Automatic Sign Language sentence recognition in the context of guided museum tours for deaf and hard-of-hearing visitors

Erion-Vasilis Pikoulis[0000−0002−2337−5142], Aristeidis Bifis[0000−0003−0246−1209], Nikolaos Arvanitis[0000−0002−2750−4669], and Dimitrios Kosmopoulos[0000−0003−3325−1247]

Computer Engineering and Informatics Department, University of Patras, Rio 26504, Greece
http://www.ceid.upatras.gr
{pikoulis,bifis,arvanitis}@ceid.upatras.gr, {dkosmo}@upatras.gr

Abstract. Sign Language (SL) translation remains an extremely challenging task despite recent breakthrough progress in the constituent fields of Computer Vision and Machine Learning, especially when tackled under a general unconstrained setup. This is mainly due to a combination of the difficulty of the task itself on the one hand, and the absence of large-scale labeled datasets that would enable using end-to-end Deep Learning based solutions, on the other. In such cases, the ability to incorporate prior information can yield a significant improvement on the translation results, by greatly restricting the search space of the potential solutions. In this work we treat the SL translation problem in the limited confinement of an interactive, guided museum tour for deaf and hard-of-hearing visitors. Prior domain knowledge enables us to compile a list of targeted questions per museum exhibit and use these lists to create an SL training dataset for the solution of the problem at hand. In our case, SL question recognition is treated as a sentence retrieval problem, whereby the goal is to predict visitor's question that best matches the available pool of possible inquiries. Our preliminary evaluation using both tailored deep architectures and traditional non-deep solutions has led to promising results for the recognition task at hand.

Keywords: sign language recognition · convolutional neural networks.

1 Introduction

The Deaf (with capital D) are defined as a group of people, with varying hearing acuity, whose primary mode of communication is a visual language, predominantly Sign Language (SL), and have a shared heritage and culture. Most Deaf prefer utilizing their native SL in their interaction with others and often avoid using writing/reading due to their rather poor written language skills [9]. This communication barrier severely impacts their interactions with the non-Deaf, significantly limiting their job opportunities, and their accessibility to healthcare services and education, among others. This situation is only made worse

by the scarcity of dedicated SL interpreters that can help alleviate the issue via their live presence or through relay services.

To help mitigate the problem, automated translation systems are recently gaining both in popularity and in performance, especially since the advent and widespread use of Deep Neural Networks. However, despite the progress, automatic SL translation (SLT) remains an open and extremely challenging task when tackled under a general unconstrained framework, requiring an interdisciplinary approach for its solution [10]. This difficulty mainly stems from the fact there are multiple information streams contributing to a sign expression, including handshapes, facial expressions, body posture, combined with the extensive use of depiction as well as epenthesis and co-articulation effects that often take place in the signing process [17].

Recent methods based on networks with self-attention (Transformers) [11], [39], that currently represent the state-of-the-art in SLT, have yielded promising results, but require large corpora for training in order to achieve their generalization potential. This aspect is especially critical for the SLT task, where the profound lack of annotated data for supervised training is well documented and it is mainly contributed to the complicated language structures of SL, and also, to the fact that almost all SLs are minority languages.

In this work, to address the translation task at hand and overcome the data scarcity issue, we follow a domain-specific approach, in the context of museum tours, utilizing a priori domain and context knowledge in order to limit the search space, thereby facilitating its solution and enhancing the quality of the obtained results. Specifically, the presented work is part of an interactive mobile application for deaf and hard-of-hearing museum visitors, developed in the framework of the SignGuide project [6].

As non-profit, open to the public institutions that acquire, conserve, research, and exhibit the material and non-material heritage of humanity, museums are among the most important institutions for fostering lifelong learning [40]. The importance of interacting with deaf users in their native SL has already been recognized by museums[5], although currently they rely mainly on utilizing specialized interpreters to fill this gap. However, efforts are increasingly being dedicated towards alleviating this communication barrier, aiming at offering deaf users automated interactive environments in the form of web pages, mobile applications, and dedicated software, utilizing tools such as video-based SLT systems, speech recognition, and avatars [18]. This is the main goal of projects such as SignGuide mentioned above, as well as ARCHES [2], and Deaf Museums [3], among others.

