Enhancing Zero-Shot Learning Based Sign Language Recognition Through Hand Landmarks and Data Augmentation

Received: 30 January 2023; Accepted: 5 March 2023 *Research Article*

Giray Sercan ÖZCAN
Department of Computer Engineering
Baskent University
Ankara, Turkey
gozcan@baskent.edu.tr
0000-0002-8770-2085

Emre SÜMER
Department of Computer Engineering
Baskent University
Ankara, Turkey
esumer@baskent.edu.tr
0000-0001-8502-9184

Yunus Can BİLGE
Image and Video Processing Group
HAVELSAN
Ankara, Turkey
ycbilge@havelsan.com.tr
0000-0003-2811-3747

***Abstract*— Sign language recognition remains a challenging area and may require a considerable amount of data to obtain satisfactory results. To overcome this, we use readily available motion text data in addition to videos for achieving recognition of unobserved classes during the training phase. Zero-Shot Sign Language Recognition (ZSSLR) with a novel technique is focused on this work, which learns a model from seen sign classes and recognizes unseen sign classes. To achieve this, the ASL-Text dataset is used which combines the video of word signs and descriptions in sign language dictionaries. Moreover, this dataset consists of sign language classes and their corresponding definitions in the sign language dictionary. In various Zero-Shot Learning (ZSL) applications, it is common for datasets to contain a limited number of examples for numerous classes across different domains. This makes the problem of sign language recognition extremely challenging. We try to overcome this by using a new approach which includes augmented data and hand landmarks. The experiment on augmented data resulted in 50.91 for top-5 accuracy. Hand landmarks are used with unaugmented data which is applied to average and LSTM deep learning layers resulting in 49.41 and 48.21 for top-5 accuracies, respectively.**

Keywords—sign language recognition, zero-shot learning

I. INTRODUCTION

The objective of Sign Language Recognition (SLR) systems that have been created is to convert sign language into either text or speech, with the goal of enabling communication between deaf people and those who can hear. This process has a significant impact on society, but the complexity of hand and finger actions make SLR very challenging. The task of developing systems for recognizing sign language remains a formidable challenge. Although sign language definitions are well-defined and organized, slight variations in body posture, hand movements, facial expressions and hand positioning can drastically alter the intended meaning of the sign language [3], [4]. Distinguishing and annotating the well-established hand shapes within sign languages can prove to be a formidable task, especially in situations where there are variations in viewpoints [31]. Moreover, akin to how natural languages transform and incorporate diversity throughout history, sign languages also undergo modifications and accept variations as time passes. Therefore, a model that can adapt to these changes is needed. The current methods used for SLR necessitate a considerable quantity of labeled information for each class. [11], [12], [13], [17]. In this study, unseen sign language classes were recognized without annotated visual data by taking advantage of sign language descriptions. In this regard, Zero-shot Sign Language Recognition (ZSSLR) has been defined in [5]. Unlike normal supervised learning, the ZSSLR method tries to predict classes that are not seen in the training phase. Compared to common ZSL studies [6], [7], [8], [9], ASL-Text [5] dataset used in this study contains significantly fewer examples per class for training which makes this task a hard zero-shot learning problem [23]. In general, ZSSLR consists of two primary components. The first component focuses on the arrangement of visual information by utilizing 3D-CNN and LSTM to examine both the temporal and spatial structure. The second one is the ZSL component which contains text data. The system developed with these two components aims to learn the closest text description of the visual data. The ASL-Text dataset was created using easily accessible and expert-prepared sign language definitions of words in the sign language dictionary. In the current study, we improved the effort introduced in [5]. We applied data augmentation and hand landmarks extraction by feeding them to the deep learning layers such as mean layer and LSTM layer. The experimental results were evaluated by top-1, top-2, and top-5 accuracies. The rest of the paper is proceeded as follows: The previous studies are examined in Section 2, the proposed approach is explained in Section 3, the applied experiments is presented in section 4, the results are discussed in Section 5 and the conclusion is addressed in Section 6.

