AUTOMATIC LIP-READING OF HEARING IMPAIRED PEOPLE

D. Ivanko, D. Ryumin, A. Karpov

St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences, SPIIRAS, Saint-Petersburg, Russian Federation – denis.ivanko11@gmail.com, dl_03.03.1991@mail.ru, karpov@iias.spb.su

Commission II WG II/5

KEY WORDS: Lip-reading, hearing impaired people, region-of-interest detection, visual speech recognition

ABSTRACT:

Inability to use speech interfaces greatly limits the deaf and hearing impaired people in the possibility of human-machine interaction. To solve this problem and to increase the accuracy and reliability of the automatic Russian sign language recognition system it is proposed to use lip-reading in addition to hand gestures recognition. Deaf and hearing impaired people use sign language as the main way of communication in everyday life. Sign language is a structured form of hand gestures and lips movements involving visual motions and signs, which is used as a communication system. Since sign language includes not only hand gestures, but also lip movements that mimic vocalized pronunciation, it is of interest to investigate how accurately such a visual speech can be recognized by a lip-reading system, especially considering the fact that the visual speech of hearing impaired people is often characterized with hyper-articulation, which should potentially facilitate its recognition. For this purpose, thesaurus of Russian sign language (TheRusLan) collected in SPIIRAS in 2018-19 was used. The database consists of color optical FullHD video recordings of 13 native Russian sign language signers (11 females and 2 males) from "Pavlovsk boarding school for the hearing impaired". Each of the signers demonstrated 164 phrases for 5 times. This work covers the initial stages of this research, including data collection, data labeling, region-of-interest detection and methods for informative features extraction. The results of this study can later be used to create assistive technologies for deaf or hearing impaired people.

1. INTRODUCTION

Automatic lip-reading, also known as visual speech recognition (VSR), has received a lot of attention in the recent years for its potential use in applications of human-machine interaction, sign language recognition, audio-visual speech recognition, biometry and biomedicine (Katsaggelos et al., 2015). Most people can understand a few unspoken words by lip-reading, and many hearing impaired individuals are quite proficient at this skill. The idea of interpreting silent speech has been around for long time. Automatic visual lip-reading was initially proposed as an enhancement to speech recognition in noisy environments (Petajan, 1984). VSR are commonly used to improve the reliability and robustness of audio speech recognition systems. However, there is a group of people for which the use of audio speech is not possible. Lip-reading plays an important role in communication by the hearing-impaired.

According to statistics reported in (Ryumin et al., 2019) over 5% of the world population – or 466 million people – has disabling hearing loss (432 million adults and 34 million children). It is estimated that by 2050 over 900 million people – or one in every ten people – will have disabling hearing loss. For the deaf and speech-impaired community, sign language (SL) serves as useful tools for daily interaction (Denby et al., 2010). Official data report about 120000 people in Russian Federation (and about 100000 people in other countries) using Russian sign language as their main way of communication.

Sign language is a structured form of hand gestures involving visual motions and signs, which is used as a communication system. SL recognition includes the whole process of tracking and identifying the signs performed and converting into semantically meaningful words and expressions (Cheok et al., 2017). Majority of sign language involves only upper part of the body from waist level upwards. Besides, the same sign can have considerably large changes in shapes when it is in different location in the sentence (Yang et al., 2010). Sign Language involves the use of different parts of the body – not only hands and fingers movements. Deaf people use lips movements as a part of a sign language in general, even though they cannot hear acoustic speech. Moreover, the lips movements of hearing impaired people are usually characterized with hyper-articulation. This fact, potentially, enables the opportunity to better recognize the visual speech of such group of people, since lip movements become more pronounced.

It is well known that hearing-impaired people and those listening in noisy acoustic conditions (noise, reverberation, multiple speakers) rely heavily on the visual input to eliminate ambiguity in acoustic speech elements. Although a significant amount of research has been devoted to the topic of visual speech decoding, the problem of lip reading for hearing impaired people remains an open issue in the field (Akbari et al., 2018).