The paper is structured as follows. In Section 2 we summarize some of the most important recent works in the field of SL translation/recognition. In Section 3 we present an overview of the recognition task addressed in this paper, highlighting the main approach used for its solution and the general setup behind the proposed tool. In Sections 4, 5 we describe in detail a non-deep and a deep learning based treatment of the problem, respectively. Our experimental evaluation of the proposed solution is presented in Section 6, and finally, Section 7 contains our conclusions.

2 Related work

Sign Language Translation has been commonly regarded as a recognition problem (see [25] [35] for details). Early approaches attempted to recognize individual and well-segmented signs by employing discriminative or generative methods within a time-series classification framework; examples include hidden Markov models (HMMs), e.g., [13] [38], [27] dynamic time warping, e.g.,[7], [29], and conditional random fields, e.g., [36], [41]. These methods used hand-crafted features; more recently, deep learning methods, such as those derived from CNNs, provided some superior representations, e.g., [34], [32].

The recognition approach, however, has rather limited real-world utility because it produces a group of words with relatively nonsensical context structure rather than a natural language output. As a result, SLT with continuous recognition is a lot more realistic framework, but it is also far more difficult to implement [26], [24], [8]. The difficulty stems from epenthesis (the incorporation of extra visual clues into signs), co-articulation (the conclusion of one sign affects the beginning of the next), and spontaneous sign generation (which may include slang, special expressions, etc.). [23] used a model comprised of a CNN-LSTM network to produce features, which are then fed to HMMs that do inference using a variation of the Viterbi method to handle the challenge. A 2D-CNN with cascaded 1D convolutional layers for feature extraction has been proposed in [22], using also a bi-directional LSTM (BLSTM) for continuous SL recognition, and utilizing the Levenshtein distance to produce gloss-level alignments. Along the same lines, the authors in [14], combine a 2D fully convolutional network with a feature enhancement module to obtain better gloss alignments. [15] employed a BLSTM fed with CNN features while [21] utilizes an adaptive encoder-decoder architecture leveraging a hierarchical BLSTM with attention over sliding windows on the decoder. A network called STMC was proposed in [42], which incorporates several cues from position and picture (hands, face, holistic) in multiple scales and feeds them to a penultimate connectionist temporal classification (CTC) layer.

The recently proposed Transformer architectures enable SLT to drastically enhance translation performance. This is amplified when SLT is combined with an SLR procedure, either as an intermediate activity or in the context of a multitask learning scheme. In particular, in [12], the authors use a Transformer network to achieve end-to-end translation. They essentially suggest an S2(G+T) architecture: They propose a Transformer network to conduct S2T, and they use the Transformer's encoder to forecast the respective gloss sequence ground-truth. The latter SLR task is carried out over all potential gloss alignments by a penultimate CTC layer [20]. Training is done collaboratively for the entire system (both tasks). The need for that intermediate step has been alleviated in later works such as [39] where a winner-takes-all activation is integrated into the Transformer architecture. In [33], the authors introduce a context-aware continuous sign language recognition using a generative adversarial network architecture. The elaborated system exploits text or contextual information to enhance the recognition accuracy, contrary to previous works that only consider

