Evaluation of Manual and Non-manual Components for Sign Language Recognition

Medet Mukushev*, Arman Sabyrov*, Alfarabi Imashev*, Kenessary Koishybay*
Vadim Kimmelman†, Anara Sandygulova*‡
*Department of Robotics and Mechatronics, School of Engineering and Digital Sciences, Nazarbayev University
Kabanbay Batyr Avenue, 53, Nur-Sultan, Kazakhstan
†Department of Linguistic, Literary and Aesthetic Studies, University of Bergen
Postboks 7805, 5020, Bergen, Norway
mmukushev@nu.edu.kz, arman.sabyrov@nu.edu.kz, alfarabi.imashev@nu.edu.kz, kenessary.koishybay@nu.edu.kz
vadim.kimmelman@uib.no, anara.sandygulova@nu.edu.kz

Abstract

The motivation behind this work lies in the need to differentiate between similar signs that differ in non-manual components present in any sign. To this end, we recorded full sentences signed by five native signers and extracted 5200 isolated sign samples of twenty frequently used signs in Kazakh-Russian Sign Language (K-RSL), which have similar manual components but differ in non-manual components (i.e. facial expressions, eyebrow height, mouth, and head orientation). We conducted a series of evaluations in order to investigate whether non-manual components would improve sign's recognition accuracy. Among standard machine learning approaches, Logistic Regression produced the best results, 78.2% of accuracy for dataset with 20 signs and 77.9% of accuracy for dataset with 2 classes (statement vs question). Dataset can be downloaded from the following website: https://krslproject.github.io/krsl20/

Keywords: Sign language Recognition, Information extraction, Statistical and machine learning methods

1. Introduction

Deaf communities around the world communicate via sign languages to express meaning and intent (Sandler and Lillo-Martin, 2006). Similar to spoken languages, each country or region has its own sign language of varying grammar and rules, leading to a few hundreds of sign languages that exist today. While automatic speech recognition has progressed to being commercially available, automatic Sign Language Recognition (SLR) is still in its infancy (Cooper et al., 2011).

Most works on sign language recognition consider manual and non-manual components separately. Manual features are features related to hands (e.g. hand configuration and motion trajectory of hands), while non-manual features include facial expressions, gaze direction, lip patterns, head and body posture. For example, signers use articulators such as facial expressions, head and body position and movement to convey linguistic information (Pfau and Quer, 2010). It has been shown that non-manual markers function at different levels in sign languages. 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). On the morphological level, facial expressions and mouth patterns are used to convey adjectival and adverbial information (e.g. indicate the size of objects or aspectual properties of events) (Crasborn et al., 2008). Non-manual markers are especially important on the sentence level and beyond. Almost universally, the negation in many sign languages is expressed by head movements (Zeshan, 2004a), while questions are distinguished from statements by eyebrow and head position (Zeshan, 2004b).

Given the important role of non-manual markers, in this paper, we evaluate whether including non-manual features improves the recognition accuracy of signs. We focus on a specific case where two types of non-manual markers play a role, namely question signs in K-RSL. Similar to question words in many spoken languages, question signs in K-RSL 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 hypothesize that the addition of non-manual markers will improve the recognition accuracy.

Sign language of Kazakhstan is closely related to Russian Sign Language (RSL) like many other sign languages within the Commonwealth of Independent States (CIS). The closest corpus is the Novosibirsk State University of Technology RSL Corpus (Burkova, 2014). However, it has been created as a linguistic corpus for studying previously unexplored fragments of RSL, thus it is inappropriate for machine learning and this research. Thus, this paper also contributes with the evaluation data which consists of 5200 videos of 10 frequently used question signs in K-RSL (Sandygulova, A., 2020). These 10 signs are 'what for', 'who', 'which, 'which-2', 'when', 'where (direction)', 'where (location)', 'what', 'how', and 'how much'. Each of this question sign is used in either statement or question that in total produce 20 classification classes.

