Towards Continuous Sign Language Recognition with Deep Learning

Boris Mocialov[1], Graham Turner[2], Katrin Lohan[3], Helen Hastie[4]

Abstract— **Humans communicate with each other using abstract signs and symbols. While the cooperation between humans and machines can be a powerful tool for solving complex or difficult tasks, the communication must be at the abstract enough level that is both natural to the humans and understandable to the machines. Our paper focuses on natural language and in particular on sign language recognition. The approach described here combines heuristics for segmentation of the video stream by identifying the epenthesis with stacked LSTMs for automatic classification of the derived segments. This approach segments continuous stream of video data with the accuracy of over 80% and reaches accuracies of over 95% on segmented sign recognition. We compare results in terms of the number of signs being recognised and the utility of various features used for the recognition. We aim to integrate the models into a single continuous sign language recognition system and to learn policies for specific domains that would map perception of a robot to its action. This will improve the accuracy of understanding the common task within the shared activity between a human and a machine. Such understanding, in turn, will foster meaningful cooperation.**

I. INTRODUCTION

Interacting with machines, users are required to use the input devices, such as remote controls, keyboards, or touch interfaces, provided with these machines. The input devices usually eliminate the uncertainty of the user input to show that the machine functions properly and reliably. While the provided input devices are reasonable communication mediums, they are neither intuitive nor natural for a human user, which leads to either an overhead of learning how to use the given input devices or even inability to use a machine, perhaps due to a disability. Such limited communication hinders opportunities for effective collaboration between humans and machines.

To eliminate this limiting factor, the machines should understand interaction that is natural to the user. This natural interaction could be achieved using natural language or gesturing, while the natural language, in turn, could be either spoken languages or sign languages [1]. We focus on the recognition of sign languages from video in this paper. Whilst much research has been done on the partially analogous problem of continuous speech recognition, few researchers have investigated sign language recognition. There is a general misconception that sign languages are simply gestures with simple rules, in fact this is not the case. A single sign, corresponding to a word or concept, is multimodal from the perspective of the producer and can have many variations within a single language. Compounding the problem is the fluid nature of signing where signs are interleaved with transitional motions called *epenthesis*, which themselves are easily confusable with signs. This is combined with synchronous facial recognition making the feature space very large and the problem complex.

In Section II, this paper identifies relevant research for the segmented, continuous, and vocabulary-based sign language recognition and divides the overall problem into three high-level sub-problems, listing the methods that are used by other authors to tackle these sub-problems. Section III describes the dataset that has been used for this paper and presents the methodology in terms of a process pipeline. Sections IV and V present results after experimenting with every component of the pipeline individually. Section VI discusses the results and their meaning in the context of the continuous sign language recognition. Finally, Section VII concludes the paper and outlines the future development of this work that aims to tackle the uncertainty in recognition for continuous signing.

II. RELATED WORK

Much previous work has focused on the recognition of signs in terms of isolated, segmented video snippets with a clear start and end time [2], [3], [4]. Alternatively, continuous sign language recognition focuses on the stream of signs in a sentence with the task to process a signed sentence and produce aligned *glosses*, which is the written form of a signed sentence in words [5], [1].

The final approach is analogous to keyword spotting in automatic speech recognition, where a finite list of signs is spotted in the video [1], [6]. This is the middle ground between the isolated sign and continuous sign language recognition and the approach that we adopt here. Our approach breaks down into 3 sub-problems, which will be discussed here in terms of previous work: 1) feature extraction; 2) detection of the movement epenthesis as a means of segmentation; and 3) classification of segmented signs.

*This work was supported by the Heriot-Watt University School of Engineering & Physical Sciences James Watt Scholarship and Engineering and Physical Sciences Research Council (EPSRC) as part of the CDT in Robotics and Autonomous Systems at Heriot-Watt University and The University of Edinburgh (Grant reference EP/L016834/1) as well as the RAEng/Leverhulme Trust Senior Research Fellowship Scheme (Hastie/LTSRF1617/13/37)

[1] Boris Mocialov is with the Department of Computer Science, the School of Engineering & Physical Sciences, and the Edinburgh Centre of Robotics, Heriot-Watt University, Edinburgh, UK bm4@hw.ac.uk

