Zero-Shot Sign Language Recognition: Can Textual Data Uncover Sign Languages?

Yunus Can Bilge[1]
yunuscan.bilge@hacettepe.edu.tr
Nazli Ikizler-Cinbis[1]
nazli@cs.hacettepe.edu.tr
Ramazan Gokberk Cinbis[2]
gcinbis@ceng.metu.edu.tr

[1] Hacettepe University
Department of Computer Engineering
Ankara, Turkey

[2] Middle East Technical University
Department of Computer Engineering
Ankara, Turkey

Abstract

We introduce the problem of zero-shot sign language recognition (ZSSLR), where the goal is to leverage models learned over the seen sign class examples to recognize the instances of unseen signs. To this end, we propose to utilize the readily available descriptions in sign language dictionaries as an intermediate-level semantic representation for knowledge transfer. We introduce a new benchmark dataset called *ASL-Text* that consists of 250 sign language classes and their accompanying textual descriptions. Compared to the ZSL datasets in other domains (such as object recognition), our dataset consists of limited number of training examples for a large number of classes, which imposes a significant challenge. We propose a framework that operates over the body and hand regions by means of 3D-CNNs, and models longer temporal relationships via bidirectional LSTMs. By leveraging the descriptive text embeddings along with these spatio-temporal representations within a zero-shot learning framework, we show that textual data can indeed be useful in uncovering sign languages. We anticipate that the introduced approach and the accompanying dataset will provide a basis for further exploration of this new zero-shot learning problem.

1 Introduction

Sign language recognition (SLR) is one of the open problems in computer vision, with several challanges remain to be addressed. For instance, while the definitions of the signs are typically clear and structural, the meaning of a sign can change based on the shape, orientation, movement, location of the hand, body posture, and non-manual features like facial expressions [52, 63]. Even the well-known hand shapes in the sign languages can be difficult to discriminate and annotate even due to viewpoint changes [40]. In addition, similar to natural languages, sign languages change and embrace variations over time. Therefore, the development of scalable methods to deal with such variations and challenges is needed.

The existing SLR approaches typically require a large number of annotated examples for each sign class of interest [3, 7, 28, 29, 53]. Towards overcoming the annotation bottleneck in scaling up SLR recognition, we explore the idea of recognizing sign language classes with

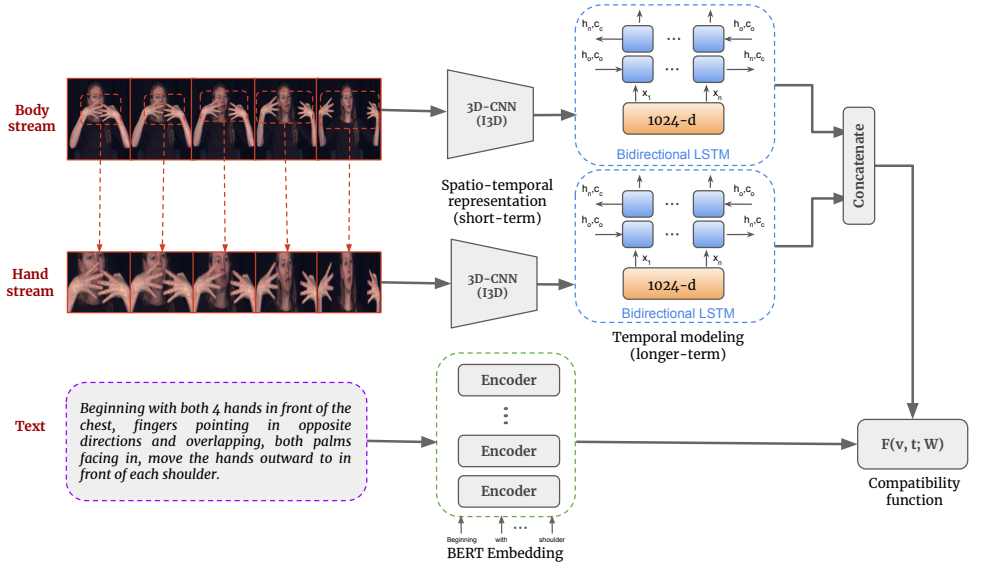


Figure 1: Overview of the proposed zero-shot sign language recognition (ZSSLR) approach.

no annotated visual examples, by leveraging their textual descriptions. To this end, we introduce the problem of *zero-shot sign language recognition* (ZSSLR). Unlike the traditional supervised SLR, where training and test classes are the same, in ZSSLR, the aim is to recognize sign classes that are not seen during training. Compared to the commonly studied ZSL problems, where most seen (training) classes have a large number of per class samples [14, 31, 43, 59], ZSSLR takes ZSL into a new extreme where most seen classes have only few training examples. This challenging situation corresponds to a *hard zero-shot learning* problem [34].

In order to realize seen to unseen class transfer, we use textual descriptions of sign classes taken from a sign language dictionary. Using a sign language dictionary for obtaining the class representations has two major advantages of being (i) readily available, and, (ii) prepared by the sign language experts in a detailed way. In this manner, we construct our ZSSLR approach on highly-informative class descriptions without requiring any ad-hoc annotations, unlike possible alternative approaches such as the attribute-based ZSL [14, 31].

