Group conversation support system for hearing impaired person

Yusuke Sarukura
Graduate School of Kanazawa University, Kanazawa, 920-1192 Japan
Tel: (+81) 90-5175-6774, Email: sarukura@blitz.ec.t.kanazawa-u.ac.jp

Hidetaka Nambo
Graduate School of Kanazawa University, Kanazawa, 920-1192 Japan
Tel: (+81) 76-234-4835, Email: nambo@blitz.ec.t.kanazawa-u.ac.jp

Shuichi Seto
Kinjo College, Hakusan, Ishikawa 924-8511 Japan
Tel: (+81) 76-276-4411, Email: seto@kinjo.ac.jp

Haruhiko Kimura
Graduate School of Kanazawa University, Kanazawa, 920-1192 Japan
Tel: (+81) 76-234-4836, Email: kimura@blitz.ec.t.kanazawa-u.ac.jp

Abstract. For hearing impaired person, visual information such as body language, facial expressions and movement of lip are very important to interpret conversation. However, in group conversation, speaker is changing dynamically. Hearing impaired person finds it is difficult to understand who is speaking in such situation and can't get visual information from speaker. This difficulty makes them to be hard to participate in meetings and makes them to be impossible to get promoted at work. Sign language interpreter or precis writer can help this situation. However, they are expensive to hire. In this paper, we propose a system to solve this problem, which uses a panoramic camera to get image of participants, then detecting speaker from the image. Then, an image of speaker is shown to user to provide visual information. We use convolutional neural networks, or deep learning, that is state of the art machine learning technique, to detect speaker from image.

Keywords: hearing impaired person, panoramic camera, convolutional neural network

1. INTRODUCTION

1.1 Background

Recently in Japan, there are revisions to "The handicapped Persons' Employment Promotion Act". Now, employers have assumed an obligation to provide rational consideration for handicapped person. It means that employers must provide equal opportunity between handicapped people and other people. Therefore, there are needs for system which enable handicapped person to do same tasks as non-handicapped person does.

Hearing impaired person utilize visual information to interpret conversation. However, in a situation where multiple people are present in a conversation, visual information is difficult to acquire because speaker is changing dynamically and hearing impaired person cannot look at speaker. This difficulty cause inequality between hearing impaired person and other people. According to Eiko (Eiko, 2014), 93.4 % of hearing impaired employees feel difficulty in work meetings and 78.9 % of them say it is hard to get promoted.

One of the solution to this problem is hiring of a sign language interpreter or a precis writer. However, its cost is very high. Therefore, several systems are developed to solve this problem.

1.2 Existing systems

FUJITSU Software Live Talk (Fujitsu, 2015) is a system which convert speech to text and support hearing impaired person to understand conversation. However, to use this system, you have to buy software license and special

microphone per speaker. Each speaker has to prepare computer and hold a microphone while speaking. Live talk is expensive and bothering to use. System must be more inexpensive and easy to use to gain popularity.

Syouo (Syouo, 2015) developed inexpensive system which use panoramic camera to capture all participants face and detect speaker by motion of their mouth. The system use Haar-like feature (Viola and Jones, 2004) to detect face region and treat lower third of it as mouth region. Though this system is preferable than live talk in point of cost and usability, there is a problem that detected region using Haar-like feature is not stable and sometimes misclassify non-speaker as speaker.

1.3 Purpose of study

Therefore, the purpose of this study is development of inexpensive and reliable system which assist hearing impaired person to understand group conversation. To establish this, we used convolutional neural network, in other words "Deep learning", state of the art machine learning technique, to detect speaker from panoramic image.

The reminder of the paper is organized as follows. In Section 2, we introduce the system we developed. In Section 3, we explain how the experiments are conducted and evaluated. The experiment results are reported in Section 4. We discuss about the results in Section 5. Finally, we conclude our work in Section 6.

2. SYSTEM

2.1 Equipment

We use KODAK PIXPRO SP360 to capture panoramic image. Which can be connected to computer via Wi-Fi, and provide 30 jpeg image per second. Provided images are 1024x1024 pixels size and contain 360° view.

Table 1 shows Specs of the computer used.

