Visual-Based Eye Contact Detection in Multi-Person Interactions

Mahmoud Qodseya, Franck Jeveme Panta, Florence Sèdes

Visual-Based Eye Contact Detection in Multi-Person Interactions

Mahmoud Qodseya
IRIT Laboratory
University of Toulouse
Toulouse, France
mahmoud.qodseya@irit.fr

Franck Panta
IRIT Laboratory
University of Toulouse
Toulouse, France
franck.panta@irit.fr

Florence Sedes
IRIT Laboratory
University of Toulouse
Toulouse, France
florence.sedes@irit.fr

***Abstract*—Visual non-verbal behavior analysis (VNBA) methods mainly depend on extracting an important and essential social cue, called eye contact, for performing a wide range of analysis such as dominant person detection. Besides the major need for an automated eye-contact detection method, existing state-of-the-art methods require intrusive devices for detecting any contacts at the eye-level. Also, such methods are completely dependent on supervised learning approaches to produce eye-contact classification models, raising the need for ground truth datasets. To overcome the limitations of existing techniques, we propose a novel geometrical method to detect eye contact in natural multi-person interactions without the need of any intrusive eye-tracking device. We have experimented our method on 10 social videos, each 20 minutes long. Experiments demonstrate highly competitive efficiency with regards to classification performance, compared to the classical existing supervised eye contact detection methods.**

***Index Terms*—eye contact detection, visual nonverbal behavior analysis, social interaction**

I. Introduction

In sociology, social interaction is a dynamic relationship of communication and information exchange between two or more individuals within a group. The analysis of this relationship provides a better understating of the human behavior in different contexts and scenarios. This analysis is not only important for social scientists, but also to those of us who want to understand better our own behavior and the behavior of our fellows.

The main purpose of social interaction analysis is to recognize and interpret human social interactions by analyzing their sensed social cues. These cues can be categorized to verbal (word) and nonverbal (wordless/visual) information [1]. The verbal behavioral cues take into the account the spoken information among persons, such as yes/no responses in answering question context. The nonverbal behavioral cues represent a set of temporal changes in neuromuscular and physiological activities, which send a message about emotions, mental state, personality, and other characteristics [2].

Nonverbal cues are accessible to our senses by sight and hearing, making them detectable through microphones, cameras or other suitable sensors (e.g., microphone, accelerometer). Nonverbal cues can be taxonomized into vocal and visual cues, where: (i) vocal cues include voice quality, silences, turn taking patterns, nonlinguistic vocalizations, and linguistic vocalizations; and (ii) visual cues include physical appearance (e.g., gender, height, ethnicity, age), face and eyes cues (e.g., facial expression, gaze direction, focus of attention), gesture and posture, and space and environment.

As shown in Figure 1, a visual non-verbal behavioral analysis schema consisting of five modules: (i) data acquisition; (ii) person detection and tracking; (iii) social cues extraction; (iv) contextual information identification; and (v) social cues analysis. Different types of sensors and devices, e.g. cameras and proximity detectors, might be used in the data acquisition module to record social interactions. Thus, one or more dedicated computer vision and image processing based (e.g. face detection) methods could be leveraged for processing the input data to detect and track person(s). The social cues extraction module takes as an input the detected person(s) to extract a feature vector (per person) describing the social cues such as head pose. The social cues understanding module deeply analyzes the primitive social cues through modeling temporal dynamics and combining signals extracted from various modalities (e.g., head pose, facial expression) at different time scales to provide more useful information and conclusions at the behavioral level of the detected persons. Indeed, this module might optionally leverage additional contextual information (e.g. type of the event, location, restaurant menu) that describe the context, in which the data is captured, to provide a precise social behavior prediction and analysis. Finally, the existence of metadata repository decouples the analysis phase from other components [3].

