Towards Real-time Sign Language Interpreting Robot: Evaluation of Non-manual Components on Recognition Accuracy

Arman Sabyrov*, Medet Mukushev*
Alfarabi Imashev, Kenessary Koishybay, Anara Sandygulova†
Nazarbayev University, Kazakhstan
arman.sabyrov@nu.edu.kz, mmukushev@nu.edu.kz
alfarabi.imashev@nu.edu.kz, kenessary.koishybay@nu.edu.kz,
anara.sandygulova@nu.edu.kz†

Vadim Kimmelman
University of Bergen
Bergen, Norway
vadim.kimmelman@uib.no

Abstract

The purpose of this work is to develop a human-robot interaction system that could be used as a sign language interpreter. The paper presents the results of the ongoing work, which aims to recognize sign language in real time. 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 2000 videos 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, 73% of accuracy for dataset with 20 signs and 80.25% of accuracy for dataset with 2 classes (statement vs question).

1. Introduction

Deaf communities around the world communicate via sign languages, which uses gestures to express meaning and intent, that include hand-shapes, arms and body, head position, facial expressions and lip-patterns [22]. 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 [3]. While automatic speech recognition has progressed to being commercially available, automatic Sign Language Recognition (SLR) is still in its infancy [7]. Currently, human signers carry out all commercial translation services, who are required to get training and to be experienced, thus are often expensive for the majority in need [25].

With the latest advancements of deep learning research, we envision that the ambitious goal of developing a complete human-robot interaction system acting as a sign language interpreter for the deaf could be achieved. The focus of the work presented in this paper is to implement a vision-based approach to be deployed on a robot such as a humanoid robot Pepper for effective automatic SLR.

However, there is a lack of sign language data for deep learning research. To date, only a handful of sign languages have their corpora [4, 6, 10, 16] for the use by general population, linguists and researchers in computer vision and machine learning. 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 [3]. Unfortunately, there are no such corpora for the majority of sign languages, which makes research on a particular sign language a very difficult and resource consuming challenge.

Sign Language used in Kazakhstan is closely related to Russian Sign Language (RSL) like many other sign languages within Commonwealth of Independent States (CIS). The closest corpus within CIS area is the Novosibirsk State University of Technology RSL Corpus [5]. However it has been created as a linguistic corpus for studying previously unexplored fragments of RSL, thus it is inappropriate for machine learning. The creation of the first K-RSL corpus will change the situation, and it can be used within CIS and beyond.

In addition, many approaches focus on the signer's hands only. However, signers use other articulators: facial expressions, and head and body position and movement to convey linguistic information, too [20]. It has been shown that non-manual markers function at different levels in sign languages. On the lexical level, signs which are manually iden-

*Joint first authors
†Corresponding author

Figure 1. Examples of each sign from our data set: 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

tical can be distinguished by facial expression or specifically by mouthing (silent articulation of a word from a spoken language) [8]. On the morphological level, facial expressions and mouth patterns are used to convey adjectival and adverbial information (e.g. indicate size of objects or aspectual properties of events) [8]. Non-manual markers are especially important on the level of sentence and beyond. Specifically, negation in many sign languages is expressed by head movements [28], and questions are distinguished from statements by eyebrow and head position almost universally [29].

Given the important role of non-manual markers, in this paper we test whether including non-manual features improves 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 addition of non-manual markers will improve recognition accuracy.

2. Related Work

Researchers dealing with monocular cameras consider manual and non-manual features separately. Manual features are features related to hands (e.g. hand configuration and motion trajectory of hands), while non-manual features are those features that do not involve hands and include facial expressions, lip patterns, head and body posture, gaze estimation. For example, the state-of-the-art performance was achieved by employing hybrid CNN-HMM approach where Language Model was used to maximize models in HMM [15]. They achieved a WER (Word Error Rate) in continuous sign language recognition of 30% for RWTH-PHOENIX-Weather 2012, 32.5% for RWTH-PHOENIX-Weather 2014 and 7.4% for SIGNUM. Cue et al. (2017) utilized Recurrent-CNN for spatio-temporal feature extraction and sequence learning. They applied their approach to a continuous sign language recognition benchmark, achieving a WER of 38.7% on RWTH-PHOENIX-Weather 2014 dataset [9]. Koller et al. (2018) [15] 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.