Table 1: Specs of the computer

OS	Ubuntu 14.04
CPU	Intel Core i5-4570 3.20Ghz
Memory	4GB

2.2 System components

System consist of four steps. First, capture participants image using SP360. Secondly, crop face regions of participants from captured image. Thirdly, detecting speaker by feeding face regions to convolutional neural network. Finally, display detected speaker image to user and back to first step. Figure 1 shows assumed use environment of our system. Details of second and third step are explained in following sections.

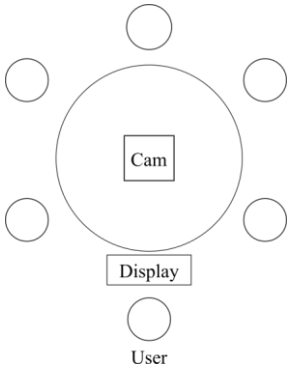

Figure 1: Assumed use environment

2.2.1 Face detection

To crop face region from image, we use Haar-like feature. Trained detector available with OpenCV (Bradski 2000) is used.

2.2.2 Speaker detection

We use convolutional neural network (CNN) to detect speech from image. CNN is type of feed-forward artificial neural network. Connectivity pattern between its neurons is inspired by the organization of the animal visual cortex. It consists of three types of layers, convolution layer, pooling layer, and fully-connected layer. In convolution layer, each neuron takes inputs from a rectangular section of the previous layer. Suppose that we have some $N \times N$ square neuron layer which is followed by convolutional layer. Output of unit Out_{ij} in convolutional layer is represented as

$$Out_{ij} = \sigma\left(\sum_{p=1}^{\omega}\sum_{q=1}^{\omega} in_{i+p,j+q} w_{p,q}\right) \qquad (1)$$

Where $w_{i,j}$ is a $\omega \times \omega$ size weight value at (i,j) and $in_{i,j}$ are input value at (i,j) from previous layer. $\sigma(x)$ is a function called activation function. Pooling layer takes small rectangular blocks from convolutional layer and subsamples it to produce single output from that block. We use max pooling which outputs maximum value in the block. Fully-connected layer takes all neurons in the previous layer and connects it to every single neuron it has, and performs classification. We use TensorFlow (Abadi et al., 2015), library for numerical computation using data flow graphs, to implement CNN.

3. EXPERIMENT

3.1 Data collection

Data are obtained from 8 subjects. 600 images are taken

for each subject. 100 images are for each Japanese vowel (a, i, u, e, o) and rest 100 images are for silent. Subjects are instructed to move around to differ angle and distance from camera while taking images. By doing this, system became capable of detecting speaker in various position. 4800 images are taken in total. Example of collected image is shown in Figure 2.

Figure 2: Example of collected data

3.2 Data preprocessing

Collected images are preprocessed before training phase. First, convert the images to grayscale. Then, crop face region using haar-like feature and resize to 64x64 pixel image. At this point, face detector is failed to detect face in some images. Therefore, such images are manually removed to improve classification. As a result, 4271 images are remained. After that, crop lower half 32x32 region of the image and treat it as mouth region. Obtained 32x32 image is used for training.

3.3 Training

After preprocessing, train convolutional neural network by feeding preprocessed data. We train neural network with a method called Adam (Ba and Kingma, 2015) which is a variation of stochastic gradient descent. Batch size is 10 and initial learning rate is 3e-6. We use the model after 1000 iteration for evaluation. Network details are shown in Table 2.Table **2**: Details of network

Table 2: Details of network

	Layer type	Layer shape	Output shape	Activation function
	Input	32x32	32x32x1	
1	Convolution	5x5x1x12	32x32x12	ReLU
2	Pooling	2x2	16x16x12	
3	Convolution	5x5x12x24	16x16x24	ReLU
4	Pooling	2x2	8x8x24	
5	FC	1536	256	ReLU
6	FC	256	6	Softmax

3.3 Evaluation

Data are divided by each subject and one subject data are used for test model trained by other seven subject data. Each division is evaluated with cross-validation. Results are evaluated by F1-score. F1 score is a harmonic mean of precision and recall. Precision is the fraction of predicted instances that are correct. Recall is the fraction of correct instance that are predicted.