At the social cues extraction level, VNBA systems mainly adopt eye contact, as an important social cue, for performing a wide range of analysis and studies such as a dominant person detection [4]. It provides multiple functions in the two-person contacts such as information seeking, establishment and recognition of social relationships, and signaling that the "channel is open for communication" [5]. Indeed, extraction of this social cue must be fully automated, accurate at detection level, and compatible with simple capturing devices such as closed-circuit television (CCTV) cameras. However, existing state-of-the-art methods require expensive special devices for

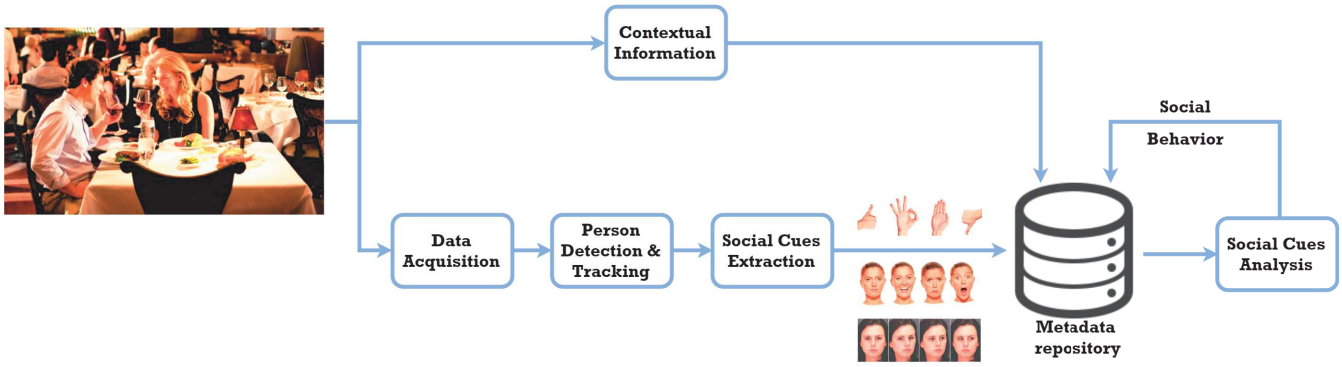

Fig. 1. A visual non-verbal behavioral analysis schema.

detecting any contacts at the eye-level. Such methods are also based on supervised machine learning techniques to produce eye contact classification models, raising the need for ground truth datasets as a difficult and time consuming task.

Eye contact detection is defined as a task of automatically detecting whether two people look at each other's eyes or face simultaneously. It is an important feature for better understanding human social behavior. Eye contact detection has numerous applications. For example, it is a key component in attentive user interfaces and it is used to analyze turn-taking, social roles, and engagement during multi-person interactions. Even more, we can deduce many things based on the eye contact [5]: (i) the topic nature, in which, there is more eye contact in case of the topic being discussed is straightforward and less personal, whereas, there is less eye contact during the hesitating passages; (ii) the relation between two persons, in which, there is more eye contact if the two persons are positively interested in each other.

In this paper, we introduce a novel geometrical method to detect eye contact in small group interactions using multiple cameras. Our method first extracts all participants' head pose from several ambient cameras and then map them to a common reference frame. After that, a check is performed for each detected person if there is an intersection between his/her gaze direction with other detected persons. Then, a temporal $Looking\_At$ square matrix is built by which we can check whether an eye contact between two participants holds or not. Our method does not require intrusive devices, which makes participants behave more naturally, and it is not limited to dyadic interactions.

The rest of paper is organized as follows. Section II reviews the related works. Section III illustrates our eye contact approach. Section IV shows a real-case experimentation that we have performed to evaluate our method for eye contact detection. Finally, in section V, we conclude with suggestions for some future directions.

II. Related Work

A. Social Behavior Analysis Methods

Various studies have been performed to detect, analyze, and asses social interactions using automatic machine learning based methods, including an automatic extraction of non-verbal social signals corresponding to multimodal (e.g., eye contact, touching, etc) nature of interactions [1]. These studies have been applied in a tremendous range of applications and domains, including role recognition [6], social interaction detection in smart meeting [7], and work environments [8], detecting deceptive behavior [9], detecting dominant people in conversations [10], and studying parent-infant interaction [11].

Social signals that have been investigated during social interactions are primitive and context independent because they are not semantic in nature and often occur unconsciously. These signals include frequency and duration of non-verbal behavioral cues occurrences such as the number of eye contact actions happened between two persons.

