MULTI-VIEW VISUAL SPEECH RECOGNITION BASED ON MULTI TASK LEARNING

HouJeung Han , Sunghun Kang and Chang D. Yoo

Korea Advanced institute of Science and Technology
School of Electrical Engineering
291 Daehak-ro, Yuseong-gu, Daejeon, Korea

ABSTRACT

Visual speech recognition (VSR), also known as lipreading is a task that recognizes word or phrase using video clip of lip movement. Traditional VSR methods are limited in that they are based mostly on VSR of frontal view facial movement. This limitation should be relaxed to include lip movement from all angles. In this paper, we propose a pose-invariant network which can recognize word spoken from any arbitrary view input. The architecture that combines convolutional neural network (CNN) with bidirectional long short-term memory (LSTM) is trained in a multi-task manner such that the pose and the word spoken are jointly classified. Here, pose classification is considered as the auxiliary task. To comparatively evaluate the performance of the proposed multi-task learning, OuluVS2 benchmark dataset is considered. The experimental results show that the deep model learned based on the proposed multi-task learning method prove its advantage compared to previous single-view VSR methods and also previous multi-view lipreading methods. This deep model achieved recognition performance of 95.0% accuracy on OuluVS2 dataset.

Index Terms— lipreading, multi view, multi task, pose-invariant, Visual Speech Recognition

1. INTRODUCTION

Speech can be represented not only as acoustic information but also as visual information. Humans, in general, understand speech by processing incoming acoustic information, but hearing-impaired people use visual information. The effectiveness of visual information is not limited to a deaf person. The use of visual information can enhance the performance of a speech recognizer. For instance, different visual cues can provide varied perception of speech, which is known as the McGurk effect[1]. Likewise, visual speech recognition (VSR) can enhance the performance of audio-visual speech recognition (AVSR), especially in noisy environment. Traditional feature extraction in method for VSR has been studied in depth in [2]. Recently, numerous methods based on using deep learning to predict phonemes or visemes have been proposed [3, 4, 5, 6], including long short-term memory (LSTM) for extracting temporal features[7, 8, 9] and both [10, 11].

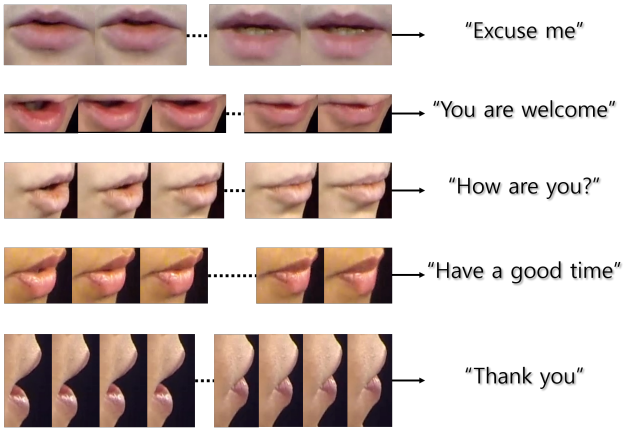


Fig. 1. Multi-view visual speech recognition.

Face can be captured at various angle and position which can make lip reading difficult. Until now, most VSR studies has been focused on frontal view, which may be impractical. In this paper, we present a model that can recognize visual speech from any arbitrary point of view as shown in Fig 1. To enhance lipreading, we present a learning method that jointly learns the position of the face and word spoken as well as the words. While mainly training with label of phrases, the model also auxiliarily train with label of facial position. We believe subsidiary information of view assist better recognition of phrases by positioning features of same class more specifically according to position.

The system is tested with the OuluVS2[12] which limits the scope of movement to yaw rotation at five points in the range of frontal to profile.

We compared our result with previous multi-view VSR task[10], and it shows improvement of 9.5% of accuracy on average of each view. We also show our method outperforms frontal view performance from other models trained with various method, presenting a performance of 95.0% accuracy. Finally, we verify the ability of auxiliary task aids recogni-