We train neural network with six classes to predict: silent and vowels (a, i, u, e, o). However, in evaluation, we treat vowels class as single speaking class, because we need just prediction for a person is speaking or not.

4. RESULTS

Confusion table for the raw results are shown in Table 3. Table 4 is the result for binary classification (silent and speaking). Prediction for silent data is multiplied by five to match data size to speaking data.

Table 3: Confusion table for raw result for classification (Row: class, Column: prediction)

	Silent	A	I	U	E	O	Score
Silent	540	49	53	65	11	24	0.78
A	11	393	95	3	171	42	0.48
I	11	84	410	8	144	35	0.60
U	65	99	28	297	21	177	0.53
E	7	229	84	3	367	29	0.51
O	5	81	2	67	9	502	0.68

Table 4: Confusion table for binary classification (Row: class, Column: prediction)

	Silent	Speaking	Score
Silent	3379	99	0.86
Speaking	1010	2700	0.83

5. DISCUSSION

Though we are using different dataset, F1 score is improved by 0.09 for silent class and 0.06 for speaking class from Syouou's work.

From Table 3, we can see that vowels F1 scores are lower than silent F1 score. Possible cause of this is that each vowels is similar to others but silent is relatively not similar to vowels. From Figure 3, we can find similarity in mouth shape between highly misclassified vowels such as A-E, I-E, and U-O. we can avoid this problem by treating vowels as same single class, but further consideration will be needed to improve robustness of the system.

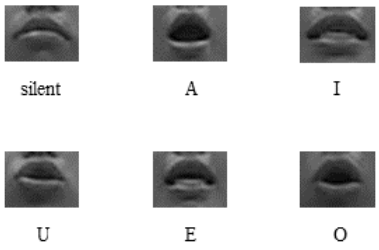

Figure 3: Comparison for mouth shapes

Table 5: Confusion matrix for the low score data division

(Row: class, Column: prediction)

	Silent	A	I	U	E	O
Silent	35	49	5	0	2	0
A	0	90	0	0	8	0
I	0	27	9	0	60	0
U	0	83	6	0	2	0
E	0	24	2	0	70	0
O	1	69	0	0	2	6

We have to mention to a data division which has very low score. The raw result for that data division is shown in Figure 3: Comparison for mouth shapes

Table **5**. For this data division, most data are classified as A. We look into the images to specify the cause of this problem. Figure 4 shows silent state image of the low score subject. We can see that there is a dark shadow near the nose. Figure 5 is a comparison of enlarged image of near this subject and other subject mouth pronouncing "A".

Figure 4: image of the low score subject

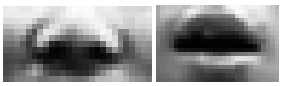

Figure 5: Comparison of low score subject nose and other subject pronouncing "A"

There is similarity in those two images in Figure 5. We can assume that this similarity makes our CNN confused and misclassified most data as "A". This can be a problem because some person might have similar shadow near nose. We think this problem is can be avoided if we use large dataset and include subject data who has similar characteristic because CNN can learn such feature.

6. CONCLUSION

In this work, we developed a group conversation support system for hearing impaired person which can detect speaker from panoramic camera accurately than existing system. The system overcome the problem of reliability using state of the art machine learning technique, convolutional neural network. There are some drawbacks in the system, but these drawbacks are considered easy to overcome by using larger dataset. We only use still image information to judge. We think a next step to improve the system is utilization of time-series information.

REFERENCES

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., ... & Ghemawat, S. (2016). Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:1603.04467.

Bradski, G. (2000). The opencv library. Doctor Dobbs Journal, 25(11), 120-126. ISO 690.

Eiko M.. (2014) The problems in office where hearing impaired or deaf person is working: based on questionnaire survey for handicapped and non-handicapped people. *Life design report*, No. 210, 4-15.

Kingma, D., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Syouo Y..(2015). The development of active learning participation support system for the hearing impaired students. *Kanazawa University Bachelor's Thesis.*

Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on (Vol. 1, pp. I-511). IEEE.