# UNSUPERVISED KEY HAND SHAPE DISCOVERY OF SIGN LANGUAGE VIDEOS WITH CORRESPONDENCE SPARSE AUTOENCODERS

*Recep Doga Siyli*[1], *Batuhan Gundogdu*[1,2], *Murat Saraclar*[1], *Lale Akarun*[1]

[1]Bogazici University, Turkey
[2]National Defense University Naval Academy, Turkey
{doga.siyli,batuhan.gundogdu,murat.saraclar,akarun}@boun.edu.tr

## ABSTRACT

Recognition of sign language is a difficult task which often requires tedious annotations by sign language experts. End-to-end learning attempts that bypass frame level annotations have achieved some success in limited datasets, but it has been shown that high quality annotations improve performance drastically. Recent unsupervised learning methods using deep neural networks have achieved successes in learning feature extraction. Yet a technique for high quality frame level classification using unsupervised techniques does not exist. In this paper, we assign labels of an isolated Sign Language(SL) dataset using end-to-end neural network architectures that have proven success in unsupervised discovery of sub-word acoustic units in speech processing. We observe that *key-hand-shapes*(*KHS*), which are meaningful visual basic parts of signs in a SL dataset can be detected using unsupervised clustering techniques. Sparse autoencoders can successfully retrieve and cluster *KHS*s used in isolated signs. In addition, using correspondent frames in an autoencoder scheme has the power to continue the learning process.

***Index Terms***— Sign language tagging, unsupervised unit discovery, correspondence sparse autoencoders, hand shape discovery

## 1. INTRODUCTION

Sign language(SL) is the native language of the hearing impaired people. Its phonology, semantics and gloss changes according to the community that uses them. SL can not be translated into its spoken version word by word and vice versa. Yet it contains Key Hand Shapes (*KHS*s) that can be seen as building blocks of the language. Some *KHS*s have common characteristics across different sign languages; while some are specific to a culture. A sign is a composition of *KHS*s, and a sentence in a SL is a composition of signs. Defining a *KHS* by only finger configuration is not sufficient as the same finger configuration has different meanings as the context given by the location around the body changes.

Although hand-shape is not the only phonological feature for SLs it can be considered as one of the key components of a sign in both algorithmic and linguistic studies [1]. The hand-shape does not reveal the whole information to be conveyed, yet is an important delimiter for annotating and breaking down a video sequence into meaningful parts. There are also other studies dealing with mouthing [2, 3, 4], motion features [5], hand skeleton features [6] and multimodal features of SL data [7]. This study focuses on frame-wise hand-shape recognition [8, 9, 10] by unsupervised clustering. [11]. Recent successes in computer vision attained by deep neural networks have relied on large annotated datasets. SLs are annotated using a set of glosses comprising a brief dictionary, which demonstrate hand-shapes and their possible meanings of that particular SL. ELAN [12], is a professional tool for the creation of complex annotations on SL videos and yet annotation of SL data-sets requires expertise and takes time. We propose an unsupervised solution which requires little expertise to discover clusters of recurrent *KHS*s. This allows the fast annotation of the dataset for subsequent supervised methods, such as deep neural networks. Rakowski and Wandzik showed in [13] that fine-tuning ASL fingerspelling and the 1 Million Hands datasets on pre-trained ImageNet dataset models outperform state-of-the-art approaches on both hand-shape classification tasks [14]. We state our problem as an unsupervised clustering problem and reach class labels through clusters found by a neural network. If the clusters are representative groupings that agree with the ground-truth class labels, we can propose that tagging the clusters is a fast and reliable way of tagging the whole dataset in a frame-wise fashion. As Farag and Brock stated in their study, publicly available SL datasets are not large and balanced enough for meaningful end-to-end learning [15]. Hence, segmenting SL videos is crucial and needed for studying the underlying structures of SLs using deep neural networks. Our study is based on RGB data whereas [15] proposes a solution depending on skeleton data.

