IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 26, 2019, accepted June 30, 2019, date of publication July 8, 2019, date of current version July 24, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2927166

Lip Reading Using Committee Networks With Two Different Types of Concatenated Frame Images

DONG-WON JANG, HONG-IN KIM, CHANGSOO JE, RAE-HONG PARK, (Senior Member, IEEE), AND HYUNG-MIN PARK, (Senior Member, IEEE)
Department of Electronic Engineering, Sogang University, Seoul 04107, South Korea

Corresponding authors: Rae-Hong Park (rhpark@sogang.ac.kr) and Hyung-Min Park (hpark@sogang.ac.kr)

This work was supported in part by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) under Grant NRF-2017R1A2B4009964, and in part by the BK21 Plus Program.

ABSTRACT This paper proposes a lip-reading method based on convolutional neural networks (CNNs) applied to two different types of concatenated frame images (CFIs), consisting of (a) full-lip images and (b) patches around lip landmarks. In addition, we introduce committee networks with the predictions obtained from the two different types of the CFIs, which provide better performance than single or committee networks using either type of the CFIs. For efficient training using a limited dataset, such as OuluVS2, we propose time-based label-preserving transform and use a quarter VGG-m in which the number of parameters is reduced compared to the VGG-m. The experimental results with the OuluVS2 dataset show that the proposed method using different types of the CFIs in committee networks outperformed the state-of-the-art methods without pre-training using a large-scale dataset.

INDEX TERMS Committee networks, concatenated frame images, convolutional neural networks, lip reading, time-based label-preserving transform, visual speech recognition.

I. INTRODUCTION

Recently, speech recognition is widely used as an effective interface in various commercial systems in moderately clean environments. However, recognition performance in noisy environments is considerably degraded because of differences between training and testing environments. Because visual information is not distorted by acoustic noise, lip reading may play an important role in automatic speech recognition in the acoustically adverse environments [1]. In contrast to features for acoustic speech recognition, visual features for lip reading are not well-established yet [1]. Conventionally, following visual features have been commonly used: local binary patterns extracted from three orthogonal planes (LBP-TOP) [2], histogram of oriented gradients (HOG) [3], discrete cosine transform (DCT) [4], and so on. Because these are rather general-purpose image features, the performance of lip reading may be improved by devising effective features.

Recently, deep learning has achieved impressive success in diverse object detection and recognition tasks [5], and, in particular, convolutional neural networks (CNNs) have been applied to visual speech recognition [6]–[10]. However, CNNs have some challenges for video recognition. Most CNNs have been trained using a single image, whereas video recognition, such as lip reading and action recognition, uses multiple frames and thus requires architectures different from conventional CNNs. Chung and Zisserman [7] reported that the CNN with multiple towers was better for lip reading than three-dimensional (3-D) CNNs which calculated the features across the whole phrase, not from a single frame. Saitoh *et al*. method [8] computed the features of the whole phrase using concatenated frame images (CFIs). Because CFIs convert video sequence to a single image, conventional CNNs such as AlexNet, VGG, and GoogleNet can work well for lip reading. More recently, long short-term memory (LSTM) worked well for speech recognition [9]–[12], but, to the best of our knowledge, the LSTM for speech recognition commonly needed a large scale dataset [9] or an additional training dataset [10] except for Lee *et al*. method [11] and Fung and Mak method [12]. Furthermore, it is well known that CNNs are highly sensitive to local structure [13], and thus the trained networks are dependent on their input representation. Nowadays, it is common to use committees [14] or ensembles [15] for improving the recognition performance.

The associate editor coordinating the review of this manuscript and approving it for publication was Junchi Yan.

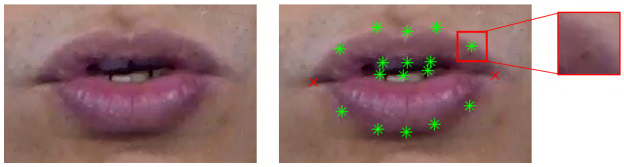


FIGURE 1. Full-lip image, its lip landmarks, and patch image.