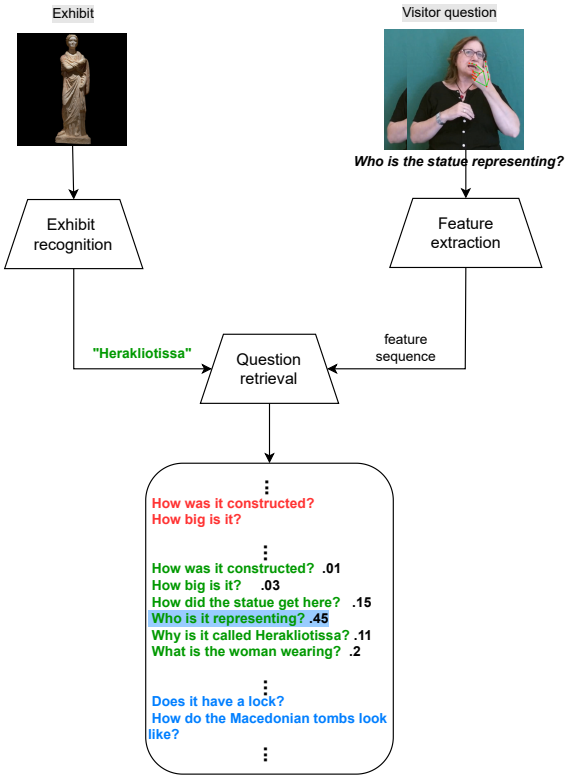

Fig. 1: Architectural overview of the SignGuide recognition system.

spatio-temporal features from video sequences. In particular, it recognizes sign language glosses by extracting these features and assess the prediction quality by modeling text information at the sentence and gloss levels.

Despite the aforementioned developments, such works still face issues in more complex real-world scenarios, mainly due to the lack of available data. On the contrary, they are most often implemented on small dictionaries relevant to certain real-world contexts, for which very labor-intensive annotation has taken place e.g., weather reports [19]. The question is how to use these advancements in real scenarios when not enough training data is available, but the structure of the conversation is more or less known, e.g., by following a protocol and can be modeled up to a certain extent a priori. To our knowledge there has been no such effort in the related literature for the SLT. This work aspires to contribute towards bridging this gap.

3 System overview

The recognition system presented in this paper is part of a museum guide app for deaf and hard-of-hearing visitors, developed for the Archaeological Museum of Thessaloniki [1], in the framework of the SignGuide project [6]. The app offers SL content for 10 selected museum exhibits, along with museum related info. Integral part of the app is also an automatic visual-based exhibit recognition system that is able to recognize the exhibit of interest as the visitor nears the exhibit area, pointing the device camera towards it. Manual exhibit selection is also at hand in case the automatic recognition system fails (e.g. due to changing lighting conditions).

Apart from being able to browse through the offered content, the visitor is also given the opportunity to pose a live question in SL, using the forward-facing camera of the device. The recorded video is sent to the SignGuide server for processing by the SL recognition system, which replies with the recognition result and the relevant content (i.e. the response to the recognized question) is displayed to the user.

Regarding SL recognition, which constitutes the main focus of this paper, we exploit our prior domain knowledge to simplify the problem to a great extent, as shown in Fig. 1. More specifically, we first compiled a comprehensive list of questions pertaining to the content of each of the 10 included exhibits, with the help of the domain experts involved in the project. Based of this list, we created a training SL dataset where by each question is signed by multiple signers. Assuming that the exhibit under consideration is known at the time the visitor's question is posed, the recognition task at hand is then treated as an exhibit-specific question retrieval problem, where the goal is to predict the question from the known list related to the exhibit of interest, that best matches the one signed by the visitor.

This domain-specific approach helped us overcome the lack of extensive training datasets (that would enable more elaborate end-to-end SL translation systems) and employ simple retrieval strategies, while still leveraging a satisfactory prediction performance, which emphasizes the importance of incorporating prior knowledge to the solution of highly complex problems such as the automatic interpretation of SL.

For the task at hand, we employ a pipeline involving landmark detection, feature extraction and matching/classification. Two such retrieval solutions are presented in subsequent sections of this paper. The first one utilizes hand-related features and employs a feature clustering approach to encode the SL video content, borrowing ideas from document retrieval tasks. On the other hand, the second approach utilizes a CNN encoder in order to transform the input sequence into a context vector, which is then fed to a classification layer for question prediction.

