K-RSL: a Corpus for Linguistic Understanding, Visual Evaluation, and Recognition of Sign Languages

Alfarabi Imashev, Medet Mukushev, Vadim Kimmelman[†], Anara Sandygulova
Department of Robotics and Mechatronics, School of Engineering and Digital Sciences,
Nazarbayev University, Kazakhstan
[†]Department of Linguistic, Literary and Aesthetic Studies,
University of Bergen, Norway
alfarabi.imashev@nu.edu.kz, mmukushev@nu.edu.kz,
vadim.kimmelman@uib.no, anara.sandygulova@nu.edu.kz

Abstract

The paper presents the first dataset that aims to serve interdisciplinary purposes for the utility of computer vision community and sign language linguistics. To date, a majority of Sign Language Recognition (SLR) approaches focus on recognising sign language as a manual gesture recognition problem. However, signers use other articulators: facial expressions, head and body position and movement to convey linguistic information. Given the important role of non-manual markers, this paper proposes a dataset and presents a use case to stress the importance of including non-manual features to improve the recognition accuracy of signs. To the best of our knowledge no prior publicly available dataset exists that explicitly focuses on non-manual components responsible for the grammar of sign languages. To this end, the proposed dataset contains 28250 videos of signs of high resolution and quality, with annotation of manual and non-manual components. We conducted a series of evaluations in order to investigate whether non-manual components would improve signs' recognition accuracy. We release the dataset to encourage SLR researchers and help advance current progress in this area toward real-time sign language interpretation. Our dataset will be made publicly available at https://krslproject.github.io/krsl-corpus

1 Introduction

There exist over 300 sign languages around the world that are native to 70 million deaf people (Bragg et al., 2019). Sign languages are comprised of hand gestures, arms and body movements, head position, facial expressions, and lip patterns (Sandler and Lillo-Martin, 2006). While automatic speech recognition has progressed to being commercially available, automatic Sign Language Recognition (SLR) is still in its infancy (Cooper et al., 2011).

To date, more than half of published vision-based research utilizes isolated sign language data with a vocabulary size of less than 50 signs (Koller, 2020). But the real-world utility of SLR solutions requires continuous recognition, which is significantly more challenging than recognising individual signs due to co-articulation (the ending of one sign affecting the start of the next), depiction (visually representing or enacting content), epenthesis effects (insertion of extra features into signs), generalization, and so on (Bragg et al., 2019). As a result, realistic, generalisable, and large datasets are necessary to advance SLR.

Current efforts in SLR do not address the complexities of sign language linguistics, and thus have a limited real-world value (Bragg et al., 2019). Chatzis et al. (2020) highlight the importance of non-manual components of sign languages. For example, they can change meaning of a verb, or differentiate between objects and people. According to Koller (2020), there is an overall lack of non-manual parameters that are included in medium and larger vocabulary recognition systems. For example, many computer vision approaches focus on the signers' hands only and tend to ignore the rich channel of information conveyed by non-manual articulators: facial expressions, mouthing, movement and position of the head and body conveying important grammatical and lexical information. In addition, many datasets allowed novice or non-native contributions (i.e. students) in addition to slower signing and simplifying the style and the vocabulary to make the computer vision problem easier but of no real value (Bragg et al., 2019). For the progress in SLR, interdisciplinary efforts are required with an involvement of native signers and sign language linguists.

Beyond targeting the local need of creating the first corpus within CIS (Commonwealth of Independent States) region suitable for machine learn-

Proceedings of the 24th Conference on Computational Natural Language Learning, pages 631–640

https://doi.org/10.18653/v1/P17

Datasets	Signers	Vocabulary	Videos
Purdue RVL-SLLL ASL (2002) (Martínez et al., 2002)	14	104	2,576
GSL Lemmas (2007) (Efthimiou and Fotinea, 2007)	2	1046	2,100
RWTH-BOSTON (2008) (Athitsos et al., 2008)	5	483	7,768
SIGNUM (2008) (Von Agris et al., 2008)	25	780	3,703
Finish S-pot (2014) (Viitaniemi et al., 2014)	5	1211	4,328
RWTH-PHOENIX-Weather 2014 T(Cihan Camgoz et al., 2018)	9	1231	45,760
Video-Based CSL (2018) (Huang et al., 2018)	50	178	25,000
KETI (2019) (Ko et al., 2019)	12	419	11,578
GSL SI (2019) (Chatzis et al., 2020)	7	310	10,290
K-RSL	**10**	**600**	**28,250**

