ICCSPA20 1570690975

Multi-angle lipreading using angle classification and angle-specific feature integration

Shinnosuke Isobe, Satoshi Tamura, Satoru Hayamizu,
Gifu University, Gifu, Japan
{isobe@asr.info.,tamura@info.,hayamizu@}gifu-u.ac.jp

Yuuto Gotoh and Masaki Nose
Ricoh Company, Ltd., Kanagawa, Japan
{yuuto.gotoh,masaki.nose}@jp.ricoh.com

***Abstract*—Recently, visual speech recognition (VSR), or namely lipreading, has been widely researched due to development of Deep Learning (DL). The most lipreading researches focus only on frontal face images. However, assuming real scenes, it is obvious that a lipreading system should correctly recognize spoken contents not only from frontal but also side faces. In this paper, we propose a novel lipreading method that is applicable to faces taken at any angles, using Convolutional Neural Networks (CNNs) which is one of key deep-learning techniques. Our method consists of three parts; the view classification part, the feature extraction part and the integration part. We firstly apply angle classification to input faces. Based on the results, secondly we determine the best combination of pre-trained angle-specific feature extraction scheme. Finally, we integrate these features followed by DL-based lipreading. We evaluated our method using the open dataset OuluVS2 dataset including multi-angle audio-visual data. We then confirmed our approach has achieved the best performance among conventional and the other DL-based lipreading schemes in the phrase classification task.**

***Index Terms*—visual speech recognition, multi-angle lipreading, deep-learning, view classification**

I. Introduction

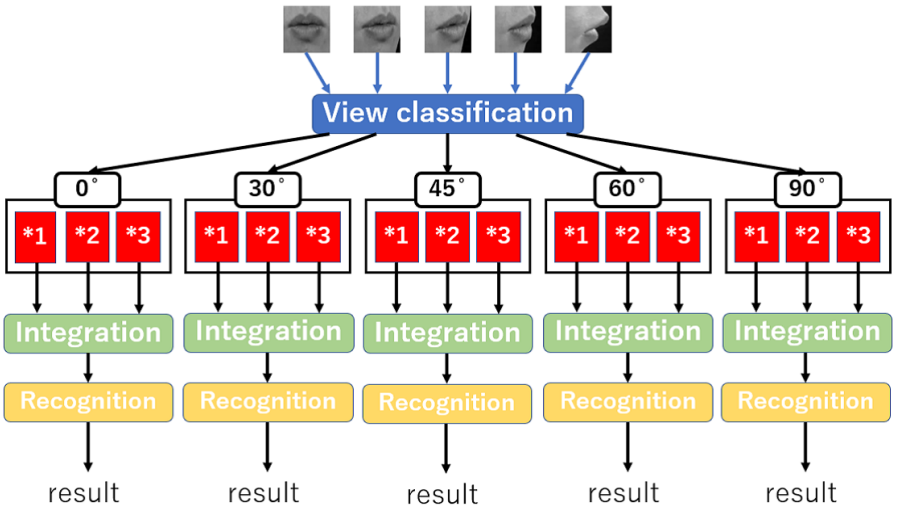
Many researchers have investigated Visual Speech Recognition (VSR), also known as lipreading, that estimates what a subject uttered only from a temporal sequence of lip images. Since lipreading technology has been utilized in Audio-Visual Speech Recognition (AVSR) also known as multimodal speech recognition, the development of lipreading directly affects AVSR, which can improve speech recognition accuracy in noisy environments. Lipreading and AVSR have a potential to be applied in various practical applications such as automatic conference minute generation and human interface on smartphones. Owing to state-of-the-art Deep Learning (DL), one of attractive Artificial Intelligence (AI) technology, recently we have achieved high performance of lipreading. However, lipreading still has several problems when we employ the technique in real-world scenes; for example, most of VSR researches have only consider frontal faces, but lipreading technology for non-frontal views is also essential for real applications. The authors thus have been developing multi-angle lipreading architecture which enables us to perform lipreading when not only frontal lip images but also non-frontal lip images are observed. There are two main approaches for multi-angle VSR. The first method is to build a VSR recognition model using training lip images captured at several angles. The second approach is to convert non-frontal lip images to frontal lip images and apply the conventional frontal lipreading technique. In this paper, we focus on the first approach, and propose a feature-integration-based multi-angle lipreading system using DL, particularly 3D Convolutional Neural Networks (CNNs), that is one kind of Deep Neural Networks (DNNs).