To study ZSSLR, we introduce a new benchmark dataset with 250 sign classes and their textual definitions. Our benchmark dataset is constructed from the ASLLVD corpus [40], where the top 250 sign classes with most in-class variance are selected and the corresponding descriptions are gathered from the Webster American Sign Language Dictionary [8][1].

We propose an embedding-based framework for ZSSLR that consists of two main components. First component models the visual data with extra attention to temporal and spatial structure, using 3D-CNNs and LSTMs. These networks operate over body and hand regions of video in conjunction, since hands are important focal points of signs. The second component, the ZSL component, learns an embedding of the visual representation to the closest text description. We rigorously evaluate our proposed approach on ASL-Text dataset and show its advantages.

[1]The dataset is available at: https://ycbilge.github.io/zsslr.html

To summarize, the main contributions of the paper are as follows: (i) we formulate the problem of zero-shot sign language recognition (ZSSLR), (ii) we define a new benchmark dataset for ZSSLR called ASL-Text, (iii) we propose a spatio-temporal representation that focuses hand and full body regions via 3D-CNNs and LSTMs and learn it in an end-to-end manner for ZSSLR, and, (iv) we present the benchmark results with detailed analyses.

2 Related Work

Sign Language Recognition (SLR). SLR has been studied more than three decades [56]. The mainstream SLR approaches can be grouped into two categories: (i) Isolated SLR [60], and, (ii) Continuous SLR [7]. Our work belongs to isolated SLR category as we target to recognize individual sign instances.

Early SLR methods mostly use hand-crafted features in combination with a classifier, such as support vector machines. Hidden Markov Models (HMM), Conditional Random Fields and neural network based approaches have also been explored to model the temporal patterns [20, 23]. Recently, several deep learning based SLR approaches have been proposed [5, 7, 9, 24, 29, 37, 38, 47, 53].

Despite the relative popularity of the topic, the problem of annotated data sparsity has been seldomly addressed in SLR research. Farhadi and Forsyth [12] is first to study the alignment of sign language video subtitles and signs in action to overcome annotation difficulty. Their approach [13] was based on transferring large amounts of labelled avatar data with few labelled human signer videos to spot words in videos. Buehler *et al*. [2] also try to reduce the annotation effort by using the subtitles of British Sign Language TV broadcasts. They apply Multiple Instance Learning (MIL) to recognize signs out of TV broadcast subtitles. Kelly *et al*. [26] and Pfister *et al*. [45] also use subtitles of TV broadcasts. Pfister *et al*. [45] differ from the two aforementioned MIL studies as they track the co-occurrences of lip and hand movements to reduce the search space for visual and textual content mapping. Nayak *et al*. [39] proposes to locate signs in continuous sign language sentences using iterated conditional modes. Pfister *et al*. [46] define each sign class with one strongly supervised example, and, train an SVM based detector out of one-shot examples. The resulting detector is then used to acquire more training samples from an another weakly-labeled data. Koller *et al*. [29] propose a combined CNN and HMM approach to train a model with large but noisy data. None of the aforementioned models approach the problem of annotated data sparsity from a zero-shot learning perspective.

Zero-Shot Learning. ZSL has been a focus of interest in vision and learning research in recent years, especially following the pioneering works of Lampert *et al*. [30] and Farhadi *et al*. [14]. The main idea is learning to generalize a recognition model for identifying unseen classes. Most of the ZSL approaches rely on transferring semantic information from seen to unseen classes. For this purpose, semantic attributes are largely used in the literature [14, 15, 18, 25, 31, 33, 42]. Semantic word/sentence vectors and concept ontologies are also studied in this context [11, 16, 32, 35, 41, 49, 51]. Label embedding models are explored to make connection between seen and unseen classes via semantic representations [1, 16, 17, 41, 48, 50, 54]. Akata *et al*. [1] propose a method to learn a compatibility function from visual to semantic feature space. As opposed to models that learn to map to a semantic space, there are also studies that learn to map to a common embedding space [17, 50].

Recently, ZSL has also been explored in the context of action recognition. Liu *et al*. [33] is first to propose attribute based model for recognizing novel actions. Jain *et al*. [25]

