Proceedings of the "PErvasive Technologies Related to Assistive Environments Conference (PETRA-2020)", Corfu, Greece, June 2020.

SL-ReDu: Greek Sign Language Recognition for Educational Applications. Project Description and Early Results

Gerasimos Potamianos
Dept. of Electrical & Computer Eng.
University of Thessaly
Volos, Greece
gpotam@ieee.org

Katerina Papadimitriou
Dept. of Electrical & Computer Eng.
University of Thessaly
Volos, Greece
aipapadimitriou@e-ce.uth.gr

Eleni Efthimiou
Inst. for Language & Speech Process.
Athena Research & Innovation Center
Marousi, Greece
eleni_e@athenarc.gr

Stavroula-Evita Fotinea
Inst. for Language & Speech Process.
Athena Research & Innovation Center
Marousi, Greece
evita@athenarc.gr

Galini Sapountzaki
Dept. of Special Education
University of Thessaly
Volos, Greece
gsapountz@sed.uth.gr

Petros Maragos
School of Electrical & Computer Eng.
National Technical Univ. of Athens
Athens, Greece
maragos@cs.ntua.gr

ABSTRACT

We present SL-ReDu, a recently commenced innovative project that aims to exploit deep-learning progress to advance the state-of-the-art in video-based automatic recognition of Greek Sign Language (GSL), while focusing on the use-case of GSL education as a second language. We first briefly overview the project goals, focal areas, and timeline. We then present our initial deep learning-based approach for GSL recognition that employs efficient visual tracking of the signer hands, convolutional neural networks for feature extraction, and attention-based encoder-decoder sequence modeling for sign prediction. Finally, we report experimental results for small-vocabulary, isolated GSL recognition on the single-signer "Polytropon" corpus. To our knowledge, this work constitutes the first application of deep-learning techniques to GSL.

CCS CONCEPTS

• **Computing methodologies** → **Neural networks**; • **Human-centered computing** → **Accessibility technologies**; • **Applied computing** → **Computer-assisted instruction**.

KEYWORDS

Greek sign language recognition, education, L2 language learning, hand tracking, convolutional neural network, encoder-decoder, Polytropon corpus

ACM Reference Format:
Gerasimos Potamianos, Katerina Papadimitriou, Eleni Efthimiou, Stavroula-Evita Fotinea, Galini Sapountzaki, and Petros Maragos. 2020. SL-ReDu: Greek Sign Language Recognition for Educational Applications. Project Description and Early Results. In *The 13th PErvasive Technologies Related to Assistive Environments Conference (PETRA '20), June 30-July 3, 2020, Corfu, Greece.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3389189.3398006

PETRA '20, June 30-July 3, 2020, Corfu, Greece

ACM ISBN 978-1-4503-7773-7/20/06…$15.00
https://doi.org/10.1145/3389189.3398006

1 INTRODUCTION

European and national policies on inclusion and accessibility, as well as the official recognition of national sign languages, have led to rapidly increasing needs for sign language (SL) education as second language (L2) among the general population [8]. Yet, non-native SL education remains a cumbersome process, demanding extensive and iterative tutor-to-learner feedback on a one-to-one basis, while also suffering from a high degree of teacher subjectivity in the evaluation of student proficiency [13, 38].

In the meantime, recent breakthroughs in the fields of computer vision and deep learning have re-ignited interest in the automatic recognition of SL from video [4, 5, 15, 17, 19, 21–23, 32, 33, 36, 37, 39, 43], especially given the fact that SL technology development lags considerably that of oral speech technologies. This is also the case for Greek SL (GSL), an under-resourced language where the so-far employed techniques in its automatic recognition have predated the deep-learning revolution [2, 30, 34, 40].

Motivated by the above, we have recently commenced an innovative project that focuses on the video-based automatic recognition of GSL, aiming at the education use-case. The project, referred to as "SL-ReDu", has as its main goal to address the need for standardized teaching and efficient self-assessment of GSL as L2, by conducting interdisciplinary research in engineering and humanities. SL-ReDu is a three-year effort, carried out in collaboration of two Departments at the University of Thessaly (Electrical and Computer Engineering, Special Education) and of the Athena Research and Innovation Center, and it is funded by the Hellenic Foundation for Research and Innovation.

