15th IAPR International Conference on Machine Vision Applications (MVA)
Nagoya University, Nagoya, Japan, May 8-12, 2017.

Two Features Combination with Gated Recurrent Unit for Visual Speech Recognition

Masaya Iwasaki, Michiko Kubokawa, Takeshi Saitoh
Kyushu Institute of Technology
680-4 Kawazu, Iizuka, Fukuoka, Japan
saitoh@ces.kyutech.ac.jp

Abstract

Recently, deep learning has attracted in the visual speech recognition (VSR) field, however, most of approaches use single feature. This paper proposes two features; image-based feature by autoencoder, and motion-based feature based on feature points, and combination approach with gated recurrent unit for VSR task. The proposed method was evaluated on a public dataset OuluVS, and the speaker independent setting was carried out. Compared with the state-of-the-art, our method obtained the highest accuracy.

1 Introduction

Recently, deep learning techniques have been successfully applied to learn features from audio-visual data for the tasks of visual speech recognition (VSR) and audio-visual speech recognition (AVSR). Ngiam et al. presented a series of tasks for multi-modal learning and showed how to train deep networks that learn features to address these tasks [1]. Hu et al. proposed a temporal multi-modal network called Recurrent Temporal Multimodal Restricted Boltzmann Machines to model audio-visual sequence in an unsupervised fashion [2]. Noda et al. proposed to apply a convolutional neural network (CNN) as the visual feature extraction mechanism for VSR [3]. Hidden Markov Models (HMM) with Gaussian mixtures were used for a task of recognizing isolated word. Takashima et al. proposed a bottleneck feature extracted from audio-visual features [4]. Most of the above mentioned methods targeted the problem of AVSR. Only [3] tackled the problem of VSR. Note that in their method, CNNs were used for visual extraction features and the classification was conducted by HMMs.

This paper defines two features; image-based feature and motion-based feature, and proposes two features combination with Gated Recurrent Unit for VSR task. We evaluate the proposed method on the public-available OuluVS database that contains short phrases and the results are significantly better than the best reported performance in speaker independent experiments.

2 Two Features

In the field of VSR tasks, the visual feature can roughly group into four categories: image-based, motion-based, geometric-feature-based, and model-based [5]. Most of VSR tasks use single feature. In this paper, we propose image-based feature and motion-based feature.

Table 1. Structure of autoencoder.

layer	N_C	S_F	S_D	S_U	S_O
input	1	—	—	—	64×64
conv1	16	3×3	2×2	—	32×32
conv2	8	3×3	2×2	—	16×16
conv3	4	3×3	2×2	—	8×8
conv4	4	3×3	—	—	8×8
conv5	4	3×3	—	2×2	8×8
conv6	8	3×3	—	2×2	16×16
conv7	16	3×3	—	2×2	32×32
output	1	3×3	—	—	64×64

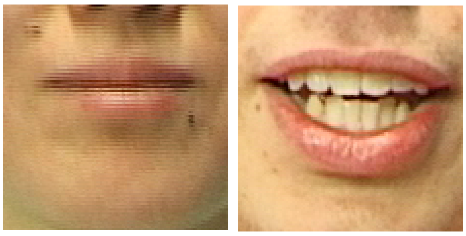

Figure 1. Sample mouth image.

2.1 Image-based feature

As for the traditional image-based features [6, 7], a gray-scale image is either used directly or after some image transformation, such as PCA and DCT, as a feature vector. In this research, an autoencoder is used for calculating image-based feature.

The autoencoder neural network is an unsupervised learning algorithm that applies back-propagation, setting the target values $\hat{x}$ to be equal to the inputs x. This is data-specific, lossy, and learned automatically from data examples. This is one of the data compression algorithm.