3.1 The SignGuide dataset

An original dataset of SL inquiries related to the 10 exhibits of interest was created in the course of the project for training and validation purposes of our SL recognition system. For the creation of the dataset, a list of potential visitor questions was compiled by the museum's archaeologists, and was subsequently signed by nine expert signers. Specifically, six were native signers, while three were experienced interpreters. There were 366 questions in the compiled list, leading to a total of 3294 questions in the SL corpus. We recorded high quality video using a machine vision camera, while the corpus is annotated at both sentence and gloss levels. The corpus size distribution per exhibit is presented in Table 1

Table 1: SL corpus size used for question recognition

Exhibit	1	2	3	4	5	6	7	8	9	10
Questions	33	15	27	28	34	46	35	70	40	38
SL corpus	297	135	243	252	306	414	315	630	360	342

3.2 Feature extraction

The first processing step in the proposed pipeline involves passing the SL videos though the "Hands" and "Pose" modules of Mepiapipe ([30], [4]) in order to infer landmark locations. Specifically, the MediaPipe tool estimates 21 landmarks per hand, and 25 upper-body landmarks for each video frame (please see Fig. 2 for illustrative examples from the available dataset).

Using these landmarks we extract rotation- and translation- invariant features in the form of pair-wise landmark distances. Specifically, regarding hand-related features, we calculate the fingertip distances corresponding to the signer's dominant hand. Similar in nature features are also extracted from the signer's body posture by means of the wrist distances from the torso and face landmarks, as well as the distance between them. This way, a total of 15 hand-related and 25 pose related features are extracted from each video frame, with the feature vectors being subsequently normalized to suppress the scale parameter.

4 Bag-of-words based SL question retrieval

In this non-deep treatment of the problem, as a dimensionality reduction step, we first use our training data to estimate latent hand-shapes by grouping the N_{frames} hand-related feature vectors into a small number of k clusters, with $k \ll N_{\text{frames}}$, where N_{frames} denotes the total number of video frames in our training corpus. We anticipate the cluster centroids to represent the fundamental hand-shapes

Fig. 2: Instances from the SignGuide dataset with hand and pose landmark annotation using the MediaPipe tool[4].

that are present in our collection of SL videos (allowing also for transitional frames).

Here we use only hand-related features since, on the one hand they represent the most informative data stream for SL interpretation, and on the other, they are by nature more easily standardized and grouped than the pose-related ones (while also limiting the dimensions of the feature space), thus facilitating the clustering task to a great extent.

SL question modelling In order to model our SL questions, we assign each video frame to the cluster its corresponding feature vector belongs to, thus transforming the SL input into a sequence of latent handshapes (namely, cluster labels).

However, when assessing the similarity between SL inputs, we must take into account that SL videos of the same question may have different number of frames, and even more importantly, that the question at hand may be signed in many different ways (e.g. by altering the order of the contained glosses). Thus a direct comparison between label (handshape) sequences becomes problematic. To overcome this obstacle, we utilize the bag-of-words concept [37] widely used in document processing tasks (here, the questions correspond to "documents" and the latent handshapes to the "words" that comprise them), and represent each input via the histogram of the latent handshapes that are present in it. Thus, question q_i is represented via the k-dimensional vector $\mathbf{h}_i$, defined as:

$$\mathbf{h}_i \equiv \left[f_1^i, f_2^i, \ldots, f_k^i\right]^{\mathrm{T}}, \tag{1}$$

where

$$f_j^i = \frac{n_j^i}{N_{\text{frames}}^i}, \tag{2}$$

with n_j^i denoting the number of appearances of the j-th latent hand-shape in the SL video of the i-th question.

Using this bag-of-words modelling, we can define a simple question retrieval system that assigns the unknown visitor question to its closest neighbor in the

training dataset, having obtained its bag-of-words representation by assigning each input feature vector to one of the feature clusters obtained from the training dataset.

5 Deep Learning based question retrieval

In this approach, instead of eliminating the time-dependencies between SL video frames, we incorporate them in our solution by utilizing deep encoders to model the time-related information present in the SL input. Although recurrent neural networks and transformers are generally considered as more suitable for sequence modelling tasks (as mentioned in Section 2), due to the limited size of our training corpus, the results obtained by such architectures for the task at hand were generally poor.