B. Eye Contact Detection Methods

Eye contact detection is a binary decision on whether someone gaze falls onto target (e.g., face, screen) or not. Many methods have been developed to handle this issue by either using a head-mounted device [12], [13] or requiring LEDs attached to the target [14]. To avoid the intrusive devices, more works focus on developing methods that do not require any intrusive device such as the work of Smith et al. [15] as they have used a classification approach to determine eye contact with a camera, but their method requires prior knowledge about the size and location of the target. Zhang et al. [16] have presented a method for eye contact detection during dyadic (two-person) interactions; however, their method works only for a single eye contact target that must be the closest object to the camera. This assumption does not hold for multi-person interactions in which multiple targets are available.

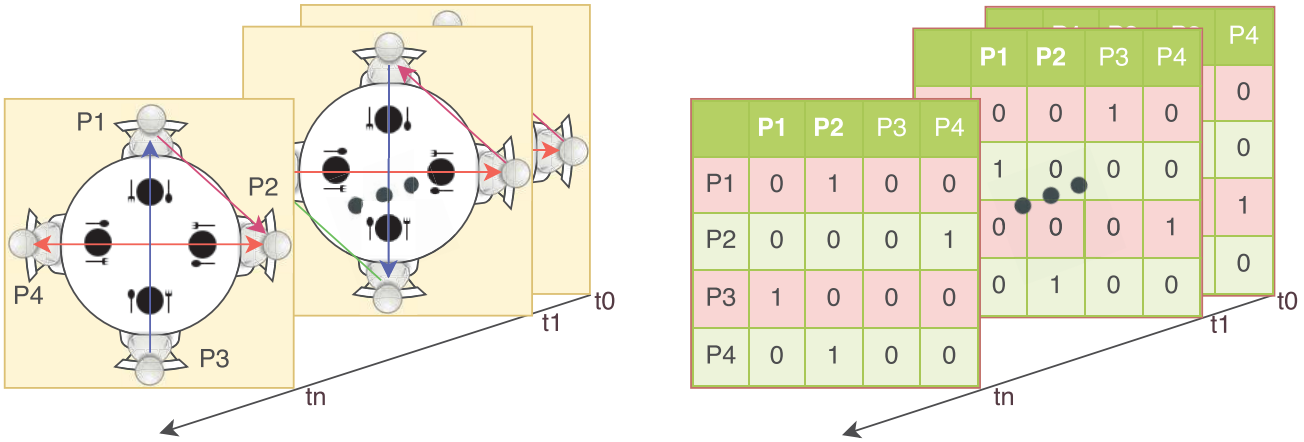

Fig. 2. *Looking_At* square matrix example. Pi is the i^{th} person; on the table, the value of (x, y) is 1 if Px is looking at Py else it is 0.