Our method consists of three parts: a view classification part, a feature extraction part and an integration part. Assume that we have a sequence of lip images to be recognized. Firstly in the view classification part, we prepare a 2DCNN that estimates the angle of input image. The model is then applied to each image in the sequence, followed by determining the angle which gets the majority in the estimation. Secondly, in the feature extraction part, we build 3DCNN models for possible combinations of angle-specific training datasets. Based on the angle obtained in the first part, we choose the best models and extract features from the models. In the last integration part, we concatenate these features followed by recognition by means of Fully Connected (FC) layer. We conducted evaluation experiments using the open dataset OuluVS2. Experiment results show that our proposed method improved recognition accuracy than conventional schemes on average. In addition, we confirm that our proposed method was not strongly affected by the view classification accuracy, because in the second part we simultaneously employ several models built using multi-angle training data.

The rest of this paper is organized as follows. In Section II, we briefly review related works on multi-angle lipreading. Section III introduces our method. Experimental setup, results, and discussion are described in Section IV. Finally Section V concludes this paper.

II. Related work

As mentioned, most conventional lipreading researches focused on frontal face images assuming that VSR systems are in front of speakers, since there are only a few datasets available having multi-angle faces. One of the public multi-angle lipreading datasets is OuluVS2 [1]. An early work of multi-angle lipreading is [2], where a system was trained using either of frontal ($0°$) or profile ($90°$) faces. According to their experimental results, the frontal view showed lower Word Error Rate (WER) than the profile view. In [3], they built a multi-angle system investigating a frontal ($0°$) view, a left profile ($90°$) view and a right profile ($-90°$) view. They



* Generates a 48-dimensional vector using the n-th best model

Fig. 1. An architecture of our proposed method.

reported they got significantly better performance when using the frontal view than the others. Saitoh et al. proposed a novel sequence image representation method called Concatenated Frame Image (CFI) [4]. Two types of data augmentation methods for CFI, and a framework of CFI-based CNN were tested.

In contrast to the above works, some researches reported that non-frontal lip images are more effective than frontal lip images. Bauman indicated that human lipreaders tend to have higher performance when slightly angled faces are available, presumably because of visibility of lip protrusion and rounding [5]. In [6], Active Appearance Model (AAM) was utilized for feature extraction at five angles, and lipreading was examined on a view-dependent system, as well as on a view-independent system using a regression method in a feature space. As a result, the view-dependent system performed the best performance at 30° than frontal view-dependent and view-independent results. Zimmermann et al. used Principal-Component-Analysis (PCA) -based convolutional networks together with Long Short-Term Memories (LSTMs) that is one of DL models, in addition to a conventional speech recognition model: Hidden Markov Models (HMMs) with Gaussian Mixture Models (GMMs) [7]. They aimed at combining multiple views by employing these techniques. They finally confirmed the highest performance was obtained at 30°. Anina et al. insisted the highest accuracy was achieved at 60° in their experiments [1]. Kumar et al. showed that profile-view lipreading provides significantly lower WERs than frontal-view lipreading [8].