We propose a method to apply CNNs to two different types of CFIs. The first type of CFI D^1 is composed of full-lip images, which is similar to Saitoh *et al*. CFI [8]. The second type of CFI D^2 is composed of patches around lip landmarks which are obtained by face alignment [16]. Multiple CNNs are trained by the two types of CFIs independently. The predictions obtained from multiple CNNs are averaged to evaluate the performance in a manner of the committee [14] during test phase. In experimental results using OuluVS2 dataset [17], the committee with different input representations of CFIs provided better accuracy than that with the same input representation.

Our contributions are as follows:

- We propose the patch-image-based CFI which results in better recognition accuracy than full-lip-image-based CFI with less computation using a quarter VGG-m with fewer trainable parameters.
- The proposed committee using different types of CFIs outperforms single or committee networks using either type of CFIs and the state-of-the-art methods trained on the OuluVS2 dataset only.
- We propose time-based label-preserving transform (TLPT) which converts an utterance video to a single image for length normalization and data augmentation.

II. PROPOSED LIP READING METHOD

As mentioned in Section I, we use two different types of CFIs for each utterance video. Generally, since the lengths of utterances are different, length normalization is needed to construct the CFIs; we propose TLPT that is simple and effective for length normalization and data augmentation.

A. TWO DIFFERENT TYPES OF THE CFIS

Fig. 1 shows a full-lip image and its lip landmarks. Also, an example of an enlarged patch-image centered at upper right lip landmark is shown. Although the face landmark detection in [16] generates 18 lip landmarks, we used only 16 landmarks to make patch-image-based CFI. Depending on speakers, the location of the lip landmarks may not be obtained correctly. Especially, the variation of patch images around the left and right corner landmarks is relatively large, which may result in performance degradation when trained on a small dataset such as OuluVS2 (see Section III-A).

Fig. 2 shows the full-lip-image-based CFI type D^1 and the patch-image-based CFI type D^2. The CFIs with the fixed

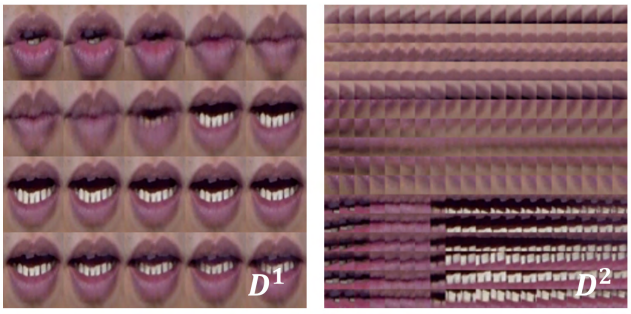


FIGURE 2. Two different types of CFIs ($L = 20$, $K = 5$, and $J = 16$).

number of frames, L, can be generated as

$$D^1 = \begin{bmatrix} I_{\phi(1)} & \cdots & I_{\phi(K)} \\ \vdots & \ddots & \vdots \\ I_{\phi(L-K+1)} & \cdots & I_{\phi(L)} \end{bmatrix},$$
$$D^2 = \begin{bmatrix} P^1_{\phi(1)} & \cdots & P^1_{\phi(L)} \\ \vdots & \ddots & \vdots \\ P^J_{\phi(1)} & \cdots & P^J_{\phi(L)} \end{bmatrix} \quad (1)$$

where I_t and K denotes the image at the t-th frame and the number of images in the horizontal direction of the representation D^1, respectively. The number of lip landmarks J was set to 16. Since most CNNs usually use square-like images, K and L were set to 5 and 20, respectively (see Fig. 2). A small patch P^j_t indicates patch image centered at the j-th lip landmark in the t-th frame I_t. The size of each patch P^j_t was set to 39×39 for the OuluVS2 dataset. The frame index $\phi(l)$ is computed by the proposed TLPT, which will be explained in Section II-B. The CFIs were resized to 224×224 when used as inputs to our networks.

B. TIME-BASED LABEL-PRESERVING TRANSFORM

We propose the TLPT to perform length normalization and data augmentation. Although training CNNs generally requires a large number of training data, datasets for visual speech recognition frequently have a small training set per class.