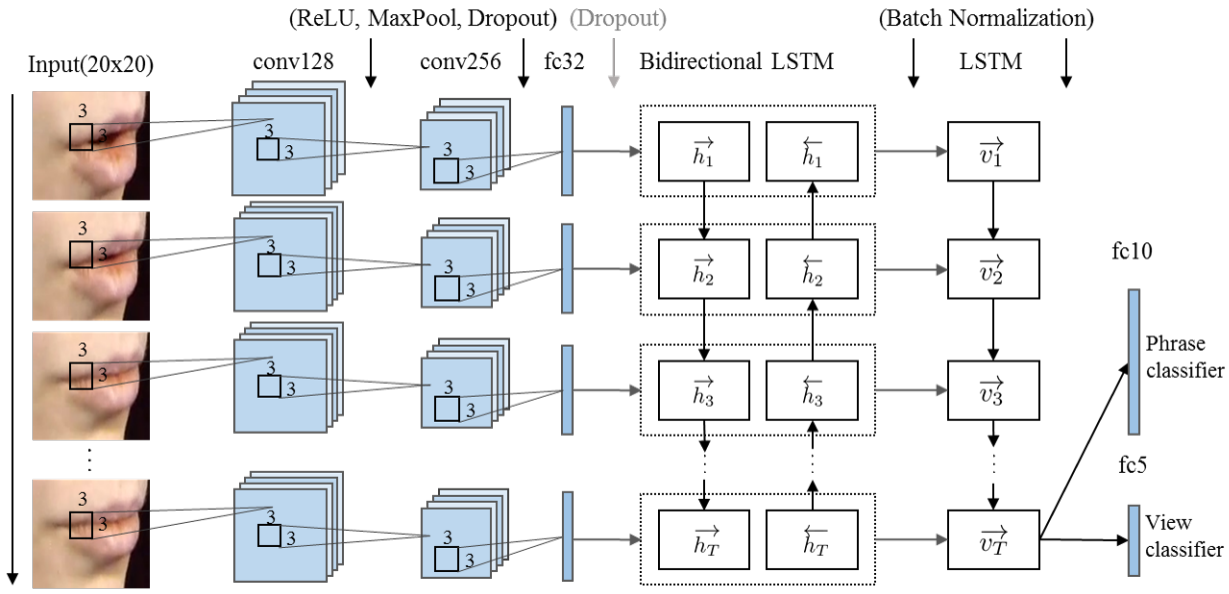


Fig. 2. Input is multi-view lip movement video. The architecture mainly consist of CNN combined with Bi-LSTM. On final layer, the features of 'fc10' are classified into phrases, and those of 'fc5' classified into facial positions.

tion of the word.

2. METHOOD

In this section, we describe a model and multi task learning algorithm. The model architecture is shown above in Fig 2. Besides classifying spoken words, this model is simultaneously trained to classify, the view. For visual feature extraction, two 3x3 CNN layer with stride 1 followed by 2x2 Max pooling and fully-connected layer is used in our architecture. After max pooling, activation and dropout are applied, and 32 unit fully-connected layer is followed by dropout. The activation function is rectified linear unit (ReLU), and the ratio of dropout is 0.4. We use two recurrent neural networks (RNN) with 128 units for temporal model, and batch normalization is operated after each RNN. Here, batch normalization is same with [13] except for using per-batch statistics to normalize the data also during testing, which helps.

2.1. Bi-directional LSTM

The first RNN layer is Bi-directional LSTM, introduced by [14], followed by another LSTM layer, as depicted in Fig 2. The system can be formulated by Eq.(1)-(5) from t = 1 to T where x_t is input of Bi-LSTM and z_t is output feature of LSTM in Fig 2, W_{xh} is the input-hidden weight matrix, W_{hh} hidden-hidden weight matrix, b_h is hidden bias vector and $\mathcal{H}$ is the hidden layer function.

$$\overrightarrow{h_t} = \mathcal{H}(W_{x\overrightarrow{h}}x_t + W_{\overrightarrow{h}\overrightarrow{h}}h_{t1} + b_{\overrightarrow{h}}) \quad (1)$$

$$\overleftarrow{h_{1t}} = \mathcal{H}(W_{x\overleftarrow{h_1}}x_t + W_{\overleftarrow{h_1}\overleftarrow{h_1}}\overleftarrow{h_{1t1}} + b_{\overleftarrow{h_1}}) \quad (2)$$

$$y_t = W_{\overrightarrow{h}y}\overrightarrow{h}_t + W_{\overleftarrow{h_1}y}\overleftarrow{h_{1t}} + b_y \quad (3)$$

$$v_t = \mathcal{H}(W_{yh}y_t + W_{hh}h_{t1} + b_h) \quad (4)$$

$$z_t = W_{hz}h_t + b_z \quad (5)$$

2.2. Multi task classification

The LSTM feature of the clip goes to two fully-connected layer. The final loss function $H(Y, V, y, v)$ of Eq.(8) is weighted sum of cross-entrpy loss of phrase $H_{phrase}(Y, y)$ of Eq 6) and cross-entropy of view $H_{view}(V, v)$ Eq.(7) where Y, V are ground truth vector of phrase and view, y, v are predicted probability of phrase and view, and β is set to 0.7 tuned on validation set. The input of the training model is a sequence of images and a pair of phrase and view label.

