Your Body Reveals Your Impressions about Others: A Study on Multimodal Impression Detection

Wang, Chen; Pun, Thierry; Chanel, Guillaume

How to cite

WANG, Chen, PUN, Thierry, CHANEL, Guillaume. Your Body Reveals Your Impressions about Others: A Study on Multimodal Impression Detection. In: 2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW). Cambridge, UK. Cambridge : IEEE, 2019. doi: 10.1109/ACIIW.2019.8925052

This publication URL: https://archive-ouverte.unige.ch//unige:123960

Publication DOI: 10.1109/ACIIW.2019.8925052

Your Body Reveals Your Impressions about Others: A Study on Multimodal Impression Detection

Chen Wang
Department of Computer Science
University of Geneva
Geneva, Switzerland
chen.wang@unige.ch

Thierry Pun
Department of Computer Science
University of Geneva
Geneva, Switzerland
thierry.pun@unige.ch

Guillaume Chanel
Department of Computer Science
University of Geneva
Geneva, Switzerland
guillaume.chanel@unige.ch

***Abstract*—Formed impressions are crucial for human-human interaction (e.g. a job interview) and an interaction with a virtual agent/robot, since they can impact people's perceptions and willingness to be involved in the interaction. There are studies on how facial features (e.g. skin color, face shape), acoustic signals and non-verbal behaviors (e.g. gestures, postures) create/leave certain impressions. However there is little research focusing on how our bodies disclose our already formed impression of someone. Forming an impression leads to emotions and behaviors which can be measured. In this paper, we investigate recognition of evoked impression of warmth and competence from the non-verbal behaviors expressed by the person forming the impression. We conducted an experiment in which participants were watching impression stimuli. We measured participant's facial expressions, eye movements and physiological reactions (electrocardiography and galvanic skin response). To recognize impressions, we tested 2 multivariate regression models with the aforementioned multimodal recordings. Our best results demonstrate the possibility to detect impressions along warmth and competence dimensions with a concordance correlation coefficient of 0.838 and 0.864. Facial expressions and eye movements are more reliable for impression detection compared with physiological signals. Finally, the higher the Berkeley emotion expressivity scores the participants have, the more accurately the impressions are detected.**

***Index Terms*—impression detection, machine learning, multimodal, LSTM**

I. Introduction

When we meet a stranger, we form our impressions by judging his/her appearance and behaviors [15], [16], [36], [40]. In this paper we call *impression prediction* (yellow arrow shown in Fig.1) the process of using the expressive signals (e.g. facial expressions, audio signals, gestures) of someone to predict what impression others will form of him/her. When we have already formed an impression of a stranger (e.g. love at first sight), our body can reflect this impression through our behaviors and physiological signals such as facial expressions and heart rate [2], [10], [11]. We call *impression detection* (blue arrow shown in Fig.1) the recognition of formed impression of others using the signals of the person forming the impression.

Formed impression (e.g. favor someone or dislike someone) is an internal state which will reflect facially and behaviorally [10], [11]. These behaviours play an important role in impression formation as it can reveal information about others' characteristics such as their sexual orientation [1], personality and interpersonal attitudes [6]. There are some works on impression prediction in term of personality traits but little study on impression detection. For building more user engaging robots/virtual agents [42], it is important to understand what user's formed impression of the agent is (impression detection) and then combined with impression prediction (what agent behavior will leave a better impression on the user) , the agent can change its behaviors accordingly.

Impression theories in human-human interaction and human-virtual agent/robot interaction have been widely studied [3]–[6], [9], [10]. Among these theories, impression represented in warmth and competence dimensions are widely accepted and are considered as the fundamental dimensions of social cognition to form an impression. Warmth represents one's intention towards others (i.e. friend or foe) whilst competence demonstrates the capacity of the one to execute his/her intention [10].

Recent studies in affective computing suggest that individuals' facial expressions and behaviors contain important information regarding their affective state and intentions [2], [10]. According to [2], when an affective state has been activated, it accounts for response tendencies including subjective feelings, physiological changes and behavioral tendencies. The response tendencies may not all be expressed due to personal or social reasons (emotion regulation). To measure the level of expressed behavioral changes associated with an emotional experience, there is a widely used questionnaire named Berkeley Expressivity Questionnaire (BEQ). It has been found that emotion expressivity is related with the Big Five personality traits [35] and that personality traits influence others' perceptions including the formation of impressions [15], [37].

Impression as an important component for social cognition and communication has not been well explored with computational models. The few existing studies in this area are mainly stereotype based prediction (e.g. women are found generally more friendly/warmer than men [10]) or take personality values, predicted from appearance based features [15] as impressions. Instead of studying the aforementioned impression

This work is supported by the Swiss National Science Foundation under Grant No.: 2000221E-164326.