The TLPT consists of two steps: generating random sections and normalizing the length of utterance. First, as in label-preserving transform [18], many random sections are generated from an utterance video. To obtain the random sections, a whole utterance or a section obtained by a simple energy-based voice activity detection (VAD) can be used as a reference section. In this paper, the whole utterance was used as the reference section not to be affected by the performance of the VAD used. The random integers α and β, which determine the start and end frames of a random section, respectively, are selected depending on the number of frames of the reference section L_R, i.e., $\alpha, \beta \in [-\lambda L_R, \lambda L_R]$ where the scale parameter λ is set to 0.1. For example, for the reference sections with 10 to 19 frames, α and β are set to -1, 0, or 1. The TLPT produces random sections using all

possible pairs of α and β values. Since the number of frames for each utterance is different, it should be resampled to a fixed number L. Let R_s and R_e denote the start and end frame indices of the reference section, respectively. Then, the start and end frame indices of a random section are expressed as $s = R_s + \alpha$ and $e = R_e + \beta$, respectively. Note that frame indices are integers. With the total number of frames of the random section equal to $T = e - s + 1$, the frame index $\phi(l)$ in CFIs is computed using the nearest-neighbor interpolation as

$$\phi(l) = s + \left\lceil \frac{lT}{L} \right\rceil - 1, \quad 1 \le l \le L, \tag{2}$$

where $\lceil x \rceil$ represents the least integer greater than or equal to x. If a frame index obtained by (2) is out of range of the utterance, it is clipped to be within the range.

Algorithm 1 summarizes the described method to generate the two proposed CFIs with the proposed TLPT. A demo of the TLPT is available online (https://github.com/dong-won-jang/TLPT).

Algorithm 1 Pseudocode of the CFI Generation With the TLPT

input : Image sequence
output: CFIs

foreach *possible pair of* α *and* β **do**
  Assign the start and end frame indices $s \leftarrow R_s + \alpha, e \leftarrow R_e + \beta, T \leftarrow e - s + 1$.
  Make the frame index $\boldsymbol{\phi}$ using (2)
  if $\boldsymbol{\phi}$ *is out of range of the utterance* **then**
    Clip $\boldsymbol{\phi}$ to be within the range
  end
  Construct two different CFIs using (1)
end

C. NETWORK ARCHITECTURE

Most CNNs were designed for the ImageNet dataset [19], which consists of a million of images. In the ImageNet dataset, 4096 neurons were used to classify 1000 classes, whereas, in our case, 10 classes in the OuluVS2 dataset should be classified [17]. Since too many neurons in the fully-connected (FC) layers may cause an overfitting problem, the number of neurons was set to 128 in our experiments; and thus we reduced the number of kernels of the VGG-m model [20] to a quarter. In the rest of the paper, we call it the quarter VGG-m model.

Table 1 shows the details of the quarter VGG-m model. The spatial sizes of kernels, strides, and paddings are the same as the conventional VGG-m model. The only difference is the number of kernels in the convolutional layers and the number of neurons in the FC layers. The first two convolutional layers were trained by local response normalization [18].

Fig. 3 shows 24 kernels at the first convolutional layer of the quarter VGG-m model for two different types of CFI representations trained on the OuluVS2 dataset. As shown in Fig. 2, D^2 was obtained by concatenation of smaller patches than D^1 so that trained kernels for D^2 included kernels to consider contrasts between adjacent small patches compared with those for D^1. Therefore, features, obtained by the kernels from D^1 and D^2 shown in Fig. 3, may extract different aspects of information useful for lip reading with committee networks.

Committee networks [14] are introduced to make a decision for classification with predictions obtained from multiple CNNs. Using the predictions corresponding to CNN outputs, the committee networks compute the arithmetic mean of the predictions for final decision. After averaged, an index of the largest prediction value is selected as final decision. In contrast to the previous committees using multiple CNNs with the same input representation [14], we use multiple CNNs with the two different representations of an utterance video as the committee members.

TABLE 1. Details of the quarter VGG-m model (from left to right: name, kernel size, stride/padding, numbers of input/output channels, input data size, output data size, and name of the input layer).