‡Corresponding author

Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020) , pages 6075–6080
Marseille, 11–16 May 2020

Figure 1: Examples of seven sign pairs from our dataset: A) "what for" statement, B) "what for" question, C) "where (direction)" statement, D) "where (direction)" question, E) "which" statement, F) "which" question, G) "where (location)" statement, H) "where (location)" question, I) "which-2" statement, J) "which-2" question, K) "what" statement, L) "what" question, M) "how" statement , N) "how" question

2. Related Work

Current SLR research is focused on continuous signing utilizing RWTH-PHOENIX-Weather 2014 (Forster et al., 2014) as a benchmark dataset. For example, Cui et al. (2017) utilized Recurrent-CNN for spatio-temporal feature extraction and sequence learning, achieving a WER of 38.7%. Zhang et al. (2019) obtained WER of 38.3% by applying transformer with reinforcement learning. Pu et al. (2019) proposed a new deep architecture for continuous SLR based on 3D-ResNet and encoder-decoder network with connectionist temporal classification with a WER result of 37.1%. Koller et al. (2018) utilized a hybrid CNN-HMM approach where the Language Model was used to maximize models in HMM. They achieved a WER (Word Error Rate) of 32.5%. Cui et al. (2019) improved WER to 22.9% by applying iterative training.

Koller et al. (2018) also provides an overview of the latest results in SLR using deep learning methods. However, their approach exploits only a single cropped hand of the signer and since it still achieves the state-of-the-art, it is hypothesized that additional modalities such as non-manual components (facial expression, eyebrow height, mouth, head orientation, and upper body orientation) might increase this performance.

Antonakos et al. (2015) presented an overview of non-manual parameters employment for SLR. Lip patterns represent the most distinctive non-manual parameter. They solve ambiguities between signs, specify expressions and provide information redundant to gesturing to support differentiation of similar signs. In addition to lip patterns, the head pose supports the semantics of a sign language. Questions, affirmations, denials, and conditional clauses are communicated, e.g., with the help of the signer's head pose. Antonakos et al. (2015) conclude that a limited number of works focused on employing non-manual features in SLR.

Freitas et al. (2017) developed models for recognition of grammatical facial expressions in Libras (Brazilian Sign Language). They used Multi-layer Perceptron and achieved F-scores over 80% for most of their experiments. One of the interesting findings of their work was that classification accuracy can vary depending on how clear the signing of the signer is.

Liu et al. (2014) developed a system that automatically detects non-manual grammatical markers. They were able to increase the recognition rate by adding high-level facial features, which are based on events such as head shake and nod, raised or lowered eyebrows. Low-level features are based on facial geometry and head pose. Combining both low-level and high-level features for recognition showed significant improvement in accuracy performance.

Kumar et al. (2017) attempted to recognize selected sign language gestures using only non-manual features. For this need, they developed a new face model with 54 landmark points. Active Appearance Model was used for extracting features of facial expressions and recognized signs using Hidden Conditional Random Field. They have used the RWTH-BOSTON-50 dataset for experiments and their proposed model achieved an 80% recognition rate.

In contrast, Yang and Lee (2013) proposed a new method that applied non-manual features, extracted from facial expressions, in addition to manual features. They used non-manual features in cases of uncertainty in decisions made based on manual features only. Facial feature points were extracted using the Active Appearance Model and then Support Vector Machines was applied for recognition of non-manual features. The highest recognition rate of 84% was achieved by their method when both manual and non-manual features were combined, which was 4% higher compared to a case when only manual features were used.

In addition to previous work, this paper aims to explicitly evaluate a particular case to emphasize the need to differentiate between similar signs that only differ in non-manual components.

