Proceeding of the IEEE
International Conference on Robotics and Biomimetics
Dali, China, December 2019

A Bi-directional Interactive System of Sign Language and Visual Speech Based on Portable Devices

Fei Wang
Faculty of Robot Science and Engineering
Northeastern University
Shenyang, 110819, China
wangfei@mail.neu.edu.cn

Shizhuo Sun, Yanjun Liu
{*School of Computer Science and Engineering,*
Faulty of Robot Science and Engineering}
Northeastern University
Shenyang, 110819, China
{*20174518,20174847*}*@stu.neu.edu.cn*

***Abstract*— At present, the natural communication between deaf and normal people is a major challenge both in theoretical research and application. In addition, in the field of Human-Machine Interaction, there is little work on bi-directional communication based on sign language and visual speech. In this paper, a portable bi-directional interactive system of sign language and visual speech was proposed to help deaf people communicate naturally under certain circumstances. In the section of modeling, Network in Network model was employed to classify sign language words, while the method using a network combined DenseNet and LSTM together was used for lip reading. By wearing the portable device, people can achieve bi-directional translation and communication of visual speech and sign language. We built our own Chinese sign language database containing more than 100 categories of words and conducted experiments in the context of the airport for experimental verification. Eventually, the average time of each round of conversation is within 15 seconds, faster than normal communication without system assistance, which verified the effectiveness and advantage of the purposed system.**

***Index Terms*— Chinese Sign Language Recognition, Visual Speech Recognition, Network in Network, DenseNet, LSTM**

I. Introduction

Although there are many hearing-aid devices currently available for people with hearing disabilities, they still find it difficult to have barrier-free conversations with people with normal hearing because neither normal people can understand sign language nor deaf people are skilled enough to read lips. Thus, a system that can overcome both problems and provide better bi-directional communication aid is of great necessity. However, there is little work on the crossover of Sign Language Recognition (SLR) and Visual Speech Recognition (VSR). In addition, traditional communication-aid machines are fixed and not convenient enough for the deaf and other disabled people to use in a variety of occasions. Hence, this paper creatively combines both technologies to develop a portable bi-directional communication system to solve the daily communication problems between deaf and normal people.

In our experiment, portable wearable devices equipped with IMU sensors and sEMG sensors were used for data acquisition to build our own sign language database, and a modified Network in Network was employed to conduct word-level sign language recognition. Also, global average pooling was added to convolutional layers to prevent samples from over fitting. Weighted Logarithmic Loss was introduced to replace Cross Entropy Loss as the cost function of the model, which was used to balance the recognition rate among different sign language words and prevent the low accuracy encountered when either the size or the level of recognition difficulty of training samples differs. In terms of VSR, we used a RGB camera for video capture and a convolutional recurrent neural network which combines biLSTM and DenseNet to process visual information of lips and translate it into structured sentences.

There are three main points in our contribution: (1) Network in Network is added to convolutional layers to prevent over-fitting and the weighting in the loss function was employed to prevent a lower recognition rate which can be caused by various sizes and difficulties in the samples in SLR; (2) A bi-directional communication system combining SLR and VSR is proposed in this paper, which contributes to better interaction between both deaf people and normal people; (3) The system can be utilized without the limitation of time and space.

This paper is organized into five sections: Section I introduces the background of the problem mentioned in this paper. Section II discusses about the works and researches related to SLR and VSR. Section III presents the system configuration and the model architecture in detail. In Section IV, we described details about the setup of the database as well as experiments and analyzed the final results. Eventually, we came to our conclusion in Section V.

II. Related Work

Many researches have been taken on Chinese sign language recognition and most existing supervised methods

∗This work was supported in part by the Fundamental Research Funds for the Central Universities of China under Grant N172608005, N182612002, N182410001, Liaoning Provincial Natural Science Foundation of China under Grant 20180520007, National Training Program of Innovation and Entrepreneurship for Undergraduates under Grant 201910145249.