Name	Ker.	Str./Pad.	Ch I/O	Input size	Output size	Input
Conv1	7×7	2/0	3/24	224×224	109×109	Image
Pool1	3×3	2/1	24/24	109×109	54×54	Conv1
Conv2	5×5	2/1	24/64	54×54	26×26	Pool1
Pool2	3×3	2/1	64/64	26×26	13×13	Conv2
Conv3	3×3	1/1	64/128	13×13	13×13	Pool2
Conv4	3×3	1/1	128/128	13×13	13×13	Conv3
Conv5	3×3	1/1	128/128	13×13	13×13	Conv4
Pool3	3×3	2/0	128/128	13×13	6×6	Conv5
Fc6	6×6	1/0	128/128	6×6	1×1	Pool3
Fc7	1×1	1/0	128/128	1×1	1×1	Fc6
Softmax	1×1	1/0	128/10	1×1	1×1	Fc7

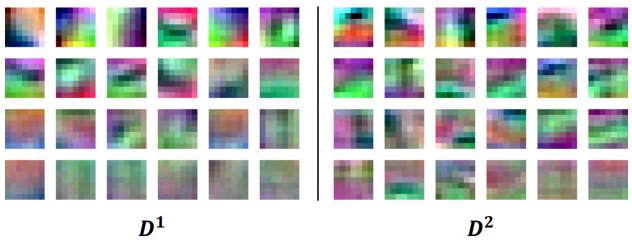


FIGURE 3. Visualization of 24 kernels at the first convolutional layer of the quarter VGG-m model for two different types of CFI representations. Red, green, and blue indicate the weights of respective color channels. The kernels were sorted by the L_1 norm in descending order from top left to bottom right.

III. EXPERIMENTAL RESULTS AND DISCUSSIONS

We have evaluated the proposed lip reading method with the OuluVS2 dataset [17]. The OuluVS2 dataset consists

TABLE 2. Averaged recognition accuracies for single networks.

Method	D^1			D^2		
	VGG	VGG128	QVGG	VGG	VGG128	QVGG
Acc. (%)	87.3	87.4	86.9	83.6	84.6	83.9

of 10 phrases, 10 digits, and 5 TIMIT sentences uttered by 52 people (40 people for training and 12 people for test). In this paper, we compared the phrase data only, where 10 phrases were uttered three times for each speaker. The phrases are ''Excuse me'', ''Good bye'', ''Hello'', ''How are you'', ''Nice to meet you'', ''See you'', ''I am sorry'', ''Thank you'', ''Have a good time'', and ''You are welcome''.

Implementation details to train CNNs are as follows: the quarter VGG-m model was used to construct CNNs for lip reading using the proposed CFI representations of an utterance video. The drop-out ratio was set to 0.5. We used the stochastic-gradient-descent method for optimization; the learning rate was initially set to 0.01, and was divided by 10 at 30 epochs. Training was terminated after 100-epoch learning. The weight decay and momentum were set to 0.0005 and 0.9, respectively. The network was initialized by zero-mean Gaussian noise with 0.01 variance. The proposed method was implemented in MatConvNet [21]. The batch size was set to 128.

In training phase, data augmentation was performed as follows: gamma correction [8], per-pixel mean subtraction, and addition of a Gaussian noise with zero mean and unit variance. We did not use random cropping or flipping. The gamma value was randomly selected from {0.6, 0.8, 1, 1.2, 1.4} which is the same as in [8].

A. ABLATION STUDY FOR SINGLE NETWORKS

Table 2 shows averaged recognition accuracies depending on network types: VGG-m model (VGG), VGG-m with 128 neurons (VGG128), and quarter VGG-m (QVGG). Each model was trained with five differently initialized networks. The VGG-m consists of 64 kernels at the first convolutional layer and the number of kernels increases in the later convolutional layers. The number of neurons in FC layers is set to 4096. The VGG-m with 128 neurons reduced the number of neurons in FC layers from 4096 to 128. In this case, the number of kernels in convolutional layers was the same as that of the VGG-m. The quarter VGG-m reduced the number of kernels at the first convolutional layer from 64 to 24. The increasing ratio of the number of kernels was the same as used in the VGG-m. The number of neurons of the quarter VGG-m in FC layers was set to 128.