[2] Prof. Graham Turner is with the Department of Languages & Intercultural Studies, Heriot-Watt University, Edinburgh, UK g.h.turner@hw.ac.uk

[3] Dr. Katrin Lohan is with the Department of Computer Science, Heriot-Watt University, Edinburgh, UK and the Edinburgh Centre of Robotics k.lohan@hw.ac.uk

[4] Prof. Helen Hastie is with the Department of Computer Science, Heriot-Watt University, Edinburgh, UK and the Edinburgh Centre of Robotics h.hastie@hw.ac.uk

Firstly, local feature extraction methods from noisy input data have recently become more precise [7], [8], [9], although, some challenges, such as tackling occlusions, still persist. The majority of the sign languages consist of manual (hands, fingers, posture) and non-manual features (facial expressions), which makes them multimodal from the signer's perspective. The features are used in parallel and tend to complement each other. Specific features, in some cases, may not be required in order to interpret the sign [5]. The common local features used for sign language recognition are body posture (shoulders, neck, waist), hands (elbows, wrists, and phalanges), and facial features (mouth and eyes).
Once the features are chosen, they should be tracked throughout the frames to get all the information that forms a sign [10]. In [11], the author questions whether all parts of the signing features are equally important during signing and how much movement and configuration variations are allowed for the sign to be recognised. In fact, [12] have shown that the index finger is the salient finger during signing and determines the speed and amplitude of signing with other fingers following the motion of the index finger. This theory is supported by [13], who shows that physiologically this should be the case that not all fingers are dominant during signing, which makes it applicable to any sign language. In the work described here, we will examine the utility of a number of different feature sets.

Secondly, regarding segmentation by means of motion epenthesis modelling, this is directed towards explicit detection of the motion between the intended signs during signing. The detection of the motion epenthesis can be achieved with dynamic programming [6], which is advantageous, because it does not require training as with machine learning approaches [1].

Thirdly, isolated signs have been previously modelled to incorporate both spatial and temporal information, such as sequential pattern mining that fuses multimodal signals [14]. The same paper uses regression, SVM and LSTM for comparisons and concludes that the models that incorporate spatial and temporal features are superior. More recent work on networks allow the network to be trained on videos of different lengths [15], which is useful because the same signs may be of different lengths due to signing speed. Most promising results are achieved with deep learning techniques, such as CNN with temporal convolution and pooling for spatio-temporal representations or RNN with long short-term memory (LSTM) to learn the mapping of feature sequences to sequences of glosses [16].

Our approach requires a large amount of quality data. In recent years, the situation regarding sign language data has improved with more readily available larger datasets that are realistic rather than simulated, and involve more complex interactions for specific tasks, such as explaining directions or story retelling [17], [18], [19].

III. Methodology

Figure 1 shows the processing pipeline for the continuous sign language recognition with the raw video data input and the recognised individual signs as output. Further, the dataset used for the training and testing of the system will be introduced. Finally, the parts of the pipeline will be discussed in detail.

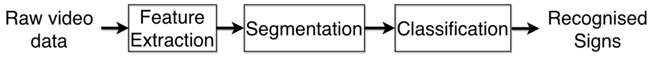


Fig. 1: Data processing pipeline for the continuous sign language recognition

A. Dataset

Due to the annotation quality, a portion of the NGT[1] corpus has been used for this project. The corpus contains approximately 100 participants telling stories, or having discussions with other Dutch sign language users.

We have chosen a part of the corpus where participants retell the Canary Row cartoon of Tweety & Sylvester by the Warner Brothers Pictures. Details about the recording setup for the corpus can be found in [17].

TABLE I: Chosen classes for training (glosses translated from Dutch with Google Translate)

Classes	Glosses	Maximum signers/class
0-10	ape, building, electricity, handwriting, look, poet, rain, run, shake, tram	7
10-20	ball, binoculars, bird, birdcage, inside, not, ready, rope, same, search	4
20-30	and, apartment, climb, corner, how, hurry, line, old, pipe, thinking	6
30-40	window, clothes, box, suitcase, contact, aunt, draw, music, funny, tighten	4

Table I lists the glosses that the models were trained on. The choice of the glosses was guided by the amount of the available instances of that particular class. The more example videos of the sign there were present in the dataset, the more likely the sign had been chosen for the training.