978-1-7281-6321-5/19//$31.00 ©2019 IEEE

focus on recognizing sign language using the Data Glove as well as Kinect [1] [2]. In lip reading, some recent works attempt to use deep neural networks to train language model end to end [3].

A. Sign Language Recognition

Currently, the data collection methods of SLR can be roughly divided into three categories:(1) using data gloves;(2) using cameras;(3) using wearable sensors.

The Data Glove integrates sensors that accurately detect the curvature of fingers. Because of its responsiveness and accuracy, it has been used in sign language recognition for some time [4]. Data Gloves have a variety of sensors which can detect changes in physical signals of hands such as angle, displacement, velocity and acceleration [5]. But the Data Glove has obvious shortcomings. The Data Glove is a data acquisition instrument with sophisticated internal structure, so it is neither easy to maintain nor convenient for the users to wear and move around. Due to these problems, SLR systems based on Data Gloves are difficult to become widespread.

The study on SLR based on Computer Vision mainly focuses on using cameras to capture sign videos and segmenting the hand areas to obtain image sequences of hand posture, thereby to understand sign language by analyzing their changes. Early visual-based SLR were achieved by using ordinary RGB or gray-scale images. Yamato et al. [6] extracted the hand contour from the sign video and segmented a ROI of 25 pixels $\times$ 25 pixels as the basic unit to recognize six sign language words.

In recent years, there are also many researches focusing on utilizing signals from surface EMG (sEMG) sensors in order to better model hand movement. M. E. Celebi et al. [7] used Hidden Markov Model (HMM) to recognize the sEMG signals generated by the movement of hands and the accuracy in identifying sign language words reached 94.6%. In addition to the application of sEMG signals, some researchers have also introduced accelerometer data into SLR research.

B. Lip Reading

At present, the existing method of lip reading mainly relies on Computer Vision. Video information can be captured with general cameras or depth cameras such as Microsoft Kinect [8] [9]. Inspired by the great success in SLR, machine learning, especially deep learning, methods has become widely used in lip reading.

In terms of automatic lip reading, HMM was the first to be employed in visual-based sentence-level lip reading with limited database [10]. Later, an HMM combined with human-engineered features was proposed and applied in the sentence-level audio visual speech recognition [11].

In recent years, with the improvement of computing capacity, deep learning methods have been used widely in lip reading. Many papers applied Convolutional Neural Network (CNN) to predict phonemes or visemes rather than the whole sentence or the full word from image frames of videos [12] [13]. But the shortage of large-scale lip reading datasets is still one of the most significant obstacles to overcome in the current researches of visual speech recognition.

III. Configuration

To address the difficulties encountered by deaf people and normal people in their conversation, in this paper we proposed a bi-directional communication system. The key technique in this system can be divided into two parts: Sign Language Recognition and Visual Speech Recognition (or lip reading).

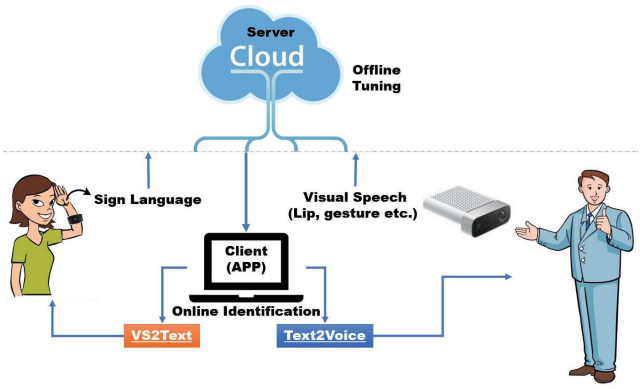

Fig. 1. System Structure. Sign Language and Visual Speech information will be translated into text and voice after they are sent to the cloud server for offline tuning and return to the client for online identification.