In this paper, we present the SL-ReDu project, along with our early GSL automatic recognition approach and results. Specifically, in Section 2, we overview the project goals, focal areas, and timeline. In Section 3, we describe our deep learning-based approach for recognizing isolated signs of GSL, followed by experimental results on the Polytropon GSL corpus [11] that are reported in Section 4. Finally, in Section 5, we conclude the paper and discuss future plans.

2 PROJECT DESCRIPTION

The SL-ReDu project is driven by three main goals that are detailed in Section 2.1, with its work focusing in the areas discussed in Section 2.2, and planned according to the timeline of Section 2.3.

2.1 Goals

The first goal of SL-ReDu is the development of innovative computer vision and machine learning algorithms for video-based automatic recognition of SL, considerably advancing the current state-of-the-art in the field. To date, the task remains challenging, due to the number of articulators involved in SL production with complex and fine motion, the scarcity of data resources covering large signing vocabularies and signer variability, and the noisy nature of visual environments in practical scenarios. Further, few only deep-learning techniques have been considered in the SL recognition literature [4, 5, 15, 17, 22, 32, 33, 36, 37, 39, 43], while all GSL recognition systems follow the traditional separate hand-crafted feature and classifier design paradigm [2, 30, 34, 40]. SL-ReDu aims to address this lag by exploiting recent deep-learning breakthroughs to the problem of GSL recognition based on 2D and 3D video data, exploiting suitable annotated corpora [11, 25] and language model, as well as collecting new data, while also expanding its recognition target to a large set of both isolated signs and continuous GSL phrases, in excess of 500 in each case, as well as to GSL finger-spelling.

The second project goal concerns the integration of the developed GSL recognizer into a prototype demonstrator system, focusing on the education use-case, namely that of L2 learning of GSL. Specifically, the system will support the educational process in two distinct pillars: (a) self-monitoring of productive learning by individual learners, and (b) objective evaluation of learning performance across multiple learners by a GSL tutor (see also Fig. 1). Concerning the former, the demonstrator is envisaged to provide the learner with a tireless observer and self-monitoring feedback until specific learning objectives are achieved, thus overcoming the bottleneck of requiring frequent SL tutor physical presence.

Concerning the latter, the developed SL-ReDu system will be used in evaluating student GSL performance at the Department of Special Education of the University of Thessaly in the context of learning and testing for the compulsory course "Introduction to Greek Sign Language" of the curriculum, thus constituting the third goal of the project. SL-ReDu aspires to greatly improve testing credibility and consistency, while significantly reducing the tutor's load. It should be noted that the application of SL recognition to education has previously involved only experimentation with a restricted number of SL articulators for a limited set of lexemes [27], while SL-ReDu is planned to address complete productions of linguistic units.

2.2 Main Focus Areas

The first area of SL-ReDu research activities concerns visual tracking and feature extraction, aiming at the detection and tracking of the visual articulators in SL video (both manual and non-manual), as well as the extraction of corresponding visual features, thus providing necessary input to the SL recognizer. Specifically for the former, both light-weight schemes and more computationally demanding approaches will be explored for tracking the signer hands,