Similar to the task defined in this paper, [16] also aims at recognition of a set of 30 hand shapes as bricks of Chinese sign language from RGB images. Aly et al. approached the task using depth images for both detection of hand shapes and signer independent clustering of American sign language videos[17]. Another relevant work by Kadir et al. approaches the task of sign language recognition in unconstrained conditions by detecting body centred descriptions of activity by coining linguistic description of these activities, and bypasses temporal modeling[18]. Recent approaches to sign language recognition focus on continuous sign language recognition [19, 20]. Huang et. al [21] used multimodal inputs and have shown success in extracting discriminitive spatio-temporal features from videos. Pigou et. al constructed a successful network that can recognize 20 Italian gestures with high accuracy [22]. For extracting more information and reaching higher accuracies on SL we need bigger annotated data-sets as in [23]. Current datasets use a combination of finger spelled signs, isolated one-handed signs, or two handed signs. Larger datasets recorded in realistic settings are needed to extract linguistic features of a sign language from datasets.

In this paper, we followed the approach of using sparse autoencoders (SAE) to discover the set of underlying *KHS*s that is assumed to compose the SL's "language model". This idea has recently been proposed for acoustic unit discovery Zerospeech challenge, in which a set of acoustic units are retrieved from untranscribed speech recordings [24] . This task is analogous to the challenge overtaken in this paper since the underlying *KHS*s could be considered as tokens of meaningful units to utter a SL sentence, just like acoustic units are used in concatenation to speak.

For this, we trained a SAE in an unsupervised fashion, which simultaneously clusters input frames on an intermediate layer after encoding, and reconstructs them at the output of the decoder layer. Therefore in sparse autoencoders, a sparsity cost is also utilized in addition to the reconstruction loss. In the next section, we describe the methodology adopted in this work and we present our findings over the experiments conducted on the Bosphorus Sign Turkish Sign Language Database (BS-TSLD).

## 2. METHODOLOGY

The main objective of this study is to discover the underlying hand-shapes that are used in Turkish Sign Language (TSL). We aim to seek the recurring hand-shapes in the TSL corpus and find the minimum number of such figures, we refer to as *KHS*s, so that the information is conveyed using them as units of meaningful tokens. As the first step, we use skeleton features of Kinect and retrieve the sub-frames that contain the active hand within each frame of the SL video. We adopt two sets of feature representations from these hand frames, meanwhile examining the representation power of hand crafted features like Histogram of Oriented Gradients (HOG) and deep learning based representation such as the bottle neck activations of resnet18. Principal component analysis (PCA) mapping to lower dimensions is also applied to each of these features to reduce variability.

### 2.1. Sparse Autoencoder (SAE)

SAE architecture is coined recently in an acoustic unit discovery task in the Zerospeech 2019 challenge, which is a similar task to that of detecting key hand shapes. In SAE, an autoencoder system with a decoder-encoder structure is trained, except in the embedding layer a *softmax* is employed such that this layer acts as a posterior probability vector over the set of underlying clusters. Hand frame representations are fed into the input layer of this autoencoder and the sparse representation is obtained through an intermediate feed-forward softmax layer. It is hoped that this embedding will yield a sequence of posterior probability vectors during prediction and enable an automated labelling of SL videos. In other words, it is expected that the activations of this layer act as the probabilities of the 'hidden' or the 'unknown' states (the *KHS* units to be discovered). To enforce sparsity of this representation, the activations of this layer is enforced to have a maximum L2-norm, since for a probability mass function maximum L2-norm is obtained for a distribution where only one of them is one and others are zero. If we denote the input hand frame representations as $\mathbf{x} \in \mathcal{R}^D$, the SAE produces the probability vector $\mathbf{p} \in [0, 1]^K$, denoting the probability of the input belonging to each of the $K$ clusters. This vector is then used as an input to the decoder layer to generate $\hat{\mathbf{x}}$ at the output, emulating $\mathbf{x}$ (1).

$$\begin{aligned} \mathbf{p} &= \texttt{softmax}\,(\texttt{encoder}\,(\mathbf{x})) \\ \hat{\mathbf{x}} &= \texttt{decoder}\,(\mathbf{p}) \end{aligned} \tag{1}$$

It is expected that similar hand-shapes will activate the same cluster node at the intermediate layer and, the resulting $\mathbf{p}$ will generate a similar enough representation, i.e. a centroid at the output so that the following combined cost is minimized:

$$J_{SAE} = \sum_{\forall t} ||\mathbf{x}_t - \hat{\mathbf{x}}_t||_2^2 - \lambda||\mathbf{p}_t||_2^2 \tag{2}$$