From the literature review, the most common sign languages recognition researches are based on American Sign Language, Indian Sign Language and Arabic Sign Language. Some early efforts on sign language recognition can be dated back to 1993, where gesture recognition techniques are adapted from speech and handwriting recognition techniques (Darrell et al., 1993). A comprehensive study of the development of the field of sign language recognition is given in the work (Cheok et al., 2017).

The focus of this research is on improving the accuracy and robustness of automatic Russian sign language recognition via adding lip-reading module to hand gesture recognition system. This paper covers the initial stages of this research, including data collection, data labeling, region-of-interest detection and methods for informative features extraction.

2. DATA AND TOOLS

The corpus (database) is mandatory for learning any modern speech recognition system based on probabilistic models and machine learning techniques. There are already a number of large commercial or free databases for learning acoustic speech recognition systems for many of the world languages. However, one major obstacle to the current research on visual speech recognition is the lack of suitable databases. In contrast to the richness of audio speech corpora, only few databases are publicly available for visual-only or audio-visual speech recognition systems (Zhou et al., 2014). Most of them include a limited number of speakers and a small vocabulary. Databases on the Russian sign language, suitable for VSR training, are practically non-existent. For this reason, in 2018, a Thesaurus of Russian sign language (TheRuSLan) (Ryumin et al. 2019) was recorded in SPIIRAS.

2.1 TheRuSLan

TheRusLan comprises recordings of 13 (11 females and 2 males) native signers of Russian sign language from “Pavlovsk boarding school for the hearing impaired” or “Deaf-Mute school” (the oldest institution for the hearing impaired in Russia). Each signer demonstrated 164 phrases for 5 times. Total number of samples in the corpora is 10660. Screenshots of the signers in the course of recording sessions are shown in Fig. 1. The corpus consists of color optical FullHD (1920×1080) video files, infrared video files (512×424) and depth video files (512×424), it also includes feature files (with 25 skeletal reference dots on each frame), text files of temporal annotation into phrases, words and gesture classes (was made manually by an expert). A brief summary of the contents of the database is presented in Table 1. All the recordings were organized into a logically structured database that comprises a file with information about all the speakers and recording parameters.

Parameter	Value
Number of signers	13
Phrases per signer	164
Number of repetitions	5
Number of samples per signer	820
Total number of samples	10 660
Recording device	MS Kinect V2
Resolution (color)	FullHD (1920×1080)
Resolution (depth sensor)	512×424
Resolution (infrared camera)	512×424
Distance to camera	1.0-2 m.
Duration (total)	7 hours 56 min.
Duration (per signer)	~36 min
Number of frames with gestures (per signer)	~30-35 000
Average age of signers	24
Total amount of data	~4 TB

Table 1. Contents of TheRuSLan database

2.2 Data labelling

Proper temporal labeling of data into viseme classes is necessary for training any practical lip-reading system in the scope of statistical methods of speech recognition. As well as a phoneme for acoustic speech, a viseme is the minimal distinguishable unit of visual speech. The number of visemes is language-dependent and for Russian there are 14-20 distinguishable speech units in different works. In the current study we used a list of 20 viseme classes (table 2), proposed in (Ivanko et. al, 2018a).

Figure 1. Example of signers during a recording session

Viseme Class	Corresponding phonemes	Viseme Class	Corresponding phonemes
V1	sil (pause)	V11	b, b’, p, p’
V2	a, a!	V12	f, f’, v, v’
V3	i, i!	V13	s, s’,z, z’,c
V4	o!	V14	sch
V5	e!, e	V15	sh
V6	y, y!	V16	j
V7	u, u!	V17	h, h’
V8	l, l’, r, r’	V18	ch
V9	d, d’,t, t’,n, n’	V19	m, m’
V10	g, g’, k, k’	V20	zh