Table 1: Datasets used for sign language recognition

ing, the motivation behind the proposed dataset is in the need to stress the importance of non-manual components present in many signs. The proposed dataset contains continuous sign language data with a focus on specifically selected cases where non-manual markers play a vital role in differentiating between similar signs or sentences. This approach of corpus creation allows researchers from different fields to conduct experiments utilising this dataset. To date, SL linguists and ML researchers were rarely able to utilize the same datasets due to limitations of both kinds. Thus, we make the following contributions:

- we release the first Kazakh-Russian Sign Language (KRSL) corpus consisting of 10 signers, 28250 continuous sentences, and vocabulary size 600 signs appropriate for ML research;
- we release raw videos appropriate for linguists and general population;
- we release isolated signs, extracted frames and features for easy and fast experiments aiming at compatibility with the formats of other SL datasets;
- we evaluate pose estimation and action recognition approaches to setup baselines on the K-RSL dataset.

Section 2 presents the background on sign languages and non-manual components followed by a brief description of other SL datasets. Section 3 outlines the proposed dataset. Section 4 details a series of baseline evaluations conducted in order to investigate whether non-manual components would improve recognition accuracy. Section 5 details our use case evaluation. Section 6 concludes the paper.

2 Related work

This section discusses related work on sign language datasets, state of the art in SLR, and the importance of non-manual features for sign languages.

2.1 Sign Language Datasets

Sign language datasets consist of videos of either isolated or continuous signing. Table 1 presents a comparison of the continuous sign language datasets commonly utilized for sign language recognition with an inclusion of the proposed K-RSL ordered by date. Bragg et al. (2019) specify that the size of the datasets, continuous signing, involvement of native signers, and signers' variety are the main concerns related to current datasets. These challenges put a limitation on the accuracy and robustness of the models developed for SLR to be deployed in the real-world applications.

2.2 Sign Language Recognition

Latest works in the area of SLR are focused on vision-based continuous sign language recognition. All the evaluations are performed on the RWTH-PHOENIX-Weather 2014 dataset (Cihan Camgoz et al., 2018). There are various approaches offering recognition frameworks utilizing deep neural networks, reinforcement learning or recurrent neural networks. For example, Zhang et al. (2019) proposed an approach that apply encoder-decoder structure to the reinforcement learning. Their method achieved competitive results when compared with other methods and has a Word Error Rate (WER) of 38.3%. Temporal segmentation creates additional challenges for continuous SLR. To address this issue, Huang et al. (2018) proposed the Hierarchical Attention Network with Latent

Phrases type	Signers	Phrases	Repetitions	Videos	Glosses
Question-Statement	5	200	10	10000	150
Emotions	5	60	10	3000	140
Emotional Question-Statement	10	30	10	9000	20
Minimal pairs	5	125	10	6250	360
K-RSL total	**10**	**415**	**10**	**28250**	**600**

Table 2: Kazakh-Russian Sign Language dataset

Space (LS-HAN). This proposed framework eliminated the preprocessing of temporal segmentation and achieved the accuracy of 0.617. Zhou et al. (2019) proposed I3D-TEM-CTC framework with iterative optimization for continuous sign language recognition. By increasing the quality of pseudo labels, the final performance of the system was improved and achieved a WER of 34.5%. However, the most promising results were achieved by combining different modalities. Cui et al. (2019) proposed recurrent convolutional neural network on the multi-modal fusion data of RGB images along with the optical flow data and achieved WER of 22.86%. Koller et al. (2019) presented approaches where they focused on the sequential parallelism to learn a sign language, mouth shape and handshape classifier. They have improved the WER to 26.0%. This clearly shows that combination of manual and non-manual features such as mouth shape could significantly improve performance of the recognition systems.