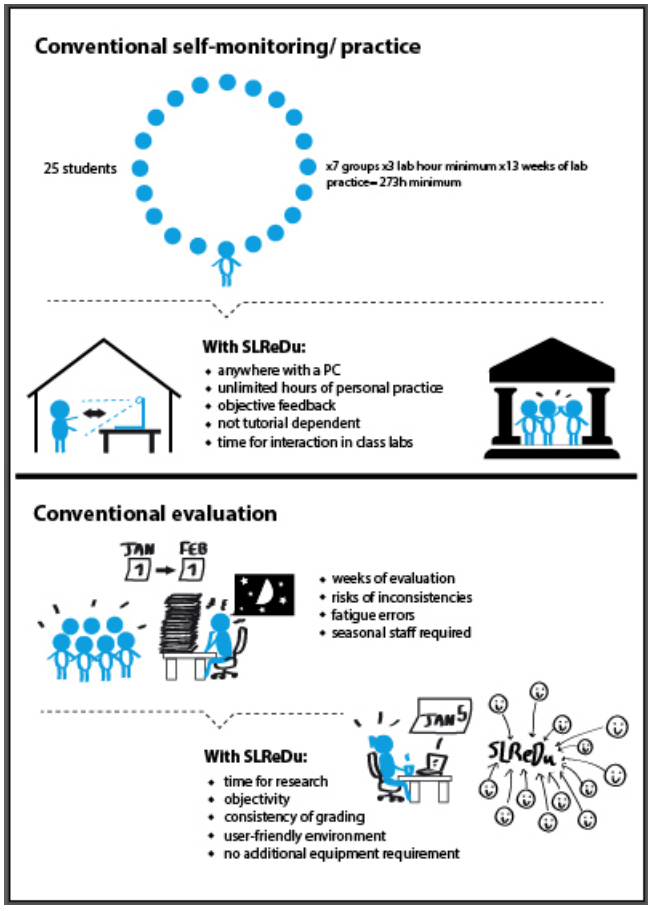

Figure 1: Schematic illustrating the SL-ReDu envisaged benefits to self-monitoring and objective evaluation for non-native GSL learning over the conventional approach.

arms, upper body, face, and facial features (primarily, the mouth, cheeks, eyes, and eyebrows). Then, traditional shape-based and/or appearance representations of the articulators will be investigated, as well as deep-learning representations employing convolutional neural networks (CNNs) and deep autoencoders.

Based on the above, appropriate classifiers for GSL recognition will be explored, both at the lower level of articulators and GSL subunits, as well as at a higher level based on the lower-level results and a language model. Articulatory actions (e.g. specific handshapes, mouthing patterns, etc.) will be recognized and the results fused, thus being able to recognize complex signs both in isolated and continuous signing. Various approaches will be investigated for this purpose, including hidden Markov models and deep learning-based techniques. GSL recognition results will be accompanied by confidence scores to assist in the education use-case, while signer adaptation techniques will also be considered.

The aforementioned activities will be supported by training data and a language model. In particular, existing GSL data resources will be harvested, as for example in this paper, where the Polytropon corpus [11] is used. Further, a new GSL dataset will be collected

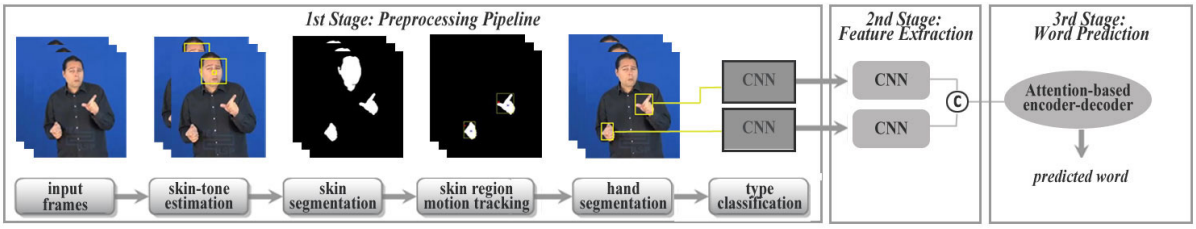

Figure 2: General architecture of the proposed three-stage system for isolated sign recognition from GSL video data.

including multiple signers, a large number of lemmas and continuous phrases (in excess of 500 in each case), as well as finger-spelled signing, relevant to the SL-ReDu use-case. The effort will be accompanied by the organization of the relevant evaluation GSL material and providing the corresponding language model.

In addition, an appropriate human-computer interface will be designed, integrating multimodal and embodied communication elements that are suitable for the L2 GSL learning task of the project. All developed components will be integrated into the SL-ReDu prototype demonstrator system, adopting a properly designed system architecture. The system will be subsequently evaluated according to the use-case scenarios of Section 2.1.

2.3 Timeline

SL-ReDu is a three-year project, commenced in January 2020. Its implementation plan foresees two phases of prototype system development and evaluation. Specifically, Phase-A of the project will produce a demonstrator for GSL recognition of isolated signs, finger-spelling, and numerals, adopting a relatively simple human-computer interface to support the education use-case, and it will be evaluated in a short one-month pilot by the end of the Summer of 2021. Phase-B of the project will incorporate lessons learned from Phase-A, and it will involve a more advanced human-computer interface and GSL recognition algorithms that will also allow recognition of GSL continuous phrases on top of the Phase-A recognition vocabulary. This will be finalized by late Summer of 2022 and evaluated in a longer, four-month evaluation campaign, allowing for system refinement and fine-tuning.