There is another strategy to conduct transformation to images or incorporating several views with DL technology. There is one work [9] converting faces viewed from various directions to frontal faces using AMMs. Experimental results showed that recognition accuracy was improved even when the face direction changes about 30° relative to a frontal view. In [10], they proposed a scheme called "View2View" using an encoder-decoder model based on CNNs. The method transformed non-frontal mouth region images into frontal ones. Their results showed that the view-mapping system worked well for VSR and AVSR. Estellers et al. introduced a pose normalization technique and performed speech recognition from multiple views by generating virtual frontal views from non-frontal images [11]. In [12], S. Petridis et al. proposed an end-to-end multi-view lipreading system based on bidirectional LSTM networks. This model simultaneously extracted features directly from the pixels and performed visual speech classification from multi-angle views. Experimental results demonstrated the combination of frontal and profile views improved accuracy over the frontal view. Zimmermann et al. also proposed another decision-fusion-based lipreading [13]; they extracted features through a PCA-based convolutional neural network, LSTM network and GMM-HMM scheme. The decision fusion succeeded by combining Viterbi paths.

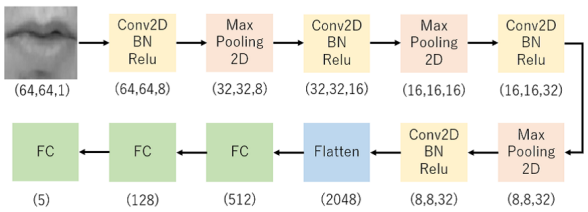
Consequently, multi-angle lipreading researches can be classified into two categories; (i) build a lipreading model which corresponds to frontal and/or non-frontal lip images; (ii) convert non-frontal lip images to frontal ones and perform frontal lipreading.

III. Methodology

Our proposed method consists of three parts: a view classification part, a feature extraction part a the integration part. Fig. 1 depicts the architecture of our method. In this section, we describe each part of our scheme.

A. View classification

Assuming real scenes, it is not guaranteed that a speaker is strictly facing to a lipreading system. One way to deal with this problem is that we prepare several models each which



* BN = Batch Normalization, FC = Fully Connected Layer

Fig. 2. A 2DCNN model for view classification.

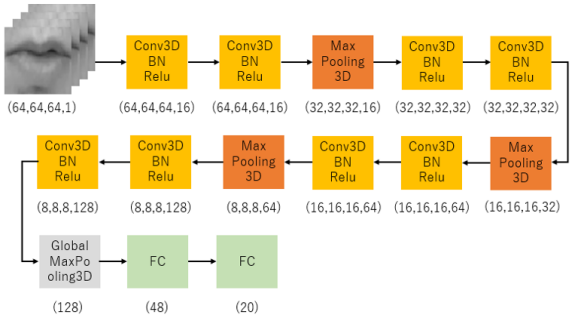


Fig. 3. A 3DCNN model for recognition.

corresponds to a certain angle, estimate at which angle face images are captured, and apply an angle-specific model. In the view classification part, we at first estimate at which angle face images were recorded among the following five candidates in this work: $0°, 30°, 45°, 60°$ and $90°$. The estimation is done for each lip image in one sequence, using a 2DCNN model illustrated in Fig 2. The 2DCNN model employs a simple and common architecture; convolutional and pooling layers are repeatedly applied followed by FC layers, to get a classification result for the five angles. After processing all input images, we determine the angle which is most often chosen.

B. Feature extraction

Before conducting this part, we prepare 3DCNN recognition models for all possible combinations of the above five angles: i.e. a model trained only using frontal images, a model from $30°$ images, $\cdots$, a model trained using both $0°$ and $30°$ data, $\cdots$, and a model using all face images. An architecture of our 3DCNN models is shown in Fig 3. The last layer has 20 outputs, each which corresponds to one class in our recognition task. As a result, we build 31 models in this case, shown in Table I. Table I also indicates preliminary experimental results: recognition accuracy to validation data at a certain angle, using a certain model chosen among those 31 models. For example, if we adopt a $30°$ model for $60°$ data, the accuracy was 87.55%.

According to the angle which we obtain in the view classification part, we select the most reliable three models for the estimated angle, which are shown as bold in Table I. For instance, we adopt 1) "$0°$+$30°$+$45°$", 2) "$0°$+$30°$+$45°$+$60°$" and 3) "$0° + 30° + 45° + 90°$" models for $45°$ data. In other words, we determine suitable angle combination patterns of training data for the estimated angle. We then utilize those models as feature extractors; we remove the last layer, resulting a new output layer generating a 48-dimensional feature vector. Finally, we obtain three 48-dimensional vectors from this part.