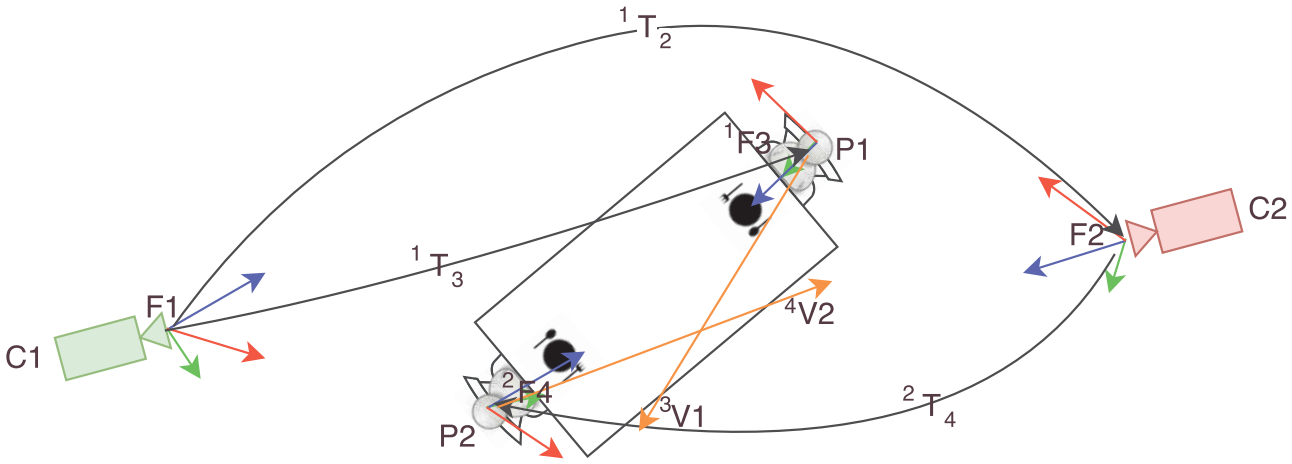

Fig. 3. Looking_At evaluation between two persons. $C1$, $C2$ are first and second cameras; $P1$, $P2$ are first and second persons; $F1$ is the reference frame of $C1$, $F2$ is the reference frame of $C2$; ${}^1F3$ is $P1$ head *pose* $w.r.t.$ $F1$, ${}^2F4$ is $P2$ head *pose* $w.r.t.$ $F2$; ${}^iT_j$ is the *pose* of Fj $w.r.t.$ Fi; ${}^3V1$ is the gaze direction of $P1$ $w.r.t.$ ${}^1F3$, ${}^4V2$ is the gaze direction of $P2$ $w.r.t.$ ${}^2F4$.

III. Materials and Method

Our eye contact detection approach uses CCTV cameras mounted at a particular height in the place where the participants set around a table. The number of cameras is conditioned by arrangement of participants around the table so that a single camera is enough if the participants set in a horizontal way and the camera covers the participants' frontal face. Then, the video streaming of the camera(s) is analyzed by processing each frame streamed from the cameras to detect eye contact between any two participants.

To detect the eye contact between the participants, we have to identify the number of participants, denoted as n, the head *pose*, and gaze direction of each person. n is given as an external information, while the participant's head *pose* and gaze direction are estimated using the OpenFace toolkit [17]. After that, as shown in Figure 2, a primitive matrix called a *Looking_At* square matrix with size of $n \times n$ is calculated, by which the eye contact between any two participants is determined. Formally, with assuming that $Looking\_At(x, y) \in \{0, 1\}$ at time t is the binary value of the participant x whether looks towards the participant y, an eye contact exists between them if $Looking\_At(x, y) = Looking\_At(y, x) = 1$. For example, in Figure 2, eye contact holds between $P2$ and $P4$.

We can calculate the values in the *Looking_At* square matrix using two different approaches. The first approach is based on supervised machine learning methods by training a classifier that takes the head pose of two participants and return whether the first one is looking at the other one. Then for all possible combinations among the participants, the trained classifier is used also to fill the data of *Looking_At* square matrix. The second approach is a geometrical one that does not require any training dataset or intransitive devices. We illustrate the second approach through an example of two participants and two cameras as follows:

1) Assign reference frames as illustrated in Figure 3, where $F1$ is the reference frame of first camera $(C1)$, $F2$ is the reference frame of second camera $(C2)$, ${}^1F3$ is the first person $(P1)$ head $pose\ w.r.t.\ F1$, and ${}^2F4$ is the second person $(P2)$ head $pose\ w.r.t.\ F2$.
2) Compute the transformation between frames, where ${}^1T_2$ is equal to the $pose$ of $C2\ w.r.t.\ F1$, ${}^1T_3$ is equal to the $pose$ of $P1$ head $w.r.t.\ F1$, and ${}^2T_4$ is equal to the $pose$ of $P2$ head $w.r.t.\ F2$. The transformation ${}^iT_j$ is used to transform a vector ${}^jV$ from Fj to Fi as

$$ {}^iV = {}^iT_j \times {}^jV \tag{1}$$

3) Check whether Pk stares at Pl. In particular, we have to check if the Pk gaze vector intersects with a sphere centered at Pl head position. Hence, both the line and the head position must be in the same reference frame. Assuming that $F1$ is the reference frame, and Pk is seen by $C1$ ($Pk = P1$) and Pl seen by $C2$ ($Pl = P2$), we transform ${}^2Vl$ to $F1$ based on equation 1 as follows:

$$ {}^1Vl = {}^1T_2 \times {}^2T_4 \times {}^4Vl \tag{2}$$

Next, we model Pk head as a sphere:

$$ \|\mathbf{x} - \mathbf{c}\|^2 = r^2 \tag{3}$$

where $\mathbf{c}$ is the sphere center, r is the sphere radius, and $\mathbf{x}$ is a point on the sphere. Geometrically, any line can be defined as:

$$ \mathbf{x} = \mathbf{o} + d\mathbf{l} \tag{4}$$

where $\mathbf{o}$ is the origin of line, $\mathbf{l}$ is the direction of the line, d is the distance along the line from the line starting point, and $\mathbf{x}$ is a point on the line.
Finally, we check the intersection through searching for points that are on the line and on the sphere. Thus, we combine equations 3 and 4, solve them for d, and substitute: (i) Pk head position (${}^1F3$) as the sphere center; (ii) the head position of $Pl\ w.r.t\ F1$ (${}^1F4 = {}^1T_2 \times {}^2F4$) as starting point of the line, and ${}^1Vl$ as the line direction:

$$ d = \frac{-\left(\mathbf{{}^1Vl}.\left(\mathbf{{}^1F4} - \mathbf{{}^1F3}\right)\right) \pm \sqrt{w}}{\left\|\mathbf{{}^1Vl}\right\|^2} \tag{5}$$

$$ w = \left(\mathbf{{}^1Vl}.\left(\mathbf{{}^1F4} - \mathbf{{}^1F3}\right)\right)^2 - \left\|\mathbf{{}^1Vl}\right\|^2 \left(\left\|\mathbf{{}^1F4} - \mathbf{{}^1F3}\right\|^2 - r^2\right)$$

If the value of $w \in \mathcal{R}^+$, then there are two intersection points crossing the sphere and Pl is looking at Pk; otherwise the line is either tangent to the sphere or not passing through the sphere at all and Pl is not looking to Pk. We need to repeat the procedure $n(n-1)$ time to fill the Looking_At square matrix.

IV. Evaluation and Results

A. Experimental Setup

Dataset and Ground-truth. The adopted dataset in performing experiments has been recorded to study multi-person social interactions. It consists of 10 videos (average recording time is 20 minutes), and four participants in each video instructed to discuss a general conversational topic. The recording has been performed in a quiet office room equipped with four cameras as shown in Figure 4. Cameras have been slightly placed above the participants to provide a near frontal view of faces of all participants taking into account turning their heads during the conversation. To obtain the participants gaze behaviour, we have asked five annotators to label the dataset with $looking\_At$ ground-truth. The annotators have identified for each participants whose face is being looked or not looked at a particular moment.

Performance Metrics. We treat the eye contact detection as a binary classification problem. Thus, we adopt various metrics to evaluate a classification model. In our work, we leverage four widely used metrics: (i) Accuracy as the ratio of number of correct predictions to the total number of input samples; (ii)Precision as the number of correct positive results divided by the number of positive results predicted by the classifier; (iii) Recall as the number of correct positive results divided by the number of all relevant samples; and (iv) F-Measure as the Harmonic Mean between precision and recall, it shows how precise the classification model is, as well as how much it is robust.

Baseline. We define a baseline to compare our method with. The baseline reflects the results obtained when applying supervised machine learning algorithms on 18 features, divided as follows: (i) head pose of person P_i; (ii) head pose of person P_j; and (iii) world frame pose w.r.t. to camera frame reference. Many learning algorithms provided by Weka tool [18]. We exploit Naive Bayes, Random Forest, J48, and Artificial Neural Network (NN) as well-known supervised learning methods to evaluate the performance of mentioned state-of-the-art features.

Parameters Setting. Our proposed method doesn't have parameters to be configured or may affect the results. Furthermore, the selected supervised learning methods in Weka tool are controlled by important parameters that may have impact on the classification performance. Thus, for the Naive Bayes method, we set the "useKernelEstimator" and "useSupervisedDiscretization" options to false value as default values set by Weka. For Random Forest, we set the option max depth to 0 (unlimited), with studying the effect of changing number of $trees \in \{20, 30, 100\}$. For J48 method, we set the minimum number of instances per leaf to 2, number of folds to 3, and confidence factor to 0.2. For neural network learning algorithm, we study the impact of having different numbers of hidden layers (from 1 to 4) each layer has 18 neurons.