978-1-7281-3891-6/19/$31.00 ©2019 IEEE

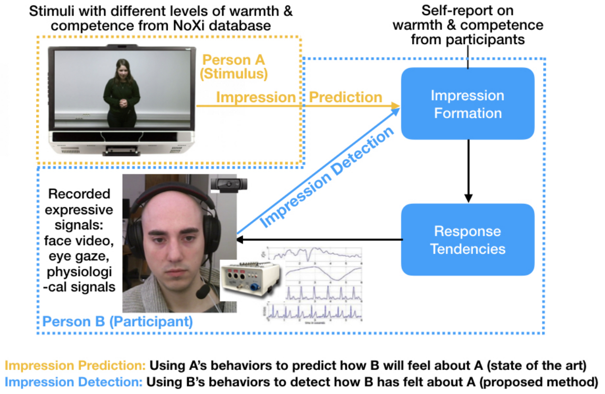
Stimuli with different levels of warmth & competence from NoXi database
Self-report on warmth & competence from participants
Person A (Stimulus)
Impression
Prediction
Impression Formation
Impression Detection
Recorded expressive signals: face video, eye gaze, physiologi -cal signals
Response Tendencies
Person B (Participant)
Impression Prediction: Using A's behaviors to predict how B will feel about A (state of the art)
Impression Detection: Using B's behaviors to detect how B has felt about A (proposed method)

Fig. 1. Impression Prediction and Detection Diagram

prediction (yellow arrow in Fig. 1), we are more interested in how to detect formed impressions, in the warmth and competence space, from the expressive behaviors of the person forming the expression (blue arrow in Fig. 1). To the best of our knowledge, there is no research on this specific topic so far. Some studies found that there is a relationship between emotion expressivity and personality traits [14], [35], [36]. However, there is no research on how expressivity influence the performance of automatic impression detection methods. In this paper, we aim to answer two research questions: (i) is it possible to reliably detect one's formed impression of others based on his/her face recording, eye gaze data and physiological signals? and (ii) is there any relation between emotion expressivity and impression detection performance?

In order to answer these research questions, we recruited participants who answered the Berkeley Expressivity Questionnaire (BEQ) and built a database of them watching impression stimuli while recording their body response and asking them to report their impression continuously. We applied machine learning models (LSTM and XGBoost) to detect participants impressions from the recorded multimodal signals (face video, eye movement and physiological signals). Finally, we compared the detection performance of the models with the BEQ reports.

II. Related Work

A. Impression Theory of Warmth and Competence

As the central dimensions of interpersonal perception, warmth and competence have been studied since 1946. Asch found that people treat warmth as relatively central in forming impressions [5]. It was then proposed that warmth and competence account for how people interpret behaviors or their impressions of others [6], [7], [11]. There are other dimensional impression theories such as self-profitable traits and other-profitable traits [8]. Although the dimensions are named differently, according to Abele and Wojciszke [9], they shared the common cores which are warmth and competence. That is the traits of warmth represent communion, collectivism and morality, while competence stand for intelligence, agency, individualism, and self-interest. Some researchers believe that warmth and competence are negatively correlated due to compensation effects (i.e when someone is perceived very competent, he/she will be more likely to be less warm) [11], [13]. Affect is a mental and bodily process that can be inferred by a human observer from a combination of contextual, behavioral and physical cues [2]. Impression which is closely related with affect [17], in principle could be observed from those cues as well. Also according to [10]–[12], judgments of warmth and competence elicit unique behavioral and emotional outcomes, which confirms that physical cues (e.g. gestures and facial expressions) could be used for inferring formed impression.

B. Impression Prediction

Impressions have been studied for prediction only (yellow arrow shown in Fig 1) to the best of our knowledge. McCurrie Mel et al. and Farnadi et al. [14], [15], [40] showed how gender, facial features (e.g. baby face), contextual information (e.g. profession) influence others' perception of trustworthiness, dominance and personality traits. There are other studies using verbal and non-verbal cues to predict service quality and hireability [38], [43], [44]. For example, Escalante et al [38] proposed a deep residual network, trained on a large dataset of short YouTube video blogs, for predicting personality impressions and whether persons are suitable for a job interview. It was found that the difference between the lowest and highest level of personality traits and interview recommendation was related with appearance and stereotype such as colors in the face, femininity-masculinity of the face and face shape [38]. In [15], a linear regression model was used to predict the interview annotation as a function of personality trait annotations in the five dimensions of the Big-Five personality model. The performance of prediction models are evaluated with different standards. For example, in [14], R squared was used and results on trustworthiness and dominance are 0.57 and 0.46 respectively. In [38], a relative mean absolute error below 0.09 was obtained on all five traits of the Big-Five personality model.