C. Integration part

In the integration part, firstly, we integrate those 48-dimensional features extracted from three angle-specific models, by simply concatenating them. Thereafter, we conduct recognition using two FC layers ($48\times3 \rightarrow 48 \rightarrow 20$). Here we apply 50% dropout between the FC layers.

IV. Experiments

A. Dataset

We chose the OuluVS2 dataset to evaluate our scheme. The database contains 10 short phrases, 10 digits sequences, and 10 TIMIT sentences uttered by 52 speakers. The corpus includes face images captured by five cameras simultaneously at $0°$ (frontal), $30°$, $45°$, $60°$, and $90°$ (profile) angles, respectively. In this paper, we adopted the phrase data and digit data, uttered three times by each speaker. In our experiment, the data spoken by 52 speakers were divided into training data by 35 speakers, validation data by 5 speakers and testing data by 12 speakers. The phrases are as follows: "Excuse me", "Goodbye", "Hello", "How are you", "Nice to meet you", "See you", "I am sorry", "Thank you", "Have a good time", "You are welcome". Each digit utterance consists of 10 digits randomly chosen.

B. Experimental Setup

We evaluated a model by utterance-level accuracy:

$$\text{Accuracy} = \frac{H}{N} \times 100\ [\%] \tag{1}$$

where H and N are the number of correctly recognized utterances and the total number of utterances, respectively. Since DNN-based model performance slightly varies depending on the probabilistic gradient descend algorithm, that is a common model training approach, we repeated the same experiment three times and the mean accuracy is calculated. In terms of DNN hyperparameters, we chose a cross-entropy function as a loss function and Adam as an optimizer. Batch size, epochs and learning rate were set to 32, 50 and 0.001, respectively. We carried out our experiments using NVIDIA GEFORCE RTX 2080 Ti. The time required to conduct one epoch when training each model is shown in Table II.

C. Preprocessing

The OuluVS2 dataset includes extracted lip images, however, their image sizes differ. In order to apply DNNs, we resized all images to 64×64. Furthermore, we normalized a frame length to 64; if the length is less than 64 we conducted upsampling, otherwise we suppressed some frames. In addition, we converted all color images to grayscale images.

TABLE III
LIPREADNIG RESULTS OF OUR PROPOSED METHOD AND CONVENTIONAL SCHEMES

Task and method \ Data		0°	30°	45°	60°	90°	Mean
20 class	Ours without view classification	**94.35**	93.89	93.98	**94.03**	**93.20**	93.89
	Ours with view classification	**94.35**	**93.98**	**94.50**	**94.03**	**93.20**	**94.01**
10 class	CNN + Data Augmentation [4]	85.6	82.5	82.5	83.3	80.3	82.84
	PCA Network + LSTM + GMM-HMM [7]	73.1	75.6	67.2	63.3	59.3	67.7
	View2View [10]	-	86.11	83.33	81.94	78.89	82.57
	End-to-end Encoder + BLSTM [12]	**91.8**	87.3	88.8	86.4	**91.2**	89.1
	End-to-End CNN-LSTM [14]	82.8	81.1	85.0	83.6	86.4	83.78
	Ours without view classification	91.02	90.56	91.20	**90.00**	88.70	90.33
	Ours with view classification	91.02	**90.74**	**92.04**	**90.00**	88.70	**90.50**

TABLE II
THE TIME FOR ONE EPOCH IN MODEL TRAINING

View classification	207s
3DCNN (1 angle)	31s
3DCNN (2 angle)	56s
3DCNN (3 angle)	82s
3DCNN (4 angle)	108s
3DCNN (5 angle)	135s
FC	0.2s

