Proceedings of Machine Learning Research 122:28–35, 2019 3rd International Workshop on Artificial Intelligence in Affective Computing

Measuring Two-People Communication from Omnidirectional Video

Yui Niibori
Faculty of Information Sciences, Hiroshima City University, 3-4-1 Ozuka-Higashi, Asaminami-Ku, Hiroshima 731-3194, Japan

Shigang Li SHIGANGLI@HIROSHIMA-CU.AC.JP
Graduate School of Information Sciences, Hiroshima City University, 3-4-1 Ozuka-Higashi, Asaminami-Ku, Hiroshima 731-3194, Japan

Abstract

In this paper we propose a method of measuring the communication between two people by analyzing their heads' information: head pose, gaze vectors and facial action units. Assuming two people are sitting around a table, an omnidirectional camera is used to observe the two people simultaneously. Next, the visual cues of the heads of the two people, including head pose, gaze vectors and facial action units, are extracted using a popular facial behavior analysis toolkit, OpenFace (Baltrusaitis et al., 2018). Then, a LSTM (Long Short Term Memory) neural network is used to learn measuring the communication between the two people from the temporal sequence of the extracted head information. The preliminary experimental results show the effectiveness of the proposed method.

Keywords: Measurement of Communication, omnidirectional video, head information and LSTM neural network

1. Introduction

Effective communication is very important for humans' daily activities including education and business. If communication is not achieved, problems may arise and cause damages for business. For example, if communication between the bank clerk and the customer for housing loan is not carried out smoothly, the business talk may be lost. It will be helpful if we know in what situations communication is achieved or not. To cope with this problem, a method is to record the whole process of the communication and check it manually; however, it will take a lot of time and cost. Therefore, it would be very helpful if we can analyze and evaluate the whole process of the communication automatically and quantitatively by analyzing recorded videos.

However, the measurement of communication is a difficult problem, which involves multiple communication channels. In this paper, we focus on head information of people, including head pose, gaze vectors and facial action units because head information contains plenty of information for measuring communication besides speaking languages, and even beyond speaking languages. Actually, we often can know whether a communication is achieved from people' head motion (nodding), gaze motion (eye contact) and facial expression (smiling or disgusting). Thus, we can measure the communication among people by analyzing their head information.

To implement such a system, we need to solve the following problems.

- Record video data of multiple people simultaneously. Since communication among people is an interactive process, synchronization of head information of different people is a necessary condition.
- Extract head information from image sequences. The head information includes head pose, gaze vectors and facial action units.
- Analyze the head information of the people in order to measure the smoothness of their communication. The communication should be measured for an interval, instead of the whole process. That is, it is better to know during the communication which part goes well and which does not.

In this paper we propose a novel method of measuring the communication between two people sitting around a table from omnidirectional videos by analyzing their head information, as shown in in Fig. 1. After extracting the head information of the two people, including head pose, gaze vectors and facial action units, a LSTM neural network is used to learn measuring the communication between the two people from the sequence of the extracted visual cues. The preliminary experiment is carried out to show the effectiveness of the proposed method.

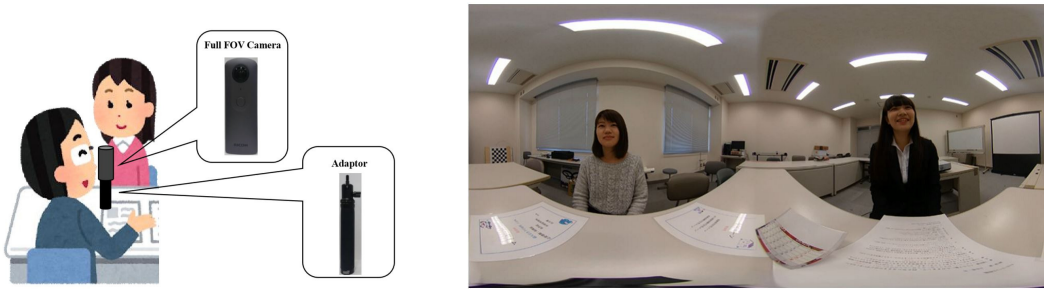

Figure 1: A sketch of the situation of two speakers sitting around a table (left image) and a sample image captured by a full field of view camera (right image).