Our autoencoder (AE) architecture is shown in Table 1. Input image of AE is a gray-scale ROI image around the lip as shown in Fig. 1, because our objective is VSR. The size of input image is set to 64×64 [pixel]. In the table, N_C is a number of channel, S_F is a filter size, S_D is a pooling (down-sampling) filter size, S_U is an up-sampling filter size, and S_O is an output size. Here, max pooling is applied as pooling process. The output of conv3, that is 256 dimensional vector, is used as the feature for recognition process.

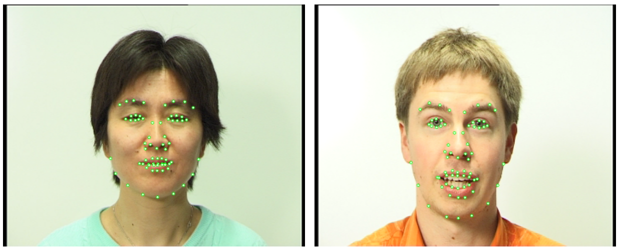

Figure 2. Facial feature points.

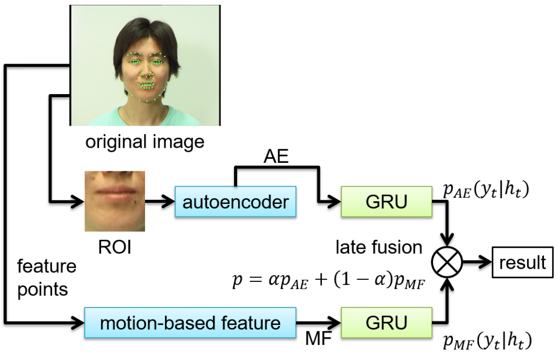

Figure 3. Processing pipeline of two features combination.

2.2 Motion-based feature

Image-based feature can contain not only lip appearance but also information of the tooth and tongue. However, this is influenced by the color difference based on the gender, race, and lighting condition. In this research, the motion-based feature is used to counter the aforementioned problem.

The typical motion-based feature is optical flow [8]. However, the optical flow is strongly sensitive to head motion, even though it has little effect in VSR task. Thus, this paper defines a new motion-based feature based on the facial feature points as shown in Fig. 2.

A subtraction value between current frame and next frame at each feature point is defined as a motion-based feature MF by the following equation:

$$d_*(i,f) = P_*(i,f) - P_*(i,f+1),$$

where i is a feature point number, f means a frame number, and $P_*(i,f)$ is a coordinate of i-th feature point. A symbol $*$ is either x or y. This feature is not a distance value; it has negative or positive value.

Before calculating MF, a normalization process is applied to the feature points. The location of the feature points is depended on the distance between the speaker and camera, and this factor needs to remove the feature. Here, a distance between both eyes is fixed, and this is not changed by the speech and facial expression. On the other hand, the distance between two lip corners or other distance are variously changed by the facial motion. Then, the distance between both two eyes is scaled to 100 pixels. Moreover, the rotation process is applied to reduce the facial angle.

3 Gated recurrent unit

Gated recurrent units (GRUs) [9] are a gating mechanism in recurrent neural networks (RNNs). A GRU has two gates, a reset gate, and an update gate. Intuitively, the former determines how to combine the new input with the previous memory, and the latter defines how much of the previous memory to keep around. If we set the reset to all 1's and update gate to all 0's, we again arrive at plain RNN model. The basic idea of using a gating mechanism to learn long-term dependencies is the same as in a Long Short-Term Memory (LSTM), but there are a few key differences: (1) A GRU has two gates, an LSTM has three gates. (2) GRUs do not possess and internal memory that is different from the exposed hidden state. They do not have the output gate that is present in LSTMs. (3) The input and forget gates are coupled by an update gate and the reset gate is applied directly to the previous hidden state. The responsibility of the reset gate in a LSTM is really split up into both reset gate and update gate. (4) We do not apply a second nonlinearity when computing the output.

In the preliminary experiment, we used both LSTM and GRU as the recognition process, and found GRU obtained higher accuracy than LSTM. Thus, in this paper, GRU is used as the recognition process.