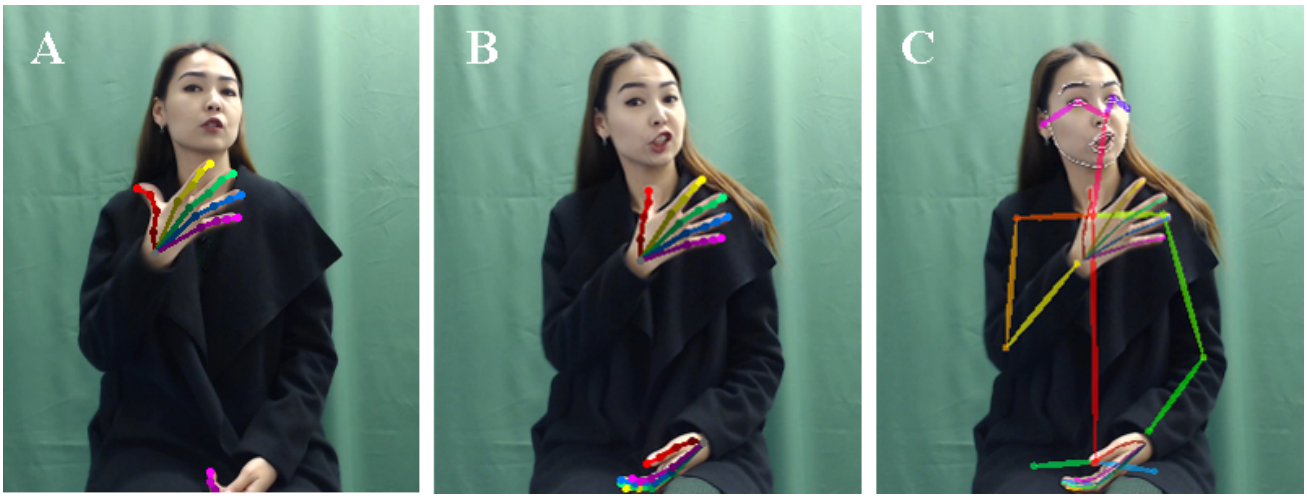

Figure 2: Examples of OpenPose features: A) 'for what' statement, only manual features, B) 'for what' question, only manual features, C) 'for what' question, with manual and non-manual features

3. Methodology

3.1. Data collection

To explore the above stated hypotheses, we collected a relatively small dataset of K-RSL similar to previously collected data (Imashev, 2017). We recorded five professional sign language interpreters. Four of them are employed as news interpreters at the national television. All signers can be considered as native signers as each has at least one deaf parent. 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.

We selected ten words (signs) and composed twenty phrases with each word (ten statements and ten questions): 'what for', 'who', 'which, 'which-2', 'when', 'where (direction)', 'where (location)', 'what', 'how', and 'how much'. We distinguish twenty classes (as each of the ten words has a realization in both statement and question form). In total, signers were asked to sign 200 phrases: 20 phrases were repeated 10 times. The reason for choosing these particular signs is that they carry different prosodic information when used in questions and statements. Also, they are similar in manual articulation but differ in non-manual articulation. Figure 1 provides examples of seven sign pairs from our dataset (out of the ten in pairs in total).

3.2. OpenPose

We utilized OpenPose in order to extract keypoints of people in the videos. OpenPose is the real-time multi-person keypoint detection library for body, face, hands, and foot estimation (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. OpenPose provides the values for each keyframe as an output in JSON format. Figure 2 presents manual and non-manual features extracted using OpenPose from each frame of the video.

3.3. Classification

Classification was performed utilizing standard machine learning approaches such as Support Vector Machines, Logistic Regression, Random Forest, Random Tree, BayesNet and others. To this end, the dataset was converted to Arff format - the format used by the Weka machine learning tool (Holmes et al., 1994), and CSV (comma separated values) format.

The classifier was trained on sequences of keyframes extracted from the OpenPose. 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: **non-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 20 classes. Logistic Regression provided the best accuracy and thus was selected to be integrated into all experiments. We used scikit-learn library for Python with default parameters as the main classification method for the experiments presented in this paper.

