Automatic Communicative Engagement Measurement and Conversational States Detection Using Visual Movement Signals

Jingguang Han
Accenture Technology Labs, Dublin, Ireland
Email: jingguang.han@accenture.com

Nick Campbell
School of Computer Science and Statistics, Trinity College Dublin, University of Dublin, Dublin, Ireland
Email: nick@tcd.ie

Abstract*—Communicative dynamics have received increasing focus from researchers in the field of social signal processing. These cognitive dynamics can enhance the experience and smoothness of both human-to-human and human-to-machine interactions dramatically. Communicative engagement [1] plays an important part among these dynamics. Promising detection and measurement results of communicative engagement have been reported with quantitative models using acoustic and linguistics signals. To our best knowledge, there is no automated system that utilizes visual signals for communicative engagement detection and measurement, nor is there any clear quantitative model. To bridge this gap, this paper presents a novel method of using multi-dimensional visual signals to automatically detect and measure communicative engagement in multi-party conversations, and a machine learning approach to automatically predict conversational states of participants: speaker vs the most engaged listener. We also present a multi-modal audio-video corpus designed and recorded by one of the authors with multiple microphones and one 360-degree video camera for three days of four people participating in natural and spontaneous social conversations. A face detection and movement measurement system based on the Viola-Jones algorithm and color differentiation algorithm was developed for quantitative analysis of visual movement signals. We applied a series of statistical methods to measure the communicative engagement in multi-party conversations with the visual movement dataset. The results are validated in comparison with the same calculations using randomly tailored signals. The comparison shows a significantly stronger correlation of the visual signals between the participants who are engaged in the communication than using pseudo signals. The result also shows a high probability of 87.3% that pairs of participants with the highest engagement coefficient containing the speakers. Furthermore, a support vector machine was trained with a 5-dimensional movement dataset and applied to predict the conversational states of the participants and distinguish the most engaged listener vs the speaker. Cross validation shows a promising 79.04% accuracy.

***Index Terms*—social signal processing, communicative engagement, image and video processing, correlation analysis, SVM**

I. Introduction

Recently, non-verbal signals for communicative dynamics studies, such as visual and acoustic signals, have been the focus of an increasing number of both human-human and human-machine communication researches. Stanford and Google AI director Prof. Fei-Fei Li has foreseen that cognitive signal understanding especially through non-verbal channels will be the next boost of AI in the recent future. Great progress has been made in longstanding areas of verbal signal processing such as speech recognition, synthesis and semantic machine understanding for the past decade, but the latest advancement in terms of detection accuracy using verbal signals alone has been slow. Research [2] has suggested that, over 60% of the information delivered in face-to-face communication is through non-verbal channels. Visual signal plays an important part among these non-verbal signals, which is similar to that we use our eyes as the primary channel to capture cognitive visual information like gestures and facial expressions in face-to-face communication. Previous experiments and studies have shown strong evidence of the significance and usefulness of using visual signals in communication studies. Han *et al.* [3] designed a robotic platform for studying human and robot interaction using visual and acoustic signals. The robot was invited for an exhibition in Science Gallery of Ireland for 3 months. A group of researchers [4] conducted a set of experiments of studying human-robot interaction by building a dialogue management system using Wikipedia as the knowledge base and implemented it in the Nao robot platform. Visual signals were shown to have a strong impact on the overall user experience and satisfaction level.

Communicative engagement between participants is a relatively new research area among non-verbal dynamics in social interactions. Most of the recent researches have

Manuscript received January 19, 2017; revised March 29, 2017.

doi: 10.18178/ijsps.5.1.44-48

been focusing mainly on acoustic signal modeling and processing, including measuring tunes similarity among different speakers in a conversation using prosodic signals [5], and measuring the agreement and disagreement level in task-based conversations using speech signal [6]. Visual signal processing and understanding has been reported to be used in human-machine platforms quite successfully. For example, the SEMAINE platform [7] studies human and computer interactions using visual signals along with others for both perceiving and delivering cognitive dynamics. It was developed by a group of researchers across different countries. However, the area of applying video and image signal processing and modeling techniques in analyzing multi-party human-human communications and making predictions, has not yet received the same level of focus. As a human, it is not difficult to recognize the cognitive visual information such as emotions and conversational timing, i.e. when to speak and when to listen. However, this still remains a challenge for computer systems.

In this paper, we present and discuss a novel method of using multi-dimensional visual signals to automatically detect and model communicative engagement in multi-parity conversations. Section II presents a multi-party conversation corpus recorded using multi-modal approaches by one author of this paper. Section III discusses data processing and feature extraction techniques using face detection and movement measurement algorithms from video sequences. In section IV, we describe a statistical method for measuring communicative engagement level of participants and a machine learning approach to predict participant's conversational states: speaker vs the "most engaged" listener.