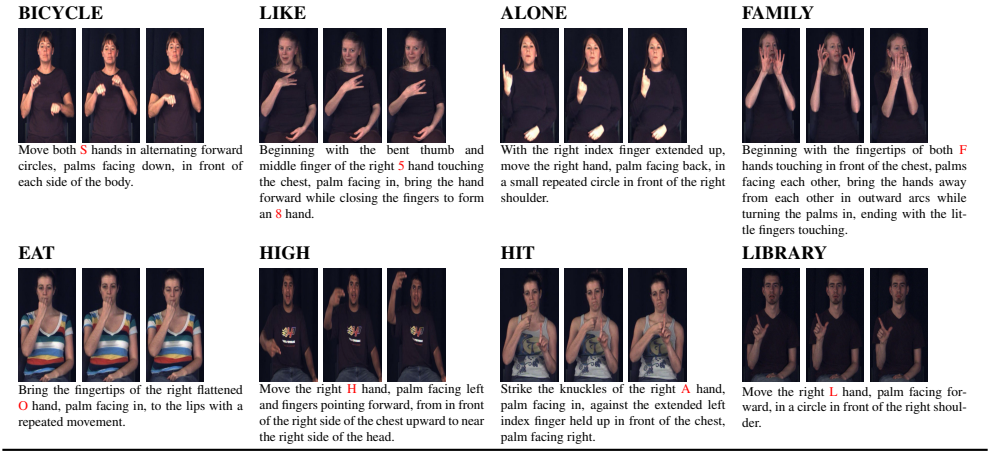


Figure 2: Example sequences and corresponding textual descriptions from the ASL-Text dataset. For visualization purposes, only the person regions of the videos are shown.

propose a semantic embedding based approach using commonly available textual descriptions, images, and object names. Xu *et al*. [65] propose a regression based method to embed actions and labels to a common embedding space. Xu *et al*. [66] also use word-vectors as a semantic embedding space in transductive settings. Wang *et al*. [61] exploit human actions via related textual descriptions and still images. Their aim is to improve word vector semantic representations of human actions with additive information. Habibian *et al*. [21] also propose to learn semantic representations of videos with freely available video and relevant descriptions. Qin *et al*. [48] use error-correcting output codes to overcome the disadvantages of attributes and/or semantic word embeddings for information transfer. Compared to action recognition, in SLR, even a subtle change in motion and/or handshape can change the entire meaning. Therefore, we argue that specialized methods are required for zero-shot recognition in SLR.

There are a couple of recent methods that introduce ZSL to gesture recognition. However, these methods are mostly limited to either robot interactions with single held-out classes ([57]), or based on attributes with limited datasets ([34]). We argue that attribute based semantic representations can be subjective and there is a high chance of missing beneficial attributes when annotating attributes manually. As also noted by [57], attribute based semantic representations are difficult to scale up as defining the attributes of even a single class can require a laborious amount of human effort. In this work, we work over an extensive dataset of classes for sign language and present an approach that does not require any manual attribute annotations.

3 ASL-Text Dataset

To facilitate ZSSLR research, we use the ASLLVD dataset [40], which is the largest isolated sign language recognition dataset available, to the best of our knowledge. We select top 250 sign classes, ranked by the number of samples per class, from ASLLVD signer variances and augment this dataset with the textual definitions of the signs from Webster American Sign Language Dictionary [8]. We refer to this new benchmark dataset as *ASL-Text*. Example

frames and their textual descriptions for the ASL-Text dataset are presented in Figure 2.

The textual descriptions include the detailed instructions of a sign with emphasis on four basic parts: hand-shape, orientation of the palms (forward, backward, etc.), movements of the hands (right, left, etc.), and the location of hands with respect to the body (in front of the chest, each side of the body, right shoulder, etc.). Moreover, some descriptions also include non-manual cues such as the facial expressions, head movements and body posture. Hand shapes are described with specialized vocabulary including the terms *F-hand*, *A-hand*, *S-hand*, *5-hand*, *8-hand*, *10-hand*, *open-hand*, *bent-v hand*, *flattened-o hand* [8]. Such a specialized vocabulary highlights the fact that ASL is a *language* on its own. From the example hand shapes shown in Figure 2, it can be seen that the textual sign language descriptions are indeed quite indicative of the ongoing gesture.

In the ASL-Text dataset, there are 1598 videos (54151 frames) in total for the 250 sign classes. The number of frames of individual videos range between 6 to 116, where the average sequence length is 33 frames. For ZSL purposes, we split the dataset into three disjoint sets (train, validation and test) based on classes. Train set includes 170, validation and test sets include 30 and 50 disjoint classes, respectively. The classes with most signer variation and in-class samples are assigned to training set. The remaining classes, which have relatively lower number of visual examples, are allocated into validation and test sets. This is done to demonstrate the real-world case; *i.e.* it is harder to train classifiers for classes that are rarely seen, therefore, we train with the classes that have relatively more examples and test on the rare classes. Overall, we have 1188, 151, and 259 video samples in training, validation, and test sets. The average length of the textual descriptions is 30 words per description, where the total vocabulary includes 154 distinct words.

The average number of instances per class is 7 for the training classes and 5 for the validation and test classes. Note that, still, the number of examples per class even for training is considerably lower than the commonly studied ZSL datasets, *e.g.* AWA-2 [54] and SUN Attribute [43], on which hundreds of per-class examples are used for training.

4 Methodology

In this section, we first give a formal definition of the problem, and then explain the components of the proposed approach, an overview of which is given in Figure 1. The implementation details can be found in Section 5.1.

Problem definition. In ZSSLR, there are two sources of information: (i) the *visual domain* $\mathbb{V}$, which consists of sign videos, and, (ii) the *textual domain* $\mathbb{T}$, which includes the textual sign descriptions. At training time, the videos, labels and the sign descriptions, are available only for the *seen* classes, $\mathbb{C}_s$. At test time, our goal is to correctly classify the examples of novel *unseen* classes, $\mathbb{C}_u$, which are distinct from the seen classes.