4 Algorithm summary

In this paper, we proposed two features of AE and MF, and these features are combined. There are several fusion schemes, input fusion, late fusion, feature fusion [10]. The proposed method uses separate GRU flows for different feature channels and apply a late decision level fusion.

The proposed processing pipeline is shown in Fig. 3. An output score p is calculated by

$$p = \alpha p_{AE} + (1-\alpha)p_{MF},$$

where p_{AE} and p_{MF} are scores of GRUs, and α is a weight value.

5 Experiment

5.1 Dataset

There are many databases available for VSR, such as AVLetters [11], CUAVE [12], Grid [13], OuluVS [14], and OuluVS2 [15]. The utterance content of each database is different, and a number of speakers is also different. We trained and tested the proposed method on OuluVS, since this database is used by several researchers and it is easy to compare with other methods.

OuluVS consists 20 subjects (17 males and three females) uttering 10 daily-use short phrases (p01: "excuse me", p02: "goodbye", p03: "hello", p04: "how are you", p05: "nice to meet you", p06: "see you", p07: "I am sorry", p08: "thank you", p09: "have a good time", and p10: "you are welcome") five times. The frame rate is 25 fps and its image resolution is 720 × 576 [pixels].

Figure 4. OuluVS database.

5.2 Experimental Protocol

There are some test protocols for VSR: speaker dependent (SD), speaker semi-dependent (SSD), and speaker independent (SI).

The SD experiment was conducted to recognize the phrases for the same speaker. In this experiment, one sample of each phrase is used as the testing set, and the remaining samples for the same speaker are considered as the training set. The same procedure was repeated for each phrase sample. The SSD experiment was conducted to recognize the phrases regardless of the user's identity. In this experiment, one sample of each phrase from all speakers is used as the testing set, and the remaining samples were considered as the training set. The same procedure was repeated for each phrase sample. The SI experiment was conducted to recognize the phrases completely independent of the speakers. In this experiment, all samples from one speaker were taken as the testing set, and the remaining samples from the other speakers were considered as the training set. The same procedure was repeated for each speaker.

The most challenge task is SI, in the experiment, we applied the leave-one-person-out cross validation.

5.3 Preprocessing

The proposed method is required feature points. We first applied an active appearance model (AAM) to detect facial feature points. This is a local search method that combines a full shape model and texture variation learnt from a training set [16]. In the experiment, we built a whole face model which contained eight eye points, four eyebrow points, 11 nose points, 12 external lip contour points, 12 internal lip contour points, and nine face outline points, as shown in Fig. 2. The number of feature points in this model is 68.

After detecting feature points by AAM, a mouth ROI is extracted based on feature point for applying autoencoder. The mouth ROI is defined as shown in Fig. 5. The size of this ROI is $0.8S \times 0.8S$ [pixel] where S is a distance between both eye corners. By considering mouth movement during utterance, the base point of the ROI is not the center of ROI but a little above the center. This ROI is fed to the autoencoder described in 2.1.

Detected feature points on the nose and lip contour points, total is 35 points, are fed to the feature calculation process described in 2.2.

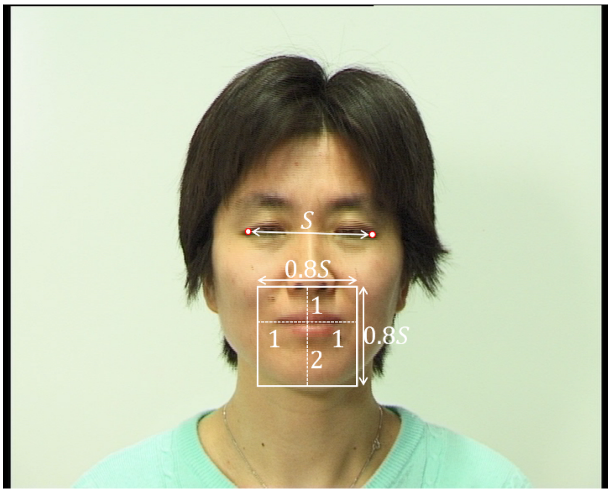

Figure 5. Mouth ROI extraction.