The mean length of a sign is 6.75 frames where one frame length is approximately 40 milliseconds. The average amount of examples per sign is approximately 11 videos and was unfortunately not enough to train our models. Therefore, for every selected class, additional data was generated using extracted features from the original data. For every video example of the real data, 200 more examples were synthesised by adding perturbation along both x and y axes to the extracted features from the original examples. For the first 100 synthesised examples, the same perturbation has been added to every extracted feature, while for the second 100 synthesised examples, different perturbations were added to every extracted features along the x and y axes. This was done to synthesise examples of a sign, where, for example, the hand is moved further from the body or the face of the signer than in the original example.

[1] Sign Language of the Netherlands - NGT (Nederlandse Gebarentaal), is the language of the deaf community in the Netherlands

B. Feature Extraction

Used features resemble the features provided by the commercial sensors, such as Microsoft Kinect. Instead of using an additional high-cost sensor such as Kinect, a standard camera is used and features are extracted with the help of the deep learning techniques, provided by the openpose library[2] [7], [8], [9].

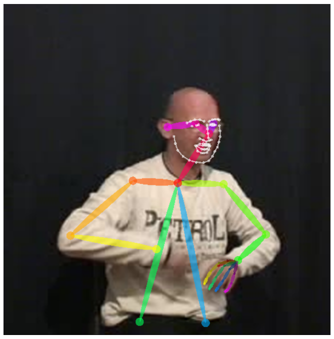


Fig. 2: NGT dataset feature extraction example with openpose. The lines are drawn between identified body features, such as shoulder, neck, phalanges, etc.

Figure 2 shows an arbitrary frame from the NGT corpus after the openpose feature extraction algorithm is applied. The algorithm provides information about the body pose, hands, and facial features. The limitation of the openpose algorithm is that it does not recover the features when occlusions are present, which is very common during signing as the hands occlude each other and the face.

C. Segmentation

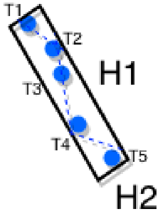
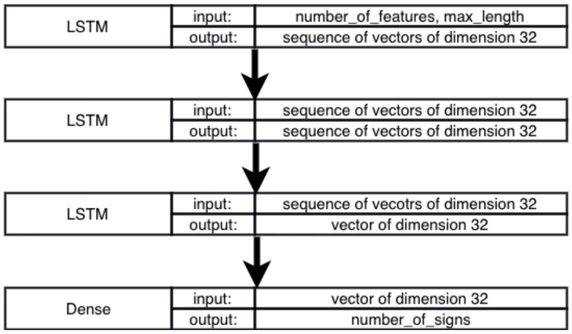


Fig. 3: Example of hand trajectory during signing that is used to decide whether the motion is epenthesis or a part of a sign. T1-T5 correspond to centroids of hand contour, acquired during feature extraction; H1 and H2 are height and width of the minimum bounding box for the T1-T5

The main assumption for the segmentation is that the hands move slower during the signing than during the motion epenthesis. Motion epentheses are identified by looking at the distance travelled by each hand an interval. In this particular experiment, 5 frames are chosen for this interval for detection of the motion epenthesis as was reported in [20]. Using the extracted features from the hands, as can be seen in Figure 2, the centroids of all the hand points are calculated and accumulated for the period of 5 frames (T1-T5 on Figure 3). Later, the minimum bounding box is calculated for the hand trajectory over 5 frames (black rectangle on the figure). At the end, the longest side of the minimum bounding box (either H1 or H2 from the figure) is taken to decide whether the segment is motion epenthesis or a part of the sign. Both H1 and H2 are considered, because the hand may travel in any direction during signing. Using similar techniques as in [20], the segment is labelled as epenthesis if the longest side of the minimum bounding box is between 18 and 60 pixels.

D. Classification



LSTM	input:	number_of_features, max_length
	output:	sequence of vectors of dimension 32
LSTM	input:	sequence of vectors of dimension 32
	output:	sequence of vectors of dimension 32
LSTM	input:	sequence of vecotrs of dimension 32
	output:	vector of dimension 32
Dense	input:	vector of dimension 32
	output:	number_of_signs

Fig. 4: Model architecture for the TensorFlow library, consisting of stacked LSTM layers and one Dense layer that outputs the sign class. Inputs and outputs specify the type of inputs and outputs for a particular layer of the network.