Instead, in this work we utilize a simple Convolutional Neural Network (CNN) structure comprised of several 1D convolutional layers. The input to the encoder is an $N_{features} \times N_{input}$ matrix, where $N_{features}$ is the dimension of the feature vector (per frame), while N_{input} is the number of input frames. As it becomes obvious, the CNN encoder requires a constant number of input frames across all inputs. Since in general SL videos have different lengths, to fulfill this requirement, we utilize downsampling and/or zero padding along the time (frame) domain as appropriate. The number of features on the other hand was $N_{features} = 40$, since in this case we make use of both hand and pose related features.

In this framework, the convolutional kernels operate along the time dimension, with all kernels having a constant size of 3 (namely, each output feature value is obtained by processing three consecutive feature vectors of the input). Following standard CNN practice, the output of each convolutional layer is followed by batch normalization layer and a max pooling operation that halves the input time resolution, while the number of kernels is doubled between consecutive layers. The output of the final conv layer is flattened and used as input to a single fully-connected layer that carries out the classification task. ReLU activations were used across all layers, while the cross entropy loss was employed for model training.

6 Experimental evaluation

In this section, we present the experimental evaluation of the proposed question retrieval systems, using the SignGuide dataset. For the bag-of-words approach, the dataset was randomly split to $80\% - 20\%$ for system training and testing, respectively, while for the CNN-based approach, the split was $80\% - 10\% - 10\%$ for training, validation, and testing, respectively. In each case, all presented results represent average values of multiple experiments involving random splits of the dataset.

As mentioned in Section 4, regarding the bag-of-words retrieval approach, question recognition is performed by assigning the unknown question to its closest neighbour in the training set, based on some pre-defined distance metric. In our evaluation, we experimented with various distance metrics inspired both from statistics (e.g. the Kolmogorov–Smirnov (KS) statistic [31]), and from document retrieval tasks (e.g., tf-idf [28] and latent semantic analysis [16]), the best results were obtained by simply considering the L^p-norms of the distance between the training and testing histograms. In our experiments we examined the L^1, L^2, and L^∞ norms for our matching task, with the best results being obtained by the L^1 norm, owing to its well-known robustness properties. Moreover, regarding the number of feature clusters (i.e. the latent handshapes used in the bag-of-words model) we experimented with numbers in the range $[50, 150]$, with the best results being obtained for values around $75 - 100$.

On the other hand, the parameter selection and architectural decisions for the CNN-based solution were as follows. Firstly, we fixed the number of input frames at 128, a value that corresponds to around 4 seconds of video and covers the vast majority of question durations in the training corpus. Moreover, regarding the used architecture, after some experimentation with the encoder depth and number of kernels per layer, we opted for a network of 5 convolutional layers, beginning with 64 kernels for the first layer, and with the number of utilized kernels being doubled at each increment in depth level. The selected input and network architecture resulted in a model with roughly 2.3 million parameters, while the context vector dimension was equal to 4096.

6.1 Results

The obtained results from the application of the two retrieval approaches with the parameter selection mentioned in the previous paragraph, are depicted in Fig. 3. As it is expected, the CNN based solution clearly outperforms its non-deep counterpart by a significant margin, having an accuracy advantage ranging from around 11% to more than 20%. In absolute terms, the top-1 accuracy of CNN-based solution ranges from around 85% for the "easier" exhibits $1 - 3$, and $9, 10$, to around 70 -75% for the more difficult cases of exhibits $4 - 8$. The top-3 values are typically $10 - 15\%$ higher than the respective top-1 ones. On the other hand, the analogous figures for the non-deep approach range from around 60% to 77%, and from around 80% to 90%, for the top-1 and top-3 measurements, respectively (with the exception of exhibit 8 where the figures are significantly lower). This difference in performance can be mainly attributed to the CNN's ability to incorporate pose-related information and (even more importantly) to model the time evolution of the feature vector in the produced encoding of the input. The combined effect of these factors is that the CNN-based retrieval solution takes into account both the signer's hand movement and handshape-related information, while the non-deep solution is based only on the latter information stream. Nevertheless, taking into account its very low computational complexity (it essentially requires only a few hundred vector comparisons), the