The remain of this paper is organized as follows. The related research is introduced in Section 2. The proposed method is explained in Section 3. After presenting the experimental results in Section 4, conclusions are given in Section 5.

2. Related Work

In this paper, we focus on the non-verbal communication, such as head motion (such as nodding), gaze vectors (eye contact) and facial expression (such as smiling) . Since human-to-human communication involves multiple people, one approach is to analyze each person's behavior from videos captured by head-mounted cameras, and then integrate the analyzed results of each person. This kind of approach is called first-person vision. The other approach is to use a remote camera, called third-person vision.

2.1. First-Person Vision Approach

In (Alletto et al., 2014) and (Alletto et al., 2015), first person perspective (ego-vision) is proposed to promptly detect people interaction in different social contexts. In (Ye et al., 2015) a wearable camera is used for the contact of eye contact. In (Fathi et al., 2012), a method is proposed for the detection and recognition of social interactions in first-person video of a social event; the location and orientation of faces are estimated and used to compute the line of sight for each face; the context provided by all the faces in a frame is used to convert the lines of sight into locations in space to which individuals attend.

As one of the closest researches, in (Yonetani et al., 2016) which deals with the problem of two-person interactions, both people are equipped with a head mounted camera so that each person always has a first-person point-of-view (POV) observation of one's self in one's own video and a second-person POV observation of the self in another video; and then, the two people's micro-actions and reactions are recognized from the paired egocentric videos.

All the above researches use a head mounted camera, and the captured first-person video is used to detect behaviors in interaction tasks.

2.2. Third-Person Vision Approach

In (Chen et al., 2011), a system is built to track the location and head pose of multiple people for discovering the space utilization and human interactions. In (Huang and Kitani, 2014), the task of activity forecasting in the context of dual-agent interactions is explored to understand how the actions of one person can be used to predict the actions of another. In (Ryoo and Aggarwal, 2009), a spatio-temporal relationship match is designed to measure structural similarity between sets of features extracted from two videos; the proposed match hierarchically considers spatio-temporal relationships among feature points, thereby enabling detection and localization of complex non-periodic activities. In (Vahdat et al., 2011), an activity is modeled with a sequence of key poses, important atomic-level actions performed by the actors, and a model is developed for recognizing human interactions – activity recognition with multiple actors.

Until now, these researches of third-person vision approach are carried out for a surveillance videos. Since various micro-actions and reactions cannot be well observed in these third-person videos, facial information, such as gaze and facial expression are not coped with.

2.3. Contributions

In this paper, we consider a setting, in which users are sitting around a table. In such a setting, an omnidirectional camera placed at center of table can observe micro-actions and reactions of people from a third-person vision video with high resolution, as shown in Fig. 1. The contribution of this paper is as follows.

- We use an omnidirectional camera to observe multiple people. From the omnidirectional video, people's micro-actions and reactions can be observed simultaneously. It is very effective and efficient in comparison with first-person vision approach, such as in (Yonetani et al., 2016) , which need to combine the videos of two people.