TABLE II
VIEW CLASSIFICATION RESULTS

Result \ Data	0°	30°	45°	60°	90°
0°	720	0	0	0	0
30°	0	678	28	0	0
45°	0	42	500	1	0
60°	0	0	192	717	0
90°	0	0	0	2	720

TABLE I
PRELIMINARY RECOGNITION RESULTS FOR VALIDATION DATA

Model \ Data	0°	30°	45°	60°	90°
0°	95.33	93.33	89.78	69.22	42.78
30°	93.78	95.89	94.67	87.55	69.00
45°	88.22	91.89	95.00	93.78	76.89
60°	66.00	80.11	88.22	95.89	90.89
90°	47.44	56.55	69.44	93.56	94.67
0° + 30°	96.00	96.56	96.22	88.67	66.56
0° + 45°	94.78	95.78	95.78	93.67	79.34
0° + 60°	92.78	94.00	93.55	95.44	88.22
0° + 90°	96.33	96.67	94.67	96.56	93.56
30° + 45°	93.56	95.56	95.22	90.89	79.00
30° + 60°	93.78	96.89	96.44	97.11	87.33
30° + 90°	94.67	97.22	96.78	95.78	**95.11**
45° + 60°	88.33	92.22	96.00	96.11	89.56
45° + 90°	89.11	93.67	94.67	96.78	94.67
60° + 90°	75.11	83.56	88.00	96.89	94.45
0° + 30° + 45°	96.89	97.55	**97.89**	96.78	76.44
0° + 30° + 60°	96.11	97.78	96.89	96.67	87.56
0° + 30° + 90°	95.11	97.89	96.78	95.89	94.45
0° + 45° + 60°	96.11	96.89	96.00	97.22	85.67
0° + 45° + 90°	94.33	95.78	95.44	95.44	93.56
0° + 60° + 90°	96.22	96.78	95.56	97.55	94.78
30° + 45° + 60°	93.78	96.89	97.44	96.55	84.11
30° + 45° + 90°	95.78	97.11	97.22	97.22	94.78
30° + 60° + 90°	95.67	97.78	97.33	97.56	94.56
45° + 60° + 90°	89.89	92.89	95.33	96.55	94.66
0° + 30° + 45° + 60°	96.67	96.89	**97.78**	97.11	86.33
0° + 30° + 45° + 90°	**97.44**	**98.33**	**97.89**	**97.67**	95.00
0° + 30° + 60° + 90°	96.44	**98.11**	97.00	**98.11**	94.22
0° + 45° + 60° + 90°	**97.67**	**98.22**	97.45	**98.22**	93.89
30° + 45° + 60° + 90°	95.22	96.78	97.11	97.22	**96.45**
0° + 30° + 45° + 60° + 90°	**96.89**	97.89	97.00	97.55	**95.89**

D. Results and Discussion

1) View classification: View classification results for the test data are shown in Table II. In the confusion matrix, for example, among 720 images at 30°, 678 images were correctly classified while 42 were wrongly recognized. The total classification accuracy was 92.64%. Our model could correctly classify 0° and 90° images, on the other hand, 30° and 60° images were slightly misclassified. We can obviously observe many errors at 45°. This may be due to the fact that 45° images looks much more similar to those at neighbor angles than the other ones.

In past multi-angle lipreading methods with image translation such as [10], misclassifications in the view classification caused a great impact to image translation and lipreading itself. In contrast, in our proposal method misclassifications in the angle classification part were expected to hardly affect the following processes, because our recognition models were built using multiple angles and we used three models simultaneously to enhance the robustness.

2) Lipreading: Recognition accuracy of our and competitive lipreading schemes is shown in Table III. The upper part of Table III indicates the results of 20-class classification task (phrase sentences and digit sequences), while the lower part shows the results of 10-class recognition (phrase sentences only).

First, the results of the 20-class recognition task are discussed. It is obviously found that our method achieved significant performance among all angle conditions. The method having view classification had the same or better performance, comparing to another one without the classification. It is interesting at 45° we found much more improvement than the other conditions, even the view classification was not sufficient. Since 45° data were used as training data in the 30° and 60° conditions, we might obtain such the improvement even if the angle classification failed.