II. RELATED WORK

SLR has been a subject of research for more than thirty years [32]. There are two main types of SLR techniques that have gained widespread popularity. These are (i) Isolated SLR [33], which focuses on recognizing individual instances of signs, and (ii) Continuous SLR [34], which aims to recognize all signs in sign language sequences. Our research falls under the category of Isolated SLR since we focus on recognizing individual sign instances. In the initial stages of SLR research, hand-crafted features were predominantly utilized in conjunction with classifiers such as support vector machines [33], [35]. Additionally, Conditional Random Fields, Hidden Markov Models (HMM), and neural network based techniques were also investigated as potential methods to model temporal patterns [36], [37]. More recently, various SLR methodologies have been proposed that leverage deep learning techniques [38], [39]. ZSL has garnered significant attention in the fields of learning and vision research in recent years, particularly after the groundbreaking works of Lampert et al. [40] and Farhadi et al. [6]. The majority of ZSL methodologies depend on transferring semantic knowledge from observed classes to those that have not been seen. In recent years, various studies have been conducted for action recognition based on semantic

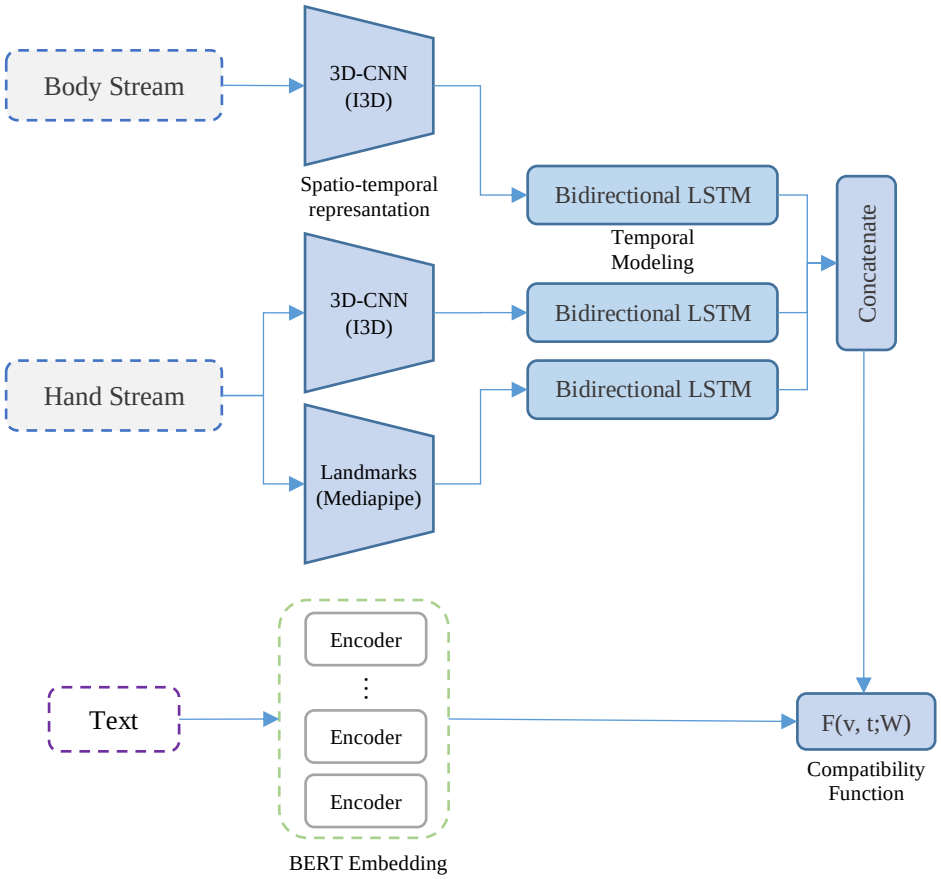


Fig. 1. General outline of the study

embedding [42], regression [43], and many others [44], [45], [46], [1], [2].