II. Video Corpus Design

Figure 1. The Tabletalk corpus recorded by 360 camera

In this section, we discuss an annotated multi-modal corpus designed and recorded by one author of this paper [8]. The "TableTalk" corpus [1] was designed for social conversational studies of four participants from different cultural backgrounds: one native English speaker (British) and three speakers of English as a second language (one Belgian, one Finn and one Japanese). The objective was to record a corpus of people talking as naturally and spontaneously as possible without any restriction [9]. The total length of this corpus is 3 hours and 30 minutes across 3 days. A multi-modal audio-video approach was applied in recording with multiple microphones and cameras. A 360-degree video camera was used in the recording which simultaneously captured the frontal faces of all the four participants around the table. It was primarly utilized to synchronize the video and audio streams from other devices. The corpus contains conversations of 31523 utterances in total and was annotated manually in terms of timing and content. Fig. 1 is a recording view of the corpus from the 360-degree camera.

[1] The TableTalk Corpus: http://sspnet.eu/2010/02/freetalk/

III. Video Signal Processing and Feature Extrction

In this section, we present an automatic video processing system utilizing two image processing techniques and algorithms: the Viola-Jones algorithm [10] for face detection and the color differentiation algorithm for head and body movement measurement.

Facial movement is an important signal for analyzing and understanding communicative engagement in multi-party conversations [11]. However, it has been very difficult to quantify the amount of movement from video sequences as features for further statistical analysis and modeling. We designed and developed an automatic image and video signal processing system to quantitatively measure visual movements (face, head and body) from the recorded videos. The Viola-Jones algorithm has been proved to work well and efficiently in object detection using Haar Cascade features, even in low resolution videos [10]. Video streams either from real-time cameras or recorded video files can be sampled into a sequence of static images. By calculating the distances of the detected faces' coordinates in the image sequence, the face movement can be measured in three dimensions: vertical, horizontal and forward/backward (distance to the camera) movement. We used OpenCV [12] which is an image processing framework containing an implementation of the Viola-Jones algorithm for detecting human faces from static images. Previous research [12] has reported an average 95% detection accuracy using this approach. For each video frame, which is a static image, the system outputs an array of 3-dimensional facial coordinates. Each coordinate group contains three parameters: horizontal position, vertical position and forward/backward position. The forward and backward position is measured using the size of face indicating the distance between the face and recording camera. When the participant moves closer to the camera, the size of detected face is bigger and vice versa. OpenCV estimates face region as a square. The coordinates represent vertical and horizontal positions and size of each detected face square. During entire recording of the corpus, a static face picture (see the top left corner of Fig. 1) has been placed beside the table to synchronize the face movement signals of all the participants in time. One example is that when people turned their faces aside and the detection system failed to capture the frontal faces, the system output will generate gaps in this continuous signal. The synchronization method described above was used to

remove the gaps and align with other people's signals in time. This approach was extended from the approach proposed by Douxchamps *et al.* [13]. We also improved the system by leveraging the latest advancement of OpenCV framework for better detection accuracy and more efficient computation.

The color differentiation algorithm [14] has been widely used as a means of quantifying the overall movement in a specific region such as head or body. When the face of a participant is detected, a region, twice as large as the detected face square and sharing the same center coordinates, is marked as the face movement activity region of interest [13]. Consequently, a square with the edge length 2.5 times of the detected face square, is placed directly under the face region and marked as the body movement activity region of interest. The total number of different pixels between two adjacent video frames is used as the index of the head or body movement. This method provides two additional dimensions to the movement features. In total, we extracted five different movement features from the automatic image/video processing system: using the Viola-Jones face detection algorithm: 1. horizontal face movement, 2. vertical face movement, 3. forward/backward face movement, and the color differentiation algorithm: 4. overall head movement 5. overall body movement. The face coordinates of the same person in two adjacent video frames can be represented as $F_1 = (x_1, y_1, z_1)$ and $F_2 = (x_2, y_2, z_2)$, where x, y and z are the horizontal, vertical and forward/backward coordinates in the 3-dimensional space. Then the movement in each dimension can be measured as $|x_1 - x_2|$, $|y_1 - y_2|$ and $|z_1 - z_2|$ respectively. Moreover, the overall movements in 2 dimensional and 3 dimensional spaces can be measured using (1) and (2) where m represents the amount of movement.