4. Experiments

We conducted a series of experiments in order to investigate whether non-manual features would improve the recognition accuracy for 20 signs. The first experiment used a k-fold cross-validation on the collected dataset of native signers (five people) where samples were divided into 2 classes (statement and questions). The second experiment used the same dataset but samples were divided into 20 classes (10 signs as statement and questions). The third 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. Each experiment was repeated 10 times with random train/test splits to avoid extreme cases. Table 1 presents mean scores and standard deviation for the first and second experiments.

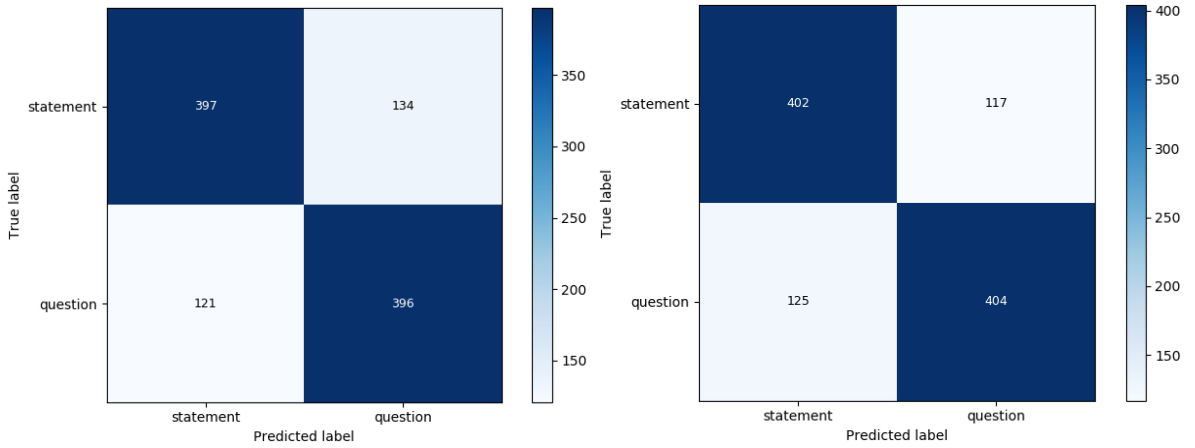

Figure 3: Confusion matrix for 2 classes (statement vs question) with manual only features (left). Accuracy is 73.9%. Confusion matrix for 2 classes with both manual and non-manual features (right). Accuracy is 77.36%.