Bilge et al. [5] are the first ones who defined the ZSSLR problem and proposed a solution. They created the ASL-Text dataset by combining sign language videos with related text descriptions in the sign language dictionary, which is also used in this study. They divided the dataset into hand and body streams. They applied the embarrassingly simple zero-shot learning (ESZSL) [21], semantic auto encoder (SAE) [22], and logistic label embedding (LLE) [5] methods to the streams and obtained various results. The experiment in which they applied the LLE method by combining the hand and body streams gave the best result of 20.9% for top-1 accuracy.

In their subsequent work [20], the authors enhanced the ASL-Text dataset by adding binary feature matrices in addition to the text descriptions used in the second component of the ZSSLR system. They also created two new benchmark datasets, MS-ZSSLR-W and MS-ZSSLR-C, and applied a shift-based CNN [24] in addition to the 3D-CNN and LSTM used in their previous work. They also introduced the problem of Generalized Zero-Shot Sign Language Recognition (GZSSLR), in which the model is trained to recognize both observed and unobserved classes. The results obtained in the study are given in two different settings: ZSL and generalized zero-shot learning (GZSL). ZSL setting achieved 31.3% and 14.7%, the GZSL setting achieved 26.9% and 34.7% for top-1 accuracies on ASL-Text and MS-ZSSLR-C datasets, respectively.

III. METHODOLOGY

The general structure of the developed architecture can be seen in Figure 1. In this section, first, the problem definition is presented and then solution methods are described.

Problem definition: ZSSLR relies on two distinct information sources: a *visual domain* that comprises sign language videos and a *textual domain* consisting of explanations for the gestures and motions performed in these videos. During the training phase videos, labels and sign language descriptions of the observed classes $\mathbb{C}_s$ are incorporated. The objective at test phase is to classify the unobserved novel classes $\mathbb{C}_u$.

The set of training samples, denoted by $S_{tr} = \{(v_i, c_i)\}_{i=1}^{N}$ contains N instances. Here, v_i represents the i-th training video, and $c_i \epsilon \mathbb{C}_s$ is the corresponding sign language video. It is assumed that there is access to the textual descriptions, denoted by $\tau(c)$ for each class. The objective is to acquire a zero-shot classifier capable of assigning each test video to a class in $\mathbb{C}_u$ based on the textual descriptions provided.

The aim is to establish zero-shot classifier model that employs label embedding. To achieve this, a compatibility function, denoted as $F(v, c)$, is defined to measure the similarity between a given input video and class pair, generating a score that reflects the degree of confidence that video v belongs to class c. Based on the compatibility function F, zero-shot classification function at test time $f: \mathbb{V} \longrightarrow \mathbb{C}_u$ is defined as:

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

Using this method, the compatibility function can classify novel unobserved classes that are encountered during the testing phase.

Short-term spatiotemporal representations were obtained with I3D [30] while longer-term dependencies were found with bi-LSTM [25]. The goal of using bi-LSTM [25] is to capture longer term dependencies as effectively as possible. Hand landmarks are extracted from streams using Mediapipe [26]. Text-based class embeddings for sign language descriptions are extracted using BERT [27], which is state-of-the-art in this area. BERT is essentially an encoder stack. The advantage of BERT over word2vec [28] and glove [29] is that the extracted representation is more sensitive to other words in the sentence. The bi-linear compatibility function utilized establishes a relation between the video and representations of class as follows:

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

$\theta(v)$ represents the d-dimensional representation of video v, while $\phi(\tau(c))$ is the m-dimensional BERT embedding of the textual descriptions, $\tau(c)$, for class c. The compability matrix, denoted by W and comprising $d \times m$ dimensions. For calculating this matrix, we use the formula given in [5].

IV. Experiments

Four experiments were conducted, these are (i) the baseline study that achieved the best results in [5], (ii) the study conducted with augmented data, (iii) the study conducted using average pooling layer on hand landmarks, and (iv) the study conducted using LSTM on hand landmarks.