$$m = \sqrt[2]{(x_1 - x_2)^2 + (y_1 - y_2)^2} \quad (1)$$

$$m = \sqrt[2]{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2} \quad (2)$$

IV. Data Analysis and Results

We conducted a series of correlational and statistical experiments for detecting and measuring communicative engagement between participants in multi-party conversations. Comparing with the manual utterances annotation, we found that there is a probability of 87.3% that the most engaged participants pair containing the speaker. Moreover, a supervised machine learning approach with the Support Vector Machine (SVM) [15] classifier was used to automatically predict the participants' conversational states: speaking vs listening, and distinguish speakers and the "most engaged" listener from the participant pair with the highest communicative engagement coefficient. The overall prediction accuracy is 79.04%.

A. Communicative Engagement Quantization and Measurement

With the visual movement data obtained from Section III, we constructed a five-dimensional data matrix for all the four participants involved in the conversation, as described in the last section. After that, we removed the random noise in the dataset such as unrealistically high values (e.g. people cannot move 100 meters per second) with a low pass filter [16]. Correlational analysis requires the movemet data of participants in a singular vector format. In this sense, we applied the Principal Component Analysis (PCA) [17] to combine the five dimensional features into one and keep most of the original variance. We selected the first component after the Singular Value Decomposition (SVD) [18] in PCA. This component column vector contains 72.4% of the data variance in the original five-dimensional movement dataset. It was used as the primary variable to represent the overall movement of participants. A customized user interface was built to align the movement signals in time. Fig. 2 is a visualization sample of the overall movement signals alignment for the participants in Fig. 1.

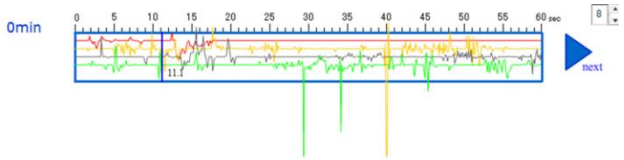

Figure 2. The visualization of motion cues: different colors represent different participants in the corpus

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 (y_i - \bar{y})^2}} \quad (3)$$

Equation (3): Pearson product-moment correlation coefficient

We measured communicative engagement coefficients using the Pearson product-moment correlation calculation (3) [19] for all the 6 possible participant pairs in the corpus (every possible pair out of the 4 participants). From the manual annotation of the dialogue utterances, we can know who is the speaker within a specific time interval. We then calculated all the three Pearson's correlation coefficients between the speaker's movement signal and the other three listeners. The average correlation coefficient of the three speaker and listener pairs is r = 0.41. The null-hypothesis significance test [20] shows an average P-value [21] of 0.015 which indicates the correlation calculation is statistically significant. For validation and comparison purpose, we segmented the speaker's overall movement vector into two halves and moved the second half before the first to generate a pseudo signal. The same Pearson's correlation calculations were applied again to calculate the coefficients using the pseudo speaker signal and real listener signals. The average coefficient was reduced to r = 0.28 (comparing with 0.41), which is a significant drop when using the real signals for both speaker and listeners. The experiments show that Pearson product-moment correlation coefficient is a valid method to quantitatively measure the visual communicative engagement in multi-party conversational interactions.

As described above, the whole corpus was annotated based on dialogue utterances and speaking timing. We

segmented the movement data into different windows according to the annotated time intervals, where the average length of the windows is 5.4 seconds. The system that we developed is able to process video data at a rate of 15 frames per second, so that 15 ∗ 5.4 = 81 frames are grouped into one window on average. We then selected the two participants whose correlation coefficient is the highest among all the 6 possible pairs for each window. We compared these two selected participants with the annotated speaker in a specific time window. The result shows that a probability of 87.3% that the highest engaged participant pairs contains the annotated speakers. It means that the speakers in multi-party conversations are more visually engaged in the communication. This finding also co-validates the result of the Pearson product-moment correlation coefficient analysis from another angle.

B. *Conversational States Detection, Speaking VS Listening*

From the experiments above, we found that 87.3% of chance that the speakers are within the highest engaged participants pair. For the next step, we aimed at automatically distinguishing the "most engaged" listener from the speaker in this pair. As mentioned in section III, we structured five movement features: horizontal, vertical and forward/backward movements, head and body overall movements. The pieces of data were labeled into two categories: speaker and listener based on the manual annotation, i.e. label each data piece 0 for listener and 1 for speaker. In this way, a labeled dataset 27439 * 6 matrix was generated. The first column of the dataset contains the categorical labels and the rest contains the five features.