B. Experimental Results

We have performed two types of experiments: (i) 10-folds cross validation at video frame level; (ii) and 10-folds cross

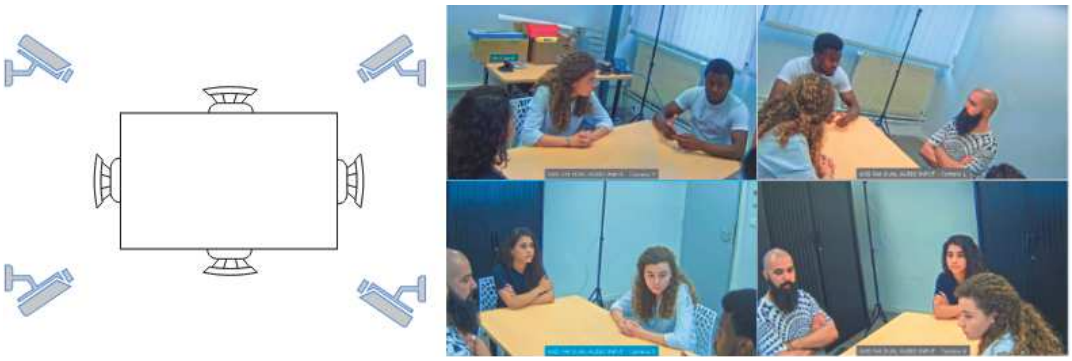

Fig. 4. Camera setup used for the data-set recording.

TABLE I
LOOKING_AT PERFORMANCE RESULTS OF OUR PROPOSED METHOD (GEOMETRICAL APPROACH) COMPARED WITH MULTIPLE SUPERVISED APPROACHES: RANDOM FOREST (RF), RANDOM TREE (RT), J48, NAÏVE BAYES, AND NEURAL NETWORK (NN), IN TERMS OF ACCURACY, PRECISION, RECALL, AND F-MEASURE FOR NOTLOOKING CLASS (0) AND LOOKING CLASS (1). RESULTS AVERAGED OVER 10 VIDEOS WHEN PERFORMING 10-FOLD VALIDATION ON EACH VIDEO.

	Accuracy	Precision 0	Precision 1	Recall 0	Recall 1	F-Measure 0	F-Measure 1
Proposed Method	78 %	85 %	59 %	85 %	59 %	85 %	59 %
RF (# Trees =10)	91 %	92 %	86 %	96 %	78 %	94 %	82 %
RF (# Trees =20)	92 %	92 %	87 %	96 %	80 %	94 %	84 %
RF (# Trees =30)	92 %	93 %	87 %	96 %	81 %	95 %	84 %
RF (# Trees =100)	92 %	94 %	87 %	95 %	83 %	95 %	85 %
RT	87 %	91 %	75 %	91 %	76 %	91 %	75 %
J48	89 %	93 %	79 %	93 %	79 %	93 %	79 %
Naïve Bayes	72 %	77 %	43 %	89 %	42 %	82 %	30 %
NN (#HL=1)	85 %	88 %	73 %	91 %	66 %	90 %	69 %
NN (#HL=2)	85 %	90 %	73 %	91 %	70 %	90 %	72 %
NN (#HL=3)	81 %	86 %	64 %	88 %	61 %	87 %	63 %
NN (#HL=4)	84 %	88 %	72 %	91 %	65 %	89 %	68 %

validation at video level. The main purpose of the first type is to study the impact of performing training a set of frames and testing on other set of frames where both sets are related to same video. At higher level, the second type of experiments give a strong indication about any possible dependency among same video frame level and different video levels. Table I reports the results of performing 10-fold cross validation at single video frame level, while Table II reports 10-folds cross validation at video level. The 10-fold cross validation is performed for each video with producing performance results in terms of the mentioned metrics. The ultimate performance result value for the first type is averaged over the entire video data-set. The results of first type of experiments show that the supervised learning based methods have generally high classification performance compared to our geometrical proposed method in terms of accuracy metric. The Random Forest learning method at different number of trees provides almost high classification performance in terms of accuracy, and other class-based metrics. These results are expected since the nature of Random Forest is in building many random trees acting as uncorrelated experts and then a voting is performed among the trees to provide the ultimate predication value. The high variation in the precision values of NotLooking class, compared to Looking one, shows that the data-set adopted in training is unbalanced at the class level, making the classification model biased towards a particular class which is NotLooking class in our case.