Although the quarter VGG-m was not the best among the three models, all of these models showed similar accuracies. However, as shown in Table 3, the VGG-m model required 94 times more parameter memory and 13 times more floating operations (FLOPs) than the quarter VGG-m model. The result demonstrated that a single large network is inefficient

TABLE 3. Parameter memory and floating operations (FLOPs) required for single networks.

Method	VGG	VGG128	QVGG
Memory (MB)	377	34	4
FLOPs (G)	1.684	1.594	0.134

J=18 J=16 J=10

FIGURE 4. Three selections of lip landmarks to construct the patch-image-based CFIs: Full landmarks ($J = 18$), landmarks with corner landmarks excluded ($J = 16$), and outer lip landmarks only ($J = 10$).

TABLE 4. Averaged recognition accuracies of the five differently initialized trials depending on lip landmark selections.

Selctions	J=18	J=16	J=10
Accuracy (%)	83.9	**89.1**	80.7

to improve the recognition accuracy on OuluVS2 dataset. In the rest of this paper, we use the quarter VGG-m model for efficiency.

Furthermore, selection of lip landmarks was crucial to improve the recognition accuracy. Fig. 4 shows the three lip landmark selections of patch image with the quarter VGG-m: full landmarks ($J = 18$), landmarks with two corner landmarks excluded ($J = 16$), and outer lip landmarks only ($J = 10$). Table 4 shows the averaged recognition accuracies of the three selections. We trained five differently initialized CNNs for each landmark selection. Experimentally, excluding two corner landmarks ($J = 16$) achieved the best performance among the three selections. Although both the models using the 18 and 16 landmarks achieved high recognition accuracies for the training dataset, the model using the full landmarks might be overfitted. The result for the model using the outer lip landmarks only ($J = 10$) gave the insight that the patch images around inner landmarks served useful information for lip reading. Therefore, we use 16 landmarks to construct the patch-image-based CFI D^2 for the remaining experiments.

B. ABLATION STUDY FOR COMMITTEE MEMBERS

Table 5 shows the recognition accuracy of the 10 differently initialized trials for each type of CFI. On average, the model using the patch-image-based CFI D^2 achieved better recognition accuracy than that using the full-lip-image-based CFI D^1.

Fig. 5 shows the averaged recognition accuracies as a function of the number of committee members. All possible combinations of the 10 trials in Table 5 were used. For example, when the number of the committee members is

TABLE 5. Recognition accuracies (%) of 10 differently initialized trials.

Trial	1	2	3	4	5	6	7	8	9	10	Avg.
D^1	88.1	86.9	85.0	88.6	85.8	88.6	86.1	85.8	87.2	88.3	87.0
D^2	88.3	89.7	90.6	88.6	88.3	90.6	89.4	87.5	89.2	87.2	88.9

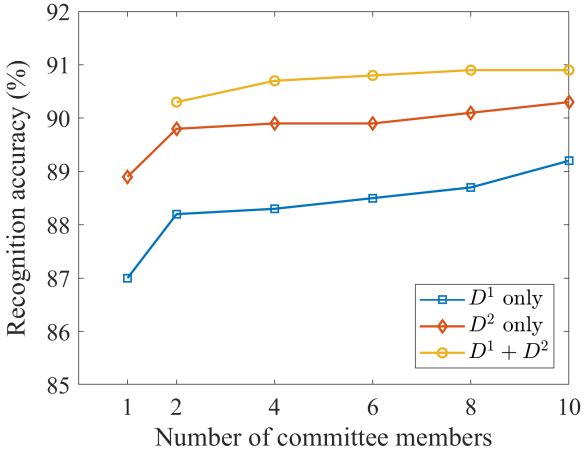


FIGURE 5. Averaged recognition accuracies depending on the number of committee members and the CFI type used.

set to 4, all possible combinations of the committee with the same types of CFIs is $\binom{10}{4} = 210$ and that of the committee with the different types of CFIs is $\binom{10}{2} \times \binom{10}{2} = 2025$. Three types of committees were computed: committee with the same types of CFIs (D^1 only and D^2 only) and committee with different types of CFIs D^1+D^2. All committees achieved better performances than a single quarter VGG-m, and the committees with different types of CFIs D^1+D^2 consistently showed higher recognition accuracies than the others.