3 SIGN LANGUAGE RECOGNITION METHOD

We next describe the proposed system for isolated sign recognition from GSL video data, developed so far as part of our SL-ReDu project activities. Our approach employs deep-learning techniques and consists of three stages, namely: (i) a pre-processing pipeline for extracting the signer hands and classifying them into left and right; (ii) an image feature extractor for each hand that is based on CNNs; and (iii) an attention-based encoder-decoder for the sign prediction task. These components are schematically depicted in Fig. 2 and are detailed next.

3.1 Hand Extraction and Type Classification

The first stage of the system adopts the pipeline of our earlier work [28] to extract the signer hands that are visible, as well as to classify them into left and right types (as viewed by the camera). The approach is hybrid, utilizing both traditional techniques for efficient detection and tracking, as well as deep learning for hand type classification, and it is based on the assumption that the signer's face is visible at frontal head pose, as is the case in SL videos.

The pipeline commences with face detection by means of the Viola-Jones algorithm [42], as well as nose region detection. The latter is used to estimate the signer's skin color range in the YCbCr color space [35], driving skin-tone based segmentation to allow detection of candidate hand regions. To address possible hand and face overlap, motion-based Kalman filtering is employed [18]. This step allows detection and tracking of the hands, as such are expected to be the only skin-tone moving objects in the SL video. As a last step, the returned object bounding boxes are fed to an AlexNet CNN [24] for final hand detection and type classification, considering three classes of interest: left, right, and no hand. An example of this process applied to GSL data is depicted in Fig. 3.

3.2 Handshape Feature Extraction

In order to extract features from the detected left and right hands, the second stage of the system applies multi-layer 2D-CNNs to their size-normalized, 224×224-pixel images, separately. Specifically, a ResNet-18 network architecture with 3×3 convolutional kernels and stride 2 is used [14], pre-trained on the ImageNet dataset [9] with the mean squared error loss function. The output of the fully-connected layer is used to yield 512-dimensional (dim) features for each hand. These are then concatenated for the two hands, resulting to 1,024-dim feature vectors (one per video frame), which

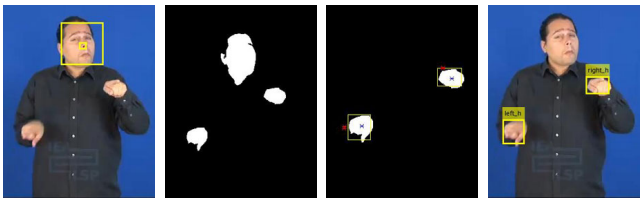

Figure 3: Example of the hand detection and type classification pipeline of [28] (first stage of the proposed system), applied on data of the Polytropon corpus. Depicted are, left to right: (a) video frame marked with a rectangular box enclosing the detected face, as well as the central square of the detected face region; (b) segmented skin region; (c) tracked hands by Kalman filtering (yellow rectangles depict detected objects, red stars the predicted object positions, and blue stars their corrected positions); (d) frame marked with rectangular boxes illustrating the signer's left and right hands.

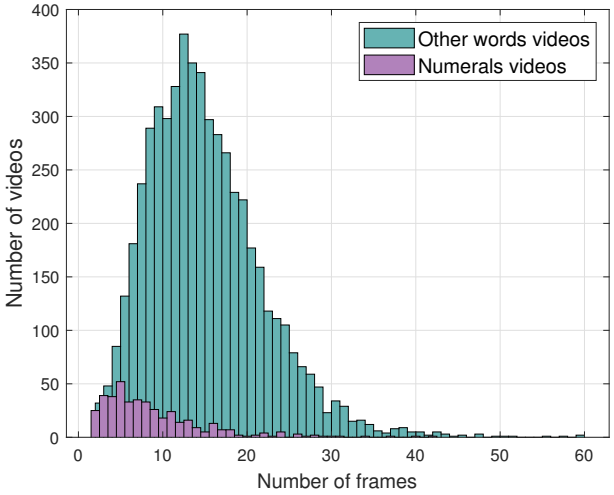

Figure 4: Duration histograms (in video frames) of signed numerals and other words in the Polytropon GSL corpus.

