Concatenated Frame Image based CNN for Visual Speech Recognition

Takeshi Saitoh[1], Ziheng Zhou[2], Guoying Zhao[2], Matti Pietikäinen[2]

Kyushu Institute of Technology, Japan[1]
University of Oulu, Finland[2]

Abstract. This paper proposed a novel sequence image representation method called concatenated frame image (CFI), two types of data augmentation methods for CFI, and a framework of CFI-based convolutional neural network (CNN) for visual speech recognition (VSR) task. CFI is a simple, however, it contains spatial-temporal information of a whole image sequence. The proposed method was evaluated with a public database OuluVS2. This is a multi-view audio-visual dataset recorded from 52 subjects. The speaker independent recognition tasks were carried out with various experimental conditions. As the result, the proposed method obtained high recognition accuracy.

1 Introduction

In the field of visual speech recognition (VSR), one of the most important problems is the extraction of visual features. All the existing approaches can be roughly grouped into four categories: (1) image-based, (2) motion-based, (3) geometric-feature-based, and (4) model-based [1, 2]. For the image-based approaches, a gray-scale image is either used directly or after some image transformation, such as PCA and DCT, as a feature vector [3, 4]. Typical motion-based methods are based on optical flow [5]. The geometric-feature-based approaches measure certain geometric features of the mouth such as the width, height, area, and aspect ratio. The model-based approaches are based on the active appearance models that jointly characterize the shapes and textures of talking mouths [6–8] and model parameters are used as visual features.

Recently, deep learning techniques have been successfully applied to learn features from audio-visual data for the tasks of VSR and audiovisual speech recognition (AVSR). Ngiam et al. [9] proposed to build a multimodal deep autoencoder consisting of stacks of the Restricted Boltzmann Machines (RBMs) for learning modality-specific information. In their work, two public datasets, AVLetters and CUAVE were used for the supervised classification. Hu et al. [10] proposed a Recurrent Temporal Multimodal Restricted Boltzmann Machines to model audio-visual sequences in an unsupervised fashion. The joint representations across the generated features of two modalities were learned using multimodal RBMs. Two public datasets, namely, the AVLetters and AVLetters2, as well as their collected dataset were used for the evaluation. Noda et al. [11]

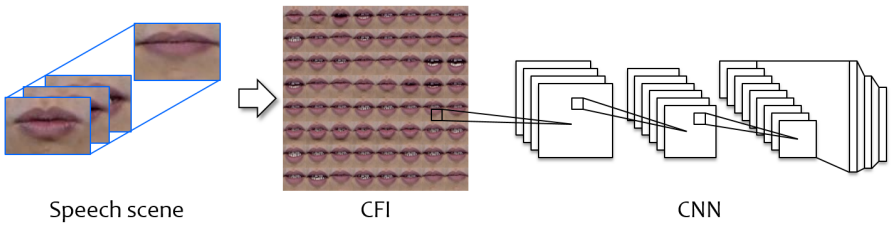


Fig. 1. Overview of proposed framework.

proposed a visual feature extraction method for VSR utilizing a Convolutional Neural Network (CNN). Hidden Markov Models (HMM) with Gaussian mixtures were used for a task of recognizing isolated word. The method was evaluated on an audio-visual speech dataset comprising 300 Japanese words uttered by six speakers. Amer et al. [12] proposed a hybrid model comprising of temporal generative and discriminative models for classifying sequential data from multiple heterogeneous modalities. Their method was evaluated on three datasets (AVEC, AVLetters, and CUAVE). Takashima et al. [13] proposed a multi-modal feature extraction method using a Convolutive Bottleneck Network (CBN), and applied to audio-visual data. Extracted bottleneck audio and visual features were used as the features input to the audio or visual HMMs and the recognition results then integrated. Their method did not use the output labels of CBN and was evaluated on a work recognition task. They used 216 words as the test data and 2,620 words as the training data. Most of the above mentioned methods targeted the problem of AVSR. Only [11] tackled the problem of VSR. Note that in their method, CNNs were used for visual extraction features and the classification was conducted by HMMs.