In this system, the Myo armband is used to collect motion signals of the fingers a nd a rms w hile a t t he s ame time translating gesture into speech and text shown on mobile devices for people with normal hearing. Microsoft Azure Kinect is used to capture videos and translate lip movement into text shown on mobile devices for people with hearing disabilities.

A. Sign Language Recognition

1) Acquisition Equipment for SLR: In this paper, sign language signals for SLR are collected by the wearable device integrating IMU and sEMG sensors shown in Figure 2. This wristband should be worn on the elbow joint of user's right arm in order to capture the data of sign language accurately.

The device includes an 8-channel sEMG sensor and a 6-channel IMU. The sEMG sensor was used to collect electrical signals of muscles in the gesture sequences, and the IMU included 3-axis accelerometer (ACC) and 3-axis gyroscope (GYR) to capture motion information. The inertial information collected from IMU can effectively be utilized to recognize the motion category of the forearm, but not

of the wrist and the palm of the hand. The movement of hands and fingers, as well as the strength information in the gesture are difficult to recognize by ACC and IMU, while sEMG signals can effectively describe the information and compensate the shortcomings of IMU. The combination of these two sensors to jointly collect the relevant signals in hand movement can better complete the SLR job. The length of data collected by the wristband was 64 in each channel after data pre-processing.

2) Classification Model: The nature of recognition of individual sign language word is the multi-classification for each gesture sequence. In this paper, the sequences here refer to time-serial signals which contain 3-channel ACC signals, 3-channel GYR signals and 8-channel sEMG signals.

As shown in the Figure 2, all three kinds of signals have demonstrated the strong local correlation, thus the convolutional layers were used to fuse different kinds of gesture information and extract features. Although in recent years numerous kinds of Convolutional Neural Network (CNN) have emerged and shown outstanding performances on recognition tasks, they are in possession of disadvantages itself. The conventional layer utilizes linear kernels to obtain feature maps, which are followed by a nonlinear activation function before fed to the next layer. However, the input data, signals for example, often live on a nonlinear manifold and in most cases linear filters are unable to abstract latent concepts properly as expected. Thus, in this paper, we proposed a novel network structure based on Network in Network (NIN) as the basic classification model for SLR. NIN emphasizes replacing the generalized linear model of convolutional filters with a "micro network" structure to better represent the nonlinear and abstract mapping between the receptive field and the feature map.

As shown in Figure 3, the input of the network had 14 channels, and the output was a word-embedding vector whose dimension is 300. We instantiated the micro neural network with a multilayer perceptron (MLP), which is a more potent nonlinear function approximator consisting of 2 fully connected layers (FC) with ReLU activation function. The resulting structure is referred as an MLP Conv layer. The overall structure of NIN proposed in our paper is the stacking of 5 MLP Conv layers. Instead of FC layers before the SoftMax classifier in conventional CNNs, Global Average Pooling (GAP) was used in this paper as it can enforce correlation between feature maps and categories with the help of MLP Conv layers. The FC layers at the end of CNNs are more prone to overfitting thus dropout regularization is often coupled with them to relieve the problems, while GAP is itself is capable of prevent overfitting for the overall structure.

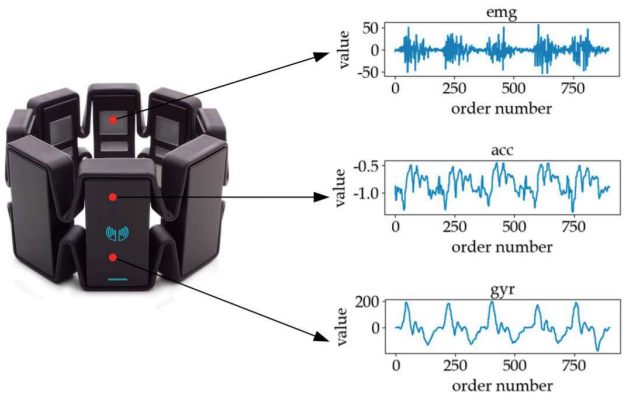

Fig. 2. Myo Wristband. The wearable device to collect signals of sign language including surface electromyogram (sEMG), accelerometer (ACC), and gyroscope (GYR).