The training set $S_{tr} = \{(v_i, c_i)\}_{i=1}^{N}$ consists of N samples where v_i is the i-th training video and $c_i \in \mathbb{C}_s$ is the corresponding sign class label. We assume that we have access to a textual description of each class c, represented by $\tau(c)$. The goal is to learn a zero-shot classifier that can correctly assign each test video to a class in $\mathbb{C}_u$, based on the textual descriptions.

In our approach, we aim to construct a label embedding based zero-shot classification model. For this purpose, we define the compatibility function $F(v,c)$ as a mapping from an input video and class pair to a score representing the confidence that the input video v belongs to the the class c. Given the compatibility function F, the test-time zero-shot classification

function $f : \mathbb{V} \rightarrow \mathbb{C}_u$ is defined as:

$$f(v) = \arg\max_{c \in \mathbb{C}_u} F(v, c). \quad (1)$$

In this way, we leverage the compatibility function to recognize novel signs at test time.

The performance of the resulting zero-shot sign recognition model directly depends on three factors: (i) video representation, (ii) class representation, and, (iii) the model used as the compatibility function F. The following three sections provide the corresponding details.

4.1 Spatio-temporal video embedding

We aim to obtain an effective video representation by extracting short-term spatio-temporal features using ConvNet features of the video snippets, and then capturing longer-term dynamics through recurrent models. We additionally improve our representation by extracting features in two separate streams: the full frames and hand regions only. The details are given in the following paragraphs.

Short-term spatio-temporal representation. We obtain our basic spatio-temporal representation by first splitting each video into 8 frames long snippets and then extracting their features using a pre-trained I3D model [5], a state-of-the-art 3D-ConvNet architecture. The I3D model is obtained by adapting a pre-trained Inception model [55] to the video domain and then fine-tuning on the Kinetics dataset. We obtain our most basic video representation by average pooling the resulting snippet features.

Modeling longer-term dependencies. Average pooling the 3D-CNN features is a well-performing technique for the recognition of non-complex (singleton) actions. Signs, on the contrary, portray more complicated patterns that are composed of the sequences of multiple basic gestures. In order to capture the transition dynamics and longer-term dependencies across the snippets of a video, we use recurrent network models that take the I3D representation sequence as input, and, provide an output embedding. For this purpose, we propose to use the bidirection LSTM (bi-LSTM) [19] model, and, compare it against the average pooling, LSTM [22] and GRU [6] models.

Two-stream video representation. Hands play a central role in expressing signs. In order to encode details of the hand-area information in a manner isolated from the the overall body movements, we detect and crop the hand regions using OpenPose [4] and form a hand-only sequence corresponding to each video snippet. We define two separate streams, including I3D and bi-LSTM networks, over these video inputs and then concatenate the resulting features to obtain the final video representation (Figure 1). When using recurrent networks, we train both streams together with the compatibility function in an end-to-end fashion.

4.2 Text-based class embeddings

We extract contextualized language embeddings from textual sign descriptions using the state-of-the-art language representation model BERT [10]. BERT architecture basically consists of a stack of encoders; specifically, multi-layer bidirectional transformers[53]. The model's main advantage over word2vec [36] and glove [44] representations is that BERT model is contextual and the extracted representations of the words change with respect to other words in a sentence.

Figure 3 shows the t-SNE visualization of all sign class BERT embeddings. A close inspection to this feature space reveals that classes that appear closer in t-SNE embeddings

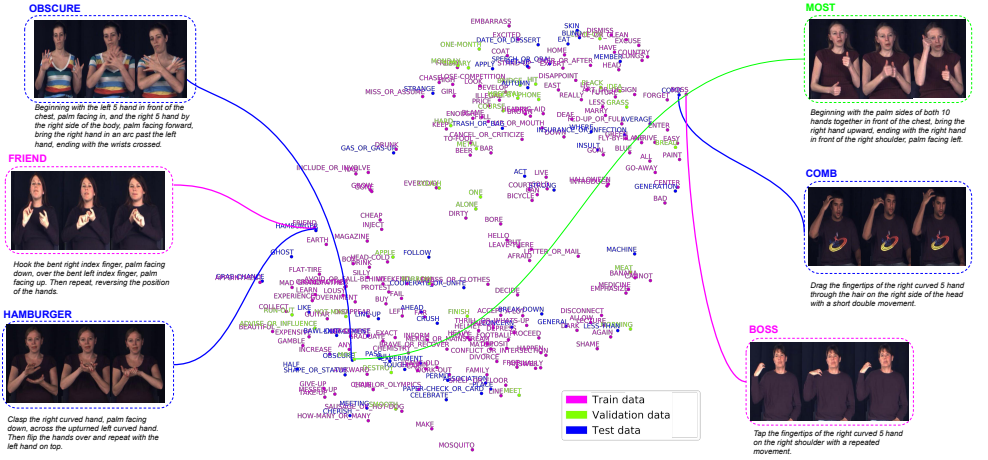


Figure 3: t-SNE visualization of sign descriptions using BERT-[10] embeddings. Nearby descriptions typically correspond to visually similar signs. Best viewed in color, with zoom.

have indeed similar descriptions. For instance, *friend* and *hamburger* signs are composed of similar motions with different hand-shapes, *obscure* and *most* signs have similar hand movements but different hand shapes and directions, and, *comb* and *boss* signs include the same repeated movement with different hand-shape and locations with respect to the body.