The difference between local structures of the two different CFIs may provide complementary information. On the other hand, quite similar accuracies were obtained regardless of the number of committee members. Because averaged recognition accuracy of the committee with different types of CFIs was saturated after eight members, we set the number of committee members to eight. If we construct the committee using eight quarter VGG-m models, the proposed committee needs approximately 1.1-GFLOPs and 32-MB parameter memory which are smaller than those of a single VGG-m as shown in Table 3. We use the proposed committee with eight members to compare with the state-of-the-art methods.

Fig. 6 shows the recognition accuracy of the proposed methods for each speaker. Trained networks with different types of CFIs generated different recognition accuracies depending on the speaker. For most speakers, the recognition accuracy of the proposed committee was higher than or comparable to the higher one between the accuracies of the two single networks. The result demonstrates that the committee networks may utilize information extractable from both representations, whereas each committee member uses only a single representation, which is expected to lack the aspect of the other representation.

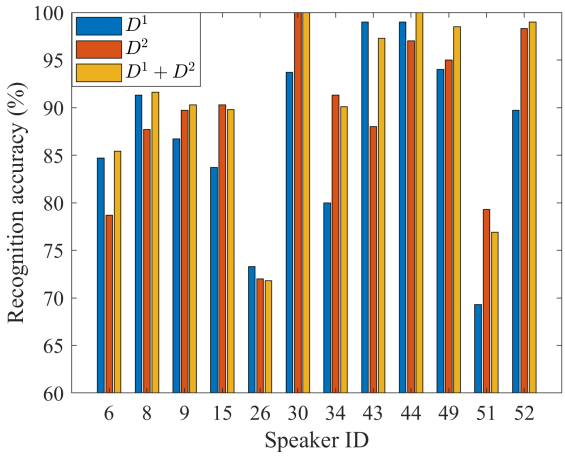


FIGURE 6. Averaged recognition accuracies for each speaker.

TABLE 6. Comparison with the state-of-the-art methods.

Method	Model		Acc. (%)
	Feature	Classifier	
Saitoh *et al.* [8]	CFI+GoogLeNet		85.6
Lee *et al.* [11]	CNN	LSTM	81.1
Fung and Mak [12]	Maxout+CNN	LSTM	87.6
Ours	CFI+QVGG+Committee		**90.9±0.7**

C. COMPARISON WITH THE STATE-OF-THE-ART METHODS

The proposed method was compared with the state-of-the-art methods trained on OuluVS2 only: Saitoh *et al.* method [8], Lee *et al.* method [11], and Fung and Mak method [12].

In Table 6, committees using two different types of CFIs provided better performance than the other state-of-the-art methods. Saitoh *et al.* used CFI with GoogLeNet, which needed approximately 2-GFLOPs with 224 × 224 RGB images, and achieved 85.6% recognition accuracy, whereas the proposed committee network, which needed approximately 1.1-GFLOPs, and achieved 90.9% recognition accuracy. The results demonstrated that the proposed committee with different types of CFIs and the quarter VGG-m is better than a single deep network, such as GoogLeNet for lip reading when trained on a small dataset.

Although Fung and Mak [12] successfully applied bidirectional LSTM to the small dataset, the proposed committee outperformed their method. When a CNN-LSTM network was trained with a large-scale dataset, such as lip reading sentences in the wild [9], it gave better recognition performance than the proposed committee. However, a similar network provided a degraded recognition accuracy with the OuluVS2 only, as shown in Table 6 [11].

The limitation of the proposed method is that the patch-image-based CFI D^2 can perform with near-frontal face only. In profile face image, the landmarks in occluded region cannot obtain the proper patch images, and thus the committee members trained on the patch-image-based CFI D^2 may degrade recognition accuracy.