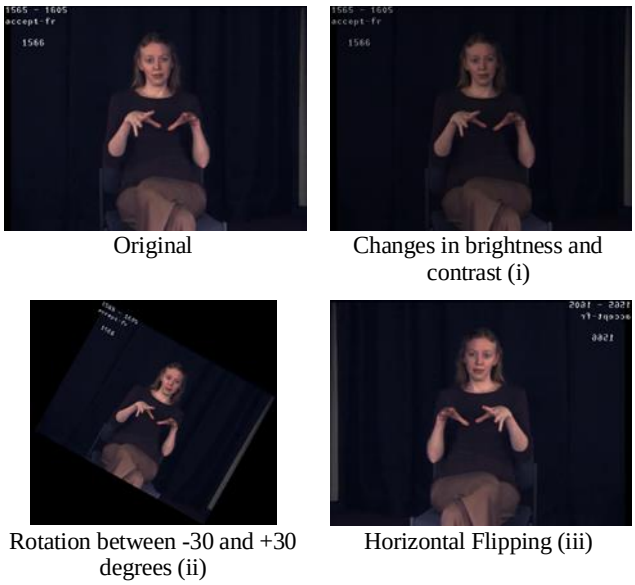
Firstly, hand streams were extracted from ASL-Text which contains signers' body streams. Therefore, two streams were worked on: Body stream and hand stream. These videos were split into 8-frame small video segments. Then we extracted short-term spatiotemporal features and longer-term dependencies from these segments.

The applied augmentations can be seen in Figure 2. These are (i) changes in brightness and contrast, (ii) rotation between -30 and +30 degrees, (iii) horizontal flipping and (iv) mix of augmentations mentioned in (i), (ii), (iii). The dataset was increased five-fold in this way, spatiotemporal representations were obtained, and longer-term dependencies were captured.

Results were obtained by extracting hand landmarks and feeding them to the average pooling layer or LSTM.

V. Results

The results can be seen in Table 1, which includes our study conducted with augmented data and landmarks. In the baseline study, a success rate of 20.38 was achieved on the validation dataset, while for the top-1, top-2, and top-5 on the test dataset, success rates of 16.94, 27.31, and 47.91 were obtained, respectively. In the study conducted with augmented data, a success rate of 19.98 was achieved on the validation dataset, while for the top-1, top-2, and top-5 on the test dataset, success rates of 19.11, 30.89, and 50.91 were found, respectively. In the study conducted using hand landmarks with average pooling, a success rate of 20.4 was obtained on the validation dataset, while for the top-1, top-2, and top-5 on the test dataset, success rates of 18.7, 28.32, and 49.41 were achieved, respectively. In the study conducted using hand landmarks with LSTM, a success rate of 19.98 was found on the validation dataset, while for the top-1, top-2, and top-5 on the test dataset, success rates of 19.31, 28.76, and 48.21 were achieved, respectively.



Original | Changes in brightness and contrast (i)

Rotation between -30 and +30 degrees (ii) | Horizontal Flipping (iii)

Fig. 1. Applied augmentations

Table 1 shows that the results obtained from the experiments are better than those obtained from the baseline study.

Experiments	Val (30 Classes)	Test (50 classes)		
	Top-1	Top-1	Top-2	Top-5
Baseline	20.38	16.94	27.31	47.91
Augmented Data	19.98	19.11	30.89	50.91
Landmarks (average pooling)	20.4	18.7	28.32	49.41
Landmarks (LSTM)	19.98	19.31	28.76	48.21

TABLE I. Experimental Results

VI. Conclusion

In this study, we aim to do sign language recognition with zero-shot learning method. We utilized techniques to increase the amount of data available for training and extracted hand landmarks by inputting them into deep learning layers like the mean layer and LSTM layer. Even with average pooling, using hand landmarks has led to an improvement in results. The best results are obtained from the study conducted using hand landmarks and LSTM. Better results can be obtained by generating more augmented data.

References

[1] J. Qin, L. Liu, L. Shao, F. Shen, B. Ni, J. Chen, and Y. Wang, "Zeroshot action recognition with error-correcting output codes," in Proc. IEEE Conf. Comput. Vis. Pattern Recog., 2017, pp. 2833–2842.

[2] M. Hahn, A. Silva, and J. M. Rehg, "Action2vec: A crossmodal embedding approach to action learning," in The British Machine Vision Conference (BMVC), September 2018.