4.3 Zero-shot learning model

In our work, we adapt a label embedding [1, 52] based formulation to tackle the ZSSLR problem. More specifically, we use bi-linear compatibility function that associates the video and class representations:

$$F(v,c) = \theta(v)^T W \phi(\tau(c)) \tag{2}$$

where $\theta(v)$ is the d-dimensional embedding of the video v, $\phi(\tau(c))$ is the m-dimensional BERT embeddings of the textual description $\tau(c)$ for the class c, and, W is the $d \times m$ compatibility matrix. In order to learn this matrix, we use the cross entropy loss with ℓ_2-regularization:

$$\min_W -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp\{F(v_i,c_i)\}}{\sum_{c_j \in \mathbb{C}_s} \exp\{F(v_i,c_j)\}} + \lambda \|W\|^2 \tag{3}$$

where λ is the regularization weight. This core formulation is also used in [52], in a completely different ZSL problem. Since the objective function is analogous to the logistic regression classifier, we refer to this approach as *logistic label embedding* (LLE).

In addition to LLE, we also adapt the *embarrassingly simple zero-shot learning* (ES-ZSL) [50] and *semantic auto-encoder* (SAE) [27] formulations as baselines. We, however, skip their formulational details here for brevity.

5 Experiments

5.1 Implementation Details

We fix the number of video frames of each sign video to 32 by either down-scaling or up-scaling. For every consecutive 8 frames, we extract 1024-d features from the last average

Table 1: Comparison of different ZSL formulations. Here, I3D [5] features are extracted over the whole frames (*i.e.* body stream) only.

Method	Val (30 Classes)	Test (50 Classes)		
	top- 1	top-1	top-2	top-5
Random	3.3	2.0	4.0	10.0
SAE	10.6	8.0	12.0	16.0
ESZSL	12.0	**16.9**	**26.0**	**44.4**
LLE	**14.1**	11.4	21.2	41.1

pooling layer of the I3D model using a stride of 4. When modeling the longer temporal context, we set LSTM's or bi-LSTM's initial hidden and cell state to average pool of each sequence during training. Hence, hidden size equals to the size of average pooled feature vector, which is 1024. For representing text, we use the $\text{BERT}_{\text{BASE}}$ model [10] and extract 768-dimensional sentence-based features. Following the description in [10], we concatenate the features from the last four layers of the pretrained Transformer of $\text{BERT}_{\text{BASE}}$ and l_2-normalize them.

We measure normalized accuracy, *i.e.* the mean accuracy per class, in all experiments. We run each experiment 5 times and report the average. Top-1, top-2, and top-5 accuracies for the random baseline are calculated by averaging over 10000 runs.

5.2 Experimental Results

We first evaluate the ZSL component of our framework. In this context, we explore three different ZSL approaches, namely SAE [27], ESZSL [50], and LLE. In these set of experiments, we have pooled the extracted 3D-CNN features over the whole frame. Table 1 shows the corresponding results, where top-1 validation accuracy, and top-1, top-2, and top-5 test accuracies are reported.

We observe that SAE [27] performs poorly with respect to other approaches. We think that this is due to auto-encoder structure of SAE method. The model learns linear embedding from video to semantic space with the purpose of reconstruction back from semantic space to video. This idea might not work well since we do not have many in-class samples for reconstruction. In addition, as stated earlier, intra-class variance is very high among signers.

Consequently, we evaluate the performance of the two-stream spatio-temporal representation of the framework. Specifically, we carry out an ablation study, where *body* denotes the full frame input stream, *hand* denotes the hand videos and *body+hand* is the case when these two streams are used in conjunction. The corresponding results are given in Table 2. Hand stream provides additional cues and increases the performance for validation classes using both methods. In test classes, ESZSL [50] does not perform well on the hand-stream; on the contrary, its performance increases when both streams are used in conjunction. Similarly, LLE benefits from the introduction of hand-stream, and outperforms ESZSL method when two streams are utilized together. Overall, we observe that proposed framework based on LLE formulation works better, especially regarding top-1 and top-2 accuracies.

We further evaluate the effect of longer temporal modeling with different RNN architectures. We experiment with three different RNN models, namely LSTM[22], GRU[6] and bi-LSTM[19] units using LLE over both hand and body streams. Table 3 presents these results. We observe that, compared to average pooling of streams, the framework benefits from the introduction of longer temporal modeling over all architectures, and performs the best with bi-LSTMs. This illustrates the importance of visual representations for ZSSLR.

Table 2: Evaluation of two-stream spatio-temporal representation. Here, *body* denotes the full frame input stream, whereas *hand* denotes the videos of hand regions and *body+hand* is the case when these two streams are used in conjunction. Here, average pooling is used in aggregating the short-term video representations.

