A Multi-view Auto Lip-reading
The Kip-reasder's (Progress Report)

Houjeung Han
School of Electrical Engineering
KAIST
Daejeon, Korea
Email: comc35@kaist.ac.kr

Jesper Loenbaek
School of Computing
KAIST
Daejeon, Korea
Email: lonbak@kaist.ac.kr

Youngsoo Jang
School of Computing
KAIST
Daejeon, Korea
Email: jys5609@kaist.ac.kr

Abstract—**Auto lip-reading (ALR) is a promising method for enhancing speech recognition, by combining the audio input with a visual. ALR is usually composed by a feature extraction part and a classification part. The feature extraction part is here often complicated by large variety in mouth shape and visual pronunciation. In this work we evaluate the usage of a combined method of the classical approach to ALR where handcrafted feature extraction is used, with the more modern approach where end-to-end neural network is used. In particular we investigate the usage of optical flow and key feature extraction in combination with neural networks, such as convolutional networks and Long short-term memory. The performance of each of these systems is compared to the current state-of-the-art architecture.**

I. Introduction

Lip-reading is a technique to understand speech by visually interpreting the movement of the lips, face and thought. This technique is not limited to deaf or hear-of-hearing people but is also used by people which have a normal hearing process. A phanomena known as the McGurk effect[1] show this relation, where the interpretation of speech for the same sound is changed with the image. Just as people use lip-reading for speech recognition it is also seen to have its application in artificial intelligence, where a higher accuracy can be obtained by combining the acoustic and visual information [2]. In artificial intelligent the combination of acoustic and visual information is know as audio-visual speech recognition (AVSR) and system with only visual information is commonly known as automatic lip-readig (ALR) or visual speech recognition (VSR). ALR also have other promising applications, beside the combination with acoustic information, such as visual password, silent speech interface and forensic video analysis.

The main challenge in ALR is duo to a large variation in visual factors both from the recording such as changes in illumination and camera angle[3], but also from factors that is person specific such as mouth shape an visual pronunciation.

In order to address each of these challenges, different experimental setup is proposed for ALR such as: Speaker dependent (SD) or speaker independent (SI), single-view, cross-view or multi-view. SD is the simplest setting where the personal variation from the speaker is removed since data from one speaker is used both for the training and evaluation. In SI the variation from the speaker such as mouth shape and visual pronunciation is included where unseen speakers is used for the evaluation. Single-view focus one the usage with an single camera angle, cross-view use one camera angle for the training and another for the testing and multi-view is including multiple camera angles for both training and testing.

Previous work in ALR can in generally be grouped in to two, one with a classical approach and a more recent approach where deep neural networks is used.

In the classical approach methods from computer vision is used for the visual feature extraction. These methods involve methods such as optical flow[4] or key feature extraction[5].

Better performing systems is later obtained by the use of neural networks, where neural networks both have its application for both the feature extraction and the temporal correlation. In [2] a deep autoencoder is used for the feature extraction and a long short-term memory (LSTM) is used in [6] to make a temporal time correlation of the features. In [7] the combination of a convolutional neural network with a LSTM is presented and is the current state-of-the-art architecture.

In this work we propose a combined solution between the classical approach and the neural network approach, by using techniques from the classical approach for the feature extraction in combination with a CNN and LSTM. We limit our scope to focus on speaker independent and with single-view, cross-view and multi-view setting, where we use the OuluVS2 database[8] to evaluate our design, in relation to the error rate of word or phrase classification. This task is related to the challenges given at the ACCV 2016 workshop, multi-view lip-reading/audio-visual challenge[1] (MLAC 2016).

II. Background

A. Convolutional Neural Network

Convolutional neural networks (CNN) is usually used in many fields of computer vision. CNN is classically applied to classification and detection task. CNN is inspired by multi-layer perceptron containing small sub-regions of a visual field called receptive field [16]. Unlike fully connected layered network, CNN has sparse connectivity and shared weights for the purpose of increasing computational efficiency and

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

global representation power. CNN is now the most popular and effective selection for learning visual features in computer vision and machine learning fields. We obtain a feature map at layer h with input x pixel at coordinates (i, j) as the following equation:

$$h_{ij} = a((W \cdot x)_{ij} + b) \tag{1}$$

, where weight matrix W and bias vector b is the filter of this feature map, a is activation function for non-linearities.