2.3 Importance of Non-manual Features

Sign languages are natural languages existing in the visual modality (Sandler and Lillo-Martin, 2006). Signs in sign languages are produced not only by using the manual articulators (the hands), but also by non-manual articulators (the body, head, facial features). The importance of the non-manual features is evidenced e.g. by the fact that signers focus their attention not on the hands of the interlocutor, but on the face (Pfau and Quer, 2010).

It has been shown that non-manual markers function at different levels in sign languages (Pfau and Quer, 2010). On the lexical level, signs which are manually identical can be distinguished by facial expression or specifically by mouthing (silent articulation of a word from a spoken language) (Crasborn et al., 2008). Signs referring to emotions are obligatorily accompanied by lexicalized facial expressions related to the corresponding emotion. Non-manual markers are especially important on the level of sentence and beyond. Specifically, negation in many sign languages is expressed by head movements (Zeshan, 2004a), and questions are distinguished from statements by eyebrow and head position almost universally (Zeshan, 2004b). Of course, signers also use the face to express their emotions, so emotional and linguistic non-manual markers can interact in complex ways (De Vos et al., 2009).

Antonakos et al. (2015) presented an overview of non-manual parameter employment for SLR and conclude that a limited number of works focused on employing non-manual features in SLR. There have been works that focused on combining both manual and non-manual features (Freitas et al., 2017; Liu et al., 2014; Yang and Lee, 2013; Mukushev et al., 2020) or non-manual features only (Kumar et al., 2017). While the importance of non-manual markers has been thoroughly demonstrated in linguistic research, their role in sign language recognition has not been investigated in detail yet.

3 The Proposed K-RSL Corpus

Given the important role of non-manual markers, in this paper we present a corpus which is motivated by the importance of both manual and non-manual features. We focus on specific cases where non-manual markers play a vital role in differentiating between similar signs or similar sentences.

3.1 Kazakh-Russian Sign Language (KRSL)

KRSL is the sign language used in the Republic of Kazakhstan. KRSL is closely related to Russian Sign Language (RSL) as centralized language policy of Soviet Union led to the spread of RSL in the Soviet republics. According to Kimmelman et al. (2020) both KRSL and RSL show a substantial lexical overlap, and are completely mutually intelligible. At the same time, it cannot be concluded that the same applies to the grammar of the two languages.

Figure 1: Examples of each sign from our dataset: A) “which one” statement, B) “which one” question, C) “which” statement, D) “which” question, E) “how” statement, F) “how” question, G) “what” statement, H) “what” question, I) “who” statement, J) “who” question, K) “when” statement, L) “when” question, M) “where(location)” statement, N) “where(location)” question, O) “where(direction)” statement, P) “where(direction)” question, Q) “where(direction)” statement, R) “where(direction)” question, S) “how much” statement, T) “how much” question.

Figure 2: Emotions: A) “happy”, B) “sad”, C) “anger”, D) “scared”, E) “pity”, F) “surprised”.

3.2 The Data

K-RSL dataset consists of videos of phrases, recorded by five professional sign language interpreters and one subset was additionally recorded by five deaf participants who are also native signers. Dataset can be divided into four subsets from the linguistic point of view: question-statement pairs, signs of emotion, emotional question-statement pairs, and phonologically similar signs (minimal pairs). They have been asked to sign 200 phrases for the first subset, 60 phrases for the second subset, 30 phrase with 3 emotional characteristics for the third subset, and 125 phrases for the fourth subset accordingly. Each phrase was repeated at least ten times in a row by each signer.

The five hearing participants are hearing native signers of KRSL, as they grew up with parents using KRSL at home. Four of them are employed as news interpreters at the national television. The setup had a green background and a LOGITECH C920 HD PRO WEBCAM. The shooting was performed in an office space without professional lighting sources. The summary of the K-RSL dataset is presented in Table 2.