### 2.2. Correspondence Sparse Auto Encoder( CoSAE)

The term *auto* in correspondence autoencoder (CAE) is in fact a misnomer. In CAE training, pairs that are assumed to be similar are fed from the input and output layers to reconstruct each other[25]. It has been shown in [26] that CAEs perform better than denoising or deblurring autoencoders in obtaining discriminative representations in an unsupervised manner. Similar to SAEs, correspondence sparse autoencoders (CoSAE) have an intermediate clustering layer to extract the cluster assignments directly. For this, the hand frames that may correspond to the same *KHS* are obtained from the dataset using dynamic time warping (DTW) alignment path obtained from the pairs sign video sequences. All pairs of videos that correspond to the same sign, $\mathbf{X} \in \mathcal{R}^{D\times N}$ and $\mathbf{Y} \in \mathcal{R}^{D\times M}$ are aligned with DTW and the alignment path $\Phi$ is used to retrieve the frames of $\mathbf{Y}$ that correspond to the frames of $\mathbf{X}$. A dropout of $p = 0.5$ was applied after the decoder. The SAE and CoSAE training architectures are demonstrated in Figure 1.

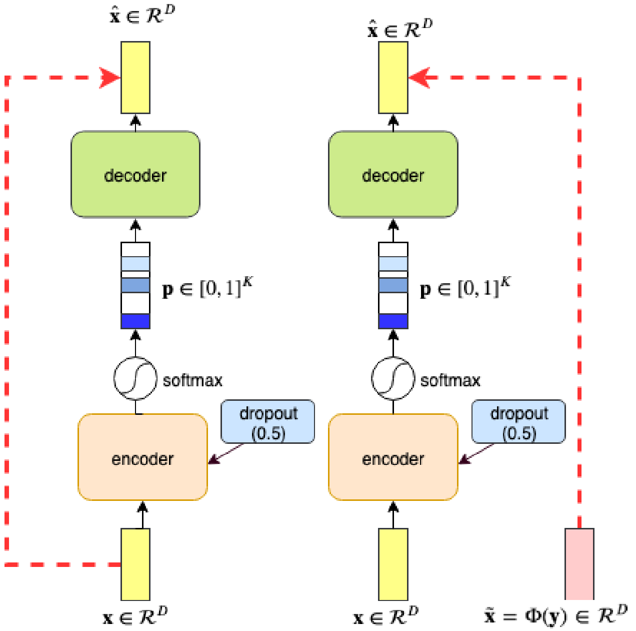
$\hat{\mathbf{x}} \in \mathcal{R}^D$
$\hat{\mathbf{x}} \in \mathcal{R}^D$
decoder
decoder
$\mathbf{p} \in [0,1]^K$
$\mathbf{p} \in [0,1]^K$
softmax
softmax
dropout (0.5)
dropout (0.5)
encoder
encoder
$\mathbf{x} \in \mathcal{R}^D$
$\mathbf{x} \in \mathcal{R}^D$
$\tilde{\mathbf{x}} = \Phi(\mathbf{y}) \in \mathcal{R}^D$

**Fig. 1**. SAE (left) and CoSAE (right) training architectures.

### 2.3. Baseline Method and the Ground Truth

We implemented the method of Deep Clustering [11] and took it as a baseline. In [11], images are fed into CNNs to obtain a discriminative representation on the penultimate layer. K-means clustering is applied to these bottleneck features prior to each epoch of neural network training and the one-hot vectors depicting the k-means clustering assignments are used to train the network. This procedure is repeated at each epoch and the normalized mutual information (NMI) between the actual class assignments are observed. Similarly, we observed the NMI scores of the SAE and CoSAE learning after each epoch as a reference. The main difference of our work from this approach is that we aim to address the two objectives simultaneously, i.e. the sparsity of the intermediate layer and the reconstruction loss.