Table 2. Viseme classes and phoneme-to-viseme mapping

State-of-the-art semi-automatic approach to viseme labeling of multimedia databases is performed using speech recognition system with subsequent expert correction of temporal annotation. In the present study, the use of this method is not possible due to the complete absence of acoustic data. Therefore, the entire temporal labeling of the database on the viseme classes was done manually by experts. The following is an example of the labeling (format: start time (sec.) - end time (sec.) – phoneme class mark – viseme class mark) of the database:

0.00	0.31	SIL	V1
0.31	0.37	m	V19
0.37	0.46	a	V2
0.46	0.53	l	V8
0.53	0.62	a	V2
0.62	0.70	k	V10
0.70	0.89	o!	V4
0.89	1.04	SIL	V1

3. PROPOSED METHOD

In general, visual speech recognition is the process of converting sequences of mouth region images into text. Every practical lip-reading system necessarily include 4 main processing stages, such as image acquisition, region of interest (ROI) localization, feature extraction and speech recognition. More detailed description of the used VSR system is described in the work (Ivanko et al., 2017).

In research works, the image acquisition stage is often replaced by the step of selecting a representative database, which is necessary for system training in the framework of the statistical approach to speech recognition. The next important step in visual speech recognition is finding the ROI on the image, since the quality of ROI has a strong influence on the recognition results. Two main approaches to solve this task are often found in the literature: Haar-like feature based boosted classification framework and the active appearance model. In this work we used both methods: Haar-like method for pixel-based features extraction and AAM for geometry-based features.

Visual data are typically stored as video sequences and contain a lot of information that is irrelevant to the pronounced speech. Such information includes the variability of the visual appearances of different speakers. The visual features should be relatively small and sufficiently informative about the uttered speech, at the same time they must show a certain level of invariance against redundant information or noise in videos. Despite many years of research, there is no any visual feature set universally accepted for representing visual speech. The most widely used features are pixel-based (image-based) (Hong et al., 2006) and geometry-based (Kumar et al., 2017). In this paper, we use the modification to the method recently proposed in (Ivanko et al., 2018b) and apply it for the task of automatic visual speech recognition of hearing impaired people.

3.1 Region-of-interest detection

TheRuSLan database contains video recordings of signers in full growth. On each frame of video data, the search for 25 key points of the human skeleton model was carried out using Kinect 2.0 in-build AAM based algorithm (Kar, 2010). An example of a detected skeleton model is shown in Fig. 2.

Based on the obtained skeleton coordinates, we searched for the ROI (face and mouth region) in the area above the neck landmark. Thus, we significantly reduced the number of false alarms of the face detection algorithm on various objects of the environment.

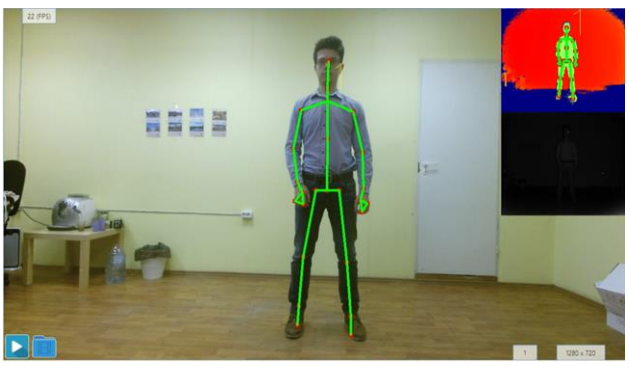

Figure 2. 25 key landmarks of Kinect 2.0 human skeleton model

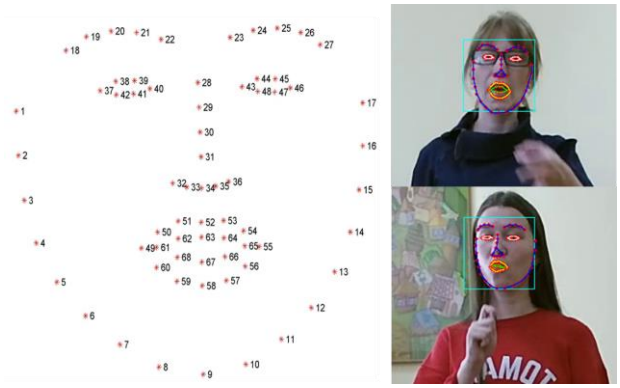

Figure 3. Full 2D face shape model (left) and the face landmarks localization algorithm results (right).