B. Long Short-Term Memory

Among numerous methodologies, recurrent neural network (RNN) and its variants are now common in handling sequential data with their promise of performance and ease of use. The fundamental neural network does not consider the dependency of all inputs and outputs. However, in various tasks, there exist dependency in inputs and outputs, such as sentence analysis. RNN recurrently use the previous computation result to compute the current output. Long-short term memory (LSTM) [9], one of the most popular RNN variants that is able to capture long-range dependencies, is commonly adopted. LSTM is RNN with gates, which is proposed to prevent the vanishing gradient problem, becoming more effective in dealing with long sequences.

The basic structure of LSTM unit consists of a cell state with three essential gates: input gate, forget gate and output gate. The cell controls the information storing for a long period via gates. Given an input vector $\mathbf{x}_t$ at time step t, the formal equation for updating gates, output and cell state are defined as follows:

$$\mathbf{i}_t = \sigma\left(\mathbf{x}_t \mathbf{U}^i + \mathbf{h}_{t-1}\mathbf{W}^i\right)$$

$$\mathbf{f}_t = \sigma\left(\mathbf{x}_t \mathbf{U}^f + \mathbf{h}_{t-1}\mathbf{W}^f\right)$$

$$\mathbf{o}_t = \sigma\left(\mathbf{x}_t \mathbf{U}^o + \mathbf{h}_{t-1}\mathbf{W}^o\right)$$

$$\mathbf{c}_t = \mathbf{c}_{t-1} \circ \mathbf{f}_t + \mathbf{i}_t \circ \tanh\left(\mathbf{x}_t \mathbf{U}^c + \mathbf{h}_{t-1}\mathbf{W}^c\right)$$

$$\mathbf{h}_t = \tanh\left(c_t\right) \circ \mathbf{o}_t$$

where $\mathbf{W}^i, \mathbf{W}^f, \mathbf{W}^o, \mathbf{W}^c \in \mathbb{R}^{N \times N}$, $\mathbf{U}^i, \mathbf{U}^f, \mathbf{U}^o, \mathbf{U}^c \in \mathbb{R}^{N \times N}$ are weight matrices, $\mathbf{h}_t$ is output vector and i, f and o represent input (i), forget (f) and output (o) gates.

We are planning to apply the bi-directional LSTM, which considers the both directional(forward and backward) sequence of data. There are some reseach about outperforming result of bi-directional LSTM([10]).

C. Optical Flow

Optical flow estimation is one of the key problems in the computer vision. In the previous approaches, Horn and Schunk suggest the original optical flow([11]). After the original optical flow, many improvements are suggested as variational model optical flow. In our approach, we are planning to apply the optical flow to feature extraction. Optical flow represents the characteristic and it can be used as a featuer of the specific part in the image.

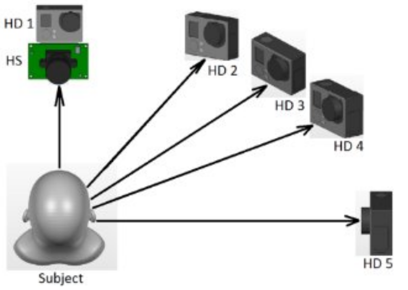

Fig. 1. Illustration of the multi-view setup used in the OuluVS2 recording

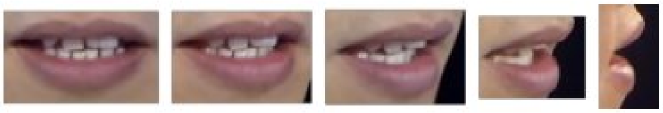

Fig. 2. Example of the preprocessed data with with the region of interest.

III. Dataset

To train and evaluate our approach we use the public available OuluVS2 database. The recorded environment used is illustrated in figure 1, where the multi-view setting can be seen. The setup make use of five cameras located at different angels in relation to the subject. For each recorded subject three different scenarios is performed, namely the pronunciation of: (1) a sequence of ten fixed digit sequences, (2) ten daily-use short English phrases and (3) five randomly selected TIMIT sentences. Examples of each of these can be seen in table I A total of 52 different test subject is used, where each subject is pronouncing each phrase three times.

A preprocessed version is available where the recording from different angles is synchronized and the region of interest (ROI) is both segmented, with the recording of interest, and cropped to only contain the mouth region. An example of the preprocessed data can be seen in figure 2.