Method	visual rep.	Val (30 Classes)	Test (50 Classes)		
		top - 1	top-1	top-2	top-5
Random	-	3.3	2.0	4.0	10.0
ESZSL	body	12.0	16.9	26.0	**44.4**
	hand	13.3	11.6	19.6	33.7
	body + hand	14.6	17.1	25.7	43.0
LLE	body	14.1	11.4	21.2	41.1
	hand	15.0	12.6	19.8	37.8
	body + hand	**16.2**	**18.0**	**27.4**	43.8

Table 3: Comparison of different RNN units with LLE method.

Temporal Representation	top-1	top-2	top-5
AvePool	18.0	27.4	43.8
LSTM [22]	18.2	28	47.2
GRU [6]	19.7	31.8	50.0
bi-LSTM [19]	**20.9**	**32.5**	**51.4**

Our overall proposed framework reaches a top-1 normalized accuracy of 20.9% and top-5 normalized accuracy of 51.4%, which is quite impressive compared to top-1 and top-5 accuracies of random baseline (2.0% and 10.0% respectively).

Figure 4 shows examples from correctly and incorrectly classified test sequences. We observe that, either the textual descriptions or the visual aspects of the classes confused with each other are very similar. This indicates that the problem domain can benefit from more detailed analyses and representations that focus on nuances, both in visual and in textual domain, which can be explored as a future direction.

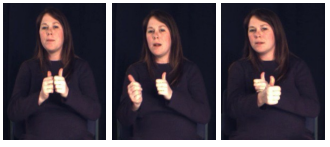
6 Conclusion

This paper introduces and explores the problem of ZSSLR. We present a benchmark dataset for this novel problem by augmenting a large ASL dataset with sign language dictionary descriptions. Our proposed framework builds upon the idea of using these auxiliary texts as an additional source of information to recognize unseen signs. We propose an end-to-end trainable ZSSLR method that focuses hand and full body regions via 3D-CNNs+LSTMs and learns a compatibility function via label embedding. Overall, the experimental results indicate that, zero-shot recognition of signs based on textual descriptions can be possible. Nevertheless, the acquired accuracy levels are quite low compared to other ZSL domains, pinpointing a substantial need for further exploration in this direction.

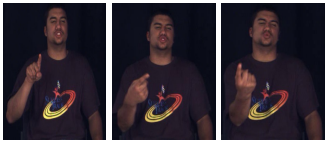
7 Acknowledgements

This work was supported in part by TUBITAK Career Grant 116E445.

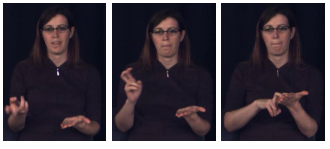
Correctly Predicted Label: STRANGE
Move the right C hand from near the right side of the face, palm facing left, downward in an arc in front of the face, ending near the left side of the chin, palm facing down.



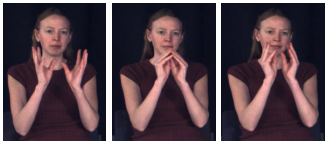
Correctly Predicted Label: AHEAD
Beginning with the palm sides of both A hands together, move the right hand forward in a small arc.



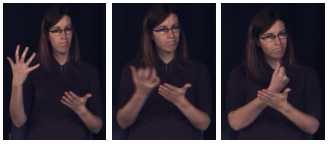
Correctly Predicted Label: INSULT
Move the extended right index finger from in front of the right side of the body, palm facing left and finger pointing forward, forward and upward sharply in an arc.



Correctly Predicted Label: GET-UP
Place the fingertips of the right bent V hand, palm facing in and fingers pointing down, on the upturned palm of the left open hand held in front of the body.



Predicted Label: BREAK-DOWN
Beginning with the fingertips of both curved 5 hands touching in front of the chest, palms facing each other, allow the fingers to loosely drop, ending with the palms facing down.
Correct Label: MEETING
Beginning with both open hands in front of the chest, palms facing each other and fingers pointing up, close the fingers with a double movement into flattened O hands while moving the hands together.



Predicted Label: AVERAGE
Brush the little-finger side of the right open hand, palm facing left, back and forth with a short repeated movement on the index-finger side of the left open hand, palm angled right.
Correct Label: GRAB-CHANCE
Bring the right curved 5 hand from in front of the right side of the body, palm facing left and fingers pointing forward, in toward the body in a downward arc while changing into an S hand, brushing the little-finger side of the right S hand across the palm of the left open hand, palm facing up in front of the chest.

Figure 4: Example predictions of our proposed model. The first four rows show examples that are correctly predicted and the last two rows show incorrect predictions, together with the textual descriptions of the predicted and ground-truth classes.

References

[1] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-embedding for attribute-based classification. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 819–826, 2013.

[2] Patrick Buehler, Andrew Zisserman, and Mark Everingham. Learning sign language by watching tv (using weakly aligned subtitles). In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2961–2968. IEEE, 2009.