To extract useful visual speech information, the first step is to locate the lips region that contains the most valuable information about pronounced speech. For face and lips regions detection we apply AAM-based algorithm implemented in the Dlib open source computer vision library (King, 2009).The main idea is to match the statistical model of object shape and appearance, containing a set of 68 facial landmarks (20 in the mouth region), to a new image. An example of used face shape model and the algorithm results depicted in Fig. 3.

This face search algorithm shows high reliability in finding the faces in the given image. One of the major drawbacks of the method is false detection, i.e. detection of faces on the furnishing, environment, clothing, etc. By reducing the search area (above the neck landmark), we managed to eliminate the error of false detection in TheRuSLan database.

3.2 Parametric representation

In the current study we used two different state-of-the-art approaches for parametric representation of visual speech: geometry-based and pixel-based.

Geometry-based features extraction was done using method similar to (Ivanko et. al, 2018c). The main idea of the method is after the normalization of the coordinates of 20 founded in the lips region landmarks, we calculate the Euclidean distances between certain landmarks (according to table 3) and save them in the feature vector. Landmarks numbers in the table correspond to the map (Howell et al., 2016), depicted in Fig. 3.

#	landmarks (№)	#	landmarks (№)
1	49 - 61	13	54 - 64
2	61 - 60	14	64 - 53
3	60 - 68	15	64 - 52
4	68 - 59	16	52 - 63
5	59 - 67	17	52 - 62
6	67 - 58	18	62 - 51
7	67 - 57	19	62 - 50
8	57 - 66	20	50 - 61
9	66 - 56	21	62 - 68
10	56 - 65	22	63 - 67
11	65 - 55	23	64 - 66
12	65 - 54	24	61 - 65

Table 3. Feature extraction lanmarks

After ROI detection the pixel-based visual features are calculated as a result of the following processing steps: greyscale conversion, histogram alignment, normalization of the detected mouth images to 32×32 pixels, mapping to a 32-dimentional feature vector using a principal component analysis (Fig. 5).

3.3 Challenges

Sign language involves a large number of hand movements and often they cover the lips region. The structure of the Russian sign language is such that the interaction of the hand with the mouth is necessary for a certain number of gestures. A similar example can be seen in Figure 4. This overlap effect is a serious problem for an automatic sign language recognition system.

In such circumstances, the algorithm for detecting lips region works with an error and either narrows the mouth area to its visible part (Fig. 4, top left), or tries to model the invisible part (Fig. 4, top right). In both cases, it is difficult to achieve optimal recognition accuracy and it is obvious that such examples should be excluded from the training set.

In this paper, we propose to use a simple method for detecting such cases. Using the coordinates of skeletal model obtained by Kinect in-build algorithm, on each frame we check how close the area of the hand to the face region. Those video frames on which the hand area intersects with the face area are automatically marked and excluded from the training set. Thus, we managed to significantly reduce the amount of irrelevant data.

3.4 Description of the parametric representation method

A general pipeline of the lips parametric representation method is presented in Figure 5. Proposed method involves sequential execution of the following steps:

1. Acquisition a video frame from the Kinect 2.0 device or video file.
2. Calculate coordinates of 25 landmarks of the human skeleton model.
3. Crop the search area above the neck landmark and use the facial landmark detection algorithm to find 68 facial key points (20 in the mouth region).
4. Check the intersection of the face and hand regions.

Figure 4. An examples of overlap effect in Russian sign language

5. Normalize the coordinates of the obtained landmarks.
6. Calculate a number of Euclidean distances between landmarks for geometry-based features extraction.
7. Calculate PCA-based visual features.
8. Save 24-dimentional geometry-based feature vector and 32-dimentional pixel-based feature vector.