$$H_{phrase}(Y, y) = -\sum_i Y(i)\ \log y(i), \quad (6)$$

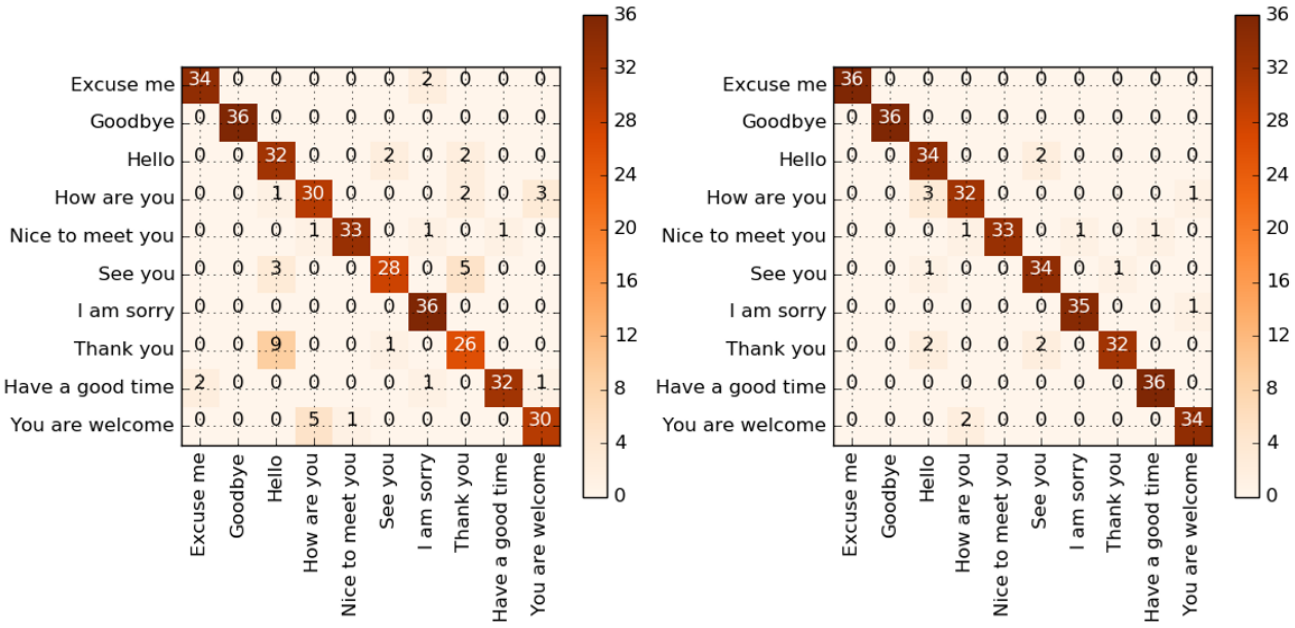
$$H_{view}(V, v) = -\sum_j V(j)\ \log v(j), \quad (7)$$

$$H(Y, V, y, v) = H_{phrase}(Y, y) + \beta H_{view}(V, v). \quad (8)$$

3. EXPERIMENT

3.1. Database

The database used in this paper is the OuluVS2[12]. The OuluVS2 contains 52 speakers saying three kinds of utterances(Digits, Phrases, TIMIT), 3 times each(except TIMIT),



(a) Single task learning

(b) Multi task learning

Fig. 3. Confusion matrix of multi-view training with single task and multi task. Vertical axis is ground-truth labels and horizontal axis is predicted labels

simultaneously recorded on five different views: {0°, 30°, 45°, 60°, 90°}, so in total there are 780 examples per utterance. In phrases, there are 10 classes: 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. The mouth region of interests (ROIs) are provided and they are downscaled to 26 by 44 in order to keep the aspect ratio constant. In this paper, we use phrase for recognition task.