One possible useful outcome of this approach is that it enables automatic labelling of the frames as a sequence of *KHS*s and transition states, a procedure that is extremely expensive to employ by human interaction. We use the *KHS* assignments that are obtained and labelled frame-by-frame by human experts as ground truth labels. Since the number of *KHS* classes are not known prior to execution, in addition to NMI, we monitor the frame accuracy rate, calculated via a learned mapping between the cluster assignments and the actual class labels.

### 2.3.1. Evaluation Criteria

The information that is shared between cluster labels and ground truth can be measured by two different criteria. The first criterion is the Normalized Mutual Information (NMI) that gives a score of accordance between two clusterings of the same dataset. NMI is expressed as:

$$NMI(C,K) = \frac{2 \times I(C;K)}{\sqrt{H(C)H(K)}} \tag{3}$$

where $C$ and $K$ are the multinomial distributions of class and cluster labels, respectively. $H$ denotes entropy and $I(C;K)$ is the mutual information between $C$ and $K$. It should be noted that this measure is independent of the number of classes and the clusters and, gives a value in the range $[0,1]$.

A directly interpretable measure is the frame error rate, or the frame accuracy rate which is the proportion of correctly classified frames. The frame accuracy rate can be considered to be analogous to the phone error rate in speech recognition. In our work, at the evaluation step, we omitted the transitional frames and considered the KHS on still images for frame error rate calculation, since human-assisted labelling is feasible only for still images. To make such an evaluation possible, we propose a many-to-one 'mapping learning' that converts the cluster assignments to their corresponding class labels using the confusion matrix. This mapping is obtained using the following formula:

$$\mathbf{MW} = \hat{\mathbf{M}} \tag{4}$$

where $\mathbf{M} \in \mathcal{R}^{C \times K}$ is the non-square confusion matrix having class labels in the rows and cluster labels in the columns. It is multiplied with $\mathbf{W} \in \mathcal{R}^{K \times C}$ the weight matrix to be learned which computes $\hat{\mathbf{M}} \in \mathcal{R}^{C \times C}$, showing the class confusion matrix. The aim is to maximize the number of samples in the diagonal of the resulting matrix $\hat{\mathbf{M}}$. There are also two regularization terms that sparsify row and column elements of the weight matrix, hence the weight matrix becomes a mapping of cluster labels to the class labels. The idea behind this maximization technique is that if the clusters can represent clusters in classes then we can label only the clusters by looking at a couple of highly activated elements and using the learned map $\mathbf{W}$ for labeling the whole dataset.

## 3. EXPERIMENTS

### 3.1. Dataset

We ran our experiments on a subset of BosphorusSign Turkish Sign Language Database (BS-TSLD)[27]. The subset of the dataset we used covers 11 signed sentences that are performed by 6 different signers and hand labelled in a frame-wise manner. The *KHS*s in the dataset represent context as well as shape. That is to say, they do not only differ in finger configuration, but also in the position of the body. Samples of these ground truth *KHS*s can be seen in Figure 2.