The results of the first type of experiments show that the supervised learning based methods are the winner in providing accurate and precise classification LookingAt model. However, the introduced results in Table II of the second type of experiments provide different conclusions that: (i) training a classification LookingAt model on a video is not necessary to performs very well on other video (social experiment), raising concerns about the degree of sensitivity when participants change their sitting/arrangement around the table; (ii) from the machine learning perspective, the decreasing in the supervised based learning methods results shows an over-fitting problem occurred, meaning that the classification models of first type experiments are not generalized enough to cover all patterns of looking among participants. Our geometrical method has same performance results since no prior training/configuration is required when processing videos. According to the results of second type experiments, our method outperforms most of supervised classification models in terms of accuracy and other class-based metrics. Indeed, the key-features of our proposed method are in: (i) no prior training dataset required and thus avoiding the annotation step as a time consuming one; (ii) it has classification performance almost at the same level with supervised based ones; (iii) and it does not required any

TABLE II
10-FOLD CROSS VALIDATION (VIDEO LEVEL) LOOKING_AT PERFORMANCE RESULTS OF OUR PROPOSED METHOD (GEOMETRICAL APPROACH) COMPARED WITH MULTIPLE SUPERVISED APPROACHES: RANDOM FOREST (RF), RANDOM TREE (RT), J48, NAÏVE BAYES, AND NEURAL NETWORK (NN), IN TERMS OF ACCURACY, PRECISION, RECALL, AND F-MEASURE FOR NOTLOOKING CLASS (0) AND LOOKING CLASS (1).

	Accuracy	Precision 0	Precision 1	Recall 0	Recall 1	F-Measure 0	F-Measure 1
Proposed Method	78 %	85 %	59 %	85 %	59 %	85 %	59 %
RF (# Trees =10)	76 %	78 %	60 %	95 %	22 %	85 %	32 %
RF (# Trees =20)	77 %	78 %	65 %	94 %	21 %	86 %	32 %
RF (# Trees =30)	77 %	78 %	66 %	96 %	23 %	86 %	33 %
RF (# Trees =100)	77 %	78 %	68 %	96 %	22 %	86 %	33 %
RT	69 %	79 %	41 %	79 %	42 %	79 %	41 %
J48	71 %	81 %	45 %	76 %	45 %	81 %	45 %
Naïve Bayes	65 %	74 %	25 %	83 %	15 %	78 %	17 %
NN (#HL=1)	76 %	80 %	56 %	90 %	36 %	85 %	43 %
NN (#HL=2)	78 %	83 %	59 %	88 %	47 %	85 %	52 %
NN (#HL=3)	77 %	83 %	58 %	87 %	51 %	85 %	53 %
NN (#HL=4)	72 %	80 %	46 %	82 %	44 %	81 %	44 %

intrusive devices.

V. CONCLUSION AND FUTURE WORK

In this paper, we have described a novel geometric-based method to detect eye contact in natural multi-person interactions without the need of eye tracking devices or any intrusive, which allows to record natural social behavior. We have evaluated our method on a recent dataset (10 social videos, where each video is 20 minutes long) of natural group interactions, which we annotated with Looking_At ground truth, and showed that it is highly efficient with regards to classification performance, and comparing to the classical supervised eye contact detection methods. Eye contact detection could be used to analyze turn-taking, social roles, and engagement during multi-person interactions. Eye contact detection is a part of the social cues extraction module which is a part of larger framework as shown in figure 1. As a future direction, we are going to consider other non-verbal cues such as facial expression and fuse them with the eye contact method to reduces the ratio of total failure.

REFERENCES

[1] Z. Akhtar and T. H. Falk, "Visual nonverbal behavior analysis: The path forward," *IEEE MultiMedia*, vol. 25, no. 2, pp. 47–60, Apr 2018.

[2] A. Vinciarelli, M. Pantic, D. Heylen, C. Pelachaud, I. Poggi, F. D'Errico, and M. Schroeder, "Bridging the gap between social animal and unsocial machine: A survey of social signal processing," *IEEE Transactions on Affective Computing*, vol. 3, pp. 69–87, 2012.