3.2. Evaluation Protocol

We followed evaluation protocol at the ACCV 2016 workshop: Multi-view lip-reading/audio-visual challenge[1]. The protocol suggested here partition 52 subjects into 40 subjects for training and validation, 12 subjects for the test. We randomly selected 28 and 12 subjects among training and validation set for training and validation purposes, respectively; therefore, we have total 4200 utterances for training and 1800 utterances for validation and testing.

3.3. Training

We use 20 by 20 pixel original RGB color image. The Adam algorithm [15] is used for the optimizer, and the latest best model is saved according to loss monitored within 300 epoch.

[1]http://ouluvs2.cse.oulu.fi/ACCVE.html

4. RESULTS

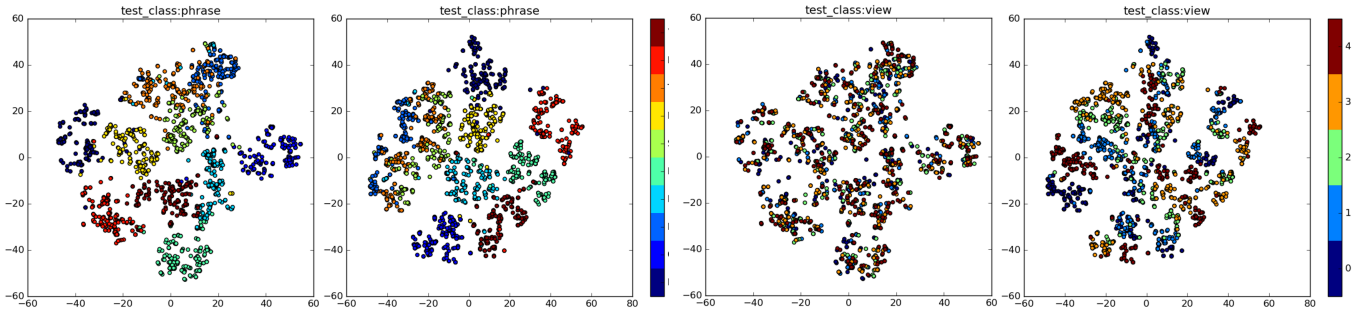
In this section, we present the result single and multi task learning of multi-view training. The single task learning refers to learning without the auxiliary task, only with classifying phrases.

We first measure all views and final accuracy by averaging each result. In table 1, we compared multi-view result of single task and the multi task with [10], which is the only experiment conducted in multi-view training (referred as "Cross-view" in this paper). Our model in both tasks outperforms baseline by 6% on average. Also, this verifies that multi task learning enhance the performance for all view, especially for frontal and 30 degree view, compared to the single task learning by 3.5% on average.

Accuracy(%)	Lee et al. [10]	Ours(single task)	Ours(multi task)
Frontal	80.6	90.3	**95.0**
30°	81.1	84.7	93.1
45°	85.0	90.6	91.7
60°	82.5	88.6	90.6
Profile	83.6	88.6	90.0
Average	82.6	88.6	92.1

Table 1. Results on Cross-view experiments.

In table 2, we compared frontal view recognition performance in each paper. First, the result of [10] is performance of model fine-tuned only frontal view after pre-training with multi-view, referred as "cross-view2" in its paper. Second,



(a) phrase label of single (left) and multi (right) task

(b) view label of single (left) and multi (right) task

Fig. 4. t-SNE plot of multi-view training with single task and multi task in test phase.

Accuracy(%)	Frontal View	training
Lee et al. [10]	82.8	multi view training
Petridis et al. [7]	84.5	single view training
Ours(single task)	90.3	multi view training
Chung and Zisserman [6]	93.2	pretrained with [6]
Chung and Zisserman [11]	94.1	pretrained with [11]
Ours(multi task)	**95.0**	multi view training

Table 2. Results of Frontal view experiments.

model of [7] is trained with single frontal view videos. Here, 40 subjects are randomly split into 30 and 10 for training and validation respectively. Third, both model of [6] and [11] are pre-trained with other dataset and fine tuned with frontal view videos. The result of our multi task method performs well without complementary dataset or training procedure.

Confusion matrix of multi-view training of our model with both single task and the multi task is shown in Fig 3. As it can be seen in the confusion matrix, three pairs of phrase { "Hello" & "Thank you" & "See you"} is one of the most challenging and confusing pair with the high error rate, as visually similar each other in the same view. As comparing both task result, multi task learning appears to reduce large amount of error with those three pairs and other pairs of three phonemes. By explicitly learn the feature of angle, auxiliary task with view classification also help cluster in view then aligned in both classes, which make more linearly separable. The auxiliary task learns to map features which are different in the main task but hidden related representations, finally more easy to be lineally separable during learning.