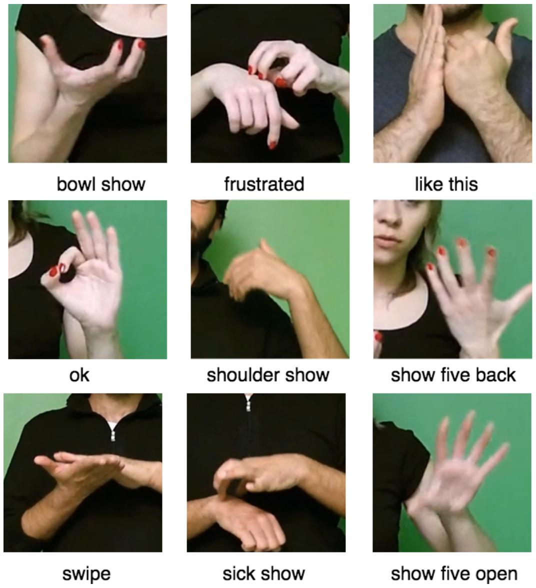

**Fig. 2**. Samples of *KHS*s from the dataset

### 3.2. Experimental set-up

We conducted our experiments using two different set of features: HOG [28] as a hand crafted feature and the pre-trained resnet18 [29] bottleneck layer activations as a deep learning-based feature. Both features are mapped to 256 dimensions with PCA. A feed-forward encoder is used with a dropout rate of $p = 0.5$ followed by a `softmax` layer to get the probability mass function $\mathbf{p}$ over $K$ clusters. The decoder layer is also a feed-forward network such that the input is reconstructed as a weighted sum of the rows of this matrix. The $\lambda$ parameter in (2) decides the trade-off between the reconstruction loss and the sparsity constraint obtained by the L2-norm of the intermediate layer. It should be noted that the intermediate layer has a maximum L2-norm when it is a one-hot vector since it is output of a `softmax` activation and sums to one. Several $K$ and $\lambda$ values are tested and compared with the baseline. Models are trained using adam optimizer using Keras toolkit with tensorflow backend[30].

### 3.3. Results

Our main aim was to obtain the cluster assignments on the intermediate sparse layer ($\mathbf{p}$) by minimizing the reconstruction loss. The intermediate layer is expected to 'classify' the input frame into one of its $K$ classes, and then in the decoder layer, this frame (or its correspondence) is reconstructed using the probability mass function observed at the $\mathbf{p}$-layer. We observed in Figure 3 that the neighboring frames have consistent class-assignments, or probability distributions. The NMI and the accuracy, obtained by treating this layer as the classification layer, is observed as the training progresses similar to the work proposed in [11].

We have experimented with various cluster numbers ($K$) and sparsity weight values ($\lambda$). We obtain good results using the parameters provided in Table 1. Since there is no means of validation and early stopping in our unsupervised setting, and the main objective of a high frame accuracy rate is different from the training loss (mse+sparsity), there is no early stopping criterion available and hence we trained all models for the same number of epochs.

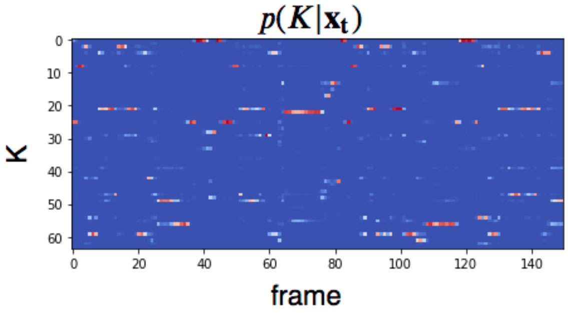

**Fig. 3**. $p$-layer activations of the network with $K = 64$, in response to frames of a sign video. Warm colors denote higher posterior probabilities, whereas blue is zero.

**Table 1**. Parameters for SAE/CoSAE training

| Parameter | value |
|---|---|
| $K$ | 256 |
| $\lambda$ | 1.0 |
| batch size | 16 |
| number of epochs | 50 |
| learning rate | 0.01 |

Table 2 gives the frame accuracy rate and the NMI scores of the proposed methods, compared with the baseline. We also included the performance of the ordinary k-means clustering with Euclidean distance and MMSE cost, since it could be considered as a natural baseline for clustering applications. It can be observed that hand crafted HOG features yield better performance than resnet18 bottleneck activations, which were trained on imagenet data. One interesting observation is that the improvement brought by the collaborative training is more significant on resnet18 bottlenecks, although it also improves the results with HOG. Overall, the best accuracy is obtained with CoSAE and HOG features. Using the NMI metric also indicates the same results: CoSAE gives the best KHS labeling results.

**Table 2**. Performance of the proposed systems

| Methodology | Accuracy | | NMI | |
|---|---|---|---|---|
| | resnet18 | HOG | resnet18 | HOG |
| k-means | 0.2995 | 0.5998 | 0.4005 | 0.6311 |
| Deep Cluster [11] | 0.4840 | - | 0.5010 | - |
| SAE | 0.4008 | 0.7026 | 0.3042 | 0.5504 |
| CoSAE | 0.7433 | **0.8307** | 0.6037 | **0.6997** |

Figure 4 demonstrates the evolution of accuracy as the training proceeds using the CoSAE method, using two different $\lambda$ values. The order of correspondence is altered in each epoch, which yields an additional increase in the accuracy. It can be observed that keeping the $\lambda$ values small results in a lower performance and higher fluctuation. It should be noted that training without correspondence (i.e. only SAE) fails to exhibit the drastic jumps in accuracy apparent in the Figure 4. Furthermore, it has been observed that alternating between CoSAE and SAE training methodologies helps to avoid local minimums. Confusion matrix between the KHS, found through CoSAE training, and human assisted ground truths, along with sample images of the whole KHS clustering, can be seen in [1].