With the video segmented, isolated sign language recognition is done by training deep learning models using TensorFlow[3] and openpose libraries. The architecture, shown in Figure 4, is composed of three stacked LSTM layers with the first two layers producing a sequence of vectors with 32 dimensions and the last LSTM layer producing a single vector, composed of 32 dimensions. At the output of the network, the dense layer outputs the likelihood of every sign. The first layer accepts a sequence of inputs (chunks) of length equals to the number of extracted features per one frame. The maximum number of chunks is set to be the longest sequence of frames for a sign and all other sequences are padded at the end with zeros. The network is trained offline with the objective function set to categorical cross entropy and the optimizer set to resilient backpropagation with the adaptive learning rate, which is a good choice for the recurrent neural networks.

Extracted features, such as posture, finger, and facial information are combined together by stacking feature vectors together for the isolated video of a single sign.

The dataset is split into training, validation, and testing sets. The training data consists of 80% of the overall dataset, validation and testing sets consists of 10% of the overall dataset each. All the dataset is shuffled before performing the split into training, validation, and testing sets. This means

[2] https://github.com/CMU-Perceptual-Computing-Lab/openpose/

[3] https://www.tensorflow.org/

that the method is not signer independent as the testing set is likely to contain some variation of the same signer from the training set. The future experiments will test the signer independent condition for a more robust solution.

IV. Results: Segmentation Accuracy

One continuous single-signer video has been used for testing the accuracy of the segmentation. The ground truth was annotated by considering the time between every gloss in the annotation file to be the epenthesis motion, with 206 motion epenthesis occurrences annotated.

The epenthesis detection returns start and end times of the epenthesis interval. To calculate the accuracy in terms of F-measure, the returned epenthesis interval is compared to the ground truth, extracted from the annotated video. As a result, the algorithm identified 201 True Positives TP that lied within the ground truth ($Predicted \in GT$). Some of the identified intervals are repeated, due to the fact that both hands are tracked and analysed for the epenthesis identification. The algorithm identified 39 False Positives FP that did not match epintheses in the ground truth ($Predicted \not\in GT$). All the intervals that were not included in the predicted TP are assumed to be True Negatives TN ($Predicted \in \neg GT$). The algorithm identified 210 TN intervals. The intervals that were considered and were not in the ground truth were assumed to be False Negatives FN ($Predicted \not\in \neg GT$). The algorithm identified 46 FN intervals.

$$F-measure =$$

$$(2 * Precision * Recall)/(Precision + Recall) = \mathbf{0.825}, \text{ where}$$

$$Precision = TP/(TP + FP) = \mathbf{0.837} \quad \text{and}$$

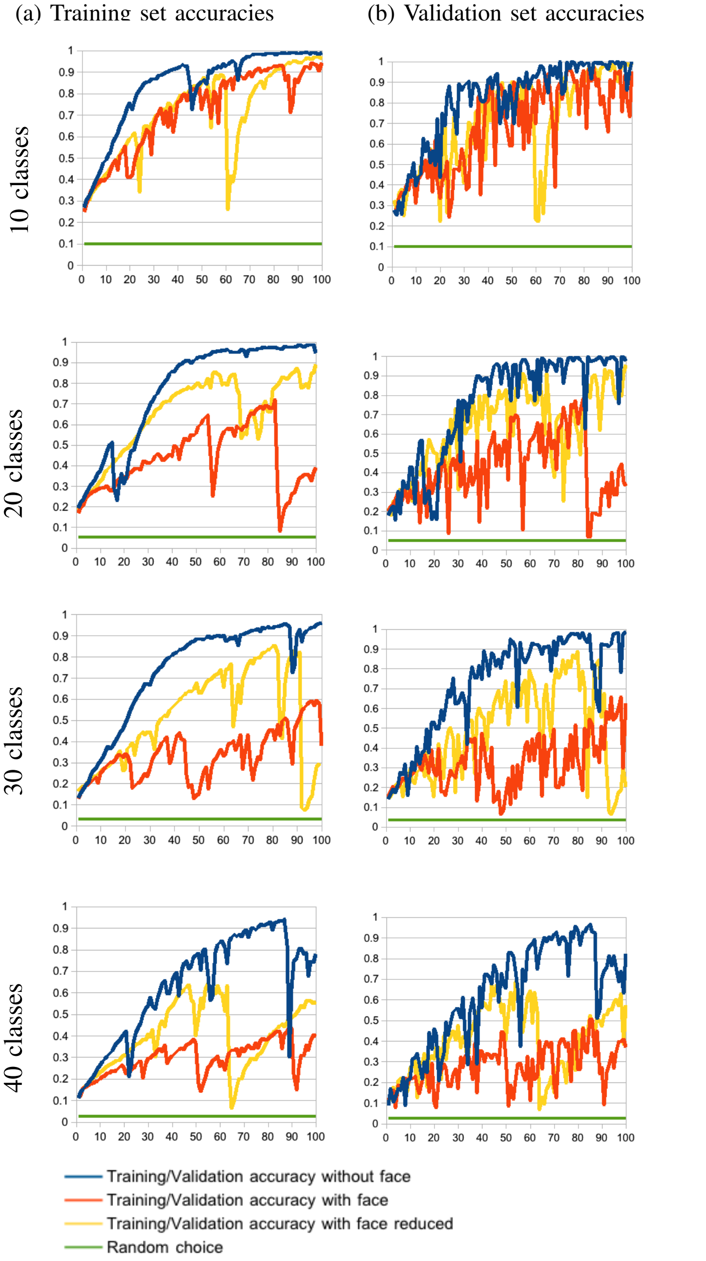
$$Recall = TP/(TP + FN) = \mathbf{0.813}$$