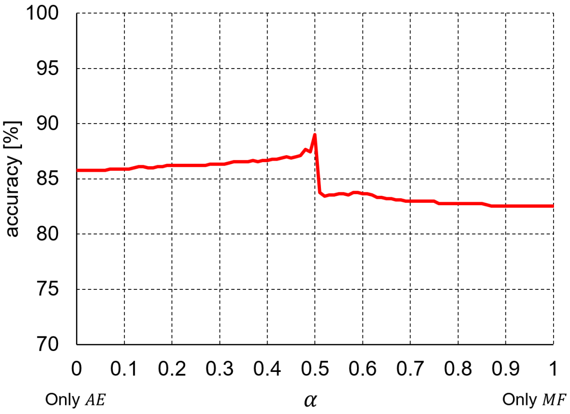

Figure 6. Recognition rates against α.

5.4 Experimental result

Experimental result of the proposed method is shown in Fig. 6. This graph shows the variation in accuracy as a function of α. When only one feature was used, AE ($\alpha = 0$) was obtained higher performance than MF ($\alpha = 1$). It can be seen that the highest performance of 89.0% was obtained when $\alpha = 0.5$. This indicates that two features combination is effective for VSR task.

Next, our proposed method was compared with the state-of-the-art method on the OuluVS dataset. Table 2 shows the results of the five different VSR approaches tested on OuluVS. The accuracy of each approach is cited from literature. It can be found that our method achieves significantly higher recognition accuracy than the other approaches.

To analysis recognition results, we calculated a confusion matrix (CM) as shown in Table 3. The confusion matrix contains information about actual and predicted categories down by the recognition task. The squares along the diagonal indicate the rate of correct recognition, whereas the squares off the diagonal indicate the rate of incorrect recognition. Several interesting results can be obtained from the CM. For example, it can be found that p01 (excuse me) and p06 (see you)

Table 2. Performance comparison with different VSR approach tested on OuluVS.

method	accuracy [%]
LBP-TOP/SVM [14]	62.4
LBP-TOP/LVM [17]	85.6
LBP/KPLS [18]	62.34
PLSD/KELM [19]	68.75
ours (AE+MF/GRU)	89.0

Table 3. Confusion matrix.

	p01	p02	p03	p04	p05	p06	p07	p08	p09	p10
p01	**73.3**	1.1	1.1	7.8	0.0	16.7	0.0	0.0	0.0	0.0
p02	0.0	**94.4**	0.0	0.0	2.2	0.0	2.2	0.0	1.1	0.0
p03	0.0	0.0	**96.7**	0.0	0.0	1.1	0.0	0.0	0.0	2.2
p04	12.2	1.1	1.1	**70.0**	0.0	13.3	0.0	0.0	2.2	0.0
p05	0.0	4.4	0.0	0.0	**95.6**	0.0	0.0	0.0	0.0	0.0
p06	15.6	1.1	0.0	4.4	0.0	**77.8**	1.1	0.0	0.0	0.0
p07	0.0	0.0	1.1	0.0	0.0	0.0	**98.9**	0.0	0.0	0.0
p08	0.0	2.2	0.0	0.0	4.4	0.0	0.0	**92.2**	1.1	0.0
p09	0.0	0.0	0.0	0.0	0.0	1.1	1.1	1.1	**96.7**	0.0
p10	0.0	0.0	2.2	0.0	0.0	2.2	1.1	0.0	0.0	**94.4**

is easy to be confused each other, p04 (how are you) is easy to be misrecognized to p01 and p06.

6 Conclusion

In this paper, we have proposed a novel two features combination with GRU for VSR task. The proposed method was evaluated on a public dataset OuluVS. In the experiments, the speaker independent setting was carried out. Compared with the state-of-the-art, our method obtained the highest accuracy in SI task.