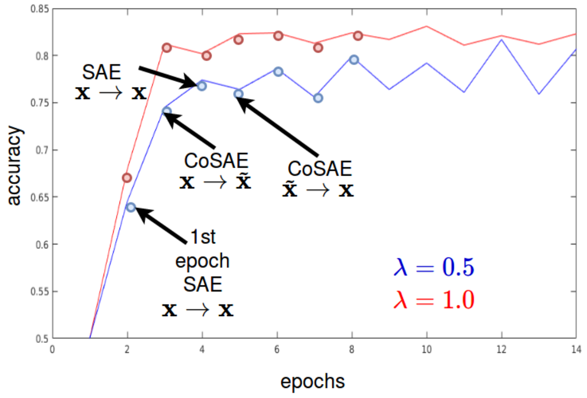

**Fig. 4**. Evolution of accuracy as a result of unsupervised training

## 4. CONCLUSION AND FURTHER WORK

In this paper we proposed an unsupervised method for detecting KHSs that convey meaningful information in Turkish sign language videos. We used correspondence sparse autoencoders that can be used to automatically label the frames of sign language videos using their sparse intermediate layer activations. Experiments conducted on a subset of BS-TID dataset show that this method can be used as a reliable tool to cluster and label sign language videos. Simple HOG features worked better than resnet18 features trained on ImageNet. Fine tuning the resnet18 on hand data may yield better features. In this work, we performed classifications on a frame basis, without temporal modeling. Our future work will incorporate temporal models such as recurrent neural networks.

## 5. ACKNOWLEDGMENTS

Authors would like to thank Mehmet Burak Kurutmaz for his invaluable help with the cluster-class mapping learning system that is used in frame error rate calculation.

## 6. REFERENCES

[1] Ashwin Thangali, *Exploiting phonological constraints for handshape recognition in sign language video*, Ph.D. thesis, Boston University, 2013.

[2] P Boyes Braem, “Functions of the mouthing component in the signing of deaf early and late learners of swiss german sign language,”,” *Foreign Vocabulary in Sign Languages: A Cross-Linguistic Investigation of Word Formation, D. Brentari, Ed. Mahwah, NJ: Erlbaum*, pp. 1–47, 2001.

[3] Rachel Sutton-Spence, “Mouthings and simultaneity in british sign language,” *AMSTERDAM STUDIES IN THE THEORY AND HISTORY OF LINGUISTIC SCIENCE SERIES 4*, vol. 281, pp. 147, 2007.

[4] Sabina Fontana, “Mouth actions as gesture in sign language,” *Gesture*, vol. 8, no. 1, pp. 104–123, 2008.

[1] http://dogasiyli.com/icassp2020-ukhs-cosae/

[5] Erdefi Rakun, Mirna Andriani, I Wayan Wiprayoga, Ken Danniswara, and Andros Tjandra, "Combining depth image and skeleton data from kinect for recognizing words in the sign system for indonesian language (sibi [sistem isyarat bahasa indonesia])," in *2013 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*. IEEE, 2013, pp. 387–392.

[6] Wenjin Tao, Ze-Hao Lai, Ming C Leu, and Zhaozheng Yin, "American sign language alphabet recognition using leap motion controller," in *Proceedings of the 2018 Institute of Industrial and Systems Engineers Annual Conference (IISE 2018)*, 2018.

[7] Sandrine Tornay, Marzieh Razavi, Necati Cihan Camgoz, Richard Bowden, and Mathew Magimai Doss, "Hmm-based approaches to model multichannel information in sign language inspired from articulatory features-based speech processing," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 2817–2821.

[8] B. Shi, A. M. Del Rio, J. Keane, J. Michaux, D. Brentari, G. Shakhnarovich, and K. Livescu, "American sign language fingerspelling recognition in the wild," in *2018 IEEE Spoken Language Technology Workshop (SLT)*, Dec 2018, pp. 145–152.