IV. CONCLUSION

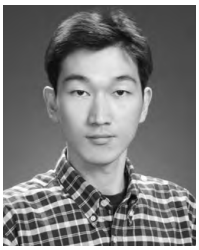
This paper proposed a method to apply CNNs to two different types of CFIs for lip reading. Multiple predictions obtained from full-lip images and patch images around lip landmarks were averaged by a committee network, which provided better performance than that using either type of CFI. The trained kernels were different depending on the CFI type, which implied that they might exploit different aspects of CFIs for lip reading. The TLPT and quarter VGG-m were effective for training using the limited dataset. The proposed committees with the two different types of CFIs provided better performance than those with the same types of CFIs consistently. Furthermore, the proposed methods gave impressive performance improvement without pre-training with a large-scale dataset. Further research will focus on continuous speech recognition and landmark selection for multi-view lip reading.

REFERENCES

[1] Z. Zhou, G. Zhao, X. Hong, and M. Pietikäinen, "A review of recent advances in visual speech decoding," *Image Vis. Comput.*, vol. 32, no. 9, pp. 590–605, Sep. 2014.
[2] Z. Zhou, X. Hong, G. Zhao, and M. Pietikäinen, "A compact representation of visual speech data using latent variables," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 1, p. 1, Jan. 2014.
[3] Y. Pei, T.-K. Kim, and H. Zha, "Unsupervised random forest manifold alignment for lipreading," in *Proc. IEEE Int. Conf. Comput. Vis.*, Sydney, NSW, Australia, Dec. 2013, pp. 129–136.
[4] G. Potamianos, C. Neti, G. Iyengar, A. W. Senior, and A. Verma, "A cascade visual front end for speaker independent automatic speechreading," *Int. J. Speech Technol.*, vol. 4, nos. 3–4, pp. 193–208, Jul. 2001.
[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778.
[6] K. Noda, Y. Yamaguchi, K. Nakadai, H. G. Okuno, and T. Ogata, "Audio-visual speech recognition using deep learning," *Appl. Intell.*, vol. 42, no. 4, pp. 722–737, Jun. 2015.
[7] J. S. Chung and A. Zisserman, "Lip reading in the wild," in *Proc. Asian Conf. Comput. Vis.*, Taipei, Taiwan, Nov. 2016, pp. 87–103.
[8] T. Saitoh, Z. Zhou, G. Zhao, and M. Pietikäinen, "Concatenated frame image based CNN for visual speech recognition," in *Proc. Asian Conf. Comput. Vis. Workshops*, Taipei, Taiwan, Nov. 2016, pp. 277–289.
[9] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, Jul. 2017, pp. 3444–3453.
[10] S. Petridis, Z. Li, and M. Pantic, "End-to-end visual speech recognition with LSTMs," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, New Orleans, LA, USA, Mar. 2017, pp. 2592–2596.
[11] D. Lee, J. Lee, and K.-E. Kim, "Multi-view automatic lip-reading using neural network," in *Proc. Asian Conf. Comput. Vis. Workshop*, Taipei, Taiwan, Nov. 2016, pp. 290–302.
[12] I. Fung and B. Mak, "End-to-end low-resource lip-reading with maxout CNN and LSTM," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Calgary, AB, Canada, Apr. 2018, pp. 2511–2515.
[13] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vis.*, Zürich, Switzerland, Jun. 2014, pp. 818–833.
[14] D. C. Ciresan, U. Meier, L. M. Gambardella, and J. Schmidhuber, "Convolutional neural network committees for handwritten character classification," in *Proc. Int. Conf. Document Anal. Recognit.*, Beijing, China, Sep. 2011, pp. 1135–1139.
[15] W. Liu, M. Zhang, Z. Luo, and Y. Cai, "An ensemble deep learning method for vehicle type classification on visual traffic surveillance sensors," *IEEE Access*, vol. 5, pp. 24417–24425, 2017.
[16] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic, "Incremental face alignment in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 1859–1866.
[17] I. Anina, Z. Zhou, G. Zhao, and M. Pietikäinen, "OuluVS2: A multi-view audiovisual database for non-rigid mouth motion analysis," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, Ljubljana, Slovenia, May 2015, pp. 1–5.
[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, Stateline, NV, USA, Dec. 2012, pp. 1097–1105.
[19] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.
[20] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," in *Proc. Conf. BMVC*, Nottingham, U.K., May 2014, pp. 1–12.
[21] A. Vedaldi and K. Lenc, "MatConvNet: Convolutional neural networks for MATLAB," in *Proc. ACM Multimedia*, Brisbane, Australia, Sep. 2015, pp. 689–692.