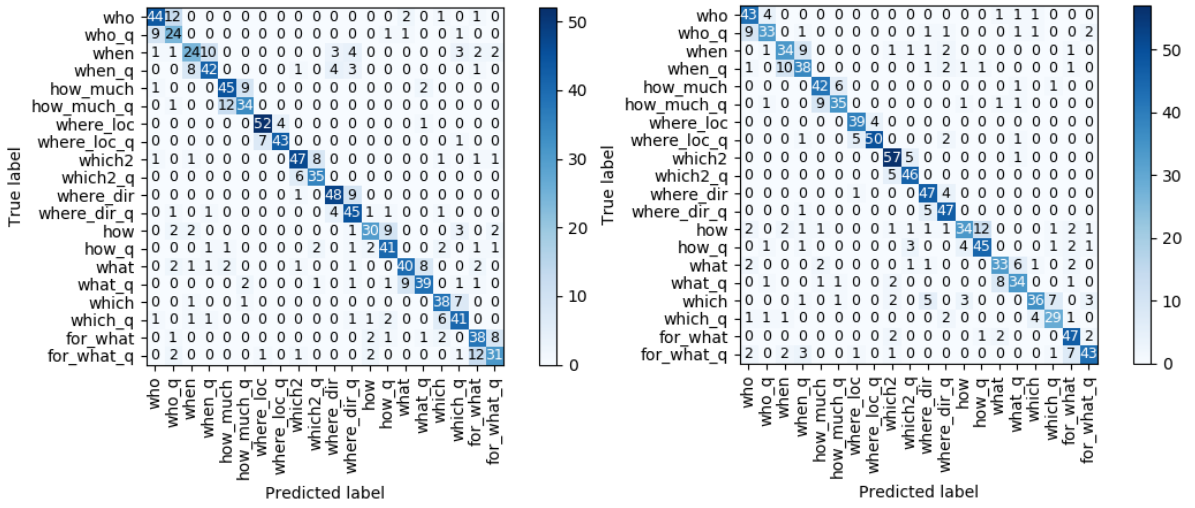

Figure 4: Confusion matrix for 20 signs with manual only features (left). Accuracy is 73.4%. Confusion matrix for 20 signs with both manual and non-manual features (right). Accuracy is 77%.

Table 1: Mean scores of accuracy for two experiments after 10 iterations with random train/test splits

	2 classes		20 classes	
Scores	Manual	Non-manual	Manual	Non-manual
Mean	73.9%	77.36%	73.4%	77%
Std Dev	0.65	0.34	0.45	0.57

4.1. A case of two classes

To experiment with all videos, the k-fold cross-validation method was applied to the classification. The whole dataset was divided into training and testing sets (80/20 split, 4160 samples for training and 1040 samples for testing). Choosing k equal to 5 (80 and 20 splits), the training and validation were performed for each fold. Figure 3 demonstrates the confusion matrices of the obtained results for the first experiment. Testing accuracies are 73.9% and 77.36% on manual-only and both manual and non-manual features respectively. A qualitative examination of the confusions in the non-manual and manual confusion matrix (Figure 3 (right)) shows that by adding non-manual features it was possible to correctly identify 8 more samples as questions and 5 more samples as statements, which were classified wrongly when using only manual features. We see that non-manual markers can be used to help distinguish different signs from each other when they are used in statements vs. questions.

4.2. A case of twenty classes

Figure 4 presents the confusion matrices of the obtained results for the second experiment. Testing mean accuracy scores are 73.4% and 77% on manual-only and both manual and non-manual features respectively.

Qualitative examination of the top confusions in manual-only confusion matrix (Figure 4 (left)) highlight confused pairs such as “who” (statement) and “who_q” (question) with 23.5% confusion, “when” (statement) and “when_q” (question) with 21.4% confusion, “how_much” (statement) and “how_much_q” (question) with 21% confusion, “For what” (statement) and “For what_Q” (question) with 22.4% confusion. Since these signs share the same hand config-