IV. Method

In this section our approach to solve the Automatic Lipreading is presented. The high-level architecture is first introduced, where the main components and their functions is described. The particular architectures which we like to evaluate is then presented.

A. High-level Architecture

In figure 3 an illustration of the high-level architecture can be seen with the three main components; *Visual model*, *temporal model* and classifier. In the figure the flow of information

(1) Digits	”1 7 3 5 1 6 2 6 6 7” ”4 0 2 9 1 8 5 9 0 4”
(2) Phrases	”Thank you” ”Have a good time”
(3) TIMIT	”Chocolate and roses never fail as a romantic gift”

TABLE I
Examples of the three different scenarios used

can also be seen, from the sequence of images inputted to the architecture and to the outputted class probabilities. The *visual model* is used to extract features from the input image. These features is then passed on to the *temporal model* where they are correlated in time. The time correlated features is then inputted to the classifier which is outputting the different class probabilities. For each of the three main components, different implementations is of interest where each of these is listed in figure 3. Each of the different methods is presented in the following sections. We would here evaluate the performance where a single method from each component is used in combination, which give a total of 16 different architectures.

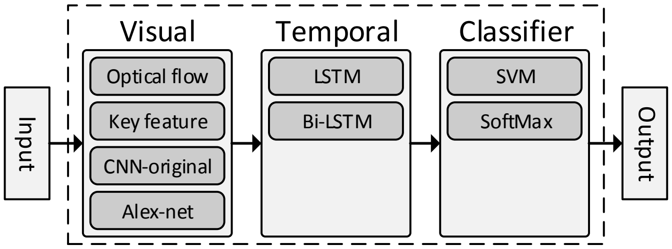

Fig. 3. High-level architecture used in this work, with the three main component and with the different method used for each of them.

B. Visual Model

For the feature extraction 4 different methods is of interest, where two of these make use of neural network's and the other two is computer vision related.

a) CNN-original: This visual model is corresponding to the one used in the state-of-the-art model, presented in [7]. An illustration of the model can be seen in figure 4. It make use of two convolutional layers with 16 to 256 filters in the shape of (3,3). Each convolutional layer has a successive max-pooling layer in the shape of (2,2). Lastly a fully connected layer with a dimension of 8 to 64 outputs is used.

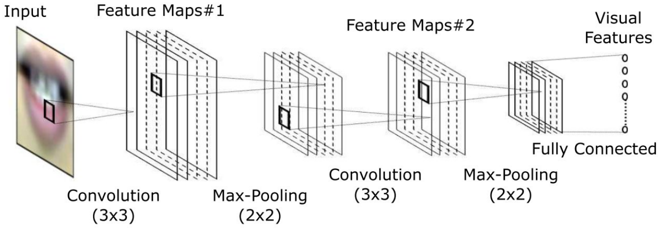

Fig. 4. CNN-original[7]

b) Alex-Net: A popular neural network architecture for image recognition is an Alex-net[12]. This architecture is a deep CNN as depicted in figure 5.

c) Optical flow: For the calculation of optical flow we expect the differential method commonly denoted as the Lucas-Kanade method[13]. The method have the assumption of the flow begin essentially constant in a local neighbourhood.

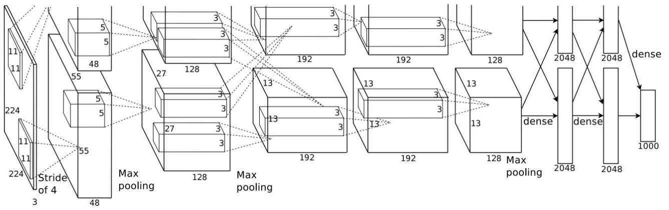

Fig. 5. Alex-net[12]

d) Key features: For the key feature extraction a method similar to the one presented in [5] is used. The mouth is divided in to four parts, lower and upper part of the upper lip and lower and upper part of the lower lip. A vertical grid is then made, where the coordinates of the intersection points is used. An illustration of this can be seen in figure 6.

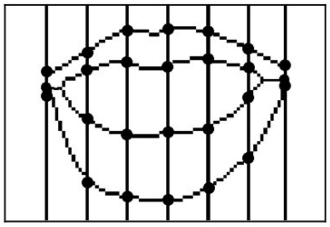

Fig. 6. Key Feature[5]

C. Temporal model

Two different temporal models is used, which is *LSTM* and *bidirectional-LSTM*. For both of these models a two layered design is used, where the size of each layer is here varying with the number of output features from the visual model.