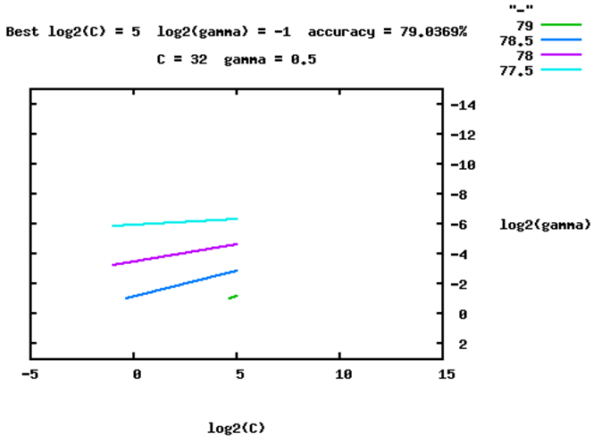

Figure 3. SVM cross validation fitting C and γ in equation (4) and (7) with prediction accuracy of 79.04%

The Support Vector Machine classifier [22] with Gaussian similarity kernel [23] was trained to predict the conversational states (speaking vs listening). We used the LibSVM [24] library for training, which was proved to be one of the most efficient SVM implementations. We utilized the five-fold cross validation method for evaluating the SVM results. The whole dataset was divided into two subsets randomly: the training dataset contains 70% of all the data; the test dataset contains the rest 30% data. We then started training and testing the SVM classifier with these datasets, during which process the classifier was also optimized with the best kernel parameters C and γ fitting the given datasets. This experiment was conducted five times with different training and tesing datasets and the best kernel parameters had been selected for the final SVM classifier construction. The overall prediction accuracy is 79.04% as shown in Fig. 3. Combing the two approaches discussed above, we are able to automatically detect/predict the speaker and the most engaged listener in a multi-party conversation.

$$f = \min_{\theta} C\sum_{i=1}^{n}[y_i \cos t_i(\theta^T x_i) + (1-y_1)\cos t_0(\theta^T x_i)] \quad (4)$$

where:

$$\cos t_0 = \log(1 - \frac{1}{1+e^{-\theta^T x}}) \quad (5)$$

$$\cos t_1 = \log(\frac{1}{1+e^{-\theta^T x}}) \quad (6)$$

Gaussian similarity kernel:

$$f_i = similarity(x, l_i) = \exp(-\frac{\left\|x - l_i\right\|^2}{2\gamma^2}) \quad (7)$$

V. Summary and Discussion

In this paper, we discussed a novel approach of measuring communicative engagement and predicting the speaker and the "most engaged" listener in multi-party conversations using visual movement signals. We firstly presented a corpus of multi-party natural and spontaneous conversations. An automatic face and body movement detection and quantization system using the Viola-Jones and color differentiation algorithms was designed and developed in order to obtain a five-dimensional movement feature dataset. We applied PCA and selected one component which contains most of the movement variances of the original dataset. Then the Pearson product moment correlation algorithm was applied to measure the communicative engagement between participants. The result was evaluated by comparing the correlation coefficient calculated using pseudo signals. It provides strong evidence that the proposed method is a valid approach to measure communicative engagement. By comparing with the manual annotation, our result shows that there is a high probability of 87.3% that the highest engaged participant pair also contains the speaker. Furthermore, a supervised machine learning method with the SVM classifier was trained to predict the conversational states in the highest engaged participants pair: speaker vs the most engaged listener with an overall prediction accuracy of 79%. Other research [5] suggests that a better accuracy could be obtained by using acoustic signals. However, they are highly restrained by the recording environment, e.g. when 1. the environment is very noisy and acoustic signals cannot be recorded clearly or 2. when the microphones are too far away to capture the sound or 3. when the listeners do not speak much during the entire conversation, the video signals could be the only channel that can be used to analyze communicative engagement.

From the experiments conducted and the results discussed above, we demonstrated strong evidence that visual signals can significantly contribute to automatic communicative engagement measurement and conversational states detection in multi-party conversations. Similar to humans use their eyes to capture vast amount of cognitive information during communication, it is obvious that visual signals contain large quantity of communicative information which can be processed and interpreted by machines. Thus, it is increasingly important to explore what types of cognitive data can be obtained from video signals, how they can be interpreted and what kind of prediction we can infer from them. The fusion of signals from different dimensions is also very promising. By combining multiple types of signals such as linguistic, acoustic and visual signals, we can glean more information and make better predictions than we use a unitary signal alone.

ACKNOWLEDGMENT

Special thanks to Science Foundation Ireland for funding this research through the "Focus on Actions in Social Talk: Network-Enabling Technology" (FASTNET) project. Grant No. (SFI) 09/IN.1/I2631.