The MLP unit is also equivalent to a convolutional layer with 1x1 filter, which makes it more straightforward to build a mlpconv layer. Inspired by the outstanding performance of ResNet, we also introduced residual connection between the input and the output of MLP Conv layers. As shown in Figure 4, in each MLP Conv layer, we added a shortcut connection parallel to the MLP model. Identity shortcut connections, which is one of the best solutions to the notorious problem of vanishing and exploding gradients, still has lower time complexity than VGG net while remains trainable without adding extra parameters or computational complexity at the same time.

B. Visual Speech Recognition

1) Acquisition Equipment for Lip Reading: The 440g portable machine shown in Figure 4 was used in this paper to capture videos of lip movement. The device contains a depth camera (1MP depth sensor with wide/narrow FOV options), a 12MP RGB video camera (an additional stream that is aligned to the depth stream), and also a motion sensor (with accelerometer and gyroscope which can capture movements in all domains). In addition, speech service SDK in Azure Kinect can capture the voice signal and convert spoken audio to text.

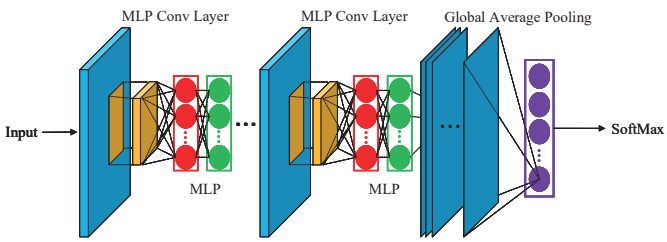

Fig. 3. Model Architecture for SLR. The model is a modified version of NIN and its input is composed of 14 channels of serial data.

2) Model Architecture: Technically speaking, different from directly analyzing gesture signals in SLR, VSR is much more intricate because it needs to obtain enough deep information from a clip of video at first, and then process the data in a supervised learning way. In addition, the alignment between word labels and image frames is also a problem which is hard to deal with. In this paper, we devised a recognition model to perform sentence-level lip reading, whose output is a sequence of pinyin corresponding to the video clip of the given sentence.

In order to extract information from video clips, we split various videos into sequences of image frames varying in length. A convolutional network architecture was used as the front-end to extract deep features. The convolutional network proposed in this paper is based on the DenseNet model, as it is more memory-efficient and faster to train with a decent classification performance.

In the front-end ConvNet, all the images were converted to gray-scale and normalized with respect to the overall mean and variance. When fed into the model, the frames in each sequence were cropped at the same position into the same size for training and centrally cropped for validation and test. In the DenseNet architecture, feature extraction was accomplished by blocks composed of multiple convolution layers using small convolution kernel. Slightly different to its counterpart in the SLR classification model, DenseNet used here was 2D instead of 1D to identify deep correlation embedded in the images. The architecture of the convolution layer is illustrated in Figure 6. Similar to SLR model, BN layers were added in the network in order to accelerate the training process and prevent overfitting.

After the ConvNet had extracted deep features, the output vectors from DenseNet would be transferred to the following model, which served as the back-end model for final recognition. In order to address vanishing gradient problem, in this paper we used bi-directional LSTM as the recurrent units. The probability distribution of the output character is generated by CTC loss over the output of LSTM layers.

LSTM (Long Short-Term Memory) network is considered to be one of the best models for sequence problems due to its sophisticated network architecture. In the lip reading problem, the challenge lies on extracting effective data from both previous hand movements and following ones.Thus, biLSTM can mine even more deep relationship from the context and better model lip image sequences.

After LSTM layers had identified the potential correlation between different frames in a sequence, the word-embedding vectors from LSTM was passed to CTC (Connectionist Temporal Classification) layer. CTC is a way to deal with sequences without the segment of single word as well as the alignment between the input and the output. To get the probability distribution of an output sentence given an input lip image sequence, CTC works by summing over the probability of all possible alignments.