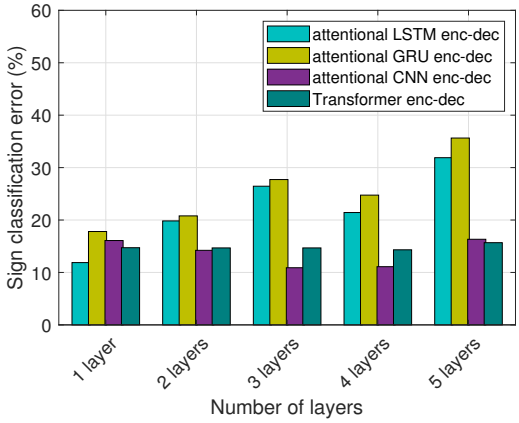

Figure 5: Sign classification error (%) of the proposed system on the Polytropon subset of highly-occurring "other words", using various numbers of layers in the four sequence learning models of Section 3.3.

are subsequently fed to the attention-based encoder-decoder of the third stage for sign prediction. Note that in the case of missing hand detections, the corresponding features are set to zeros.

3.3 Sequence learning model

The task of isolated sign recognition from SL videos can be viewed as a sequence learning problem that can be addressed by an attention-based encoder-decoder [3]. In the typical form of such model, the encoder receives latent-representation sequential data and outputs a sequence of hidden states, while the decoder maps the latter to the desired output (sign IDs). The attention mechanism performs alignment between the input and output, attending to the most relevant information in the source sequence. There exist a variety of such models in the literature, mostly recurrent neural network (RNN) based ones. In this paper, four models are investigated, namely the:

- *Attentional LSTM encoder-decoder*, where a long short-term memory (LSTM) [16] is employed as the RNN. Specifically, a one-layer LSTM encoder-decoder is used with hidden dimensionality equal to 256.
- *Attentional GRU encoder-decoder*, where gated recurrent units (GRUs) [6] are used. In particular, a one-layer GRU encoder-decoder with 128 hidden units is employed.
- *Attentional CNN encoder-decoder*, which has the advantage over RNNs of allowing parallelization, as CNNs do not depend on previous time computations. Specifically, a multi-step attention-based, 3-layer CNN encoder-decoder is used with kernel width 5 and 128 hidden units, as in [29].
- *Transformer encoder-decoder*, a more recently introduced sequence learning model [41] that differs from the above by substituting recurrent layers with multi-head attention ones, incorporating position encoding, and applying layer normalization. Here, a 4-layer transformer is employed, with 8 heads for transformer self-attention, 2048-dimension hidden transformer feed-forward, and 512 hidden units.

3.4 Implementation details

All aforementioned models were implemented in PyTorch [31], and their training was carried out exploiting GPU acceleration.

Further, for fine-tuning the hand-type classification model of Section 3.1, the hand dataset of [26] was employed. During its training, stochastic gradient descent with momentum was used, with an initial learning rate of 0.004 decayed by a factor of 0.5 and a mini-batch of 128 images.

For fine-tuning the handshape feature extractor of Section 3.2, the corpus of Section 4.1 was employed. For this purpose, stochastic gradient descent with momentum was used, with an initial learning rate of 0.001 decayed by a factor of 5 every 20 epochs. Dropout with a rate of 0.5 was added.

For attentional RNN training, the Adam optimizer [20] was employed with initial learning rate of 0.001 decayed by a factor of 0.3. Beam search was used for decoding with beam-width 5, and dropout was added at a rate of 0.3. The attentional CNN encoder-decoder training was conducted using the Adagrad optimizer [10] with an initial learning rate of 0.003, which was decreased by a factor of 0.3. Dropout of 0.8 and beam search of width 5 were employed. Finally, for the Transformer encoder-decoder training, the Adam optimizer was used with an initial learning rate of 0.001 decreased by a factor of 2.0 and dropout 0.8. Parameter initialization was carried out by the Xavier process [12].

4 GSL RECOGNITION EXPERIMENTS

4.1 Dataset and Experimental Framework

Our experiments are conducted on the Polytropon GSL corpus [11]. This contains three repetitions of 3,600 sentences performed by a single signer, recorded by two frontal-view cameras, a Kinect and an RGB one. Here, the video data of the RGB camera are used, which are available at a 25 Hz frame-rate and 848 × 480-pixel resolution. Corpus annotations are based on ELAN [1, 7] and are provided at both the signed sentence and signed word levels, containing labels and time-stamps for proper nouns, verbs, adverbs, and numerals. The corpus signed vocabulary consists of 39 unique numerals and