[3] W. 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.

[4] Y. Wu and T. S. Huang, "Vision-based gesture recognition: A review. In International Gesture Workshop," pp.103–115. Springer, 1999.

[5] Y. C. Bilge, N. İ. Cinbiş, R. G. Cinbiş, "Zero-Shot Sign Language Recognition: Can Textual Data Uncover Sign Languages?," in British Machine Vision Conference (BMVC), 2019

[6] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. "Describing objects by their attributes," in Proc. IEEE Conf. Comput. Vis. Pattern Recog., pp.1778–1785. IEEE, 2009.

[7] C. H. Lampert, H. Nickisch, and S. Harmeling, "Attribute-based classification for zero-shot visual object categorization," IEEE Transactions on Pattern Analysis and Machine Intelligence, 36(3):453–465, 2014.

[8] G. Patterson and J. Hays, "Sun attribute database: Discovering, annotating, and recognizing scene attributes," in Proc. IEEE Conf. Comput. Vis. Pattern Recog., pp.2751–2758. IEEE, 2012

[9] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. "The caltechucsd birds-200-2011 dataset," 2011.

[10] P. Kumar, H. Gauba, P.P. Roy , D.P. Dogra, "A multimodal framework for sensor based sign language recognition," Neurocomputing, vol.259, pp.21-38, 2017.

[11] O. Koller, J. Forster,H. Ney, "Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers," in Computer Vision and Image Understanding, vol.141, pp.108-125, 2015.

[12] S. Tamura and S. Kawasaki, "Recognition of sign language motion images," Pattern recognition, vol. 21, no. 4, pp. 343–353, 1988.

[13] M. B. Waldron and S. Kim, "Isolated asl sign recognition system for deaf persons," IEEE Transactions on rehabilitation engineering, vol. 3, no. 3, pp. 261–271, 1995.

[14] M. W. Kadous et al., "Machine recognition of auslan signs using powergloves: Towards large-lexicon recognition of sign language," in Proc. Workshop on the Integration of Gesture in Language and Speech, vol. 165, 1996

[15] M. Zahedi, D. Keysers, T. Deselaers, and H. Ney, "Combination of tangent distance and an image distortion model for appearancebased sign language recognition," in Joint Pattern Recognition Symposium, 2005, pp. 401–408

[16] H. Cooper and R. Bowden, "Sign language recognition using boosted volumetric features," in Proc. IAPR Conference on Machine Vision Applications, 2007, pp. 359–362

[17] O. Koller, O. Zargaran, H. Ney, and R. Bowden, "Deep sign: hybrid cnn-hmm for continuous sign language recognition," In British Machine Vision Conference, 2016.

[18] K. Grobel and M. Assan, "Isolated sign language recognition using hidden markov models," In IEEE International Conference on Systems, Man, and Cybernetics. Computational Cybernetics and Simulation, vol. 1, pp. 162–167. IEEE, 1997.

[19] C. L. Huang and W.Y. Huang. "Sign language recognition using model-based tracking and a 3d hopfield neural network," Machine vision and applications, 10(5-6):292–307, 1998

[20] Y. C. Bilge, N. İ. Cinbiş, R. G. Cinbiş, "Towards Zero-Shot Sign Language Recognition," Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 1, pp. 1217-1232, 2023.

[21] B. Romera-Paredes and P. Torr, "An embarrassingly simple approach to zero-shot learning," in Proc. Int. Conf. Mach. Learn., 2015, pp. 2152–2161.

[22] E. Kodirov, T. Xiang, and S. Gong, "Semantic autoencoder for zero-shot learning," in Proc. IEEE Conf. Comput. Vis. Pattern Recog., 2017, pp. 3174–3183.

[23] N. Madapana and J. P. Wachs, "Hard zero shot learning for gesture recognition," in IAPR International Conference on Pattern Recognition, 2018, pp. 3574–3579.

[24] J. Lin, C. Gan, and S. Han, "Tsm: Temporal shift module for efficient video understanding," in Proc. IEEE International Conference on Computer Vision, 2019, pp. 7083–7093.