Non-manual features bring a significant meaning to sign language recognition as these parameters are essential for recognition of sign language, since they carry grammatical and prosodic information. Despite that facial features have been crucial for humans to grasp and understand sign language quite for a long time, the examination of their significance for automatic SLR was proved only in 2008 by Ulrich von Agris et al. [26].

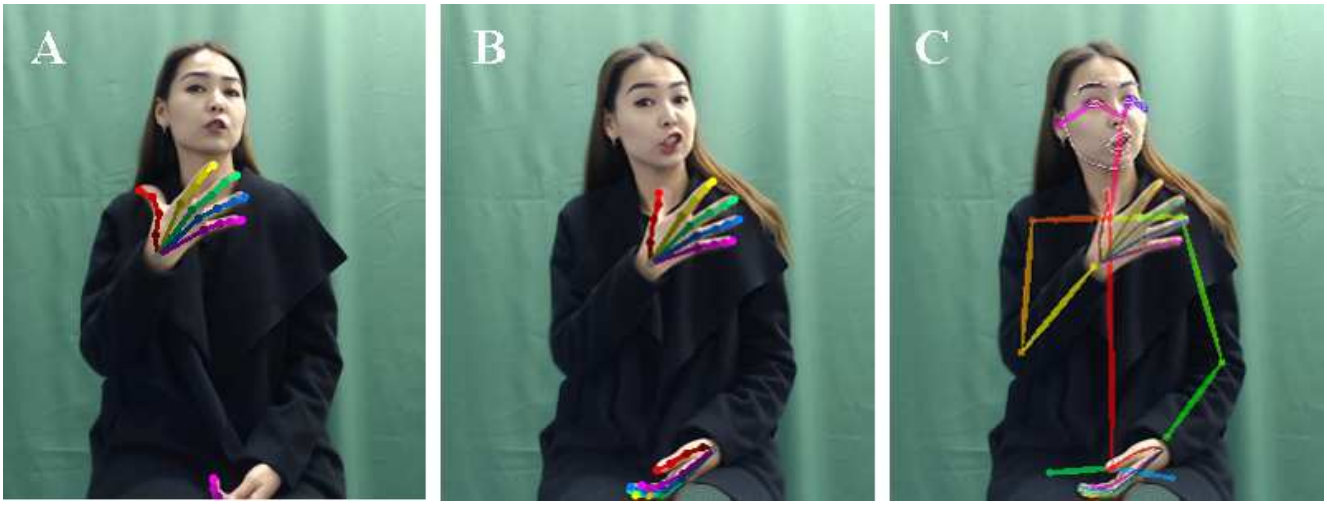

Figure 2. OpenPose resulting video screenshots: A) 'for what' statement, only manual features, B) 'for what' question, only manual features, C) 'for what' question, with manual and non-manual features

Antonakos et al. (2015) [2] 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 signers head pose. Antonakos et al. (2015) [2] conclude that limited number of works focused on employing non-manual features in SLR.

Freitas et al. (2017) [11] developed models for recognition of Grammatical Facial Expressions in Libras Sign Language. They have 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 empathetic the signer is.

Liu et al. [19] developed a system that automatically detected non-manual grammatical markers. They were able to increase 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. [18] attempted to recognize selected sign language gestures using only non-manual features. For this needs they developed 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 RWTH-BOSTON-50 dataset for experiments and their proposed model achieved 80% recognition rate.

In contrast, Yang and Lee [27] proposed a new method which 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 Active Appearance Model and then Support Vector Machines was applied for recognition of non-manual features. 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 the case when only manual features were used.