In this paper, we propose a novel image sequence representation, called the concatenated frame image (CFI) and the data augmentation method for CFI. As shown in Fig. 1, VSR is tackled by the CFI-based CNNs. We evaluate our approach on the newly collected public audiovisual database, OuluVS2 [14] and the results show that our approach performs well in a speaker independent setting.

The rest of this paper is organized as follows: in Section 2 the proposed concatenated frame image is described. Section 3 provides details of the constructed CNN. In Section 4, the OuluVS2 database and experimental results are described. This paper concludes in Section 5.

2 Concatenated Frame Image

Let I_f denotes the f-th frame of a video sequence that records a certain utterance and I'_f a resized image of I_f. Let the sequence length be F and the image sizes of I_f and I'_f be $W \times H$ [pixels] and $W' \times H'$ [pixels], respectively. The proposed

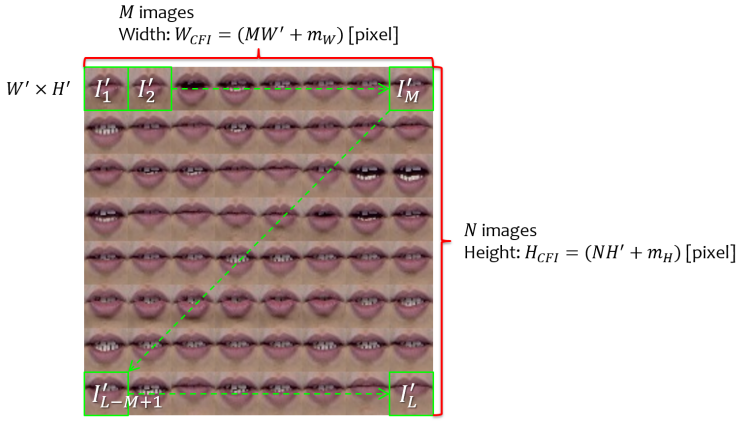


Fig. 2. Overview of base CFI.

concatenated frame image is an image that is formed by concatenating $\{I'_f\}$ following a specific rule that sample images with uniform intervals from $\{I'_f\}$ and re-organize them into M rows and N columns:

$$\text{CFI} = \begin{pmatrix} I'_1 & \cdots & I'_M \\ \vdots & \ddots & \vdots \\ I'_{L-M+1} & \cdots & I'_L \end{pmatrix}, \tag{1}$$

where $L = M \times N$ is the number of sub-images in CFI. Here, F may be different across different videos. Each CFI is an image with a size of $W_{CFI} \times H_{CFI} = (MW' + m_W) \times (NH' + m_H)$ pixels, where m_W and m_H are the number of pixels surround sub-images. Figure 2 shows an overview of the construction of a CFI. In this CFI, $W' = 32$, $H' = 32$, $M = 8$, $N = 8$, $m_W = 0$, and $m_H = 0$. The left-top I'_f is the first frame image, and the right-bottom I'_f is the last frame image.

Figure 3 shows three samples of CFI. These CFIs are generated by the same speech scene. Five parameters of Fig. 3 are the same: $W = 228$ pixel, $H = 150$ pixel, $F = 146$, $W_{CFI} = 256$ pixel, and $H_{CFI} = 256$ pixel. However, the values of each L are different. L of the left-side CFI, middle-side CFI, and right-side CFI of Fig. 3 are 49, 64, and 81, respectively. The larger L indicates that CFI has the high time resolution.

2.1 Data augmentation

Data augmentation (DA) techniques are effective for reducing overfitting on training datasets and therefore, improving generalization of the trained neural networks [15]. Typical DA methods for CNNs include applying some translation,

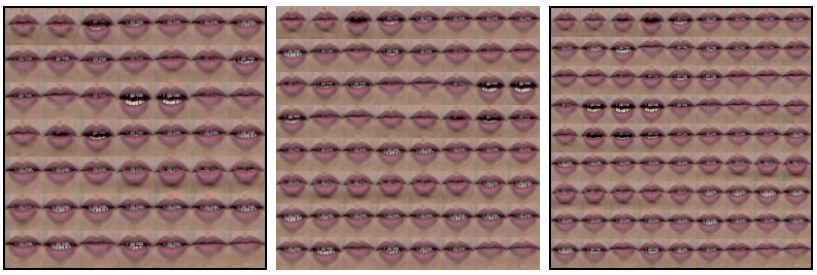


Fig. 3. CFI samples (left:$L = 49$, middle:$L = 64$, right:$L = 81$).