3.2.1 Question vs Statement

Similar to question words in many spoken languages, question signs in KRSL can be used not only in questions (*Who came?*) but also in statements (*I know who came*). Thus, each question sign can occur either with non-manual question marking (eyebrow raise, sideward or backward head tilt), or without it. In addition, question signs are usually accompanied by mouthing of the corresponding Russian/Kazakh word (e.g. *kto/kim* for ‘who’, and *chto/ne* for ‘what’). While question signs are also distinguished from each other by manual features, mouthing provides extra information, which can be used in recognition. Thus, the two types of non-manual markers (eyebrow and head position vs. mouthing) can play a different role in recognition: the former can be used to distinguish statements from questions, and the latter can be used to help distinguish different question signs from each other. To this end, we selected ten words and composed twenty phrases with each word (ten statements and ten questions): ‘what’, ‘who’, ‘which’, ‘which one’, ‘when’, ‘where (direction)’, ‘where (location)’, ‘why’, ‘how’, and ‘how much’. We distinguish them to twenty classes (as ten words have a pair in both statement and question form).

Figure 3: Examples of facial expressions in neutral, surprised and angry state of mind: A) neutral statements, B) neutral question, C) surprised statement, D) surprised question, E) angry statement, E) angry question.

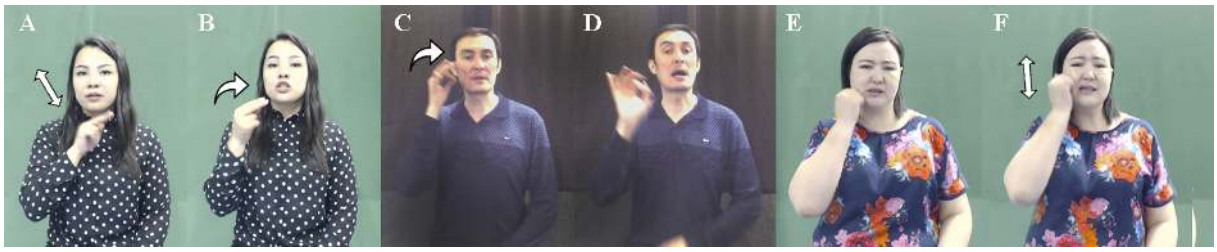

Figure 4: Examples of three phonological minimal pairs: A) "tea", B) "Thursday", C) "orange", D) "October", E) "Moscow" F) "old".

3.2.2 Emotion signs

In KRSL, as in other sign languages, the signs for emotions, such as ANGRY, SAD, SURPRISED, SCARED, PITY, HAPPY are accompanied with facial expressions corresponding to the emotion named by the sign. Therefore, we collected phrases containing the six signs for basic emotions. We hypothesized that, since facial expressions in this signs are lexically associated with them, inclusion of non-manual components can improve recognition of these signs.

3.2.3 Emotional questions vs. emotional statements

De Vos et al. (2009) analyzed interaction of emotional facial expressions and grammatical non-manual markers in Sign Language of the Netherlands (NGT). They elicited polar and content questions in NGT, as well as sentences with topic marking signed neutrally, with anger, or with surprise. Polar questions and topics are normally accompanied with raised eyebrows, while content questions with furrowed eyebrows; the emotion of anger causes eyebrow furrowing, and the emotion of surprise causes eyebrow raise. Therefore, in some of the contexts emotions and grammar were in agreement (e.g. surprised polar questions), while in others in competition (e.g. angry polar questions). The researchers found that emotional and grammatical non-manuals interact in complex ways.

We created a similar dataset for KRSL. The signers were asked to sign ten sentences as either a statement (no eyebrow movement expected), a polar questions (eyebrow raise expected) or wh-questions (adding single question sing), and with three different emotions: neutral, surprise (eyebrow raise expected), and anger (eyebrow furrowing expected). We hypothesized that emotions and grammatical markers would interact in complex ways, and that these interactions might negatively influence recognition accuracy when recognizing sentence types (questions vs statements).