The proposed method allows determining speaker's lip region in the video stream during articulation and it represents the region of interest as a feature vector with the dimensionality of 56 components (24 for geometry-based features and 32 for pixel-based features). The method also allows automatically detect the intersection between hand and face regions, and exclude inappropriate data from the training set. This parametrical representation can be used for the training of an automatic system of Russian sign language recognition and it is planned for further studies to augment the performance of existing baseline.

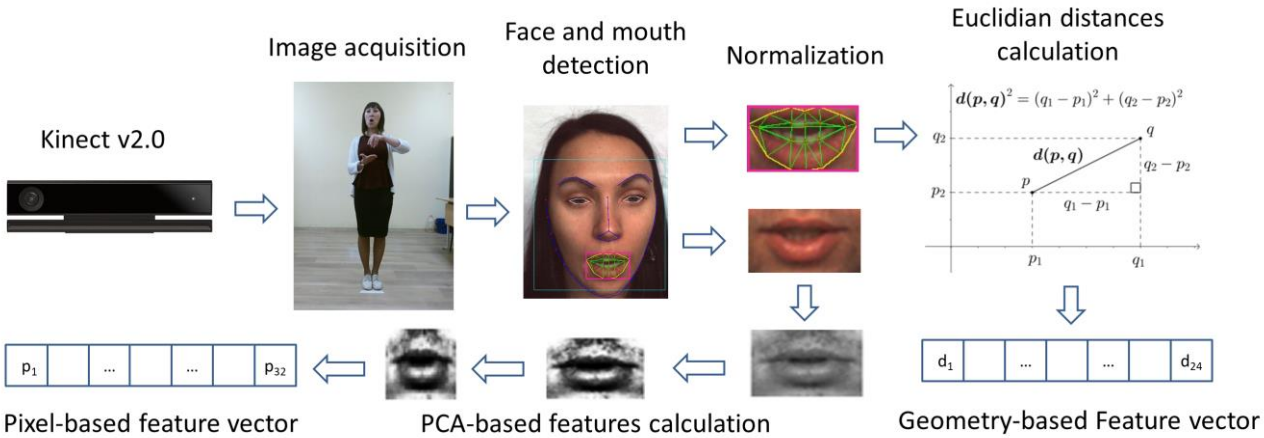

Figure 5. General pipeline of the parametric representation method

4. CONCLUSIONS AND FUTURE WORK

The paper discusses the possibility of use additional modality (lip-reading) to improve the accuracy and robustness of automatic Russian sign language recognition system. Such studies have not been conducted before and it is of great practical interest to investigate how good visual speech of hearing impaired people can be recognized. The present study is the first step towards this direction and covers following stages: data collection, data labeling, region-of-interest detection and methods for informative features extraction.

We also proposed method for parametric representation of visual speech with adaptation to overlapping of hand and mouth regions. The feature vectors obtained using this method will later be used to train the lip-reading system. Due to its versatility, this method can also be used for different tasks of biometrics, computer vision, etc.

In further research, we are going to use statistical approaches, e.g. based on some types of Hidden Markov Models or deep neural networks for development of a robust and accurate Russian sign language recognition system. The results of this study can later be used to create assistive technologies for deaf or hearing impaired people.

ACKNOWLEDGEMENTS

This research is financially supported by the Ministry of Science and Higher Education of the Russian Federation, agreement No. 14.616.21.0095 (reference RFMEFI61618X0095).

REFERENCES

Akbari et al., 2018. H. Akbari, H. Arora, L. Cao, N. Mesgarani, LIP2AUD-SPEC: Speech reconstruction from silent lip movements video. *Proceedings of ICASSP2018*, pp. 2516-2520.

Cheok et al., 2017. M. J. Cheok, Z. Omar, M. H. Jaward, A review of hand gesture and sign language recognition techniques. *International Journal of Machine Learning and Cybernetics*, vol. 10, no. 1, pp. 131-153.