rotation, mirror reverse, and color change to images. Here we propose two DA strategies for CFI to the tackle the problem of VSR:

The first strategy is to generate CFIs by applying Gamma correction for brightness changes. Since the skin color is different depending on the gender and the race, Gamma correction is effective to this problem. Given an input image I, its pixel intensities are first scaled from the range $[0, 255]$ to $[0, 1]$. We then obtain a gamma corrected image I_{gamma} through:

$$I_{gamma}(r, c) = I(r, c)^{1/\gamma}, \tag{2}$$

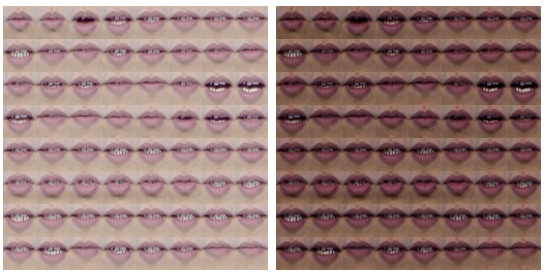
where (r, c) denotes the pixel value and γ is the gamma value. Finally, I_{gamma} is scaled back to the range $[0, 255]$. Here, $\gamma < 1$ shifts an input image towards the darker end of the spectrum while $\gamma > 1$ makes the image appear lighter.

The second strategy is applied to the temporal domain. Utterance speed is different depending on individuals, and utterance time is different depending on the utterance content. Then, temporal shift is applied. The standard CFI as previously defined is built by L frames samped from a video sequence of F frame length. For DA, $(L-\alpha)$ frames are sampled and temporal shifted CFI is generated with $(L - \alpha)$ frames. This means that the sampling interval $(= F/(L - \alpha))$ of the temporal shifted CFI is shorter than the sampling interval $(= F/L)$ of the standard CFI. To obtain L images for making the CFI, the first or last frame is replicated α times on the left-top or right-bottom of the CFI.

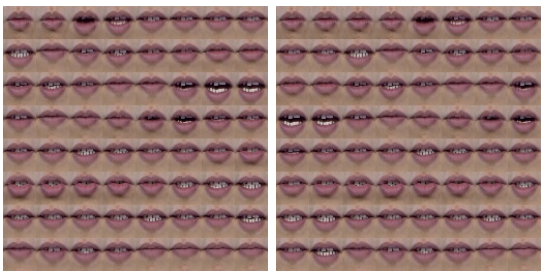
Figure 4 demonstrates the two DA strategies. Two samples of Fig. 4(a) are CFIs with the Gamma correction applied. Two of samples of Fig. 4(b) are CFIs which the temporal shift applied.

3 Convolutional neural network

Recent developments in deep learning technologies have greatly advanced the performance of the state of the art of many visual recognition tasks. In particular, convolutional neural networks (CNNs) have been established as a powerful class of models for image recognition tasks [15]. CNNs consist of alternating convolutional layers and pooling layers. Convolution layers take inner product of the



(a) Gamma correction (left: $\gamma = 0.6$, right: $\gamma = 1.4$)



(b) temporal shift (left: $\alpha = -2$, right: $\alpha = 2$)

Fig. 4. Data augmentation samples.

linear filter and the underlying receptive field followed by a nonlinear activation function, such as rectifier, sigmoid, tanh, at every local portion of the input.

There have been a number of pre-trained CNN models available. In this research, we build our CNNs for VSR based on three well-known models: Network In Network (NIN) [16], AlexNet [15], and GoogLeNet [17].