D. Classifier

There are many kinds of classifier to determine the output. In our approach, we will apply two different classifier, namely a support vector machine (SVM) and softmax function. SVM classify the data points with some vectors that can divide whether correct or not. Softmax function normalizes the probabilities and reconstruct it with distinct difference.

V. Experiment and Result

A. Single-view Lip-Reading

Fig 7 shows the accuracy result of single - view on word phrase test data. The results are similar to the one presented in [7]. We have best accuracy on Profile view.

Fig 8 is Confusion matrix of profile view, which is the best in single view test accuracy. The x-axis is prediction classes and the y-axis is true classes. We can see that 'Thank you' and 'See you' is the most confusing word phrase.

B. Cross-view Lip-Reading

In Cross view experiment, we conduct it as two seperate stages. At first stage, as we refer it to 'Cross View(CV)', we train 5 total view data all together simultaneously and test each

	Our Results (Accuracy of Test Data)					
Training Data	(1)Frontal	(2)30°	(3)45°	(4)60°	(5)Profile	Average
(1)Frontal	81.1 %					
(2)30°		80 %				
(3)45°			76.9 %			77.9 %
(4)60°				69.2 %		
(5)Profile					82.2 %	

	Single-View Baseline Results† (Accuracy of Test Data)					
Method	(1)Frontal	(2)30°	(3)45°	(4)60°	(5)Profile	Average
DCT-PCA-HMM†	63%	62%	62%	63%	57%	61%
DCT-HiLDA-HMM†	74%	72%	73%	73%	68%	72%
RAW-PLVM†	73%	75%	76%	75%	70%	74%

Fig. 7. Single-view test accuray of word phrases.

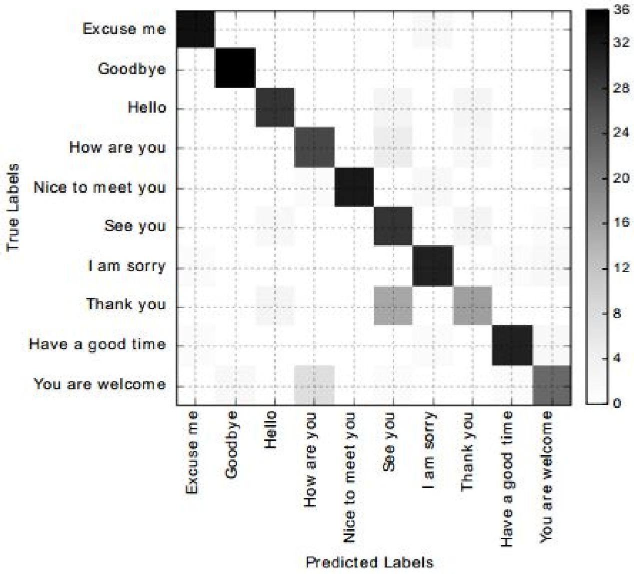

Fig. 8. Confusion matrix of best of Single Views

view seperately. In this approach, we get the average accuracy 82.6%, and all of the results are better than preliminary results. At Second stage, we call this stage as ”Cross-view2(CV2)”, and here, we finetune the first stage result to a certain view and also test with the certain view. Here, we can see slightly better performance on each view and also on average compare to CV.

		Accuracy of Test Data					
Training Data		(1)Frontal	(2)30°	(3)45°	(4)60°	(5)Profile	Average
CV	All	80.6 %	81.1 %	85 %	82.5 %	83.6 %	82.6 %
	All+(1)Frontal	82.8 %					
	All+(2)30°		81.1 %				
CV2	All+(3)45°			85 %			83.8 %
	All+(4)60°				83.6 %		
	All+(5)Profile					86.4 %	

Fig. 9. Cross-view test accuracy. CV : Cross View, CV2 : Cross View 2

In CV, Profile view shows the best performance(83.6%), and in CV2, 60-degree view shows the best performance on total Cross view experiment.

The confusion matrices of test accuracies of profile view(CV) and 60-degree view (CV2). The axis lable share the same label in fig 8. It shows that a gradual improvement from single-view to cross-view.