DONG-WON JANG received the B.S., M.S., and Ph.D. degrees in electronic engineering from Sogang University, Seoul, South Korea, in 2012, 2015, and 2019, respectively, where he is currently a Post Doctor of electronic engineering. His current research interests include deep learning-based computer vision, image restoration/enhancement, and lip reading.

HONG-IN KIM received the B.S., M.S., and Ph.D. degrees in electronic engineering from Sogang University, Seoul, South Korea, in 2013, 2015, and 2019, respectively. He is currently a Research Engineer with the AI center, SK Telecom Company Ltd. His current research interests include deep learning-based computer vision, image processing, and human-computer interaction.



CHANGSOO JE received the B.S. degree in physics, and the M.S. and Ph.D. degrees in media technology from Sogang University, Seoul, South Korea, in 2000, 2002, and 2008, respectively, where he is currently a Research Professor of electronic engineering. He was a Research Professor, from 2009 to 2010, and a Postdoctoral Research Fellow, from 2008 to 2009, of computer science and engineering with Ewha Womans University. His research interests include computer vision, computer graphics, and image processing. He is currently serving as an Associate Board Member for *International Journal of Sensors* and *Wireless Communications and Control*. He received an Outstanding Paper Award at the Korea Computer Graphics Society Conference, in 2008, and a Samsung Human-tech Thesis Prize Award from Samsung Electronics Company Ltd., in 2004. He is currently an Editor of *Journal of Engineering and Computer Innovations* (JECI).

RAE-HONG PARK (S'76–M'84–SM'99) received the B.S. and M.S. degrees in electronics engineering from Seoul National University, Seoul, South Korea, in 1976 and 1979, respectively, and the M.S. and Ph.D. degrees in electrical engineering from Stanford University, Stanford, CA, in 1981 and 1984, respectively.

In 1984, he joined the faculty of the Department of Electronic Engineering, Sogang University, Seoul, where he is currently an Emeritus Professor with the Department of Electronic Engineering and a Distinguished Professor with the ICT Convergence Disaster/Safety Research Institute. In 1990, he was a Visiting Associate Professor with the Computer Vision Laboratory, Center for Automation Research, University of Maryland, College Park. In 2001 and 2004, he was with the Digital Media Research and Development Center (DTV image/video enhancement), Samsung Electronics Company Ltd., Suwon, Korea, where he was with Digital Imaging Business (R&D Team) and Visual Display Business (R&D Office), in 2012. His current research interests include video communication, computer vision, and pattern recognition.

Dr. Park was a recipient of a 1990 Post-doctoral Fellowship presented by the Korea Science and Engineering Foundation (KOSEF), the 1987 Academic Award presented by the KITE, the 2000 Haedong Paper Award presented by the Institute of Electronics Engineers of Korea (IEEK), the 1997 First Sogang Academic Award, and the 1999 Professor Achievement Excellence Award presented by Sogang University. He is a co-recipient of the Best Student Paper Award of the IEEE International Symposium, Multimedia (ISM 2006), and IEEE International Symposium Consumer Electronics (ISCE 2011). He served as an Editor for the *Korea Institute of Telematics and Electronics (KITE) Journal of Electronics Engineering*, from 1995 to 1996.

HYUNG-MIN PARK (M'08–SM'12) received the B.S., M.S., and Ph.D. degrees in electrical engineering and computer science from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 1997, 1999, and 2003, respectively. From 2003 to 2005, he was a Postdoctoral with the Department of Biosystems, KAIST. From 2005 to 2007, he was with the Language Technologies Institute, Carnegie Mellon University. In 2007, he joined the Department of Electronic Engineering, Sogang University, Seoul, South Korea, and is currently a Professor. His main research interests include robust speech recognition and computer vision.