In contrast to the latest related work in SLR this paper aims to include more modalities than just the hands' features for real-time performance within human-robot interaction. The motivation behind this work lies in the need to differentiate between similar signs that only differ in non-manual components.

3. Methodology

3.1. Data collection

To explore the above stated hypotheses, we have collected a relatively small dataset of K-RSL similar to previously collected data [1, 14].

To explore current research questions, we recorded three professional sign language interpreters. Two of them are employed as news interpreters at the national television. Each signer can be considered as a native signer as they all have at least one deaf parent. They have been asked to sign 200 phrases, which contain 10 signs both used in statements and questions. Each phrase was repeated ten times in a row. 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 and composed twenty phrases with each word (ten statements and ten questions): 'what for', 'who', 'which, 'which-2', 'when', 'where (direction)', 'where (location)', 'why', 'how', and also 'how much'.We distinguish them to twenty classes (as ten words have a pair in both statement and question form). The reason for choosing these particular signs is that they carry different prosodic information as can be used in questions and statements. Also, they are similar but different in manual articulation. Figure 1 provides examples of 14 sign pairs from our data set.

3.2. OpenPose

We utilized OpenPose 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 [23]. It detects 2D information of 25 keypoints (joints) in a body and feet, 2x21 keypoints in both hands and 70 keypoints in a face. It also provides a 3D single-person keypoint detection in real time. OpenPose provides the values for each keyframe as an output in JSON format. Since dataset we use consists of RGB videos, we only consider 2D keypoints in this work. Figure 2 presents OpenPose resulting video screenshots with keypoints for manual and non-manual features.

The reason for choosing OpenPose instead of other image processing techniques for extracting manual and non-manual features is in its high accuracy and reliability, since we aim for the real-time performance in real-world condition to be used as a complete human-robot interaction system.

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 [12], and CSV (comma separated values) formats.

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. The classifier was trained on sequences of keyframes extracted from the OpenPose. The sequence of keyframes holds the frames of each sign video. Consequentially, one datapoint holds concatenated keypoints of each video and has a maximum of 30 frames * 274 keyframes = 8220 manual plus non-manual features for one datapoint and 30 frames * 84 keyframes = 2520 only manual features for each of 20 classes.

3.4. Human-Robot Interaction

We implemented a simple scenario where a humanoid robot Pepper would encourage people to repeat the signs displayed on the screen. As the user performed each sign in front of the camera, OpenPose extracted the features that were tested on corresponding Logistic Regression model to output the predicted sign in real-time for the robot to pronounce it with Aldebaran Robotics NaoQi's text-to-speech engine. To this end, we used Robotic Operating System (ROS) for integration between camera's feed and OpenPose engine. The system's hardware and software components are presented in Figure 3.