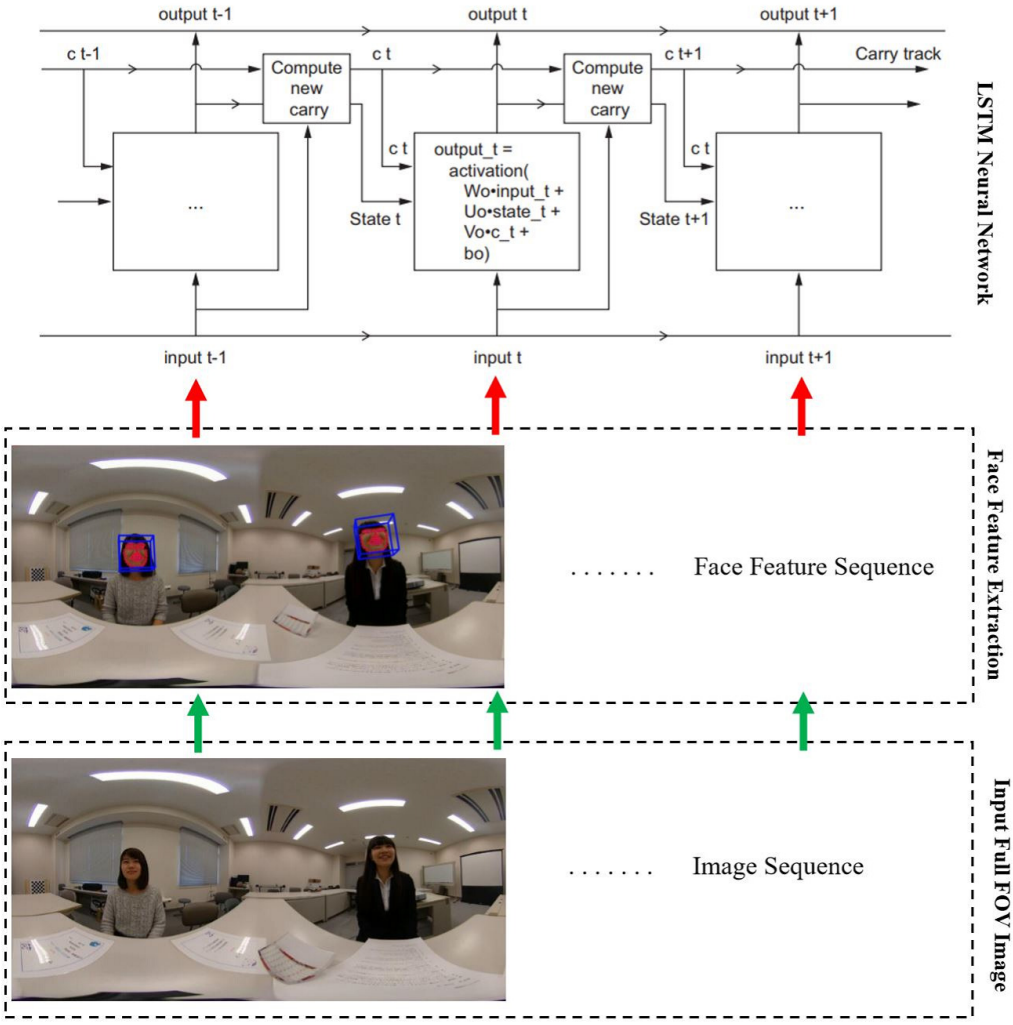

Figure 2: The block diagram of the proposed method. At the bottom is the captured omnidirectional image sequence; at the middle is the face feature sequence; at the top is a LSTM neural network.

- We propose a all-in-one method to measure communication between two people. Instead of focus on some specific behavior, such nodding, eye contact or smiling, we use the head pose, gaze vectors, and facial action units of two people at each instant simultaneously as the input of a LSTM (Long Short-Term Memory) neural network (Hochreiter and Schmidhuber, 1977) to measure their communication.

3. Proposed Method

In this section, we explain the proposed method according to the process shown in Fig. 2.

3.1. Acquisition of omnidirectional Videos

In the setting of this paper, two people sits at the opposite sides of a table, one person starring a bank clerk and the other a customer. A script which mimics a talk between a bank clerk and a customer about housing loan is prepared beforehand. An omnidirectional camera is placed around the center of table and records the conversation lasting about one and a half minutes, as shown in Fig. 2. Totally, 20 pieces of videos are collected.

3.2. Face Feature Extraction

We use a popular facial behavior analysis toolkit, called OpenFace (Baltrusaitis et al., 2018), to extract face features for every frame of the videos. The output of OpenFace includes landmarks (the coordinates of face feature points), head pose, gaze vectors of eyes, and facial action units corresponding facial expression. Since head pose and gaze vectors are estimated from landmarks, in this paper we use head pose, gaze vectors and facial action units as head features for measuring the smoothness of communication.