[3] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, and Richard Bowden. Subunets: End-to-end hand shape and continuous sign language recognition. In *Proc. IEEE Int. Conf. on Computer Vision*, pages 3075–3084. IEEE, 2017.

[4] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. In *arXiv preprint arXiv:1812.08008*, 2018.

[5] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6299–6308, 2017.

[6] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *EMNLP*, 2014.

[7] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. Neural sign language translation. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7784–7793, 2018.

[8] Elaine Costello. *Random House Webster's Concise American Sign Language Dictionary*. Random House, 1999.

[9] Runpeng Cui, Hu Liu, and Changshui Zhang. Recurrent convolutional neural networks for continuous sign language recognition by staged optimization. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7361–7369, 2017.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[11] Mohamed Elhoseiny, Babak Saleh, and Ahmed Elgammal. Write a classifier: Zero-shot learning using purely textual descriptions. In *Proc. IEEE Int. Conf. on Computer Vision*, pages 2584–2591, 2013.

[12] Ali Farhadi and David Forsyth. Aligning asl for statistical translation using a discriminative word model. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, volume 2, pages 1471–1476. IEEE, 2006.

[13] Ali Farhadi, David Forsyth, and Ryan White. Transfer learning in sign language. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1–8. IEEE, 2007.

[14] Ali Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. Describing objects by their attributes. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1778–1785. IEEE, 2009.

[15] Vittorio Ferrari and Andrew Zisserman. Learning visual attributes. In *Proc. Adv. Neural Inf. Process. Syst.*, pages 433–440, 2008.

[16] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, et al. Devise: A deep visual-semantic embedding model. In *Proc. Adv. Neural Inf. Process. Syst.*, pages 2121–2129, 2013.

[17] Yanwei Fu, Timothy M Hospedales, Tao Xiang, Zhenyong Fu, and Shaogang Gong. Transductive multi-view embedding for zero-shot recognition and annotation. In *Proc. European Conf. on Computer Vision*, pages 584–599. Springer, 2014.

[18] Yanwei Fu, Timothy M Hospedales, Tao Xiang, and Shaogang Gong. Learning multimodal latent attributes. *IEEE transactions on pattern analysis and machine intelligence*, 36(2):303–316, 2014.

[19] Alex Graves and Jürgen Schmidhuber. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 18(5-6):602–610, 2005.

[20] Kirsti Grobel and Marcell Assan. Isolated sign language recognition using hidden markov models. In *IEEE International Conference on Systems, Man, and Cybernetics. Computational Cybernetics and Simulation*, volume 1, pages 162–167. IEEE, 1997.

[21] Amirhossein Habibian, Thomas Mensink, and Cees GM Snoek. Video2vec embeddings recognize events when examples are scarce. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(10):2089–2103, 2017.

[22] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

[23] Chung-Lin Huang and Wen-Yi Huang. Sign language recognition using model-based tracking and a 3d hopfield neural network. *Machine vision and applications*, 10(5-6):292–307, 1998.

[24] Jie Huang, Wengang Zhou, Houqiang Li, and Weiping Li. Sign language recognition using 3d convolutional neural networks. In *IEEE Int. Conf. on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2015.

[25] Mihir Jain, Jan C van Gemert, Thomas Mensink, and Cees GM Snoek. Objects2action: Classifying and localizing actions without any video example. In *Proc. IEEE Int. Conf. on Computer Vision*, pages 4588–4596, 2015.

[26] Daniel Kelly, John Mc Donald, and Charles Markham. 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.

[27] Elyor Kodirov, Tao Xiang, and Shaogang Gong. Semantic autoencoder for zero-shot learning. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3174–3183, 2017.

[28] Oscar Koller, Jens Forster, and Hermann Ney. Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *Comput. Vis. Image Understand.*, 141:108–125, 2015.

[29] Oscar Koller, O Zargaran, Hermann Ney, and Richard Bowden. Deep sign: hybrid cnn-hmm for continuous sign language recognition. In *British Machine Vision Conference*, 2016.

[30] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 951–958. IEEE, 2009.

[31] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(3):453–465, 2014.

[32] Jimmy Lei Ba, Kevin Swersky, Sanja Fidler, et al. Predicting deep zero-shot convolutional neural networks using textual descriptions. In *Proc. IEEE Int. Conf. on Computer Vision*, pages 4247–4255, 2015.

[33] Jingen Liu, Benjamin Kuipers, and Silvio Savarese. Recognizing human actions by attributes. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3337–3344. IEEE, 2011.

[34] Naveen Madapana and Juan P Wachs. Hard zero shot learning for gesture recognition. In *IAPR International Conference on Pattern Recognition*, pages 3574–3579. IEEE, 2018.

[35] Thomas Mensink, Efstratios Gavves, and Cees GM Snoek. Costa: Co-occurrence statistics for zero-shot classification. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2441–2448, 2014.

[36] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Proc. Adv. Neural Inf. Process. Syst.*, pages 3111–3119, 2013.

[37] Pavlo Molchanov, Xiaodong Yang, Shalini Gupta, Kihwan Kim, Stephen Tyree, and Jan Kautz. Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural network. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4207–4215, 2016.