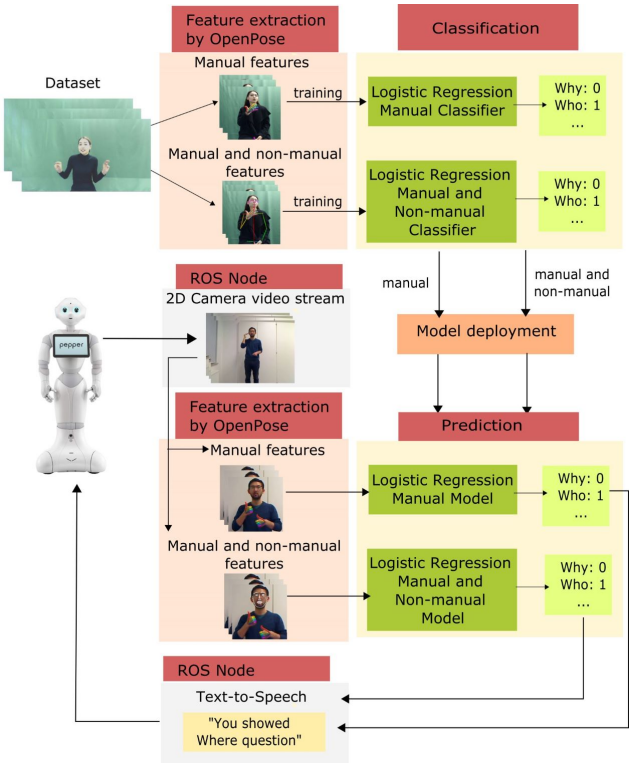

Figure 3. System's hardware and software components

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 singers (three 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.

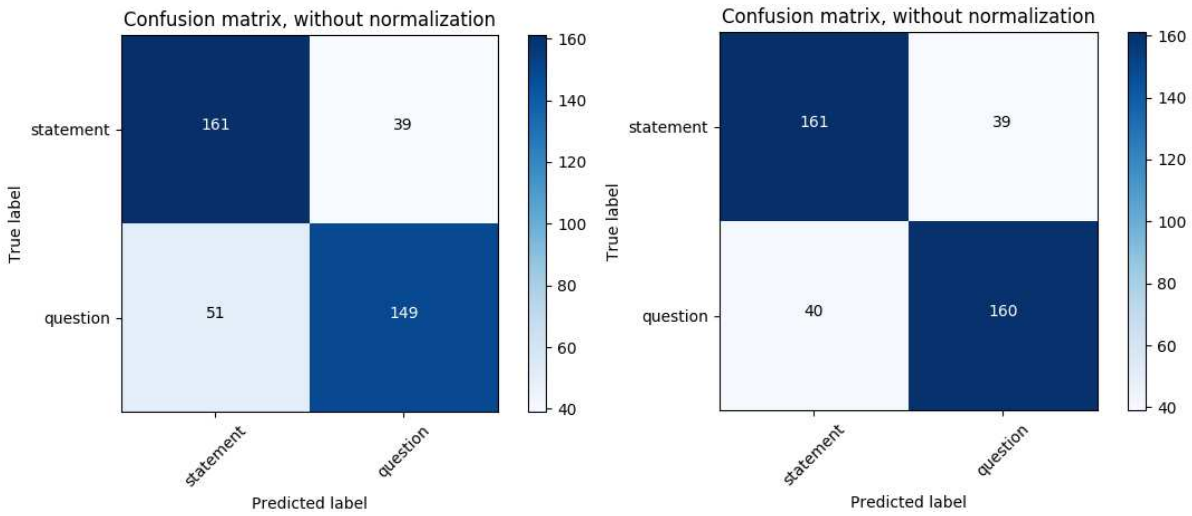

Figure 4. k-fold cross-validation experiment on statement-question dataset. Confusion matrix for 2 classes (statement vs question) with manual only features (left). Accuracy is 77.5%. Confusion matrix for 2 classes (statement vs question) with both manual and non-manual features (right). Accuracy is 80.25%.