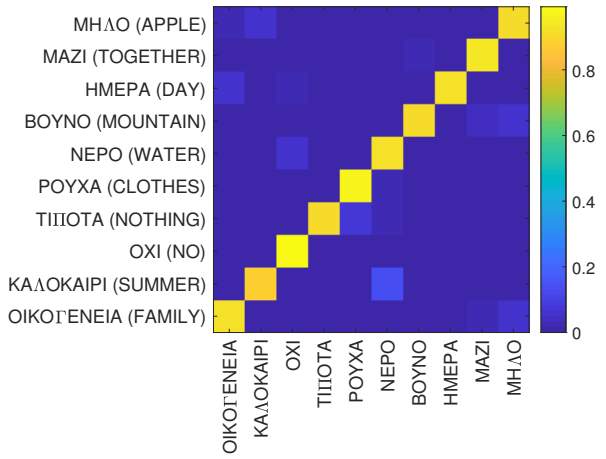

Figure 6: Confusion matrix of ten "other words" for the attentional CNN encoder-decoder. The horizontal axis depicts predicted words, while the vertical one the ground truth.

(a) (b)

Figure 7: Examples of a frame of signed words (a) "summer" and (b) "water" in Polytropon. These are frequently confused by our model due to their similar handshapes.

2,664 other words, with their signing duration statistics varying significantly, as also depicted in Fig. 4. Among those, words with a sufficient number of occurrences are selected for recognition, to allow enough data for deep-learning model training. Specifically, two isolated small-vocabulary sign recognition tasks are built, the first concerning ten unique numerals that appear between 20 and 140 times in the corpus (45.2 times on average) and the second 103 unique "other words" (i.e., non-numerals) that appear between 30 to 110 times (52.6 on average). These yield 422 and 5,414 video snippets of numerals and other words, respectively, which are obtained by "cutting" the longer video database files based on the ELAN annotation time-stamps of the words of interest. All experiments on the two resulting datasets of numerals and "other words" are conducted using ten-fold cross-validation, where 80% of each fold is allocated to training, 10% to validation, and 10% to testing.

4.2 Results

The performance of our proposed approach to GSL recognition is reported in Table 1. There, the sign classification error (%) achieved by all four sequence models of Section 3.3 is shown on both constructed Polytropon subsets of Section 4.1, namely that of highly-occurring numerals and that of "other words". It is apparent that the best results are achieved by the attention-based CNN encoder-decoder, while the worst by the attentional GRU encoder-decoder. It is also interesting to note that performance is comparable for both recognition tasks and consistent across the four models. Bearing in mind that the vocabulary size of the "other words" task is an order of magnitude larger than that of numerals (103 vs. 10), the similar performance attained may be due to the fact that numerals have consistently shorter durations, as shown in Fig 4.

Table 1: Sign classification error (%) of the proposed isolated GSL recognition system on the Polytropon corpus subsets of highly-occurring numerals and other words (see Section 4.1), using the four sequence learning models of Section 3.3.

Encoder-decoder model	numerals	other words
Attentional LSTM	11.47	11.88
Attentional GRU	18.14	17.82
Attentional CNN	**11.06**	**10.90**
Transformer	14.45	14.32

Next, in Fig. 5, we investigate the performance of the four sequence models for different numbers of encoder-decoder layers. It can be observed that the lowest errors for attentional RNNs are achieved for one layer, which may be due to the limited data size of Polytropon. On the contrary, the attentional CNN and transformer methods prove less sensitive to the number of layers.

Finally, in Fig. 6, we visualize the confusion matrix for a subset of ten "other words" selected at random. The bright yellow diagonal demonstrates the successful classification achieved. Among the confusable pairs of this matrix, we depict in Fig. 7 one video frame example of signed words in GSL for "summer" and "water". Obviously the signing handshapes look very similar, although their positioning (and track), as well as their non-manual articulation differ. Since however our proposed system only encodes handshapes, it faces difficulties in discriminating between the two signs.

5 SUMMARY AND FUTURE WORK

In this paper, we provided a summary of the SL-ReDu project that aims to advance the automatic recognition of GSL and exploit such in its teaching as a second language. Further, we introduced a deep learning-based approach for isolated sign recognition of GSL, which we successfully evaluated on small-vocabulary, single-signer data.

Moving forward, we plan to commence a large-scale data collection effort for GSL, including multiple signers, a large number of lemmas (in excess of 500), as well as finger-spelled signing, relevant to the curriculum of the introductory GSL class at the Department of Special Education of the University of Thessaly. Further, we plan to extend our proposed deep-learning GSL recognition system, by incorporating information concerning manual articulation positioning, as well as non-manual articulation.

