applying a late fusion strategy combining pose and handshape based KWS models. Our findings in gloss-specific analysis suggests that handshape and pose based KWS models excel at retrieving different glosses. In future, we aim applying fusion at earlier stages of processing to learn which feature we should rely on for each specific keyword.

9 Acknowledgements

This study was supported in part by the Scientific and Technological Research Council of Turkey (TUBITAK) under Project 117E059.

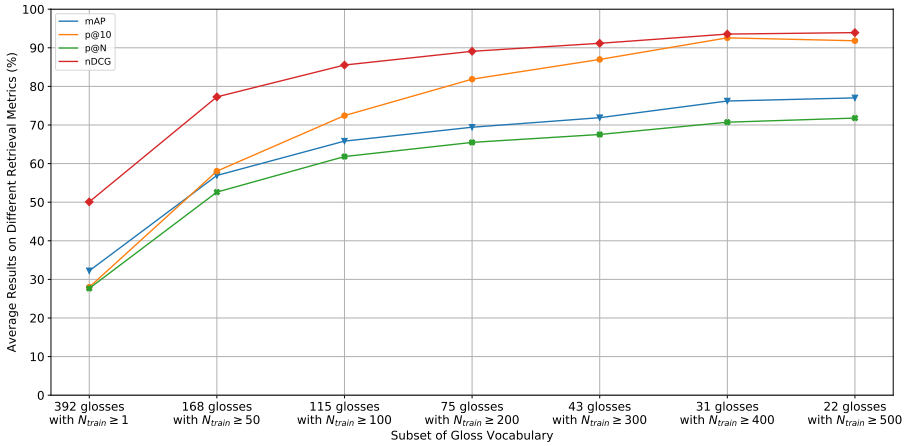

Fig. 4. Gloss search results of the MultiTask+Pose1 fusion model on different vocabulary subsets. N_{train} denotes the number of training utterances labeled with the keyword. In general, the retrieval performance during test time is higher for keywords with more weakly labeled utterances in the training partition.

References

1. de Amorim, C.C., Macêdo, D., Zanchettin, C.: Spatial-temporal graph convolutional networks for sign language recognition. In: International Conference on Artificial Neural Networks. pp. 646–657. Springer (2019)
2. Ari, I., Uyar, A., Akarun, L.: Facial feature tracking and expression recognition for sign language. In: International Symposium on Computer and Information Sciences. pp. 1–6. IEEE (2008)
3. Buehler, P., Zisserman, A., Everingham, M.: Learning sign language by watching tv (using weakly aligned subtitles). In: Proc. CVPR. pp. 2961–2968 (2009)
4. Camgoz, C., Hadfield, S., Koller, O., Ney, H., Bowden, R.: Neural sign language translation. In: Proc. CVPR. pp. 7784–7793 (2018)
5. Camgoz, N.C., Hadfield, S., Koller, O., Bowden, R.: Subunets: End-to-end hand shape and continuous sign language recognition. In: Proc. ICCV. pp. 3075–3084. IEEE (2017)
6. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: Proc. CVPR (2017)
7. Cooper, H., Bowden, R.: Learning signs from subtitles: A weakly supervised approach to sign language recognition. In: Proc. CVPR. pp. 2568–2574 (2009)
8. Farhadi, A., Forsyth, D.: Aligning asl for statistical translation using a discriminative word model. In: Proc. CVPR. vol. 2, pp. 1471–1476. IEEE (2006)
9. Gattupalli, S., Ghaderi, A., Athitsos, V.: Evaluation of deep learning based pose estimation for sign language recognition. In: Proceedings of the 9th ACM International Conference on Pervasive Technologies Related to Assistive Environments. pp. 1–7 (2016)
10. Kelly, D., Mc Donald, J., Markham, C.: Weakly supervised training of a sign language recognition system using multiple instance learning density matrices. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) **41**(2), 526–541 (2011)
11. Koller, O., Ney, H., Bowden, R.: Read my lips: Continuous signer independent weakly supervised viseme recognition. In: Proc. ECCV. pp. 281–296. Springer (2014)
12. Koller, O., Ney, H., Bowden, R.: Deep hand: How to train a cnn on 1 million hand images when your data is continuous and weakly labelled. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3793–3802 (2016)
13. Luong, T., Pham, H., Manning, C.D.: Effective approaches to attention-based neural machine translation. In: Proc. EMNLP. pp. 1412–1421 (2015)
14. Orbay, A., Akarun, L.: Neural sign language translation by learning tokenization. In: 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020). pp. 9–15
15. SignDict: Signdict, https://signdict.org/
16. Siyli, R.D.: Hospisign : A framewise annotated turkish sign language dataset, http://dogasiyli.com/hospisign/
17. Tamer, N.C., Saraçlar, M.: Cross-lingual keyword search for sign language. In: Proc. LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives. pp. 217–223. European Language Resources Association (ELRA), Marseille, France (2020), https://www.aclweb.org/anthology/2020.signlang-1.35

18. Viitaniemi, V., Jantunen, T., Savolainen, L., Karppa, M., Laaksonen, J.: S-pot–a benchmark in spotting signs within continuous signing. In: Proc. Language Resources and Evaluation Conference (LREC), ISBN 978-2-9517408-8-4 (2014)
19. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Thirty-second AAAI conference on artificial intelligence (2018)
20. Yang, H.D., Sclaroff, S., Lee, S.W.: Sign language spotting with a threshold model based on conditional random fields. IEEE transactions on pattern analysis and machine intelligence **31**(7), 1264–1277 (2008)