This work was also supported by the Major Program of the National Social Science Fund of China (13&ZD189)

REFERENCES

[1] C. Gallois, H. Giles, E. Jones, A. C. Cargile, and H. Ota, "Accommodating intercultural encounters: Elaborations and extensions," 1995.

[2] I. N. Engleberg, "Communication principles and strategies," 2006.

[3] J. G. Han, J. Dalton, B. Vaughan, C. Oertel, C. Dougherty, C. D. Looze, and N. Campbell, "Collecting multi-modal data of human-robot interaction," in *Proc. IEEE 2nd International Conference on Cognitive Infocommunications*, 2011, pp. 1–4.

[4] J. G. Han, N. Campbell, K. Jokinen, and G. Wilcock, "Investigating the use of non-verbal cues in humanrobot interaction with a nao robot," in *Proc. IEEE 3rd International Conference on Cognitive Infocommunications*, 2012, pp. 679–683.

[5] C. D. Looze and S. Rauzy, "Measuring speakers' similarity in speech by means of prosodic cues: Methods and potential," in *Proc. Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[6] B. Vaughan, "Prosodic synchrony in co-operative taskbased dialogues: A measure of agreement and disagreement," 2011.

[7] S. Petridis and M. Pantic, "Fusion of audio and visual cues for laughter detection," in *Proc. International Conference on Content-Based Image and Video Retrieval*, 2008, pp. 329–338.

[8] N. Campbell and A Tabeta, "A software toolkit for viewing annotated multimodal data interactively over the web," in *Proc. LREC*, 2010.

[9] N. Campbell, "An audio-visual approach to measuring discourse synchrony in multimodal conversation data," *Linguistic Theory and Raw Sound*, vol. 40, p. 199, 2009.

[10] P. Viola, M. J. Jones, and D. Snow, "Detecting pedestrians using patterns of motion and appearance," *International Journal of Computer Vision*, vol. 63, no. 2, pp. 153–161, 2005.

[11] A. Csapo, E. Gilmartin, J. Grizou, J. G. Han, R. Meena, D. Anastasiou, K. Jokinen, and G. Wilcock, "Multimodal conversational interaction with a humanoid robot," in *Proc. IEEE 3rd International Conference on Cognitive Infocommunications*, 2012, pp. 667–672.

[12] G. Bradski, "The opencv library," *Doctor Dobbs Journal*, vol. 25, no. 11, pp. 120–126, 2000.

[13] D. Douxchamps and N. Campbell, "Robust real time face tracking for the analysis of human behaviour," *Machine Learning for multimodal Interaction*, pp. 1–10, 2008.

[14] C. F. Lam and M. C. Lee, "Video segmentation using color difference histogram," in *Multimedia Information Analysis and Retrieval*, Springer, 1998, pp. 159–174.

[15] J. A. K. Suykens and J. Vandewalle, "Leastsquares support vector machine classifiers," *Neural Processing Letters*, vol. 9, no. 3, pp. 293–300, 1999.

[16] J. F. Kaiser and W. A. Reed, "Data smoothing using lowpass digital filters," *Review of Scientific Instruments*, vol. 48, no. 11, pp. 1447–1457, 1977.

[17] H. Hotelling, "Analysis of a complex of statistical variables into principal components," *Journal of Educational Psychology*, vol. 24, no. 6, p. 417, 1933.

[18] C. Eckart and G. Young, "The approximation of one matrix by another of lower rank," *Psychometrika*, vol. 1, no. 3, pp. 211–218, 1936.

[19] T. R. Derrick, B. T. Bates, and J. S. Dufek, "Evaluation of time-series data sets using the pearson product-moment correlation coefficient," *Medicine and Science in Sports and Exercise*, vol. 26, pp. 919–919, 1994.

[20] W. W. Rozeboom, "The fallacy of the nullhypothesis significance test," *Psychological Bulletin*, vol. 57, no. 5, p. 416, 1960.

[21] I. Lawrence and K. Lin, "A concordance correlation coefficient to evaluate reproducibility," *Biometrics*, pp. 255–268, 1989.

[22] S. Amari and S. Wu, "Improving support vector machine classifiers by modifying kernel functions," *Neural Networks*, vol. 12, no. 6, pp. 783–789, 1999.

[23] Y. Chen, X. S. Zhou, and T. S. Huang, "One-class svm for learning in image retrieval," in *Proc. International Conference on Image Processing*, 2001, vol. 1, pp. 34–37.

[24] C. C. Chang and C. J. Lin, "Libsvm: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, p. 27, 2011.