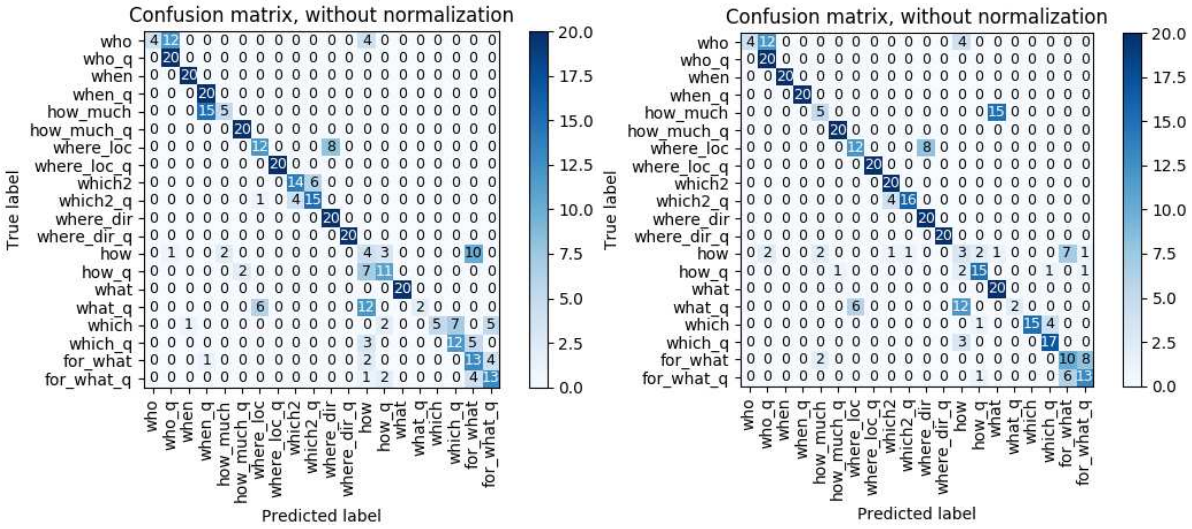

Figure 5. k-fold cross-validation experiment on distinct signs dataset. Confusion matrix for 20 signs with manual only features (left). Accuracy is 66.75%. Confusion matrix for 20 signs with both manual and non-manual features (right). Accuracy is 73%.

4.1. A case of two classes

In order 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, 1600 samples for training and 400 samples for testing). Choosing k equal to 5 (80 and 20 split), the training and validation were performed for each fold. Figures 4 (left) and 4 (right) show the confusion matrices of the obtained results for the first experiment. Mean averages are 81% and 86% for validation accuracy on manual-only and both manual and non-manual features respectively. Testing accuracy is 77.5% and 80.25% on manual-only and both manual and non-manual features respectively. Qualitative examination of the confusions in non-manual and manual confusion matrix (Figure 4 (right) shows that by adding non-manual features it was possible to correctly identify 11 samples as questions, that were classified as statements 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 statement vs questions.

4.2. A case of twenty classes

Figures 5 (left) and 5 (right) show the confusion matrices of the obtained results for the second experiment. Mean averages are 93.9% and 94.9% for validation accuracy on manual-only and both manual and non-manual fea-