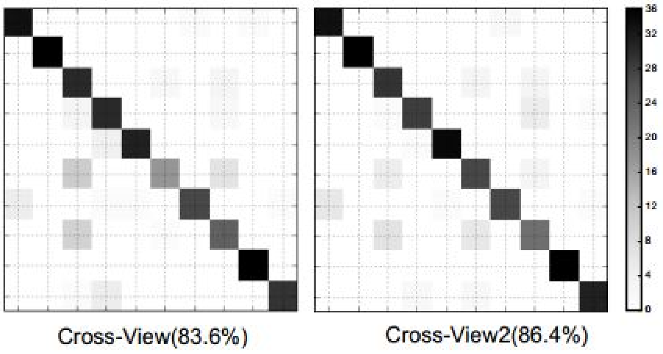

Fig. 10. Confusion matrices of best of Cross Views

C. Multi-view Lip-Reading

In Multi-view experiment, we conduct slightly different type of architecture as inputs are five times larger then single or cross view experimetn. Here, we conduct Merge Image method as this method of multi-view is best in our previous paper.

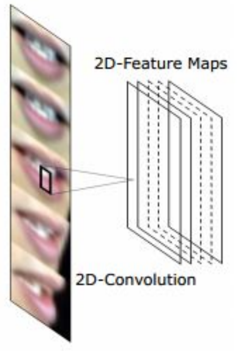

Fig. 11. Merge Image model architencture for the multi-view setting

Merge Images: As shown in fig 11, we append five images from the different view at the same time into a single image as an input of the visual model. In this architecture, we expect to learn all the five view feature by 2D-CNN. While out of our experiment, a more elaborate configuration is that all five images avoid convolving each other along the edges.

Despite it has more feature (give times more inputs), the result is worse than cross-view. Fig 12 shows the summary of all experiment through this paper.

VI. Conclusion

The conclusion goes here.

VII. Current Project State and Future

We have currently been able to redo the current state-of-the-art results, by evaluating the architecture explained in [7]. From this we will expand upon this architecture and implement together with evaluate the proposed architecture explain in section IV.

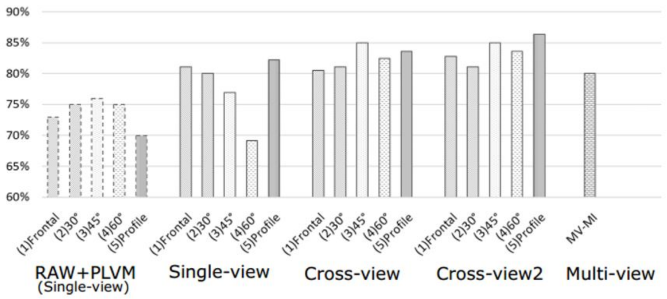

Fig. 12. The summary of the accuraciese. MV-MI is refers to Merge Images.

References

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

[2] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal Deep Learning," *Proceedings of The 28th International Conference on Machine Learning (ICML)*, pp. 689–696, 2011.

[3] G. Potamianos and C. Neti, "Audio-visual speech recognition in challenging environments." *Interspeech*, pp. 1293–1296, 2003.

[4] A. A. Shaikh, D. K. Kumar, W. C. Yau, and J. Gubbi, "Lip reading using optical flow and support vector machines," *3rd International Congress on Image and Signal Processing*, pp. 327–330, 2010.

[5] M. Li and Y. M. Cheung, "A novel motion based lip feature extraction for lip-reading," *Proceedings - 2008 International Conference on Computational Intelligence and Security, CIS 2008*, vol. 1, pp. 361–365, 2008.

[6] M. Wand, J. Koutnik, and J. Schmidhuber, "Lipreading with long short-term memory," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, mar 2016, pp. 6115–6119.

[7] D. Lee, J. Lee, and K.-e. Kim, "Multi-View Automatic Lip-Reading using Neural Network," pp. 1–14.

[8] I. Anina, Z. Zhou, G. Zhao, and M. Pietikainen, "OuluVS2: A multi-view audiovisual database for non-rigid mouth motion analysis," *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, FG 2015*, 2015.

[9] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[10] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional lstm and other neural network architectures," *Neural Networks*, pp. 5–6, 2005.

[11] B. Horn and B. Schunck, "Determining optical flow: A retrospective," *Artificial Intelligence*, vol. 59, no. 1-2, pp. 81–87, 1993.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Advances In Neural Information Processing Systems*, pp. 1–9, 2012.

[13] B. D. Lucas and T. Kanade, "An Iterative Image Registration Technique with an Application to Stereo Vision," *Imaging*, vol. 130, no. x, pp. 674–679, 1981.