NIN is proposed by Lin et al. [16]. NIN consists of mlpconv layers which use multilayer perceptrons to convolve the input and a global average pooling layer as a replacement for the fully connected layers in conventional CNN. NIN used in this research consists of four mlpconv layers, and the mlpconv layers are followed by a spatial max pooling layer which down-samples the input image by a factor of three. To reduce overfitting in the fully connected layers, regularization method called dropout is applied on the outputs of the last mlpconv layers.

AlexNet is proposed by Krizhevsky et al. [15]. This model consists of five convolutional layers, some of which are followed by max pooling layers, and three fully connected layers.

GoogLeNet is proposed by Szegedy et al. [17]. This model is based on using a sparsely connected architecture in order to avoid computational bottlenecks and improve computational efficiency over the entire network as they go deeper

Table 1. Lists of utterance content.

Phase 1: digit sequences		Phase 2: phrases	
p1-01	1735162667	p2-01	Excuse me
p1-02	4029185904	p2-02	Good bye
p1-03	1907880328	p2-03	Hello
p1-04	4912118551	p2-04	How are you
p1-05	8635402112	p2-05	Nice to meet you
p1-06	2390016764	p2-06	See you
p1-07	5271613670	p2-07	I am sorry
p1-08	9744435587	p2-08	Thank you
p1-09	6385398565	p2-09	Have a good time
p1-10	7324019950	p2-10	You are welcome

and wider. The sparsely connected architecture is called inception modules that construct a sparser representation of the convolution networks by clustering the neurons with the highest correlation and uses an extra 1×1 convolutional layer as dimensionality reduction. This model though much deeper than AlextNet.

4 Experiments and Results

4.1 Dataset

The OuluVS2 database[1] [14] is one of the largest dataset for VSR. It is a multi-view audio-visual dataset for non-rigid mouth motion analysis. The dataset contains video recording from 52 subjects (39 males and 13 females) speaking three types of utterances: continuous digit sequences, short phrases and TIMIT sentences. The lists of the utterance content of first two types are shown in Table 1. In phase 1, a subject was asked to utter continuously ten fixed digit sequences. Each sequence consisted of ten randomly generated digits and was repeated three times during recording. In phase 2, the subject was asked to speak ten daily-use short English phrases. The same set of phrases was used in the OuluVS dataset [18]. Every phrase was uttered three times.

In OuluVS2, each utterance were filmed with six cameras placed around a subject. The six cameras included five GoPro Hero3 Black Edition cameras and a PuxeLink PL-B774U camera. The former five cameras are called HD cameras. The image size of these cameras is 1920×1080 pixels, and its frame rate is 30fps. On the other hand, the image size recorded by the latter camera is 640×480 pixels, and its frame rate is 100fps. As regards the five HD cameras, HD1, HD2, HD3, HD4, and HD5 are located in the following positions: 0° (frontal view), 30°, 45°, 60°, and 90° (profile view) to the subject's right hand side. The recording was made in an ordinary office environment with three extra lights placed behind the camera to illuminate the subject's face.

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

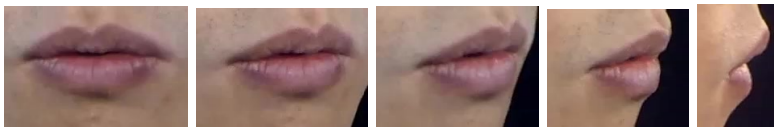


Fig. 5. ROI images (HD1, HD2, HD3, HD4, and HD5).

Table 2. Statistics of OuluVS2 (ROI).

		digit sequences			phrase sequences		
		min	max	ave	min	max	ave
HD1	width [pixel]	162	294	208	158	262	201
	height [pixel]	64	218	125	64	196	116
HD2	width [pixel]	138	240	183	130	214	175
	height [pixel]	64	190	122	64	186	114
HD3	width [pixel]	130	230	180	124	210	171
	height [pixel]	64	198	123	64	184	114
HD4	width [pixel]	100	204	146	94	184	137
	height [pixel]	64	186	118	64	168	109
HD5	width [pixel]	64	164	97	64	136	89
	height [pixel]	86	190	133	78	188	126
frame number		83	297	161	8	20	36

The original images of OuluVS2 are covered subject's whole face, however, OuluVS2 provides the ROI image around the talking mouth. Figure 5 shows the examples of the provided ROI images by extracting from five HD cameras at each frame. The ROI sizes of each scene are different. The statistics of OuluVS2 are summarized in Table 2. The ROI of near frontal view is horizontally long shape, and the ROI of near profile view is vertically long shape. As seen in Table 2, the number of frame of phrase sequences is smaller than digit sequences.