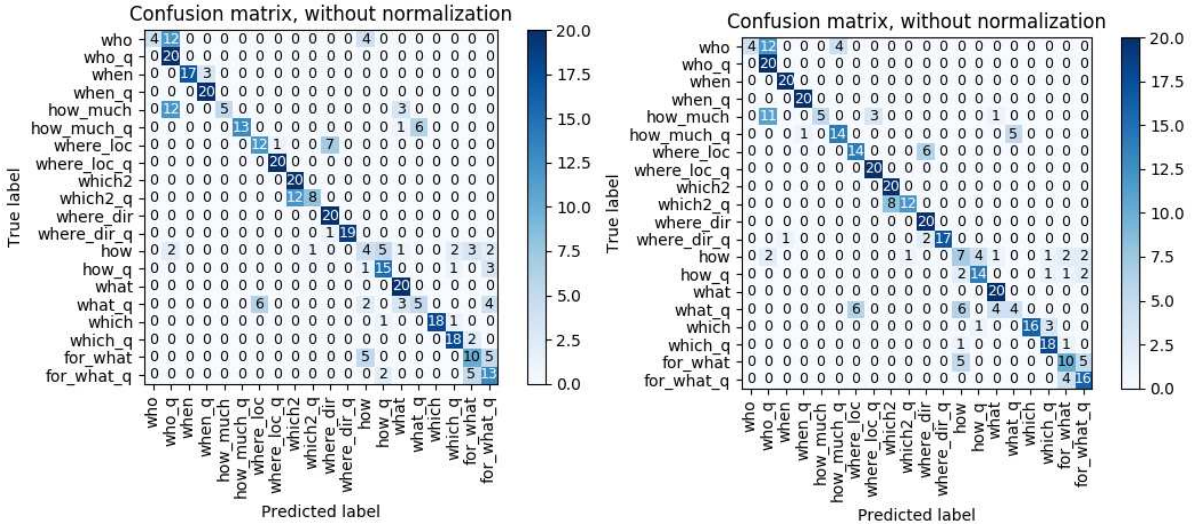

Figure 6. k-fold cross-validation experiment on different combinations of non-manual features. Confusion matrix for 20 signs with manual and non-manual (faceline, eyebrows, eyes, mouth) features (left). Accuracy is 73.75%. Confusion matrix for 20 signs with manual and non-manual (eyebrows, eyes, mouth) features (right). Accuracy is 72.75%.

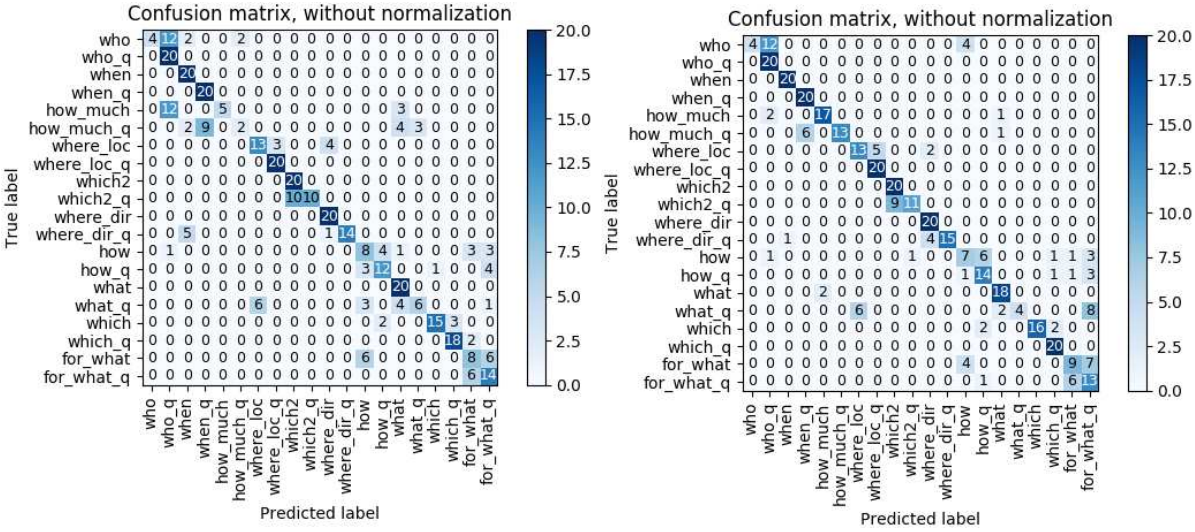

Figure 7. k-fold cross-validation experiment on different combinations of non-manual features. Confusion matrix for 20 signs with manual and non-manual (only eyebrows and eyes) features (left). Accuracy is 67.25%. Confusion matrix for 20 signs with manual and non-manual (only mouth) features (right). Accuracy is 73.5%.

tures respectively. Testing accuracies are 66.75% and 73% on manual-only and both manual and non-manual features respectively.

Qualitative examination of the top confusions in manual-only confusion matrix (Figure 5 (left) highlight confused pairs such as “Which2” (statement) and “Which2_Q” (question) with 27.5% confusion, “Which” (statement) and “Which_Q” (question) with 57.5% confusion, “How” (statement) and “How_Q” (question) with 75% confusion, “For what” (statement) and “For what_Q” (question) with 23% confusion. Since these signs share the same hand configurations and only the facial expression changes, it is expected that manual-only features caused such an error. And as expected, non-manual features improved recognition accuracy by 6% on average (from 66.75% accuracy to 73% accuracy) mainly between these signs (“Which2” pair had a decrease to 10% confusion, “Which” pair decreased to 20% confusion, “How” pair decreases its confusion to 55%).