3.2.4 Minimal pairs

Similar to words in spoken languages, signs can form minimal pairs: one can find signs that are minimally different in their manual component (Sandler and Lillo-Martin, 2006). For instance, the KRSL signs "Moscow", "old", and "grandmother" all have the same handshape (the fist) and location (the cheek), but different movements. It is possible to find signs which are distinguished by handshape only or by location only as well.

We hypothesized that minimal pairs of signs are potentially difficult for recognition, as they are quite similar in shape. However, these signs are additionally distinguished by mouthing (see above). Therefore, including non-manual components can improve sign recognition for such pairs of signs. We thus created a dataset with 15 minimal pairs of

signs signed as parts of phrases.

3.3 Openpose Feature Extraction

We utilized OpenPose library (Cao et al., 2017; Wei et al., 2016) in order to extract the keypoints of the person in the videos. OpenPose is the real-time multi-person keypoint detection library for body, face, hands, and foot estimation provided by Carnegie Mellon University (Simon et al., 2017). It detects 2D information of 25 keypoints (joints) on the body and feet, 2x21 keypoints on both hands and 70 keypoints on the face. It also provides a 3D single-person keypoint detection in real time on multi-camera videos. OpenPose provides the values for each keyframe as an output in JSON format. Since the dataset we use consists of RGB videos, we only consider 2D keypoints in this work.

4 Baseline methods

Signing recognition can be considered as a variation of action recognition or human pose estimation tasks. Keypoint detection library OpenPose (Cao et al., 2017; Wei et al., 2016) enables us to evaluate both manual (hand keypoints) and non-manual features (face and pose keypoints). One of the latest works in action recognition (Tran et al., 2018) introduces a new spatiotemporal convolutional block R(2+1)D that achieves state-of-the-art results. In order to analyze and classify collected dataset we employ both approaches as a baseline models for isolated sign recognition. We have extracted isolated clips from the statement-question subset of following signs: 'what', 'who', 'which', 'which one', 'when', 'where (direction)', 'where (location)', 'why', 'how', and 'how much'. We distinguish them to twenty classes (as ten words have a pair in both statement and question form).

4.1 Pose estimation baseline

Our subsets mainly imply classification problems and have sequential features. Generally, we extract features in each frame of videos using OpenPose (Cao et al., 2017; Wei et al., 2016) library and then feed it to the classification algorithm. Therefore, we exploit classical machine learning techniques, namely Logistic regression by concatenating sequences of keypoints into one sample. The sequence of keyframes holds the frames of each sign video. Since we aim to compare performances of non-manual features, we prepared two conditions: **manual only** and **manual and non-manual features combined**. Consequentially, in the first case, one datapoint consists of concatenated keypoints of each video and has a maximum of 30 frames * 84 keypoints = 2520 **manual only** features, while in the second case, one datapoint consists of 30 frames * 274 keypoints = 8220 **manual and non-manual features** for each of the twenty classes. We used the *scikit-learn* library for Python as the keypoints classification method for the experiments presented in this paper.

4.2 Action recognition baseline

Latest works in action recognition either employ Two-Stream Inflated 3D ConvNet (I3D) (Carreira and Zisserman, 2017) or spatiotemporal convolutional block R(2+1)D (Tran et al., 2018). Both architectures are usually trained on ImageNet (Russakovsky et al., 2015) and fine-tuned on Kinetics dataset (Kay et al., 2017).

In this paper, we employ R(2+1)D (Ghadiyaram et al., 2019) model which is highly accurate and significantly faster than other approaches. It is additionally pre-trained on over 65 million videos. Also, it uses as input only video frames, which makes it faster comparing to other approached that require optical flow fields as additional input. In order to recognize signs from our dataset we fine-tuned R(2+1)D on the statement-questions subset. Since we have a different number of classes in our subset, only the last fully connected of the model is re-trained.