4.2 Experimental Settings

We tested our system in a speaker-independent VSR setting. We used 12 subjects (s06, s08, s09, s15, s26, s30, s34, s43, s44, s49, s51, and s52, 10 males and 2 females) for testing, and remaining 40 subjects for training. Note that digit strings and phrases were recognized as a whole instead of being modelled by visemes.

Moreover, we generated eight types of CFIs for DA by choosing four γ values (0.6, 0.8, 1.2, and 1.4) with $\alpha = 0$ as well as four α values (-2, -1, 1, and 2) with $\gamma = 1.0$. As a result, the test data included 360 (= 10 phrases $\times$ 12 subjects$\times$3 samples) CFIs. The training data contained 1,200 (= 10 phrases$\times$ 40 subjects $\times$ 3 samples) CFIs without DA and the number increased to 10,800 (= 10 phrases $\times$ 40 subjects $\times$ 3 samples $\times$ 9 types) CFIs after DA. To generate

Table 3. Three conditions and parameters for generating CFIs.

$M \times N$	L [frame]	W_{CFI} [pixel]	H_{CFI} [pixel]	W' [pixel]	H' [pixel]	m_W [pixel]	m_H [pixel]
7×7	49	256	256	36	36	4	4
8×8	64	256	256	32	32	0	0
9×9	81	256	256	28	28	4	4

CFIs, some parameters are required. In the experiment, we chose three sets of parameters as shown in Table 3.

As for the CNN model, three well-known models: NIN [16], AlexNet [15], and GoogLeNet [17], were used. We used Chainer [2], a flexible framework of deep learning for creating CNN model and training. We trained all models using stochastic gradient descent with 0.9 momentum, mini-batches of size 32, and the learning rate is initialized at 0.01. The softmax with the cross-entropy loss was used as a classifier. In the experiment, we used a personal computer with an Intel Core i7-3770 processor (3.4GHz), 16 GB RAM, and a single NVIDIA GeForce GTX970 graphic processing unit with 6 GB on-board graphics memory.

4.3 Single-view Lip-reading

In our experiments, training and test data were recorded by the same video camera. Since the OuluVS2 database has five camera views (HD1, HD2, HD3, HD4, and HD5), we carried out the recognition experiment for each of the five views.

Experimental results for recognizing digit strings are shown in Table 4. Table 5 shows the experimental results for recognizing short phrases. In either table, the upper half of the table shows the recognition results without DA, and the lower half the recognition results with DA. It is clear that our system achieved higher recognition accuracy when using DA than when not using DA. Regarding the three pre-trained models, there was no large difference among them in terms of system performance. Considering the parameter L, it can be seen that high recognition accuracy was obtained at $L = 49$ when recognizing digit strings, and at $L = 64$ when recognizing phrases.

Next, we selected the parameter and model settings with the highest recognition rates (those marked as bold in Table 4 and Table 5) and have a closer look at the system performance for each digit string and short phrase. Tables 6 and 7 show the recognition results. Regarding viewing angle, frontal or near frontal viewing angle obtained high recognition accuracy. When recognizing the digit strings, the lowest recognition accuracy was 80.6%. It was 58.3% when recognizing short phrases. It shows that the system tends to recognize longer video sequences of a talking mouth better.

[2] http://chainer.org/

Table 4. Recognition results by various conditions (digit).