[38] Pradyumna Narayana, Ross Beveridge, and Bruce A Draper. Gesture recognition: Focus on the hands. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5235–5244, 2018.

[39] Sunita Nayak, Sudeep Sarkar, and Barbara Loeding. Automated extraction of signs from continuous sign language sentences using iterated conditional modes. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2583–2590. IEEE, 2009.

[40] Carol Neidle, Ashwin Thangali, and Stan Sclaroff. Challenges in development of the american sign language lexicon video dataset (asllvd) corpus. In *Proc. 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon, Language Resources and Evaluation Conference (LREC) 2012*, 2012.

[41] Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Greg S Corrado, and Jeffrey Dean. Zero-shot learning by convex combination of semantic embeddings. *Proc. Int. Conf. Learn. Represent.*, 2014.

[42] Devi Parikh and Kristen Grauman. Relative attributes. In *Proc. IEEE Int. Conf. on Computer Vision*, pages 503–510. IEEE, 2011.

[43] Genevieve Patterson and James Hays. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2751–2758. IEEE, 2012.

[44] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proc. of conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[45] Tomas Pfister, James Charles, and Andrew Zisserman. Large-scale learning of sign language by watching tv (using co-occurrences). In *British Machine Vision Conference*, 2013.

[46] Tomas Pfister, James Charles, and Andrew Zisserman. Domain-adaptive discriminative one-shot learning of gestures. In *Proc. European Conf. on Computer Vision*, pages 814–829. Springer, 2014.

[47] Lionel Pigou, Mieke Van Herreweghe, and Joni Dambre. Sign classification in sign language corpora with deep neural networks. In *International Conference on Language Resources and Evaluation (LREC) Workshop*, pages 175–178, 2016.

[48] Jie Qin, Li Liu, Ling Shao, Fumin Shen, Bingbing Ni, Jiaxin Chen, and Yunhong Wang. Zero-shot action recognition with error-correcting output codes. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2833–2842, 2017.

[49] Marcus Rohrbach, Michael Stark, and Bernt Schiele. Evaluating knowledge transfer and zero-shot learning in a large-scale setting. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1641–1648. IEEE, 2011.

[50] Bernardino Romera-Paredes and Philip Torr. An embarrassingly simple approach to zero-shot learning. In *Proc. Int. Conf. Mach. Learn.*, pages 2152–2161, 2015.

[51] Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. Zero-shot learning through cross-modal transfer. In *Proc. Adv. Neural Inf. Process. Syst.*, pages 935–943, 2013.

[52] William C Stokoe Jr. Sign language structure: An outline of the visual communication systems of the american deaf. *Journal of deaf studies and deaf education*, 10(1):3–37, 2005.

[53] Stephanie Stoll, Necati Cihan Camgoz, Simon Hadfield, and Richard Bowden. Sign language production using neural machine translation and generative adversarial networks. In *British Machine Vision Conference*. British Machine Vision Association, 2018.

[54] Gencer Sumbul, Ramazan Gokberk Cinbis, and Selim Aksoy. Fine-grained object recognition and zero-shot learning in remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 56(2):770–779, 2018.

[55] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1–9, 2015.

[56] Shinichi Tamura and Shingo Kawasaki. Recognition of sign language motion images. *Pattern recognition*, 21(4):343–353, 1988.

[57] Wil Thomason and Ross A Knepper. Recognizing unfamiliar gestures for human-robot interaction through zero-shot learning. In *International Symposium on Experimental Robotics*, pages 841–852. Springer, 2016.

[58] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proc. Adv. Neural Inf. Process. Syst.*, pages 5998–6008, 2017.

[59] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[60] Hanjie Wang, Xiujuan Chai, Xiaopeng Hong, Guoying Zhao, and Xilin Chen. Isolated sign language recognition with grassmann covariance matrices. *ACM Transactions on Accessible Computing (TACCESS)*, 8(4):14, 2016.

[61] Qian Wang and Ke Chen. Alternative semantic representations for zero-shot human action recognition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 87–102. Springer, 2017.

[62] Jason Weston, Samy Bengio, and Nicolas Usunier. Large scale image annotation: learning to rank with joint word-image embeddings. *Machine learning*, 81(1):21–35, 2010.

[63] Ying Wu and Thomas S Huang. Vision-based gesture recognition: A review. In *International Gesture Workshop*, pages 103–115. Springer, 1999.

[64] Yongqin Xian, Bernt Schiele, and Zeynep Akata. Zero-shot learning-the good, the bad and the ugly. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4582–4591, 2017.

[65] Xun Xu, Timothy M. Hospedales, and Shaogang Gong. Semantic embedding space for zero-shot action recognition. *2015 IEEE International Conference on Image Processing (ICIP)*, pages 63–67, 2015.

[66] Xun Xu, Timothy Hospedales, and Shaogang Gong. Transductive zero-shot action recognition by word-vector embedding. *International Journal of Computer Vision*, 123(3):309–333, 2017.

[67] Yi Zhu, Yang Long, Yu Guan, Shawn Newsam, and Ling Shao. Towards universal representation for unseen action recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 9436–9445, 2018.