4.3 Implementation details

The action recognition baseline is implemented in PyTorch (Paszke et al., 2019) and uses a R(2+1)D pre-trained model (Ghadiyaram et al., 2019). Model input size (number of consecutive frames) is set to 8 and batch size is 16. We train the model for 20 epochs with a starting learning rate of 0.0001. All frames are scaled to a resolution of 112 112 and keeping original ratio. Also, during the training process frames are randomly cropped with scale between 0.6 and 1. The pose estimation baseline is implemented using scikit-learn library (Pedregosa et al., 2011) and takes as an input sequence of keypoints extracted using the OpenPose library (Cao et al., 2017; Wei et al., 2016). We train Logistic Regression classifier using the 'lbfgs' solver and L2 penalty.

4.4 Suggested Train-Test Splits

As stated in Table 2, each subset has 5 signers, which were assigned an approximately equal number of videos. The only exception is the Emotional Question-Statement subset which has 10 signers. We assign all videos performed by 4 signers in the train set and videos with the remaining signer into the test set. In addition, we choose the remaining signer for each class randomly, to diversify train and test data. Validation set is randomly chosen from the train set and has 20% length of the train set.

4.5 Data augmentation

The main problem of developing sign language recognition algorithm is that data is usually not big and/or diverse enough for generalization. Thus, we suggest a simple method to augment image sequences of fixed length from videos with a variable amount of frames. The only constraint is that a video has to be longer than a chosen fixed length.

Given a sign video $V = (f_1, f_2, ..., f_m)$ that contains m frames, which satisfies condition $m \geq n$, where n is the chosen fixed sequence length, we pick equally distanced frames from videos with a random initial frame. By distance between the frames, we mean the difference between their indexes, let's call it s.

$$s = \left\lfloor \frac{m}{n} \right\rfloor$$

The initial frame is picked among all possible candidates which are first s frames with k leftover frames after them. Here, $k = m \mod n$. Therefore, the augmented fixed sized sequence is $S = (f_i, f_{i+s}, f_{i+2s}, ..., f_{i+ns})$, where i is a random integer from 1 to $s + k$.

5 Experimental Results

A series of experiments was conducted in order to investigate whether non-manual features would improve recognition accuracy. All experiments were performed on isolated signs extracted from the Question-Statement subset and divided into 20 classes (10 signs as statement and questions). The first experiment was the classification of 20 classes. For this reason we trained two baseline models: a logistic regression model using only manual features and with non-manual features as an input, and a R(2+1)D model on full frames as an input. Evaluation of each model was repeated 10 times with random train/test splits to avoid extreme cases. Table 3 presents the mean scores and standard deviations for the first experiment. The second experiment used the same dataset with 20 classes to compare and contrast the accuracy in terms of its improvement with different combinations of non-manual components. Table 4 presents the accuracy scores for each combination of features.

	R(2+1)D	Logistic regression	
Features	Full frame	Manual	Non-manual
Mean	86%	73.4%	77%
Std Dev	1	0.45	0.57

Table 3: Mean scores of accuracy for the question-statement subset after 10 iterations with random train/test splits

5.1 Question vs. Statement

Our first experiment used the Question-Statement subset divided into 20 classes (10 signs used in statements and questions). We have extracted manual and non-manual features for the isolated signs of the Question-Statement subset. The highest accuracy was achieved by the R(2+1)D model and was 86%, which is 9% higher comparing to the Logistic regression model. For the Logistic regression model trained on sequence of keypoints testing mean accuracy scores are 73.4% and 77% on manual-only and both manual and non-manual features respectively. As expected, non-manual features improved the results by 3.6% on average (from 73.4% accuracy to 77% accuracy). At the same time, improvement was not very high. The reason for that could be that the number of non-manual features is bigger than the number of manual features.

5.2 A case of combining different modalities

In this experiment different combinations of non-manual markers (eyebrow and head position vs. mouthing) were compared and their role in recognition was analyzed.

The lowest testing accuracy was 73.25% for the combination of manual features and eyebrows keypoints. Eyebrows without any other non-manual feature did not provide valuable information for recognition. Only when they were used in combination with other features, the accuracy was im-

proved. The highest testing accuracy was 78.2% for the combination of manual features and faceline, eyebrows, and mouth keypoints. When only mouth keypoints were used in combination with the manual features, the accuracy also increased by 0.5% compared to the baseline of 77%. Thus, we see that mouthing provides extra information, which can be used in recognition, because signers usually articulate words while performing corresponding signs. Eyebrows and head position provide additional grammatical markers to differentiate statements from questions.