	CNN model (L)	HD1	HD2	HD3	HD4	HD5	ave.
without DA	NIN (49)	45.8	10.1	13.5	44.6	73.5	37.5
	NIN (64)	52.1	11.3	13.8	63.2	57.9	39.6
	NIN (81)	61.7	9.7	13.5	56.1	70.1	42.2
	AlexNet (49)	12.1	9.6	45.1	70.0	59.3	39.2
	AlexNet (64)	10.7	10.6	10.4	9.6	44.9	17.2
	AlexNet (81)	9.9	9.2	11.5	42.9	33.8	21.4
	GoogLeNet (49)	11.5	9.4	10.3	47.2	52.8	26.3
	GoogLeNet (64)	9.9	11.9	9.7	8.5	14.2	10.8
	GoogLeNet (81)	12.2	8.8	8.9	10.3	28.3	13.7
with DA	NIN (49)	85.6	74.4	73.1	89.7	85.0	81.6
	NIN (64)	87.2	85.6	85.6	85.6	86.4	86.1
	NIN (81)	80.3	70.0	78.1	88.1	81.1	79.5
	AlexNet (49)	89.2	82.2	**90.8**	**91.7**	**87.5**	88.3
	AlexNet (64)	10.3	90.6	**90.8**	85.3	11.9	57.8
	AlexNet (81)	85.0	68.9	9.4	89.2	80.6	66.6
	GoogLeNet (49)	**89.4**	**92.5**	86.7	87.5	86.4	**88.5**
	GoogLeNet (64)	**89.4**	91.7	83.1	89.4	84.7	87.7
	GoogLeNet (81)	86.9	90.6	85.3	85.3	85.6	86.7

Table 5. Recognition results by various conditions (phrase).

	CNN model (L)	HD1	HD2	HD3	HD4	HD5	ave.
without DA	NIN (49)	19.3	25.3	41.9	63.8	68.6	43.8
	NIN (64)	11.9	18.8	21.4	69.7	64.4	37.3
	NIN (81)	34.0	11.0	46.7	43.8	65.6	40.2
	AlexNet (49)	36.5	8.9	9.7	61.4	36.8	30.7
	AlexNet (64)	8.9	25.1	10.0	36.5	72.4	30.6
	AlexNet (81)	9.3	11.3	27.5	63.1	69.3	36.1
	GoogLeNet (49)	68.5	12.8	10.4	61.8	62.9	43.3
	GoogLeNet (64)	63.5	24.9	49.4	59.0	58.5	51.1
	GoogLeNet (81)	66.3	17.4	25.7	58.8	57.8	45.2
with DA	NIN (49)	77.5	77.5	78.9	71.1	74.7	75.9
	NIN (64)	81.1	79.7	**82.5**	81.9	74.7	80.0
	NIN (81)	75.3	81.7	77.8	77.8	73.3	77.2
	AlexNet (49)	82.8	75.6	80.6	80.8	79.2	79.8
	AlexNet (64)	81.7	**82.5**	81.9	**83.3**	75.3	80.9
	AlexNet (81)	73.6	74.4	75.8	76.9	75.3	75.2
	GoogLeNet (49)	83.6	81.7	81.9	78.3	76.7	80.4
	GoogLeNet (64)	**85.6**	79.7	80.8	**83.3**	**80.3**	**81.9**
	GoogLeNet (81)	83.1	76.7	78.6	79.7	78.6	79.3

Table 6. Recognition results in detail (digit).

phrase	HD1	HD2	HD3	HD4	HD5
p1-01	86.1	88.9	86.1	83.3	80.6
p1-02	97.2	94.4	91.7	91.7	80.6
p1-03	91.7	97.2	94.4	91.7	91.7
p1-04	83.3	83.3	86.1	91.7	86.1
p1-05	100.0	94.4	97.2	100.0	91.7
p1-06	80.6	97.2	88.9	91.7	83.3
p1-07	80.6	94.4	94.4	86.1	91.7
p1-08	97.2	94.4	94.4	97.2	97.2
p1-09	91.7	86.1	86.1	91.7	88.9
p1-10	86.1	94.4	88.9	91.7	83.3
ave.	89.4	92.5	90.8	91.7	87.5

Table 7. Recognition results in detail (phrase).