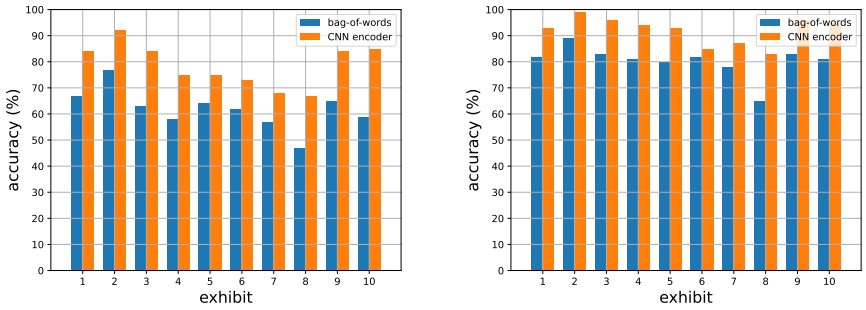

Fig. 3: Experimental performance evaluation of the non-deep and CNN-based question retrieval approaches presented in the paper.

clustering based solution still manages to leverage an acceptable performance (especially concerning its top-3 accuracy).

7 Conclusions and future work

In this work, we presented a system for the automatic retrieval of the visitor's question in the context of a guided museum tours for deaf and HoH visitors. The recognition task at hand is treated as a question retrieval problem, having determined a pool of potential questions per exhibit, with the help of the museum's archaeologists. An original SL dataset was also created based on the compiled list of questions. Two retrieval strategies, namely, a non-deep bag-of-words approach involving feature clustering, as well as a deep learning based approach utilizing CNN encoders, were presented here. The followed domain-specific approach helped us overcome the lack of extensive training datasets and employ simple retrieval strategies, while still leveraging a satisfactory prediction performance, as demonstrated by the presented experimental results. Based on these preliminary findings, it appears that this research direction seems promising for alleviating the need for more annotated data in low-resource languages like the SLs. On the downside, the treatment of the SL translation problem as a question retrieval task (namely, where the potential user-questions are pre-defined), may prove too restrictive in real use. Towards mitigating this issue, we are currently developing an SL recognition tool aiming at predicting the likelihood of a gloss being present in the signed question (thus defining a probability distribution over the available dictionary), given the recorded video and the exhibit of interest. Utilizing this knowledge, the aim is to develop an SL-based retrieval system whereby both the query and content are in SL form (represented by the respective gloss distributions). This will broaden the scope and usability of the system, while at the same time, by omitting the challenging SL translation step, we are expecting that it will also increase its real world performance.

Acknowledgements This research has been co-financed by the European Regional Development Fund of the European Union and Greek national funds through the operational program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH–CREATE–INNOVATE (project code: T1EDK-2-01392).