Features combination	Accuracy
Manual only	73.4%
Manual & Non-manual all	77%
Manual & Face, eyebrows, mouth	**78.2%**
Manual & Eyebrows, mouth	77.2%
Manual & Only mouth	77.5%
Manual & Only eyebrows	73.25%

Table 4: Comparison of results of features combinations

6 Conclusion

This paper presents the K-RSL dataset motivated by the need to create SL datasets for interdisciplinary purposes e.g. for computer vision and computational linguistics research. Due to the challenging nature of SLR, the proposed dataset aims to attract the attention of the computer vision community with the K-RSL dataset being linguistically rich. The data was carefully selected to find various cases when manual gestures will not provide good performance and will stress the need to include non-manual components into consideration. In addition to computer vision community, this dataset can be utilized by the linguistics community to explore research questions and computationally prove their hypotheses. Future work will include expanding the vocabulary of the corpus in addition to diversifying and increasing the number of signers recorded in noisy environmental conditions (e.g. outside of the office environment).

Acknowledgment

This work was supported by the Nazarbayev University Faculty Development Competitive Research Grant Program 2019-2021 "Kazakh Sign Language Automatic Recognition System (K-SLARS)". Award number is 110119FD4545.

References

Epameinondas Antonakos, Anastasios Roussos, and Stefanos Zafeiriou. 2015. A survey on mouth modeling and analysis for sign language recognition. *11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, 1:1–7.

Vassilis Athitsos, Carol Neidle, Stan Sclaroff, Joan Nash, Alexandra Stefan, Quan Yuan, and Ashwin Thangali. 2008. The american sign language lexicon video dataset. *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–8.

Danielle Bragg, Oscar Koller, Mary Bellard, Larwan Berke, Patrick Boudreault, Annelies Braffort, Naomi Caselli, Matt Huenerfauth, Hernisa Kacorri, Tessa Verhoef, et al. 2019. Sign language recognition, generation, and translation: An interdisciplinary perspective. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*, pages 16–31. ACM.

Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7291–7299.

Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308.

Theocharis Chatzis, Andreas Stergioulas, Dimitrios Konstantinidis, Kosmas Dimitropoulos, and Petros Daras. 2020. A comprehensive study on deep learning-based 3d hand pose estimation methods. *Applied Sciences*, 10(19):6850.

Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018. Neural sign language translation. pages 7784–7793.

Helen Cooper, Brian Holt, and Richard Bowden. 2011. Sign language recognition. *Visual Analysis of Humans*, pages 539–562.

Onno A Crasborn, Els Van Der Kooij, Dafydd Waters, Bencie Woll, and Johanna Mesch. 2008. Frequency distribution and spreading behavior of different types of mouth actions in three sign languages. *Sign Language & Linguistics*, 11(1):45–67.

Runpeng Cui, Hu Liu, and Changshui Zhang. 2019. A Deep Neural Framework for Continuous Sign Language Recognition by Iterative Training. *IEEE Transactions on Multimedia*, 21(7):1880–1891.

Connie De Vos, Els Van der Kooij, and Onno Crasborn. 2009. Mixed signals: Combining linguistic and affective functions of eyebrows in questions in sign language of the netherlands. *Language and speech*, 52(2-3):315–339.

Eleni Efthimiou and Stavroula-Evita Fotinea. 2007. Gslc: creation and annotation of a greek sign language corpus for hci. *International Conference on Universal Access in Human-Computer Interaction*, pages 657–666.

Fernando A Freitas, Sarajane M Peres, Clodoaldo AM Lima, and Felipe V Barbosa. 2017. Grammatical facial expression recognition in sign language discourse: a study at the syntax level. *Information Systems Frontiers*, 19(6):1243–1259.

Deepti Ghadiyaram, Du Tran, and Dhruv Mahajan. 2019. Large-scale weakly-supervised pre-training for video action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12046–12055.