In this way, the model of hybrid neural network proposed in this paper is able to finish jobs of processing videos and modeling sentences, which means conducting VSR completely end to end.

IV. Experiment

A. Dataset

1) Sign Language Dataset: Currently there is no existing dataset to support our experiment on Chinese Sign Language Recognition. Thus, we constructed a Chinese Sign Language database based on sEMG and IMU signals on our own. We collected multiple sets of words and phrases in different scenarios. Each set of data contains 70 categories of sign language words, and each category contains more than 100 samples.

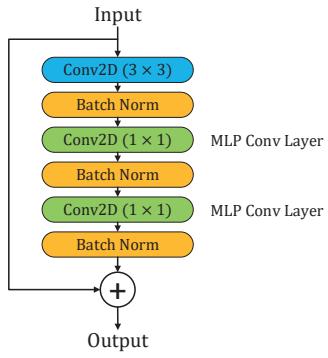

Fig. 4. Architecture of MLP Conv Layers. In this paper we introduced residual connections into them to enable the NIN model perform better.

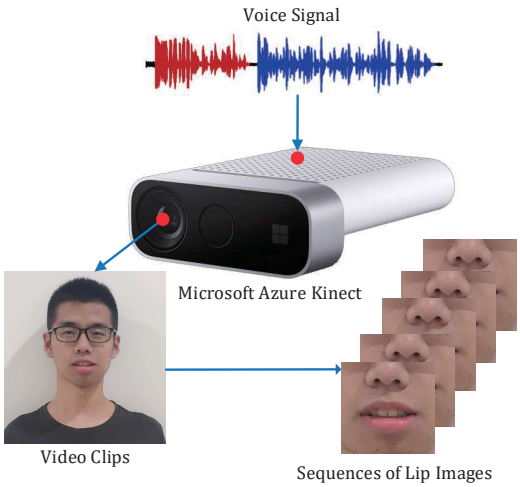

Fig. 5. Microsoft Azure Kinect. The portable device we used to collect the video information which is needed in lip reading.

Considering the randomness of sampling, more than 70 volunteers between the age of 19 and 24 who had experience of learning sign language were invited to help us collect data. Each sign language word was collected 10 times by one volunteer each time.Chinese Sign Language dataset we constructed contains 138 words. And the length of each sign language data varies from 164 to 207 points.

2) Visual Speech Dataset: In order to achieve better recognition accuracy in lip reading, we used LRW-1000 from Institute of Computing Technology, Chinese Academy of Sciences as our training set. LRW-1000 is a naturally distributed large-scale benchmark for Visual Speech Recognition in the wild, which contains 1,000 classes with 718,018 samples from more than 2,000 individual speakers. To the best of the authors' knowledge, it is currently the largest sentence-level lip reading dataset and also the only public large-scale Mandarin lip reading dataset.

B. Experimental Setup

As shown in Table I, we extracted 70 categories of Chinese sign language words in the scene of the airport from the database.

Similarly, we selected sentences containing the same categories of words from the visual speech database. Under the circumstance of the airport we set up, ten pairs of volunteers were asked to have a short conversation with each other. The short conversation contains five sentences shown in Table II. Concerning most normal people are unable to understand sign language and unable to communicate with deaf people without any assistance, we divided volunteers into two groups evenly. In Group A, deaf volunteers were asked to wear Myo wristband and hold Azure Kinect in the conversation. In Group B, deaf volunteers were accompanied by sign language translators in order to talk normally with volunteers without hearing disabilities. When normal volunteers in Group B began to speak, deaf volunteers in this group could merely guess what their partners were saying.

Sign language signals were obtained by the wristband in Figure 2, and the IMU and sEMG signals were concatenated and sent into the classification model. Video information of 25 fps and voice signal were also captured by the Azure Kinect shown in Figure 5 and OpenCV was used to divide the collected video to images for further processing as well as online recognition. In the end, we collated experiment data and obtained feedback from each group.