References

1. Archaelogical Museum of Thessaloniki. https://www.amth.gr/en/
2. ARCHES project. https://www.arches-project.eu/
3. Deaf Museums project. https://www.arches-project.eu/
4. MediaPipe. https://google.github.io/mediapipe/
5. The Met. https://www.metmuseum.org/events/programs/access/visitors-who-are-deaf
6. The SignGuide project. http://signguide.gr/
7. Alon, J., Athitsos, V., Yuan, Q., Sclaroff, S.: A unified framework for gesture recognition and spatiotemporal gesture segmentation. IEEE transactions on pattern analysis and machine intelligence **31**(9), 1685–99 (Sep 2009). https://doi.org/10.1109/TPAMI.2008.203
8. Aloysius, N., Geetha, M.: Understanding vision-based continuous sign language recognition. Multimedia Tools and Applications **79**(31-32), 22177–22209 (May 2020). https://doi.org/10.1007/s11042-020-08961-z, https://doi.org/10.1007/s11042-020-08961-z
9. Babcock, R.D.: Interpreted writing center tutorials with college-level deaf students. Linguistics and Education **22**(2), 95–117 (2011)
10. Bragg, D., Koller, O., Bellard, M., Berke, L., Boudreault, P., Braffort, A., Caselli, N., Huenerfauth, M., Kacorri, H., Verhoef, T., et al.: Sign language recognition, generation, and translation: An interdisciplinary perspective. In: The 21st international ACM SIGACCESS conference on computers and accessibility. pp. 16–31 (2019)
11. Camgoz, N.C., Koller, O., Hadfield, S., Bowden, R.: Sign language transformers: Joint end-to-end sign language recognition and translation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10023–10033 (2020)
12. Camgöz, N.C., Koller, O., Hadfield, S., Bowden, R.: Sign language transformers: Joint end-to-end sign language recognition and translation. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020. pp. 10020–10030. IEEE (2020). https://doi.org/10.1109/CVPR42600.2020.01004, https://doi.org/10.1109/CVPR42600.2020.01004
13. Chatzis, S.P., Kosmopoulos, D.I., Varvarigou, T.A.: Robust sequential data modeling using an outlier tolerant hidden markov model. IEEE transactions on pattern analysis and machine intelligence **31**(9), 1657–69 (Sep 2009). https://doi.org/10.1109/TPAMI.2008.215
14. Cheng, K.L., Yang, Z., Chen, Q., Tai, Y.W.: Fully convolutional networks for continuous sign language recognition. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) Computer Vision – ECCV 2020. pp. 697–714. Springer International Publishing, Cham (2020)

15. Cui, R., Liu, H., Zhang, C.: Recurrent convolutional neural networks for continuous sign language recognition by staged optimization. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1610–1618 (2017). https://doi.org/10.1109/CVPR.2017.175
16. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. Journal of the American society for information science **41**(6), 391–407 (1990)
17. Dudis, P.G.: Depiction of events in ASL: Conceptual integration of temporal components. University of California, Berkeley (2004)
18. Enríquez, R.A.H., Cáceres, J.R.R., Robles, T.d.J.Á.: Accessible interactive systems for deaf users in museums: systematic mapping review. In: 2022 International Conference on Inclusive Technologies and Education (CONTIE). pp. 1–5. IEEE (2022)
19. Forster, J., Schmidt, C., Koller, O., Bellgardt, M., Ney, H.: Extensions of the sign language recognition and translation corpus rwth-phoenix-weather. In: LREC. pp. 1911–1916 (2014)
20. Graves, A., Fernández, S., Gomez, F., Schmidhuber, J.: Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In: Proceedings of the 23rd International Conference on Machine Learning. p. 369–376. ICML '06, Association for Computing Machinery, New York, NY, USA (2006). https://doi.org/10.1145/1143844.1143891, `https://doi.org/10.1145/1143844.1143891`
21. Huang, S., Ye, Z.: Boundary-adaptive encoder with attention method for chinese sign language recognition. IEEE Access **9**, 70948–70960 (2021)
22. Koishybay, K., Mukushev, M., Sandygulova, A.: Continuous sign language recognition with iterative spatiotemporal fine-tuning. In: 2020 25th International Conference on Pattern Recognition (ICPR). pp. 10211–10218 (2021). https://doi.org/10.1109/ICPR48806.2021.9412364
23. Koller, O., Camgoz, N.C., Ney, H., Bowden, R.: Weakly supervised learning with multi-stream cnn-lstm-hmms to discover sequential parallelism in sign language videos. IEEE Transactions on Pattern Analysis and Machine Intelligence **42**(9), 2306–2320 (2020). https://doi.org/10.1109/TPAMI.2019.2911077
24. Koller, O., Zargaran, S., Ney, H.: Re-sign: Re-aligned end-to-end sequence modelling with deep recurrent cnn-hmms. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3416–3424 (2017). https://doi.org/10.1109/CVPR.2017.364
25. Koller, O.: Quantitative survey of the state of the art in sign language recognition, arXiv:2008.09918v2 [cs.cv] (2020)
26. Koller, O., Forster, J., Ney, H.: Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. Computer Vision and Image Understanding **141**, 108 – 125 (2015). https://doi.org/https://doi.org/10.1016/j.cviu.2015.09.013, `http://www.sciencedirect.com/science/article/pii/S1077314215002088`, pose & Gesture
27. Lang, S., Block, M., Rojas, R.: Sign language recognition using kinect. In: International Conference on Artificial Intelligence and Soft Computing. pp. 394–402. Springer (2012)
28. Leskovec, J., Rajaraman, A., Ullman, J.D.: Mining of Massive Datasets. Cambridge University Press, USA, 2nd edn. (2014)
29. Lichtenauer, J.F., Hendriks, E.A., Reinders, M.J.T.: Sign language recognition by combining statistical dtw and independent classification. IEEE transactions on pattern analysis and machine intelligence **30**(11), 2040–6 (Nov 2008). https://doi.org/10.1109/TPAMI.2008.123