As shown in Fig. 2, although the facial behavior analysis toolkit OpenFace is developed for the processing of perspective images. We find that it is not a problem for detecting faces from an equirectangular image when the faces are close to the equator (vertical center line) of an equirectangular image . In fact, we can also generate perspective images from an equirectangular image firstly, and then, apply OpenFace to the generated perspective images to extract head features if necessary.

3.3. LSTM Neural Network

To measure the smoothness of communication, we need an observation obtained during a period of time. This means that the input should be a temporal sequence of head features. Therefore we use a LSTM neural network (Hochreiter and Schmidhuber, 1977) in our task, as shown in Fig. 2. LSTM neural network is one of recurrent neural networks. By adding a way of carry information across many time steps, LSTM can prevent older signals from gradually vanishing during processing, that is, eliminate the vanishing gradient problem. (Chollet, 2017).

Concretely, in this paper a temporal sequence of head features of 90 image frames is used to determine the state of communication occurring at next time instant.

4. Experiments

4.1. Data Set

In this paper, 20 pieces of videos imitating a talk between a bank clerk and a customer about housing loan are collected. Every frame of the videos is labeled manually with binary values corresponding to whether the communication is achieved. Data i, $\mathbf{d}_i$, is composed of head features of two people, including head pose $(\mathbf{p}_1, \mathbf{p}_2)$, gaze vectors $(\mathbf{g}_1, \mathbf{g}_2)$ and facial action units $(\mathbf{a}_1, \mathbf{a}_2)$ and a label (c) for frame i.

$$\mathbf{d}_i = \{(\mathbf{p}_{1i}, \mathbf{g}_{1i}, \mathbf{a}_{1i}, \mathbf{p}_{2i}, \mathbf{g}_{2i}, \mathbf{a}_{2i}), c_i\}$$

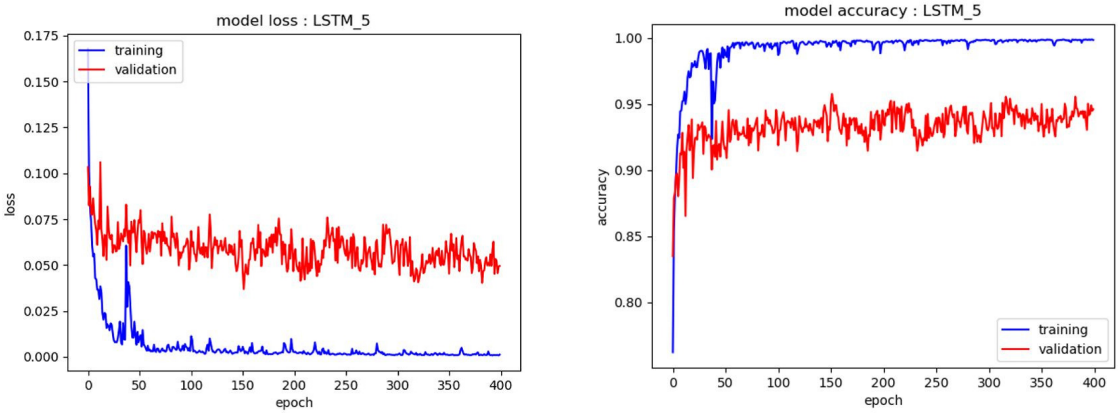

Figure 3: Loss (left image) and accuracy (right image) in training and validation.

. The data set, $\mathbf{D}$, used in this research is composed of 60432 data, totally. That is,

$$\mathbf{D} = \{\sum \mathbf{d}_i | i = 1, ..., 60432\}.$$

4.2. Network Training and Test

We divide the data set $\mathbf{D}$ including 60432 data into three parts: 52096 data is used for training, 5789 data for validation, and 2546 data for test. Note that the temporal continuity of these data is maintained when they are input to the LSTM neural network. The LSTM neural network is implemented using Keras (Chollet, 2017).

In Fig. 3, the left image shows the loss for the training and validation, and the right image shows the accuracy for the training and validation. The loss of the training and validation decrease to 0 and 0.06, respectively, and the accuracy of training and validation approach to 1 and 0.93, respectively.