phrase	HD1	HD2	HD3	HD4	HD5
p2-01	88.9	91.7	94.4	94.4	80.6
p2-02	97.2	94.4	94.4	94.4	91.7
p2-03	80.6	66.7	72.2	69.4	69.4
p2-04	83.3	80.6	80.6	86.1	83.3
p2-05	100.0	94.4	91.7	91.7	88.9
p2-06	83.3	80.6	66.7	69.4	66.7
p2-07	94.4	88.9	100.0	91.7	88.9
p2-08	58.3	75.0	66.7	69.4	63.9
p2-09	91.7	86.1	94.4	97.2	91.7
p2-10	77.8	66.7	63.9	69.4	77.8
ave.	85.6	82.5	82.5	83.3	80.3

At last, we discuss the recognition result for each test subject. Figure 6 shows the average recognition results at each viewing angle. The blue and red bars are average recognition accuracies for the digit strings and short phrases, respectively. The recognition accuracies for speakers s06 and s51, especially from near profile viewing angle, are lower than the accuracies for other speakers. Speakers s30 and s44 obtained the highest recognition accuracy among all.

4.4 Comparison with other methods

The OuluVS2 database is a newly collected dataset, and there are only a few baseline recognition results provided in [14]. In their work, for feature extraction, 2D DCT features from each image were computed and PCA applied to reduce the feature dimension to 100. For recognition, a whole-word HMM was constructed for classification. In their experiment, leave-one-speaker-out cross validation was applied to the performance evaluation. The best recognition rates of 47% was obtained from HD4 camera image. This evaluation protocol is not the same as the protocol of this paper. It is clear that our method outperformed their baseline system by a large margin.

5 Conclusion

In this paper, we proposed a novel sequence image representation, namely, CFI, and a CFI-based CNN for VSR. The proposed system was evaluated on a public dataset, OuluVS2. In the experiments, the speaker independent setting was carried out with various parameter settings evaluated. In the current experiment, we did not compare the recognition accuracy with other state-of-the-art methods, such as 3D-ConvNet+LSTM [19] in the same protocol. In future, we plan to evaluate our result with other methods. We also consider to add further experiments with other datasets.

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

References

1. Dupont, S., Luettin, J.: Audio-visual speech modeling for continuous speech recognition. IEEE Trans. Multimedia **2** (2000) 141–151
2. Zhou, Z., Zhao, G., Hong, X., Pietikainen, M.: A review of recent advances in visual speech decoding. Image and Vision Computing **32** (2014) 590–605
3. Bregler, C., Konig, Y.: "eigenlips" for robust speech recognition. In: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP1994). (1994) 669–672
4. Lucey, P.J., Potamianos, G., Sridharan, S.: Patch-based analysis of visual speech from multiple views. In: Proc. of International Conference on Auditory-Visual Speech Processing (AVSP2008). (2008) 69–73