ACKNOWLEDGMENTS

The research work was supported by the Hellenic Foundation for Research and Innovation (H.F.R.I.) under the "First Call for H.F.R.I. Research Projects to support Faculty members and Researchers and the procurement of high-cost research equipment grant" (Project Number: 2456).

REFERENCES

[1] 2019. ELAN (Version 5.8) [Computer software]. Nijmegen: Max Planck Institute for Psycholinguistics, The Language Archive. https://archive.mpi.nl/tla/elan.

[2] Epameinondas Antonakos, Vassilis Pitsikalis, and Petros Maragos. 2014. Classification of extreme facial events in sign language videos. *EURASIP Journal on Image and Video Processing* 14 (2014).

[3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *Computing Research Repository* (2014). arXiv:abs/1409.0473v7.

[4] Kshitij Bantupalli and Ying Xie. 2018. American sign language recognition using deep learning and computer vision. In *Proc. IEEE International Conference on Big Data*. 4896–4899.

[5] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowen. 2018. Neural sign language translation. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 7784–7793.

[6] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 1724–1734.

[7] Onno Crasborn and Han Sloetjes. 2008. Enhanced ELAN functionality for sign language corpora. In *Proc. Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora*. 39–43.

[8] Maartje De Meulder. 2016. *The Power of Language Policy: The Legal Recognition of Sign Languages and the Aspirations of Deaf Communities*. Ph.D. Thesis, Faculty of Humanities, University of Juväskylä, Finland.

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 248–255.

[10] John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research* 12 (2011), 2121–2159.

[11] Eleni Efthimiou, Kiki Vasilaki, Stavroula-Evita Fotinea, Anna Vacalopoulou, Theodoros Goulas, and Athanasia-Lida Dimou. 2018. The POLYTROPON parallel corpus. In *Proc. International Conference on Language Resources and Evaluation (LREC)*.

[12] Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proc. International Conference on Artificial Intelligence and Statistics (AISTATS)*, Vol. PMLR 9. 249–256.

[13] Tobias Haug, Wolfgang Mann, Eveline Boers-Visker, Jessica Contreras, Charlotte Enns, Ros Herman, and Katherine Rowley. 2016. *Guidelines for Sign Language Test Development, Evaluation, and Use*. Unpublished document (upd. 2018), retrieved from http://www.signlang-assessment.info/.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 770–778.

[15] Siming He. 2019. Research of a sign language translation system based on deep learning. In *Proc. International Conference on Artificial Intelligence and Advanced Manufacturing (AIAM)*. 392–396.

[16] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computing* 9 (1997), 1735–1780.

[17] Jie Huang, Wengang Zhou, Houqiang Li, and Weiping Li. 2015. Sign language recognition using 3D convolutional neural networks. In *Proc. IEEE International Conference on Multimedia and Expo (ICME)*.

[18] Jong-Min Jeong, Tae-Sung Yoon, and Jin-Bae Park. 2014. Kalman filter based multiple objects detection-tracking algorithm robust to occlusion. In *Proc. SICE Annual Conference*. 941–946.

[19] Byeongkeun Kang, Subarna Tripathi, and Truong Q. Nguyen. 2015. Real-time sign language fingerspelling recognition using convolutional neural networks from depth map. In *Proc. IAPR Asian Conference on Pattern Recognition (ACPR)*. 136–140.

[20] Diederik P. Kingma and Jimmy Lei Ba. 2014. Adam: A method for stochastic optimization. *Computing Research Repository* (2014). arXiv:abs/1412.6980v9.

[21] Oscar Koller, Jens Forster, and Hermann Ney. 2015. Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *Computer Vision and Image Understanding* 141 (2015), 108–125.

[22] Dimitrios Konstantinidis, Kosmas Dimitropoulos, and Petros Daras. 2018. A deep learning approach for analyzing video and skeletal features in sign language recognition. In *Proc. IEEE International Conference on Imaging Systems and Techniques (IST)*.

[23] Ioannis Koulierakis, Georgios Siolas, Eleni Efthimiou, Stavroula-Evita Fotinea, and Andreas-Georgios Stafylopatis. 2019. Gesture recognition using keypoints detection in the context of sign language translation. In *Proc. Workshop on Sign Language Translation and Avatar Technologies (SLTAT)*.

[24] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NeurIPS) 25*. 1097–1105.