Next, we discuss about the results of the 10-class recognition task in which only phrase sentences were used. Focusing

on the average of recognition accuracy, our proposed method achieved the highest accuracy regardless of the presence or absence of the angle classification part, comparing to the conventional lipreading systems. It is observed that our method is particularly effective at the middle-angle (30°, 45° and 60°) conditions, while the end-to-end system got higher accuracy for frontal and profile images.

Finally, we focus on the recognition performance in the 10- and 20-class recognition tasks. We can generally agree that the larger the number of classes becomes, the more difficult the classification task is. Nevertheless, as shown in Table III, the 20-class task looks easier. We consider this is because the digit recognition task is much more easier; compared to phrase sentences, the length tends to be longer and we can easily find cues for classification.

V. CONCLUSION

In this paper, we proposed a multi-angle lipreading system in which feature extraction were conducted using angle-specific models based on view classification result, followed by feature integration and lipreading. We employed DNNs in our system, to perform view classification, feature extraction and recognition. Experiments were conducted in two tasks using OuluVS2 corpus. Then we found our scheme could significantly work well compared to past works, as well as we clarified the effectiveness of view classification and feature extraction from pre-trained angle-specific models. In this paper, we prepared five angle-specific models of which angles were employed in OuluVS2. We are planning to conduct experiments using different angle settings.

REFERENCES

[1] I. Anina, Z. Zhou, G. Zhao, and M. Pietikäinen, "OuluVS2: A multi-view audiovisual database for non-rigid mouth motion analysis," Proc. FG2015, 2015.

[2] P. Lucey and G. Potamianos, "Lipreading using profile versus frontal views," Proc. MMSP2006, pp.24–28, 2006.

[3] P. Lucey, S. Sridharan, and D. Dean, "Continuous pose invariant lipreading," Proc. INTERSPEECH2008, pp.2679–2682, 2008.

[4] T. Saitoh, Z. Zhou, G. Zhao, and M. Pietikäinen, "Concatenated frame image based CNN for visual speech recognition," Proc. ACCV2016, 2016.

[5] S.L. Bauman and G. Hambrecht, "Analysis of view angle used in speech reading training of sentences," American Journal of Audiology, vol.4, no.3, pp.67–70, 1995.

[6] Y. Lan, B. J. Theobald, and R. Harvey, "View independent computer lip-reading," Proc. Multimedia and Expo, pp.432–437, 2012.

[7] M. Zimmermann, M. Mehdipour Ghazi, H. K. Ekenel, and J.-P. Thiran, "Visual speech recognition Using PCA networks and LSTMs in a tandem GMM-HMM system," Proc. ACCV2016, 2016.

[8] K. Kumar, T. Chen, and R. Stern, "Profile view lip reading," Proc ICASSP2007, pp.429–432, 2007.

[9] Y. Komai,N. Yang,T. Takiguchi, and Y. Ariki, "Robust AAM based audio-visual speech recognition against face direction changes," Proc. ACM Multimedia 2012, pp.1161–1164, 2012.

[10] A. Koumparoulis and G. Potamianos, "Deep view2view mapping for view-invariant lipreading," Proc. SLT2018, pp.588–594, 2018.

[11] V. Estellers and J.-P. Thiran, "Multipose audio-visual speech recognition," Proc. EUSIPCO2011, pp.1065–1069, 2011.

[12] S. Petridis, T. Stafylakis, P. Ma, F. Cai, G. Tzimiropoulos, and M. Pantic, "End-to-end multiview lip reading," Proc. ICASSP2018, pp.6548–6552, 2018.

[13] M. Zimmermann, M. MehdipourGhazi, H.K.Ekenel, and J.-P. Thiran, "Combining multiple views for visual speech recognition," Proc. AVSP2017, 2017.

[14] D. Lee, J. Lee, and K. E. Kim. "Multi-view automatic lip-reading using neural network," Proc. ACCV2016, 2016.