[25] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional lstm and other neural network architectures," Neural Networks, vol. 18, no. 5-6, pp. 602–610, 2005.

[26] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.L. Chang, M. Yong, J. Lee, W.T. Chang. "Mediapipe: A framework for perceiving and processing reality." InThird Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR) 2019 Jun (Vol. 2019).

[27] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pretraining of deep bidirectional transformers for language understanding," in NAACL, 2019, pp. 4171–4186.

[28] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean,"Distributed representations of words and phrases and theircompositionality," in Proc. Adv. Neural Inf. Process. Syst., 2013, pp.3111–3119.

[29] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in Proc. of conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532–1543.

[30] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in Proc. IEEE Conf. Comput. Vis. Pattern Recog., 2017, pp. 6299–6308.

[31] C. Neidle, A. Thangali, and S. 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

[32] S. Tamura and S. Kawasaki, "Recognition of sign language motion images," Pattern recognition, vol. 21, no. 4, pp. 343–353, 1988.

[33] H. Wang, X. Chai, X. Hong, G. Zhao, and X. Chen, "Isolated sign language recognition with grassmann covariance matrices," ACM Transactions on Accessible Computing (TACCESS), vol. 8, no. 4, p. 14, 2016.

[34] N. Cihan Camgoz, S. Hadfield, O. Koller, H. Ney, and R. Bowden, "Neural sign language translation," in Proc. IEEE Conf. Comput. Vis. Pattern Recog., 2018, pp. 7784–7793.

[35] M. B. Waldron and S. Kim, "Isolated asl sign recognition system for deaf persons," IEEE Transactions on rehabilitation engineering, vol. 3, no. 3, pp. 261–271, 1995.

[36] K. Grobel and M. Assan, "Isolated sign language recognition using hidden markov models," in IEEE International Conference on Systems, Man, and Cybernetics. Computational Cybernetics and Simulation, vol. 1, 1997, pp. 162–167.

[37] C.-L. Huang and W.-Y. Huang, "Sign language recognition using model-based tracking and a 3d hopfield neural network," Machine vision and applications, vol. 10, no. 5-6, pp. 292–307, 1998.

[38] D. Li, C. Rodriguez, X. Yu, and H. Li, "Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison," in The IEEE Winter Conference on Applications of Computer Vision, 2020, pp. 1459–1469.

[39] B. Saunders, N. C. Camgoz, and R. Bowden, "Progressive transformers for end-to-end sign language production," in European Conference on Computer Vision (ECCV), 2020.

[40] C. H. Lampert, H. Nickisch, and S. Harmeling, "Learning to detect unseen object classes by between-class attribute transfer," in Proc. IEEE Conf. Comput. Vis. Pattern Recog., 2009, pp. 951–958.

[41] J. Liu, B. Kuipers, and S. Savarese, "Recognizing human actions by attributes," in Proc. IEEE Conf. Comput. Vis. Pattern Recog., 2011, pp. 3337–3344.

[42] M. Jain, J. C. van Gemert, T. Mensink, and C. G. Snoek, "Objects2action: Classifying and localizing actions without any video example," in Proc. IEEE Int. Conf. on Computer Vision, 2015, pp. 4588–4596.

[43] X. Xu, T. M. Hospedales, and S. Gong, "Semantic embedding space for zero-shot action recognition," 2015 IEEE International Conference on Image Processing (ICIP), pp. 63–67, 2015.

[44] X. Xu, T. Hospedales, and S. Gong, "Transductive zero-shot action recognition by word-vector embedding," International Journal of Computer Vision, vol. 123, no. 3, pp. 309–333, 2017.

[45] Q. Wang and K. Chen, "Alternative semantic representations for zero-shot human action recognition," in Joint European Conference on Machine Learning and Knowledge Discovery in Databases, 2017, pp. 87–102.

[46] A. Habibian, T. Mensink, and C. G. Snoek, "Video2vec embeddings recognize events when examples are scarce," IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 10, pp. 2089–2103, 2017.