Jie Huang, Wengang Zhou, Qilin Zhang, Houqiang Li, and Weiping Li. 2018. Video-based sign language recognition without temporal segmentation. In *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*, pages 2257–2264. AAAI press.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.

Vadim Kimmelman, Alfarabi Imashev, Medet Mukushev, and Anara Sandygulova. 2020. Eyebrow position in grammatical and emotional expressions in kazakh-russian sign language: A quantitative study. *PloS one*, 15(6):e0233731.

Sang-Ki Ko, Chang Jo Kim, Hyedong Jung, and Choongsang Cho. 2019. Neural sign language translation based on human keypoint estimation. *Applied Sciences*, 9(13):2683.

Oscar Koller. 2020. Quantitative survey of the state of the art in sign language recognition. *arXiv preprint arXiv:2008.09918*.

Oscar Koller, Cihan Camgoz, Hermann Ney, and Richard Bowden. 2019. 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*, pages 1–1.

Sunil Kumar, Manas Kamal Bhuyan, and Biplab Ketan Chakraborty. 2017. Extraction of texture and geometrical features from informative facial regions for sign language recognition. *Journal on Multimodal User Interfaces*, 11(2):227–239.

Jingjing Liu, Bo Liu, Shaoting Zhang, Fei Yang, Peng Yang, Dimitris N Metaxas, and Carol Neidle. 2014. Non-manual grammatical marker recognition based on multi-scale, spatio-temporal analysis of head pose and facial expressions. *Image and Vision Computing*, 32(10):671–681.

Aleix M Martínez, Ronnie B Wilbur, Robin Shay, and Avinash C Kak. 2002. Purdue rvl-slll asl database for automatic recognition of american sign language. *Proceedings. Fourth IEEE International Conference on Multimodal Interfaces*, pages 167–172.

Medet Mukushev, Arman Sabyrov, Alfarabi Imashev, Kenessary Koishybay, Vadim Kimmelman, and Anara Sandygulova. 2020. Evaluation of manual and non-manual components for sign language recognition. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 6073–6078.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Roland Pfau and Josep Quer. 2010. Nonmanuals: Their prosodic and grammatical roles. *Sign languages*, pages 381–402.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252.

Wendy Sandler and Diane Lillo-Martin. 2006. *Sign language and linguistic universals*. Cambridge University Press.

Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. 2017. Hand keypoint detection in single images using multiview bootstrapping. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1145–1153.

Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. 2018. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459.

Ville Viitaniemi, Tommi Jantunen, Leena Savolainen, Matti Karppa, and Jorma Laaksonen. 2014. S-pot–a benchmark in spotting signs within continuous signing. European Language Resources Association (LREC).

Ulrich Von Agris, Moritz Knorr, and Karl-Friedrich Kraiss. 2008. The significance of facial features for automatic sign language recognition. *8th IEEE International Conference on Automatic Face & Gesture Recognition*, pages 1–6.

Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. 2016. Convolutional pose machines. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4724–4732.

Hee-Deok Yang and Seong-Whan Lee. 2013. Robust sign language recognition by combining manual and non-manual features based on conditional random field and support vector machine. *Pattern Recognition Letters*, 34(16):2051–2056.

Ulrike Zeshan. 2004a. Hand, head and face-negative constructions in sign languages. *Linguistic Typology*, 8(1):1–58.

Ulrike Zeshan. 2004b. Interrogative constructions in signed languages: Crosslinguistic perspectives. *Language*, pages 7–39.

Zhihao Zhang, Junfu Pu, Liansheng Zhuang, Wengang Zhou, and Houqiang Li. 2019. Continuous Sign Language Recognition via Reinforcement Learning. pages 285–289. Institute of Electrical and Electronics Engineers (IEEE).

Hao Zhou, Wengang Zhou, and Houqiang Li. 2019. Dynamic Pseudo Label Decoding for Continuous Sign Language Recognition. pages 1282–1287. Institute of Electrical and Electronics Engineers (IEEE).