V. Results:Classification

Figure 6 shows the training progress of the model, trained for classifying 10-40 classes of individual signs from the NGT corpus. The figures suggest that the training can produce effective model for the recognition of the signs. However, the training is not stable, the accuracy fluctuates between the epochs and occasionally drops down to the random choice accuracy level. When the model is trained with facial features, the performance degrades, because the input feature vector is increased in size, which makes it more difficult for the model to generalise. When the number of features is reduced from full facial to reduced facial information, the accuracy increases, but does not surpass the accuracy of the model without the facial features. Generally, the more classes the model is trained to distinguish, the more challenging the recognition task.

Table II shows the accuracies, achieved on the testing data for models, trained for 100 epochs on different amount of classes. It is worth noting that not the best, but the last trained model has been used on the training data.

The table shows that the best accuracy is achieved with the lowest number of classes and that the accuracy degrades with addition of more extracted features. This result could arise due to the amount of features used for the recognition, some of which could be perceived only as noise during the training and the recognition, as they do not convey any meaning for the chosen signs.


(a) Training set accuracies

(b) Validation set accuracies

Fig. 6: Individual sign language classification model training for 10, 20, 30, and 40 classes. The graphs in the right column correspond to the training accuracy and the graphs in the left column correspond to the validation accuracy. X-axes correspond to the number of epochs the model was trained on, while the Y-axes correspond to the accuracy of the model on the validation set.

VI. Discussion

Segmentation accuracy indicates that the approach that uses heuristics to detect epenthesis can achieve sufficient results. Varying parameters of the segmentation may yield

TABLE II: Testing accuracies

	Without face information	**With face reduced features**	**With face full features**
10 classes	0.999	0.992	0.955
20 classes	0.972	0.951	0.344
30 classes	0.983	0.207	0.625
40 classes	0.807	0.572	0.378

different results by manipulating the threshold of the H1 and/or H2 and simultaneously changing the number of frames for which H1 and H2 are computed. By allowing more frames, it would be more likely for the H1 or H2 to increase, because the epenthesis will become a part of the segment. Therefore, it is important to use the information about the average sign length and choose the number of frames to be fewer than the average number of frames per sign.

The classification results suggest that for the selected signs, listed in Section III-A, the inclusion of the facial features degrades the classification accuracy, whether all the features are chosen or the reduced amount. More experiments will be required to identify whether these results are consistent for the different signs, even those that are heavily dependent on the facial features. Additional consultation with a linguist will be needed to identify which signs are heavily dependent on the facial features and which are not in the NGT dataset. Obtained results support the claim that not all extracted features are necessary for the successful classification of signs.