[3] M. Qodseya, "Visual non-verbal social cues data modeling," in *Advances in Conceptual Modeling - ER 2018 Workshops EmpER, MoBiD, MREBA, QMMQ, SCME, Xi'an, China, October 22-25, 2018, Proceedings*, 2018, pp. 82–87. [Online]. Available: https://doi.org/10.1007/978-3-030-01391-2_16

[4] Y. Fukuhara and Y. Nakano, "Gaze and conversation dominance in multiparty interaction," in *2nd Workshop on Eye Gaze in Intelligent Human Machine Interaction*, vol. 9, 2011, pp. 9–16.

[5] M. Argyle and J. Dean, "Eye-contact, distance and affiliation," *Sociometry*, pp. 289–304, 1965.

[6] W. Dong, B. Lepri, F. Pianesi, and A. Pentland, "Modeling functional roles dynamics in small group interactions," *IEEE Transactions on Multimedia*, vol. 15, no. 1, pp. 83–95, Jan 2013.

[7] M. Poel, R. Poppe, and A. Nijholt, "Meeting behavior detection in smart environments: Nonverbal cues that help to obtain natural interaction," in *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on*. IEEE, 2008, pp. 1–6.

[8] Y. Kizumi, K. Kakusho, T. Okadome, T. Funatomi, and M. Iiyama, "Detection of social interaction from observation of daily living environments," in *Future Generation Communication Technology (FGCT), 2012 International Conference on*. IEEE, 2012, pp. 162–167.

[9] M. Abouelenien, V. Prez-Rosas, R. Mihalcea, and M. Burzo, "Detecting deceptive behavior via integration of discriminative features from multiple modalities," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 5, pp. 1042–1055, May 2017.

[10] D. B. Jayagopi, H. Hung, C. Yeo, and D. Gatica-Perez, "Modeling dominance in group conversations using nonverbal activity cues," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 3, pp. 501–513, 2009.

[11] M. Avril, C. Leclre, S. Viaux, S. Michelet, C. Achard, S. Missonnier, M. Keren, D. Cohen, and M. Chetouani, "Social signal processing for studying parentinfant interaction," *Frontiers in Psychology*, vol. 5, p. 1437, 2014. [Online]. Available: https://www.frontiersin.org/article/10.3389/fpsyg.2014.01437

[12] E. Chong, K. Chanda, Z. Ye, A. Southerland, N. Ruiz, R. M. Jones, A. Rozga, and J. M. Rehg, "Detecting gaze towards eyes in natural social interactions and its use in child assessment," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 1, no. 3, pp. 43:1–43:20, Sep. 2017. [Online]. Available: http://doi.acm.org/10.1145/3131902

[13] M. Aghaei, M. Dimiccoli, and P. Radeva, "With whom do i interact? detecting social interactions in egocentric photo-streams," in *Pattern Recognition (ICPR), 2016 23rd International Conference on*. IEEE, 2016, pp. 2959–2964.

[14] J. D. Smith, R. Vertegaal, and C. Sohn, "Viewpointer: Lightweight calibration-free eye tracking for ubiquitous handsfree deixis," in *Proceedings of the 18th Annual ACM Symposium on User Interface Software and Technology*, ser. UIST '05. New York, NY, USA: ACM, 2005, pp. 53–61. [Online]. Available: http://doi.acm.org/10.1145/1095034.1095043

[15] B. A. Smith, Q. Yin, S. K. Feiner, and S. K. Nayar, "Gaze locking: Passive eye contact detection for human-object interaction," in *Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology*, ser. UIST '13. New York, NY, USA: ACM, 2013, pp. 271–280. [Online]. Available: http://doi.acm.org/10.1145/2501988.2501994

[16] X. Zhang, Y. Sugano, and A. Bulling, "Everyday eye contact detection using unsupervised gaze target discovery," in *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology, UIST 2017, Quebec City, QC, Canada, October 22 - 25, 2017*, 2017, pp. 193–203. [Online]. Available: https://doi.org/10.1145/3126594.3126614

[17] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "Openface: an open source facial behavior analysis toolkit," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2016, pp. 1–10.

[18] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: An update," *SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 10–18, Nov. 2009. [Online]. Available: http://doi.acm.org/10.1145/1656274.1656278