In Fig 4, we plot t-distributed stochastic neighbor embedding plot (t-SNE plot)[16] for verification. The Fig 4(a) is labeled with phrases and Fig 4(b) is labeled with the view. The plot in single task appear to be arbitrary scattered in class of phrase; on the other hand, the difference can be observed the plot in multi task that elements are clustered not only in class of phrase but also in that of view, and most importantly, more aligned with both class of phrase and view, especially on "Thank you" & "Hello" &"See you" pairs. This mainly enhances the performance of phrase classification.

5. CONCLUSION

In this paper, we propose a multi task based approach that gives competitive performance with multi-view training on the OuluVS2 dataset. While recognizing phrases from any arbitrary view input, the auxiliary task that also classify view assist classification of phrases. We developed an architecture that combines CNN with Bidirectional LSTM. We achieved 95% accuracy in frontal view test and 92.1% in multi-view test . As conducting the experiment with both single task and multi task with the same architecture and multi-view dataset, we verify the ability of multi task learning.

6. ACKNOWLEDGEMENTS

This work was partly supported by the ICT R&D program of MSIP/IITP [2016-0-00563, Research on Adaptive Machine Learning Technology Development for Intelligent Autonomous Digital Companion] and partly supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Science, ICT & Future Planning(NRF-2017R1A2B2006165). The authors are with the School of Electrical Engineering under the bk21plus program, Korea Advanced Institute of Science and Tech-nology, Daejeon 305-701, South Korea.

7. REFERENCES

[1] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," vol. 264, pp. 746–748, Dec. 1976.

[2] Ziheng Zhou, Guoying Zhao, Xiaopeng Hong, and Matti Pietikäinen, "A review of recent advances in visual speech decoding," *Image and vision computing*, vol. 32, no. 9, pp. 590–605, 2014.

[3] Kuniaki Noda, Yuki Yamaguchi, Kazuhiro Nakadai, Hiroshi G Okuno, and Tetsuya Ogata, "Lipreading using convolutional neural network.," in *INTERSPEECH*, 2014, pp. 1149–1153.

[4] Oscar Koller, Hermann Ney, and Richard Bowden, "Deep learning of mouth shapes for sign language," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2015, pp. 85–91.

[5] Yiting Li, Yuki Takashima, Tetsuya Takiguchi, and Yasuo Ariki, "Lip reading using a dynamic feature of lip images and convolutional neural networks," in *Computer and Information Science (ICIS), 2016 IEEE/ACIS 15th International Conference on*. IEEE, 2016, pp. 1–6.

[6] JS Chung and A Zisserman, "Lip reading in the wild," in *Asian Conference on Computer Vision*, 2016.

[7] Stavros Petridis, Zuwei Li, and Maja Pantic, "End-to-end visual speech recognition with lstms," .

[8] Stavros Petridis and Maja Pantic, "Deep complementary bottleneck features for visual speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 2304–2308.

[9] Michael Wand, Jan Koutník, and Jürgen Schmidhuber, "Lipreading with long short-term memory," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 6115–6119.

[10] Daehyun Lee, Jongmin Lee, and Kee-Eung Kim, "Multi-view automatic lip-reading using neural network," in *ACCV 2016 Workshop on Multi-view Lip-reading Challenges*. Asian Conference on Computer Vision (ACCV), 2016.

[11] Joon Son Chung and Andrew Zisserman, "Out of time: automated lip sync in the wild," in *Workshop on Multi-view Lip-reading, ACCV*, 2016.

[12] Iryna Anina, Ziheng Zhou, Guoying Zhao, and Matti Pietikäinen, "Ouluvs2: A multi-view audiovisual database for non-rigid mouth motion analysis," in *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*. IEEE, 2015, vol. 1, pp. 1–5.

[13] Sergey Ioffe and Christian Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[14] Alex Graves and Jürgen Schmidhuber, "Framewise phoneme classification with bidirectional lstm and other neural network architectures," *Neural Networks*, vol. 18, no. 5, pp. 602–610, 2005.

[15] Diederik Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[16] Laurens van der Maaten and Geoffrey Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.