[25] Silke Matthes, Thomas Hanke, Anja Regen, Jakob Storz, Satu Worseck, Eleni Efthimiou, Athanasia-Lida Dimou, Annelies Braffort, John Glauert, and Eva Safar. 2012. Dicta-Sign – Building a multilingual sign language corpus. In *Proc. Workshop on the Representation and Processing of Sign Languages: Interactions Between Corpus and Lexicon*.

[26] Arpit Mittal, Andrew Zisserman, and Philip H. S. Torr. 2011. Hand detection using multiple proposals. In *Proc. British Machine Vision Conference (BMVC)*.

[27] Jill P. Morford and Martina L. Carlson. 2011. Sign perception and recognition in non-native signers of ASL. *Language Learning and Development* 7 (2011), 149–168.

[28] Katerina Papadimitriou and Gerasimos Potamianos. 2018. A hybrid approach to hand detection and type classification in upper-body videos. In *Proc. European Workshop on Visual Information Processing (EUVIP)*.

[29] Katerina Papadimitriou and Gerasimos Potamianos. 2019. End-to-end convolutional sequence learning for ASL fingerspelling recognition. In *Proc. Annual Conference of the International Speech Communication Association (Interspeech)*. 2315–2319.

[30] Vassilia N. Pashaloudi and Konstantinos G. Margaritis. 2004. A performance study of a recognition system for Greek sign language alphabet letters. In *Proc. International Conference on Speech and Computer (SPECOM)*. 545–551.

[31] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch. In *Proc. Neural Information Processing Systems Workshops (NeurIPS-W)*.

[32] Lionel Pigou, Sander Dieleman, Pieter-Jan Kindermans, and Benjamin Schrauwen. 2015. Sign language recognition using convolutional neural networks. In *Proc. European Conference on Computer Vision Workshops (ECCVW)*, Vol. LNCS 8925. 572–578.

[33] G. Anantha Rao, K. Syamala, P. V. V. Kishore, and A. S. C. S. Sastry. 2018. Deep convolutional neural networks for sign language recognition. In *Proc. Conference on Signal Processing and Communication Engineering Systems (SPACES)*. 194–197.

[34] Anastasios Roussos, Stavros Theodorakis, Vassilis Pitsikalis, and Petros Maragos. 2013. Dynamic-affine invariant shape-appearance handshape features and classification in sign language videos. *Journal of Machine Learning Research* 14 (2013), 1627–1663.

[35] Khamar Basha Shaik, P. Ganesan, V. Kalist, B. S. Sathish, and J. Merlin Mary Jenitha. 2015. Comparative study of skin color detection and segmentation in HSV and YCbCr color space. *Procedia Computer Science* 57 (2015), 41–48.

[36] Bowen Shi and Karen Livescu. 2017. Multitask training with unlabeled data for end-to-end sign language fingerspelling recognition. In *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. 389–396.

[37] Bowen Shi, Aurora Martinez Del Rio, Jonathan Keane, Jonathan Michaux, Diane Brentari, Greg Shakhnarovich, and Karen Livescu. 2018. American sign language fingerspelling recognition in the wild. In *Proc. IEEE Spoken Language Technology Workshop (SLT)*. 145–152.

[38] David H. Smith and Jeffrey E. Davis. 2014. Formative assessment for student progress and program improvement in sign language as L2 programs. In *Teaching and Learning Signed Languages*, David McKee, Russell S. Rosen, and Rachel McKee (Eds.). Palgrave Macmillan, London, 253–280.

[39] Wenjin Tao, Ming C. Leu, and Zhaozheng Yin. 2018. American sign language alphabet recognition using convolutional neural networks with multiview augmentation and inference fusion. *Engineering Applications of Artificial Intelligence* 76 (2018), 202–213.

[40] Stavros Theodorakis, Vassilis Pitsikalis, and Petros Maragos. 2014. Dynamic-static unsupervised sequentiality, statistical subunits and lexicon for sign language recognition. *Image and Vision Computing* 32 (2014), 533–549.

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS) 30*. 5998–6008.

[42] Paul Viola and Michael Jones. 2001. Rapid object detection using a boosted cascade of simple features. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*.

[43] Zhaoyang Yang, Zhenmei Shi, Xiaoyong Shen, and Yu-Wing Tai. 2019. SF-Net: Structured feature network for continuous sign language recognition. *Computing Research Repository* (2019). arXiv:abs/1908.01341v1.