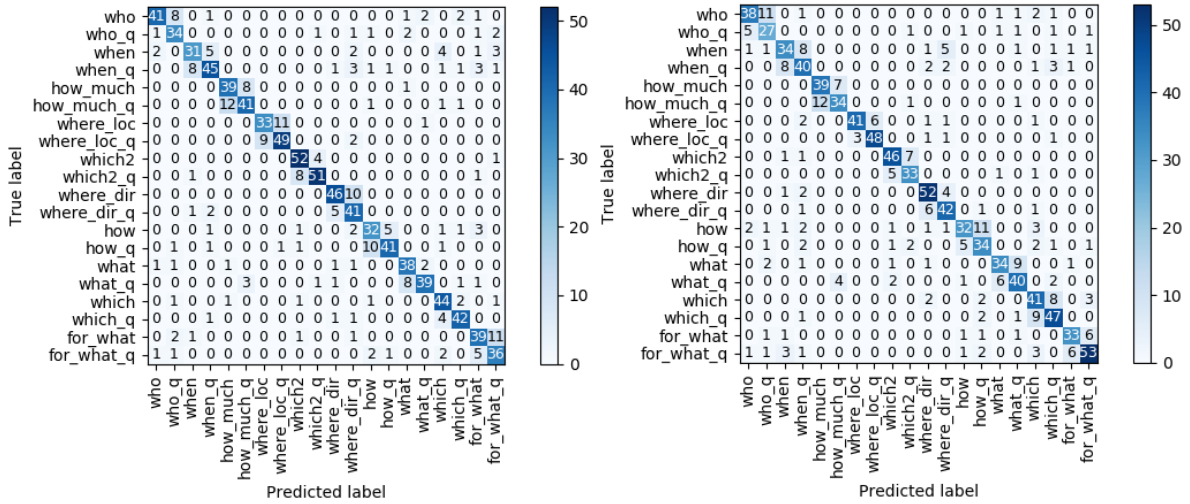

Figure 5: Confusion matrix for 20 signs with manual and non-manual (faceline, eyebrows, eyes, mouth) features (left). Accuracy is 78.2%. Confusion matrix for 20 signs with manual and non-manual (eyebrows, eyes, mouth) features (right). Accuracy is 77.2%.

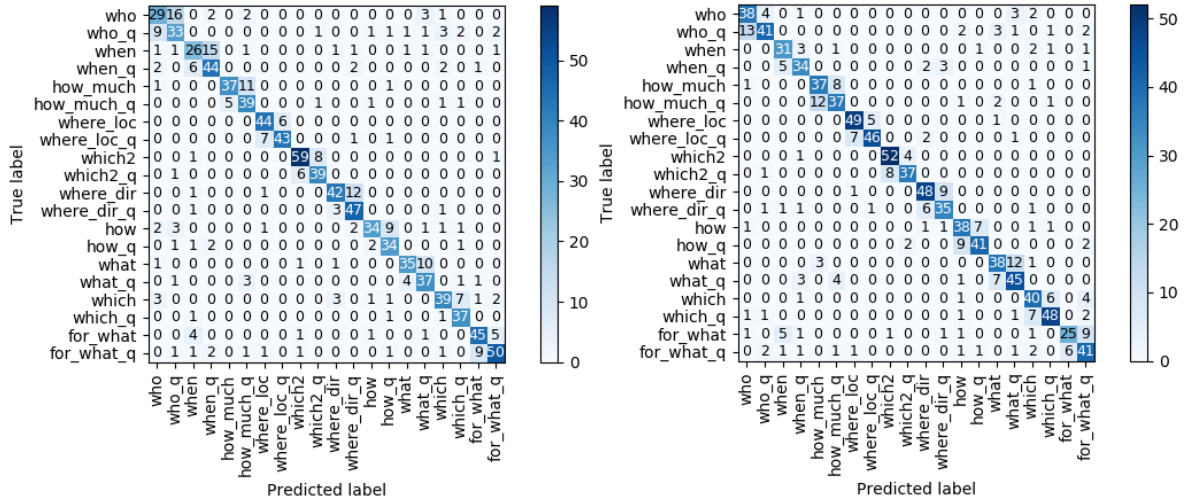

Figure 6: Confusion matrix for 20 signs with manual and non-manual (only eyebrows and eyes) features (left). Accuracy is 73.25%. Confusion matrix for 20 signs with manual and non-manual (only mouth) features (right). Accuracy is 77.5%.

urations and only the facial expression changes, it is expected that manual-only features caused such an error. And as expected, non-manual features improved the results by 3.6% on average (from 73.4% accuracy to 77% accuracy) for mainly these signs ("who" pair had a decrease to 14.6% confusion, "how_much" pair decreased to 16.3% confusion, "for_what" pair decreases its confusion to 9%).

4.3. A case of combining different modalities

Figures 5 and 6 show the confusion matrices of the obtained results for the third experiment. 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 combination of manual features and eyebrows keypoints. Eyebrows without any other non-manual feature did not provide valuable information for recognition. Only when they are used in combination with other features, the accuracy was improved. The highest testing accuracy was 78.2% for 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.