Next, the learned model is applied to the test data set. A sample is shown in In Fig. 4. The left image of Fig. 4 shows the ground truth and predicted values by orange and blue colors, respectively. The right image of Fig. 4 show the difference between the ground truth and predicted values; the difference, i.e., the error, mainly appears at the boundary where the state of communication changes. The 90.2% accuracy is achieved for the entire test data set.

5. Conclusions

In this paper we propose a method of measuring communication between two people by observing their head information from omnidirectional videos. The head information used includes head pose, gaze vectors and action units, which are extracted using a facial behavior analysis toolkit, called OpenFace (Baltrusaitis et al., 2018). A temporal sequence of face information is input to a LSTM recurrent neural network to learn the measurement of communication. As shown in the experimental results, for the specific setting of this paper, as high as 90.0% accuracy of measurement of two-people communication is achieved.

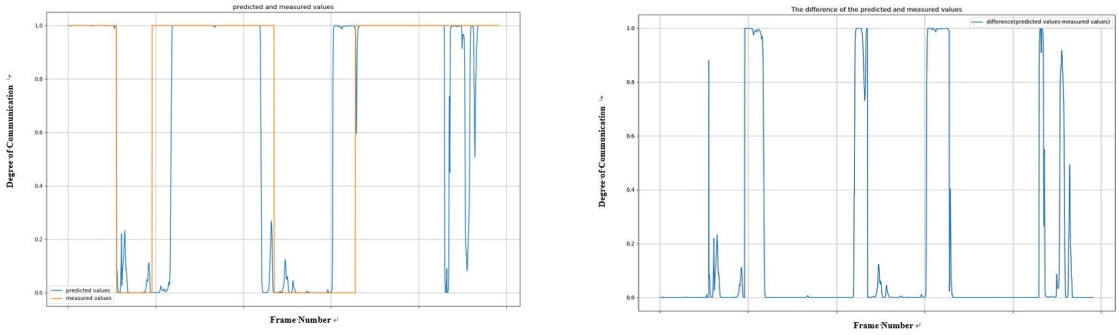

Figure 4: Test results. The left image shows the ground truth and predicted values by orange and blue colors, respectively. The right image show the difference between the ground truth and predicted values.

In this paper, we focus on two people and non-verbal head information. How to cope with the problem of more than three people and the inclusion of verbal information is our future work to do.

References

S. Alletto, G. Serra, S. Calderara, F. Solera, and R. Cucchiara. From ego to nos-vision: Detecting social relationships in first-person views. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, page 4082 – 4096, 2014.

S. Alletto, G. Serra, S. Calderara, and R. Cucchiara. Understanding social relationships in egocentric vision. *Pattern Recognition*, 48(12):594–599, 2015.

T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency. Openface 2.0: Facial behavior analysis toolkit. *IEEE International Conference on Automatic Face and Gesture Recognition*, 2018.

C.-W. Chen, R. Ugarte, C. Wu, and H. Aghajan. Discovering social interactions in real work environments. *Proceedings of the IEEE International Conference on Automatic Face Gesture Recognition and Workshops (FG)*, page 933–938, 2011.

Francois Chollet. Deep learning with python. *MANNING*, pages 178–224, 2017.

A. Fathi, J. K. Hodgins, and J. M. Rehg. Social interactions: A first-person perspective. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, page 1226–1233, 2012.

S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8): 1735–1780, 1977.

D.-A. Huang and K. M. Kitani. Action-reaction: Forecasting the dynamics of human interaction. *Proceedings of the European Conference on Computer Vision (ECCV)*, page 489–504, 2014.

M. Ryoo and J. Aggarwal. Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, page 1593–1600, 2009.

A. Vahdat, B. Gao, M. Ranjbar, and G. Mori. A discriminative key pose sequence model for recognizing human interactions. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, page 1729–1736, 2011.

Z. Ye, Y. Li, Y. Liu, C. Bridges, A. Rozga, and J. M. Rehg. Detecting bids for eye contact using a wearable camera. *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–8, 2015.

Ryo Yonetani, Kris M. Kitani, and Yoichi Sato. Recognizing micro-actions and reactions from paired egocentric videos. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.