VII. Conclusion and Future Work

The paper presented the continuous sign language recognition pipeline that uses heuristic approach for epenthesis detection and deep learning for isolated signs recognition in a continuous stream of video data. The methods show adequate results when tested individually, while more resources need to be invested for an integrated continuous sign language recognition system. The paper investigated the utility of the extracted features for the sign language recognition model. The results suggest that, for the selected signs from the NGT dataset and the chosen stacked LSTM model, not all the features are necessary to perform relatively accurate sign language recognition. Our primary goal is to support continuous natural interaction between the user and the machine as we focus on sign languages as means for communication. Sole segmentation and recognition of the perceived signs is not enough to achieve the understanding of sign language between the human and the machine in terms of dialogue. To cope with the occasional misclassifications, we propose to learn policies for the specific domains (i.e. navigation domain) that map perception to action and reduce the classification confusion, as the choices of actions available to the machine will be restricted by the current state.

References

[1] G. Fang, W. Gao, and D. Zhao, "Large-vocabulary continuous sign language recognition based on transition-movement models," *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, vol. 37, no. 1, pp. 1–9, Jan 2007.

[2] 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.

[3] Y. Jiang, J. Tao, W. Ye, W. Wang, and Z. Ye, "An isolated sign language recognition system using rgb-d sensor with sparse coding," in *2014 IEEE 17th International Conference on Computational Science and Engineering*, Dec 2014, pp. 21–26.

[4] K. M. Lim, A. W. Tan, and S. C. Tan, "A feature covariance matrix with serial particle filter for isolated sign language recognition," *Expert Systems with Applications*, vol. 54, pp. 208–218, 2016.

[5] O. Koller, J. Forster, and H. Ney, "Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers," *Computer Vision and Image Understanding*, vol. 141, no. Supplement C, pp. 108 – 125, 2015, pose & Gesture. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1077314215002088

[6] R. Yang, S. Sarkar, and B. Loeding, "Handling movement epenthesis and hand segmentation ambiguities in continuous sign language recognition using nested dynamic programming," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 3, pp. 462–477, 2010.

[7] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[8] T. Simon, H. Joo, I. Matthews, and Y. Sheikh, "Hand keypoint detection in single images using multiview bootstrapping," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[9] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[10] H. Cooper, B. Holt, and R. Bowden, *Sign Language Recognition*. London: Springer London, 2011, pp. 539–562.

[11] A. Gineke, M. J. Reinders, E. A. Hendriks, H. de Ridder, and A. J. van Doorn, "Influence of handshape information on automatic sign language recognition," in *International Gesture Workshop*. Springer, 2009, pp. 301–312.

[12] S. Ojala, T. Salakoski, and O. Aaltonen, "Coarticulation in sign and speech," in *workshop Multimodal Communication, from Human Behaviour to Computational Models*, 2009.

[13] J. Ann, "On the relation between ease of articulation and frequency of occurrence of handshapes in two sign languages," *Lingua*, vol. 98, no. 1, pp. 19 – 41, 1996, sign Linguistics Phonetics, Phonology and Morpho-syntax. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0024384195000313

[14] H. Cate, F. Dalvi, and Z. Hussain, "Sign language recognition using temporal classification," *CoRR*, vol. abs/1701.01875, 2017. [Online]. Available: http://arxiv.org/abs/1701.01875

[15] N. C. Camgöz, S. Hadfield, O. Koller, and R. Bowden, "Subunets: End-to-end hand shape and continuous sign language recognition," *ICCV 2017 Proceedings*, 2017.

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

[17] O. A. Crasborn and I. Zwitserlood, "The corpus NGT: an online corpus for professionals and laymen," 2008.

[18] R. Nishio, S.-E. Hong, S. König, R. Konrad, G. Langer, T. Hanke, and C. Rathmann, "Elicitation methods in the DGS (german sign language) corpus project," in *Poster presented at the 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies, following the 2010 LREC Conference in Malta, May 22.-23., 2010.*

[19] G. Quinn, A. Merrison, B. Davies, K. Pollitt, and G. Turner, "Task-oriented discourse between british sign language (BSL) users," in *British Association for Applied Linguistics 41st Annual Meeting*, 2008.

[20] A. Choudhury, A. K. Talukdar, M. K. Bhuyan, and K. K. Sarma, "Movement epenthesis detection for continuous sign language recognition," *Journal of Intelligent Systems*, 2017.