4.3. A case of combining different modalities

Figures 6 and 7 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 67.25% 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 73.75% 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 73%. Thus, we see that mouthing provides extra information, which can be used in recognition, because signers usually articulate signs while performing it. Eyebrows and head position provide additional grammatical markers to differentiate statements from questions.

5. Conclusion and Future Work

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. [11], Yang and Lee [27] we saw an improvement in 6% for the k-folded performance. Table 1 compares our results obtained from the experiments:

Table 1. Comparison of results

Method	5-fold	80/20 split
Manual only	93.9%	66.75%
Manual & Non-manual full	94.9%	73%
Manual & Face, eyebrows, mouth	88%	**73.75%**
Manual & Eyebrows, mouth	89%	72.75%
Manual & Only mouth	88%	73.5%
Manual & Only eyebrows	88%	67.25%

This 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 into the perception system. In addition, we deployed an architecture of automatic SLR onto a humanoid robot in order to conduct a real-world signer independent experiment in real time. Future work will involve expanding the K-RSL dataset and conducting a real-world experiment with native signers and a robot.

References

[1] Aitpayev K. (2015). Kazakh Sign Language Corpus. Retrieved from https://kslc.kz/.

[2] Antonakos E., Roussos A., Zafeiriou S., A survey on mouth modeling and analysis for Sign Language recognition, Automatic Face and Gesture Recognition (FG) 2015 11th IEEE International Conference and Workshops on, vol. 1, pp. 1-7, 2015.

[3] Aran, O.: Vision based sign language recognition: modeling and recognizing isolated signs with manual and non-manual components. PhD dissertation, Bogazii University, 2008.

[4] British Sign Language Corpus Project (2010). Retrieved from www.bslcorpusproject.org.

[5] Burkova, S., I., Russian Sign Language Corpus Project (2014). Retrieved from http://rsl.nstu.ru/site/project.

[6] Chai X., Wang H., and Chen X., The DEVISIGN Large Vocabulary of Chinese Sign Language Database and Baseline Evaluations. Technical report, Key Lab of Intelligent Information Processing of Chinese Academy of Sciences, 2014.00000.

[7] Cooper, H., Holt, B., & Bowden, R. (2011). Sign language recognition. In Visual Analysis of Humans (pp. 539-562). Springer, London.

[8] Crasborn, O., van der Kooij, E., Waters, D., Woll, B., & Mesch, J. (2008). Frequency distribution and spreading behavior of different types of mouth actions in three sign languages. Sign Language & Linguistics, 11(1), 4567. https://doi.org/10.1075/sll.11.1.04cra

[9] Cui, R., Liu, H., & 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 (pp. 7361-7369).

[10] Forster J., Schmidt C., Koller O., Bellgardt M., and Ney H., Extensions of the Sign Language Recognition and Translation Corpus RWTH-PHOENIX-Weather. In International Conference on Language Resources and Evaluation, pages 19111916, Reykjavik, Island, May 2014.

[11] Freitas, F.A., Peres, S.M., Lima, C.A.M. et al., Grammatical facial expression recognition in sign language discourse: a study at the syntax level, Information Systems Frontiers, (2017) 19:1243. https://doi.org/10.1007/s10796-017-9765-z

[12] Holmes, G., Donkin, A., & Witten, I. H. (1994). Weka: A machine learning workbench.

[13] Hong, S., Setiawan, N., Lee, C.: Real-time vision based gesture recognition for human-robot interaction. In: Procs. of Int. Conf. on Knowledge-Based and Intelligent Information & Engineering Systems : Italian Workshp : Neural Networks, LNCS, vol. 4692, p. 493. Springer, Vietri sul Mare, Italy (2007)

[14] Imashev, A., Sign Language Static Gestures Recognition Tool Prototype. Application of Information and Communication Technologies (AICT), 2017, 11th IEEE International Conference on. IEEE, 2017. pp. 1-4.