[9] Bowen Shi, Aurora Martinez Del Rio, Jonathan Keane, Diane Brentari, Greg Shakhnarovich, and Karen Livescu, "Fingerspelling recognition in the wild with iterative visual attention," 2019.

[10] Virender Ranga, Nikita Yadav, and Pulkit Garg, "American sign language fingerspelling using hybrid discrete wavelet transform-gabor filter and convolutional neural network," *Journal of Engineering Science and Technology*, vol. 13, no. 9, pp. 2655–2669, 2018.

[11] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze, "Deep clustering for unsupervised learning of visual features," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 132–149.

[12] Onno Crasborn and Han Sloetjes, "Enhanced elan functionality for sign language corpora," in *6th International Conference on Language Resources and Evaluation (LREC 2008)/3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora*, 2008, pp. 39–43.

[13] Nicolas Pugeault and Richard Bowden, "Spelling it out: Real-time asl fingerspelling recognition," in *2011 IEEE International conference on computer vision workshops (ICCV workshops)*. IEEE, 2011, pp. 1114–1119.

[14] Alexander Rakowski and Lukasz Wandzik, "Hand shape recognition using very deep convolutional neural networks," in *Proceedings of the 2018 International Conference on Control and Computer Vision*, New York, NY, USA, 2018, ICCCV '18, pp. 8–12, ACM.

[15] Iva Farag and Heike Brock, "Learning motion disfluencies for automatic sign language segmentation," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 7360–7364.

[16] Xianwei Jiang, "Isolated chinese sign language recognition using gray-level co-occurrence matrix and parameter-optimized medium gaussian support vector machine," in *Frontiers in Intelligent Computing: Theory and Applications*, pp. 182–193. Springer, 2020.

[17] Walaa Aly, Saleh Aly, and Sultan Almotairi, "User-independent american sign language alphabet recognition based on depth image and pcanet features," *IEEE Access*, vol. 7, pp. 123138–123150, 2019.

[18] Timor Kadir, Richard Bowden, Eng-Jon Ong, and Andrew Zisserman, "Minimal training, large lexicon, unconstrained sign language recognition.," in *BMVC*, 2004, pp. 1–10.

[19] Oscar Koller, Sepehr Zargaran, and Hermann Ney, "Re-sign: Re-aligned end-to-end sequence modelling with deep recurrent cnn-hmms," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3416–3424, 2017.

[20] R. Cui, H. Liu, and C. Zhang, "Recurrent convolutional neural networks for continuous sign language recognition by staged optimization," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 1610–1618.

[21] Jie Huang, Wengang Zhou, Houqiang Li, and Weiping Li, "Sign language recognition using 3d convolutional neural networks," in *2015 IEEE International Conference on Multimedia and Expo (ICME)*, June 2015, pp. 1–6.

[22] Lionel Pigou, Sander Dieleman, Pieter-Jan Kindermans, and Benjamin Schrauwen, "Sign language recognition using convolutional neural networks," 03 2015, vol. 8925, pp. 572–578.

[23] Oscar Koller, Hermann Ney, and Richard Bowden, "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*, 2016, pp. 3793–3802.

[24] Bolaji Yusuf, Alican Gok, Batuhan Gundogdu, Oyku Deniz Kose, and Murat Saraclar, "Temporally-aware acoustic unit discovery for zerospeech 2019 challenge," *Proc. Interspeech 2019*, pp. 1098–1102, 2019.

[25] Herman Kamper, Micha Elsner, Aren Jansen, and Sharon Goldwater, "Unsupervised neural network based feature extraction using weak top-down constraints," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5818–5822.

[26] Daniel Renshaw, Herman Kamper, Aren Jansen, and Sharon Goldwater, "A comparison of neural network methods for unsupervised representation learning on the zero resource speech challenge," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[27] "Bosphorussign," `https://www.cmpe.boun.edu.tr/pilab/BosphorusSign/home_en.html`, Accessed: 2019-10-20.

[28] Navneet Dalal and Bill Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*. IEEE, 2005, vol. 1, pp. 886–893.

[29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[30] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.