5. Conclusion

Automatic SLR poses many challenges since each sign involves various manual and non-manual components and varies from signer to signer. Since deep learning methods require a lot of data and it is quite challenging to collect the data from native signers, many datasets are not balanced and have only limited vocabulary. We decided to investigate whether improvement in recognition accuracy would

be due to the addition of non-manual features. Similarly to related works by Freitas et al. (2017), Yang and Lee (2013) we saw an improvement in 5% for the experiments with the addition of non-manual features. Table 2 compares our results obtained from the experiments:

Table 2: Comparison of results

Features	Test 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%

The aim of this paper was not in achieving the best accuracy in the literature of automatic SLR, nor in utilizing a large dataset of continuous signs for the prediction, but rather to compare and contrast the accuracies in terms of improvement when non-manual components are integrated as an additional modality for recognition.

6. Acknowledgements

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.

7. Bibliographical References

Antonakos, E., Roussos, A., and Zafeiriou, S. (2015). A survey on mouth modeling and analysis for sign language recognition. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 1, pages 1–7. IEEE.

Burkova, S. I. (2014). Russian sign language corpus project. *Retrieved from http://rsl.nstu.ru/site/project*.

Cooper, H., Holt, B., and Bowden, R. (2011). Sign language recognition. In *Visual Analysis of Humans*, pages 539–562. Springer.

Crasborn, O. A., Van Der Kooij, E., Waters, D., Woll, B., and Mesch, J. (2008). Frequency distribution and spreading behavior of different types of mouth actions in three sign languages. *Sign Language & Linguistics*, 11(1):45–67.

Cui, R., Liu, H., and Zhang, C. (2017). Recurrent convolutional neural networks for continuous sign language recognition by staged optimization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7361–7369.

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

Forster, J., Schmidt, C., Koller, O., Bellgardt, M., and Ney, H. (2014). Extensions of the sign language recognition and translation corpus rwth-phoenix-weather. In *LREC*, pages 1911–1916.

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

Holmes, G., Donkin, A., and Witten, I. H. (1994). Weka: A machine learning workbench. In *Proceedings of ANZIIS'94-Australian New Zealand Intelligent Information Systems Conference*, pages 357–361. IEEE.

Imashev, A. (2017). Sign language static gestures recognition tool prototype. In *2017 IEEE 11th International Conference on Application of Information and Communication Technologies (AICT)*, pages 1–4. IEEE.

Koller, O., Zargaran, S., Ney, H., and Bowden, R. (2018). Deep sign: enabling robust statistical continuous sign language recognition via hybrid cnn-hmms. *International Journal of Computer Vision*, 126(12):1311–1325.

Kumar, S., Bhuyan, M. K., and Chakraborty, B. K. (2017). Extraction of texture and geometrical features from informative facial regions for sign language recognition. *Journal on Multimodal User Interfaces*, 11(2):227–239.

Liu, J., Liu, B., Zhang, S., Yang, F., Yang, P., Metaxas, D. N., and Neidle, C. (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.

Pfau, R. and Quer, J. (2010). Nonmanuals: Their prosodic and grammatical roles. *Sign languages*, pages 381–402.

Pu, J., Zhou, W., and Li, H. (2019). Iterative Alignment Network for Continuous Sign Language Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4165–4174.

Sandler, W. and Lillo-Martin, D. (2006). *Sign language and linguistic universals*. Cambridge University Press.

Simon, T., Joo, H., Matthews, I., and Sheikh, Y. (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.

Yang, H.-D. and Lee, S.-W. (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.

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

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

Zhang, Z., Pu, J., Zhuang, L., Zhou, W., and Li, H. (2019). Continuous sign language recognition via reinforcement learning. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 285–289. IEEE.

8. Language Resource References

Sandygulova, A. (2020). *Kazakh-Russian Sign Language Statement Question Dataset (KRSL20)*. K-SLARS Project, 1.0, ISLRN 634-887-032-200-8.