[15] Koller, O., Zargaran, S., Ney, H., & 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.

[16] Ko, S. K., Kim, C. J., Jung, H., Cho, C. (2018). Neural Sign Language Translation based on Human Keypoint Estimation. arXiv preprint arXiv:1811.11436.

[17] Lang, S., Block, M., & Rojas, R.: Sign language recognition using kinect. In Artificial Intelligence and Soft Computing. Springer Berlin Heidelberg, pp. 394-402 (2012)

[18] Kumar, S., Bhuyan, M.K., Chakraborty, B.K., Extraction of texture and geometrical features from informative facial regions for sign language recognition, Journal on Multimodal User Interfaces, (2017) 11: 227. https://doi.org/10.1007/s12193-017-0241-3

[19] Liu, J., Liu, B., Zhang, S., Yang, F., Yang, P., Metaxas, D.N., Neidle, C., Non-manual grammatical marker recognition based on multi-scale, spatio-temporal analysis of head pose and facial expressions, Image and Vision Computing, Volume 32, Issue 10, 2014, Pages 671-681, ISSN 0262-8856, https://doi.org/10.1016/j.imavis.2014.02.009.

[20] Pfau, R. & Quer, J. (2008). Nonmanuals: their prosodic and grammatical roles. In D. Brentari (Ed.), Sign Languages (pp. 381402). Cambridge: Cambridge University Press.

[21] Sahoo, A. K., Mishra, G.S., Ravulakollu K.K.,: Sign Language Recognition: State of the Art. ARPN Journal of Engineering and Applied Sciences 9, no. 2 (2014): pp. 116-134.

[22] Sandler, W., & Lillo-Martin, D. C. (2006). Sign language and linguistic universals. Cambridge: Cambridge University Press.

[23] Simon, T., Joo, H., Matthews, I. A., & Sheikh, Y. (2017, July). Hand Keypoint Detection in Single Images Using Multiview Bootstrapping. In CVPR (Vol. 1, p. 2).

[24] Starner, T., Pentland, A.: Real-time american sign language recognition from video using hidden markov models. In: Procs. of Int. Symposium on Computer Vision, pp. 265 270 (1995). DOI 10.1109/ISCV.1995.477012

[25] Tazhigaliyeva, N., Kalidolda, N., Imashev, A., Islam, S., Aitpayev, K., Parisi, G. I., Sandygulova, A. (2017, May). Cyrillic manual alphabet recognition in RGB and RGB-D data for sign language interpreting robotic system (SLIRS). In 2017 IEEE International Conference on Robotics and Automation (ICRA) (pp. 4531-4536). IEEE.

[26] Von Agris, U., Knorr, M., Kraiss, K. F.: The significance of facial features for automatic sign language recognition //Automatic Face & Gesture Recognition, 2008.FG'08.8th IEEE International Conference on. IEEE, 2008. . 1-6.

[27] Yang, H.D., Lee, S.W., Robust sign language recognition by combining manual and non-manual features based on conditional random field and support vector machine, Pattern Recognition Letters, Volume 34, Issue 16, 2013, Pages 2051-2056, ISSN 0167-8655, https://doi.org/10.1016/j.patrec.2013.06.022.

[28] Zeshan, U. (2004a). Hand, head, and face: Negative constructions in sign languages. Linguistic Typology, 8(1), 158. https://doi.org/10.1515/lity.2004.003

[29] Zeshan, U. (2004b). Interrogative Constructions in Signed Languages: Crosslinguistic Perspectives. Language, 80(1), 739.

[30] Zieren, J., Kraiss, K.: Non-intrusive sign language recognition for human computer interaction. In: Procs. of IFAC/IFIP/IFORS/IEA symposium on analysis, design and evaluation of human machine systems (2004).