TABLE I
SIGN LANGUAGE CATEGORIES

today	toilet	you	key	flight
yesterday	return	me	miss	extension
tomorrow	ticket	we	search	support
morning	think	what	lighter	doctor
noon	luggage	how	gratitude	make
afternoon	worry	why	late	consign
evening	friend	borrow	airport	time
endorse	watch	able	cigarette	knife
connection	communicate	confiscate	direction	restaurant
recharge	China	empty	deaf	Beijing
help	Liaoning	terminal	world	counter
ID card	security	take off	deposit	floor
gate	seat	children	passport	food
water	sorry	suitcase	pack	[]

TABLE II
TEST SENTENCES

-Excuse me, where is the luggage stored?
-Sorry, I don't know.
-Where is the ticket change?
-Next to the toilet.
-Thanks for your help.
-No thanks.

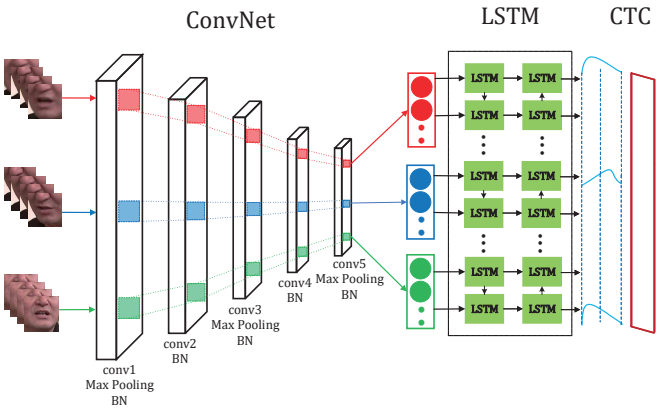

Fig. 6. Model architecture for VSR. The network contains a ConvNet Unit (DenseNet) and a Recurrent Unit (Bi-directional LSTM) as well as CTC loss to identify sentences end to end.

C. Training Strategy

In this paper, both the sign language database and visual speech database were divided into training set, validation set and test set according to the proportion of 98%, 1% and 1%. In order to better evaluate the model performance, these sets came from the same word/sentence distribution and were shuffled before the training started. The data size of different sets is listed in Table III.

During the beginning of the training, He initialization, a initialization method that works well for networks with ReLU activation, was used to accelerate the training time and attain better accuracy.

While the DenseNet was more prone to converge to an expected optima, the LSTM converged very slowly when the

TABLE III

DATASET DIVISION

	Train Set	Dev Set	Test Set
Sign Language	686,000	7,000	7,000
Visual Speech	49,000	500	500

number of timesteps is large due to a hard time the network encountered when initially extracting deep information from all the input steps. Thus, the hyperparameters for our model were tuned on the development set for each recognition task.

D. Results

As shown in Table IV, we reported the identification results of our system. It can be seen that in the part of SLR, we had an around 80% average recognition rate and in the section of VSR, we achieved an average recognition rate of about 35%.

TABLE IV

ACCURACY OF SENTENCES

Sentence	SLR Acc	VSR Acc
I	78.3%	31.2%
II	84.6%	36.5%
III	79.2%	32.6%
IV	83.3%	37.8%
V	81.2%	38.4%

During the experiment, we also recorded the overall time of both groups. While the average time of each conversation for Group A (with assistance of the purposed system) was less than 20 seconds, the average time for Group B (with merely sign language translation from the deaf people to normal people) reached up to 50 seconds.The shorter duration of time in this experiment verified the effectiveness of the communication-aid system. Compared with the current mode of communication which only relies on sign language, our system greatly reduces the time of communication and improve the efficiency. And portable devices can be carried around in different scenes, which meets the needs of most hearing-impaired people.