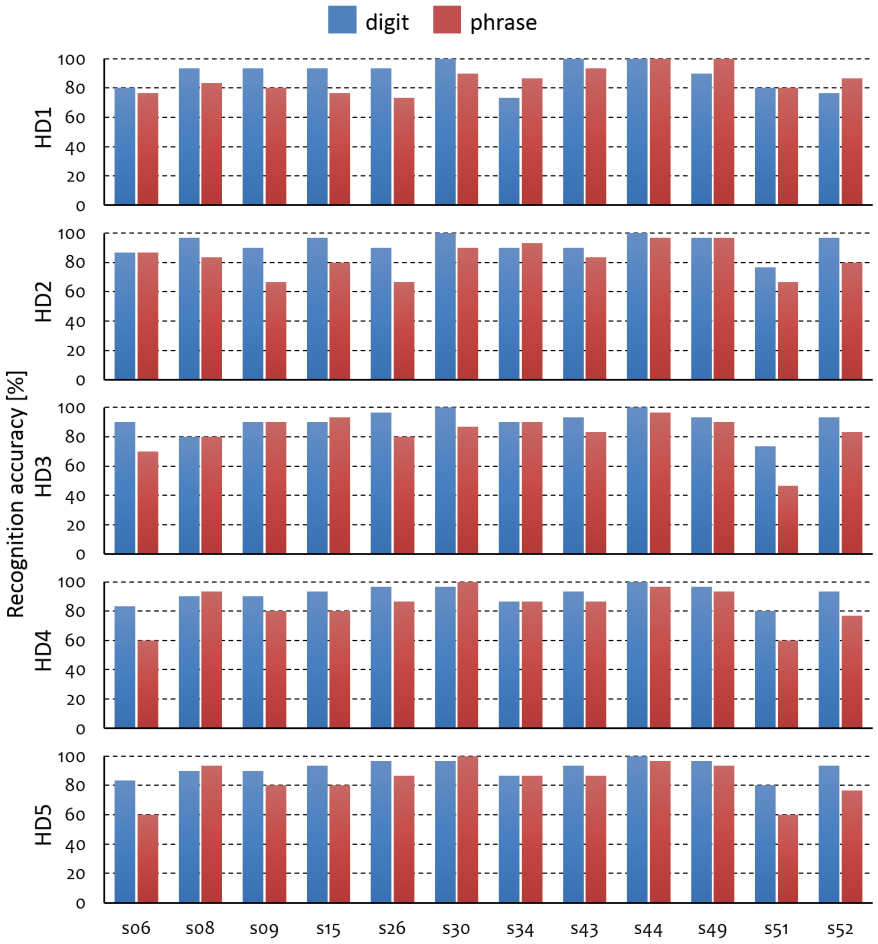


Fig. 6. Individual recognition results.

5. Shiraishi, J., Saitoh, T.: Optical flow based lip reading using non rectangular ROI and head motion reduction. In: 11th IEEE International Conference on Automatic Face and Gesture Recognition (FG2015). (2015)
6. Matthews, I., Cootes, T.F., Bangham, J.A., Cox, S., Harvey, R.: Extraction of visual features for lipreading. IEEE Trans. Pattern Anal. & Mach. Intell. **24** (2002) 198–213
7. Shin, J., Lee, J., Kim, D.: Real-time lip reading system for isolated Korean word recognition. Pattern Recognition **44** (2011) 559–571
8. Saitoh, T.: Efficient face model for lip reading. In: International Conference on Auditory-Visual Speech Processing (AVSP). (2013) 227–232
9. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A.Y.: Multimodal deep learning. In: 28th International Conference on Machine Learning. (2011) 689–696
10. Hu, D., Li, X., Lu, X.: Temporal multimodal learning in audiovisual speech recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2016) 3574–3582
11. Noda, K., Yamaguchi, Y., Nakadai, K., Okuno, H.G., Ogata, T.: Lipreading using convolutional neural network. In: INTERSPEECH. (2014) 1149–1153
12. Amer, M.R., Siddiquie, B., Khan, S., Divakaran, A., Sawhney, H.: Multimodal fusion using dynamic hybrid models. In: IEEE Winter Conference on Applications of Computer Vision (WACV). (2014) 556–563
13. Takashima, Y., Kakihara, Y., Aihara, R., Takiguchi, T., Araki, Y., Mitani, N., Omori, K., Nakazono, K.: Audio-visual speech recognition using convolutive bottleneck networks for a person with severe hearing loss. IPSJ Transaction on Computer Vision and Applications **7** (2015) 64–68
14. Anina, I., Zhou, Z., Zhao, G., Pietikainen, M.: OuluVS2: a multi-view audiovisual database for non-rigid mouth motion analysis. In: IEEE International Conference on Automatic Face and Gesture Recognition (FG). (2015)
15. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems. (2012)
16. Lin, M., Chen, Q., Yan, S.: Network in network. In: International Conference on Learning Representations (ICLR). (2014)
17. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S.: Imagenet classification with deep convolutional neural networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2015)
18. Zhao, G., Barnard, M., Pietikainen, M.: Lipreading with local spatiotemporal descriptors. IEEE Transactions on Multimedia **11** (2009) 1254–1265
19. Baccouche, M., Mamalet, F., Wolf, C., Garcia, C., Baskurt, A.: Sequential deep learning for human action recognition. In: International Workshop on Human Behavior Understanding (HBU2011). (2011)