Our future plan is to add further experiments with other datasets. Furthermore, we consider to research not only isolated phrases but also the continuous speech to improve the quality of VSR.

Acknowledgement

This work was supported by JSPS KAKENHI Grant Number 15K12601 and 16H03211.

References

[1] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng: "Multimodal deep learning." *28th International Conference on Machine Learning (ICML)*, pp.689–696, 2011.

[2] D. Hu, X. Li, and X. Lu: "Temporal multimodal learning in audiovisual speech recognition." *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.3574–3582, 2016.

[3] K. Noda, Y. Yamaguchi, K. Nakadai, H. G. Okuno, and T. Ogata: "Lipreading using convolutional neural network," *INTERSPEECH*, pp.1149–1153, 2014.

[4] Y. Takashima, Y. Kakihara, R. Aihara, T. Takiguchi, Y. Araki, N. Mitani, K. Omori, and K. Nakazono: "Audio-visual speech recognition using convolutive bottleneck networks for a person with severe hearing loss," *IPSJ Trans. on Computer Vision and Applications*, vol.7, pp.64–68, 2015.

[5] Z. Zhou, G. Zhao, X. Hong, and M. Pietikainen: "A review of recent advances in visual speech decoding," *Image and Vision Computing*, vol.32, pp.590–605, 2014.

[6] C. Bregler and Y. Konig: "Eigenlips for robust speech recognition," *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP1994)*, 1994.

[7] P. J. Lucey, G. Potamianos, S. Sridharan: "Patch-based analysis of visual speech from multiple views," *International Conference on Auditory-Visual Speech Processing (AVSP2008)*, 2008.

[8] J. Shiraishi and T. Saitoh: "Optical flow based lip reading using non rectangular ROI and head motion reduction," *11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, 2015.

[9] J. Chung, C. Gulcehre, K.-H. Cho, and Y. Bengio: "Empirical evaluation of gated recurrent neural networks on sequence modeling," arXiv preprint arXiv: 1412.3555, 2014.

[10] C. Zhang and Y. Tian: "Automatic video description generation via LSTM with joint two-stream encoding," *International Conference on Pattern Recognition*, 2016

[11] I. Matthews, T.Cootes, J. Bangham, S. Cox, and R. Harvey: "Extraction of visual features for lipreading," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol.24, no.2, pp.198–213, 2002.

[12] E. K. Patterson, S. Gurbuz, Z. Tufekci, and J. N. Gowdy: "Moving-talker, speaker-independent feature study, and baseline results using the CUAVE multimodal speech corpus," *EURASIP Journal on Applied Signal Processing*, vol.2002, no.11, pp.1189–1201, 2002.

[13] M. Cooke, J. Barker, S. Cunningham, and X. Shao: "An audio-visual corpus for speech perception and automatic speech recognition," *Journal of the Acoustical Society of America*, vol.120, no.5, pp.2421–2424, 2006.

[14] G. Zhao, M. Barnard, and M. Pietikainen: "Lipreading with local spatiotemporal descriptors," *IEEE Trans. on Multimedia*, vol.11, no.7, pp.1254–1265, 2009.

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

[16] T. F. Cootes, G. J. Edwards, and C. J. Taylor: "Active appearance models." *European Conference on Computer Vision*, no.2, pp.484–498, 1998.

[17] Z. Zhou, G. Zhao, and M. Pietikainen: "Unsupervised random forest manifold alignment for lipreading," *International Conference on Computer Vision (ICCV)*, 2013.

[18] A. Bakry and A. Elgammal: "MKPLS: Manifold kernel partial least squares for lipreading and speaker identification," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.684–691, 2013.

[19] L. Lu, X. Zhang, X. Xu, and D. Shang: "Video analysis using spatiotemporal descriptor and kernel extreme learning machine for lip reading," *Journal of Electronic Imaging*, vol.24, no.5, 2015.