V. CONCLUSION

In this paper, we proposed a bi-directional communication system based on visual speech and sign language. The application of the portable devices in this paper makes it more convenient to be utilized not limited to time and place in practical circumstances.

Due to the effectiveness of the proposed method in the experiment, the system in this paper can expect to be applied widely to the scenarios such as airports, libraries, hospitals, banks and vast majority of other social scenes.

But the system still has shortcomings on the volume of the identifiable words/sentences. In the future work, the recognized pinyin will be translated into Chinese characters in VSR section and we will build our own visual speech database. In addition, we will continuously enlarge our sign language database in order to include more Chinese sign language words. As for recognition machines, a more convenient machine such as digital glasses will be considered to conduct lip reading.

REFERENCES

[1] M. Ahmed, M. Idrees, Z. ul Abideen, R. Mumtaz, and S. Khalique, "Deaf talk using 3D animated sign language: A sign language interpreter using Microsoft's kinect v2," in *2016 SAI Computing Conference (SAI)*. IEEE, 2016, pp. 330-335.

[2] D. Neoh, K. S. Mohamed Sahari, and W. Z. F. Wan Ibrahim, "A Dataglove Hardware Design and Real-Time Sign Gesture Interpretation," in *2018 Joint 10th International Conference on Soft Computing and Intelligent Systems (SCIS) and 19th International Symposium on Advanced Intelligent Systems (ISIS)*. IEEE, 2018, pp. 946-949.

[3] Y. M. Assael, B. Shillingford, S. Whiteson, and N. de Freitas, "LipNet:End-to-End Sentence-level Lipreading," arXivpreprint arXiv:1611.01599, 2016.

[4] K. W. Kim, M. S. Lee, B. R. Soon, M. H. Ryu, and J. N. Kim, "Recognition of sign language with an inertial sensor-based data glove," *Technology and health care: official journal of the European Society for Engineering and Medicine*, vol. 24, no. s1, p. S223, 2015.

[5] W. Gao, X. Chen, and J. Ma, "The communication system between deaf and normal people based on multi-mode interface technology," *Chinese Journal of Computers*, vol. 23, no. 12, pp. 1253-1260, 2000.

[6] J. Yamato, J. Ohya, and K. Ishii, "Recognizing human action in time-sequential images using hidden Markov model," in *Proceedings 1992 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 1992, pp. 379-385.

[7] M. E. Celebi, N. Codella, and A. Halpern, "Dermoscopy Image Analysis: Overview and Future Directions," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 2, pp. 474-478, 2019.

[8] J. Wang, J. Zhang, K. Honda, and J. Wei, "Audio-visual speech recognition integrating 3D lip information obtained from the Kinect," *Multimedia Systems*, vol. 22, no. 3, p. 315, 2016.

[9] A. Czyżewski, B. Kostek, M. Szykulski, and T. E. Ciszewski, "Building Knowledge for the Purpose of Lip Speech Identification," in *Advances in Intelligent Systems and Computing*. Springer, 2016, pp. 3-14.

[10] G. Papandreou, A. Katsamanis, V. Pitsikalis, and P. Maragos, "Multimodal Fusion and Learning with Uncertain Features Applied to Audiovisual Speech Recognition," in *2007 IEEE 9th Workshop on Multimedia Signal Processing*. IEEE, 2007, pp. 264–267.

[11] G. Papandreou, A. Katsamanis, V. Pitsikalis, and P. Maragos, "Adaptive multimodal fusion by uncertainty compensation with application to audiovisual speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 3, pp. 423-435, 2009.

[12] K. Noda, Y. Yamaguchi, K. Nakadai, H. G. Okuno, and T. Ogata, "Lipreading using convolutional neural network," in *2014 Annual Conference of the International Speech Communication Association (INTERSPEECH)*. ISCA, 2014, pp. 1149–1153.

[13] O. Koller, H. Ney, and R. Bowden, "Deep learning of mouth shapes for sign language," in *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*. IEEE, 2015, pp. 85–91.