Darrell et al., 1993. T. Darrell, A. Pentland, Space-time gestures, *Proceedings on Computer Vision and Pattern Recognition, IEEE computer society conference*, pp. 335-340.

Denby et al., 2010. B. Denby, T. Schultz, K. Honda, T. Hueber, J. M. Gilbert, J. S. Brumberg, Silent speech interfaces, *Journal on Speech Communication*, vol. 52, pp. 270-287.

Hong et al., 2006. S. Hong, H. Yao, Y. Wan, R. Chen, A PCA based visual DCT feature extraction method for lip-reading. In: *Proceedings of the intelligent informatics, hiding multimedia and signal processing*, pp 321-326.

Howell et al., 2016. D. Howell, S. Cox, B. Theobald, Visual units and confusion modelling for automatic lip-reading. In: *Image and Vision Computing*, vol. 51, pp. 1-12.

Ivanko et al., 2018a. D. Ivanko, A. Karpov, D. Fedotov, I. Kipyatkova, D. Ryumin, Dm. Ivanko, W. Minker, M. Zelezny, Multimodal speech recognition: increasing accuracy using high speed video data. In: *Journal of Multimodal User Interfaces*, vol. 12, no. 4, pp. 319-328.

Ivanko et al., 2017. D. Ivanko, A. Karpov, D. Ryumin, I. Kipyatkova, A. Saveliev, V. Budkov, M. Zelezny, Using a high-speed video Camera for robust audio–visual speech recognition in acoustically noisy conditions. In: *SPECOM 2017, LNAI 10458*, pp 757-766.

Ivanko et al., 2018b. D. Ivanko, D. Ryumin, I. Kipyatkova, A. Karpov, Lip-Reading Using Pixel-based and Geometry-based Features for Multimodal Human-robot Interfaces. In: *Zavalishin Readings 2019, in press*.

Ivanko et al., 2018c. D. Ivanko, D. Ryumin, A. Axyonov, M. Zelezny, Designing advanced geometric features for automatic Russian visual speech recognition. In: *Proceedings of the 20th International Conference on Speech and Computer (SPECOM 2018)*, pp. 245-255.

Kar, 2010. A. Kar, Skeletal tracking using Microsoft Kinect, In: *Methodology*, vol. 1, pp. 1-11.

Katsaggelos et al., 2015. K. Katsaggelos, S. Bahaadini, R. Molina, Audiovisual fusion: challenges and new approaches. In: *Proceedings of the IEEE*, vol. 103, no. 9, pp 1635-1653.

King, 2009. D. E. King, Dlib-ml: A Machine Learning Toolkit, *Journal of Machine Learning Research*, vol. 10, pp. 1755-1758

Kumar et al., 2017. S. Kumar, MK. Bhuyan, B. Chakraborty, Extraction of texture and geometrical features from informative facial regions for sign language recognition. In: *Journal of Multimodal User Interfaces (JMUI)*, vol. 11, no. 2, pp. 227-239.

Petajan, 1984. E. D. Petajan, Automatic lipreading to enhance speech recognition. In: *IEEE Communications Society Global Telecommunications Conference*, Atlanta, USA.

Ryumin et al., 2019. D. Ryumin, D. Ivanko, A. Axyonov, A. Karpov, I. Kagirov, M. Zelezny, Human-robot interaction with smart shopping trolley using sign language: data collection. In: *IEEE the 1st International Workshop on Pervasive Computing and Spoken Dialogue Systems Technology (PerDial 2019)*, in press.

Yang et al., 2010. R. Yang, S. Sarkar, B. Loeding, Handling movement epenthesis and hand segmentation ambiguities in continuous sign language recognition using nested dynamic programming, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no.3, pp. 462-477.

Zhou et al., 2014. Z. Zhou, G. Zhao, X. Hong, M. Pietikainen, A review of recent advances in visual speech decoding. *Image and Vision Computing*, vol. 32, 590-605.