30. Lugaresi et. al., C.: MediaPipe: A framework for building perception pipelines. arXiv preprint (2019), https://arxiv.org/pdf/1906.08172.pdf
31. Massey Jr, F.J.: The kolmogorov-smirnov test for goodness of fit. Journal of the American statistical Association **46**(253), 68–78 (1951)
32. Neverova, N., Wolf, C., Taylor, G., Nebout, F.: Moddrop: Adaptive multi-modal gesture recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **38**(8), 1692–1706 (2016). https://doi.org/10.1109/TPAMI.2015.2461544
33. Papastratis, I., Dimitropoulos, K., Daras, P.: Continuous sign language recognition through a context-aware generative adversarial network. Sensors **21**(7) (2021). https://doi.org/10.3390/s21072437, https://www.mdpi.com/1424-8220/21/7/2437
34. Pigou, L., Dieleman, S., Kindermans, P.J., Schrauwen, B.: Sign language recognition using convolutional neural networks. In: Agapito, L., Bronstein, M.M., Rother, C. (eds.) Computer Vision - ECCV 2014 Workshops. pp. 572–578. Springer International Publishing, Cham (2015)
35. Rastgoo, R., Kiani, K., Escalera, S.: Sign language recognition: A deep survey. Expert Systems with Applications **164**, 113794 (2021)
36. Ruiduo Yang, Sarkar, S.: Detecting coarticulation in sign language using conditional random fields. In: 18th International Conference on Pattern Recognition (ICPR'06). vol. 2, pp. 108–112 (2006). https://doi.org/10.1109/ICPR.2006.431
37. Salton, G., McGill, M.J.: Introduction to modern information retrieval. mcgraw-hill (1983)
38. Vogler, C., Metaxas, D.: Handshapes and movements: Multiple-channel american sign language recognition. In: Camurri, A., Volpe, G. (eds.) Gesture-Based Communication in Human-Computer Interaction. pp. 247–258 (2004)
39. Voskou, A., Panousis, K.P., Kosmopoulos, D., Metaxas, D.N., Chatzis, S.: Stochastic transformer networks with linear competing units: Application to end-to-end sl translation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11946–11955 (2021)
40. Wakatsuki, D., Kobayashi, M., Miyagi, M., Kitamura, M., Kato, N., Namatame, M.: Survey for people with visual impairment or hearing loss on using museums in japan. In: Computers Helping People with Special Needs: 17th International Conference, ICCHP 2020, Lecco, Italy, September 9–11, 2020, Proceedings, Part II 17. pp. 209–215. Springer (2020)
41. Yang, H., Lee, S.: Robust sign language recognition with hierarchical conditional random fields. In: Pattern Recognition, International Conference on. pp. 2202–2205. IEEE Computer Society, Los Alamitos, CA, USA (aug 2010). https://doi.org/10.1109/ICPR.2010.539, https://doi.ieeecomputersociety.org/10.1109/ICPR.2010.539
42. Zhou, H., Zhou, W., Zhou, Y., Li, H.: Spatial-temporal multi-cue network for continuous sign language recognition. In: The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020. pp. 13009–13016. AAAI Press (2020), https://aaai.org/ojs/index.php/AAAI/article/view/7001