C. Multimodal Affect Detection

As far as we know, there is no existing research on multimodal impression detection (blue arrow shown in 1). However, as shown in Section II-A, forming an impression is associated with emotions. Studies on emotion detection are thus relevant for impression detection. Emotion annotations from human annotators usually contains noises. There are studies on getting a set of reliable affect annotations from multiple annotators by applying smooth windows and dimensionality reduction [28], [29]. With reliable annotations, previous works in automatic emotion detection have explored a variety of models which can be generally classified as non-temporal and temporal models based on whether temporal information is used. The non-temporal models usually require contextual features (e.g. semantic associations) while temporal models emphasize the dynamic information in the model directly. Other than the models, the modalities used for training the models also differ enormously. For example, Brady et al. [18] derived high-level features from acoustic, visual and physiological modalities using sparse coding and deep learning method. Povolny et al. [19] focused on the audio modality only extracting bottleneck acoustic features. With multimodal features, fusion can be applied at early (input feature), late (prediction output) or middle (intermediate presentation) level [20]. In many cases, it has been shown that the multimodal methods outperform unimodal methods [22]–[24]. Besides, single task and multitask learning are both explored for emotion recognition. Chen et al. [28] tested different regression models on predicting emotion in arousal and valence dimensions using single task and multitask framework. Their work showed improvements of multitask learning for both dimensions. There are also studies indicating that multitask learning improves visual feature based emotion recognition but does not improve for audio based systems [19].

III. Data Acquisition

To answer our research questions, we designed an experiment to collect multimodal signals and continuous self-reported impressions in the warmth/competence space. Fig 1 shows the data collected in our experiment and how it relates to the formation of impressions. Our experimental design and data recording was approved by the ethic committee from University of Geneva. Before the experiment, a consent form was provided and signed by the participant. Berkeley Expressivity Questionnaire (BEQ) and demographic questionnaire were filled as well. During the experiment, participants watched stimuli (13 short video clips) from the Noxi database [25] with physiological sensors (electrocardiography (ECG) and galvanic skin response (GSR) sensors) attached on their skin. While watching the stimuli, the participants reported their formed impression in warmth and competence continuously by pressing the keyboard: up and down arrows for warmth; left and right arrows for competence. At the same time a upper body video (Logitech webcam C525 & C920, sample rate 30 fps), the eye movements (Tobii TX300 & T120, recording at 300Hz and 120Hz respectively) and physiological signals (ECG and GSR using a Biosemi amplifier, sample rate 512 Hz) of participants were recorded.

A. Berkeley Expressivity Questionnaire

The Berkeley Expressivity Questionnaire (BEQ) was used to measure the level of emotion expressivity before watching stimuli. BEQ is a self-reported measure of emotional expressivity, which is widely used for affective related experiments. Emotion expressivity refers to the strength of behavioral (e.g. facial, vocal, postural) changes associated with emotional experiences [35]. There are 3 distinct facets measuring emotion expressivity which are: impulse strength, negative expressivity and positive expressivity. The 3 facets have their corresponding questions in BEQ and the question sequence were mixed. The impulse strength facet represents individual differences in the intensity of emotional response tendencies. Questions representing this facet are 6 in total including ”People often do not know what I am feeling”. The negative expressivity facet captures the expression of negative feelings (e.g. anger, fear and nervous) as well as socially inappropriate leakage of negative emotions. This factor is defined with 6 questions such as ”People often do not know what I am feeling.” The positive expressivity concerns expressions of positive emotions, for example, amusement and happiness. This facet contains 4 questions (e.g. When I am happy, my feelings show.). A 7 point Likert scale was applied for the expressivity measurement. The BEQ score calculation is presented in [35]. The 3 facets scores are computed from the corresponding questions and an overall score is the average of the 3 facets. The statistic of the BEQ scores reported by participants are shown in I. Overall BEQ scores ranged from 2.2 to 6.52 with a standard deviation of 1.15 showing that we managed to collect data from participants with high and low expressivty.

TABLE I
Participant BEQ Score

BEQ Score	Mean	STD
Negative Expressivity	3.72	1.14
Positive Expressivity	5.22	1.35
Impulse Strength	4.67	1.42
Overall	4.54	1.15

B. Impression Stimuli

13 stimuli were selected from the Noxi database [25]. Noxi is a database of natural dyadic novice-expert interactions. It recorded screen-mediated face-to-face interactions discussing a wide range of topics. Audio and upper body videos of novice and experts were recorded with a Kinect. The videos were annotated by more than 30 annotators with annotations of voice activities, gestures, head movement, and warmth/competence [25]. Compared with the data we collected for this study, Noxi does not include physiological signals, eye movement recordings and impression self-report.

The 13 stimuli for our experiment were selected from different expert videos in which an expert explains a topic of his/her interest. Each stimulus was cut in around 2 minutes (mean = 1.92, std = 0.22) with different levels of warmth (mean = 0.56, std = 0.18) and competence (mean = 0.52, std = 0.28) and as many gestures as possible based on the Noxi annotations [26].

C. Data Collection

In total we recorded multimodal data of 62 participants (23 female and 39 male). English proficiency levels were requested to be over B2 in the Common European Framework of Reference, to guarantee that they were able to understand and follow experiment instructions. Participants with epilepsy history were excluded.

Participants were given time to get familiar with physiological sensors, eye tracker as well as the annotation tool. Participants were also well explained to the meaning of warmth and competence. To help participants better annotate their impressions, a paper copy of warmth/competence traits and corresponding annotating keyboard keys was provided to them.

Participants were requested to report warmth and competence by pressing the keyboard whenever they felt their impression was changing while watching the stimuli (2 dimensions can be annotated at the same time by pressing the keys). This lead to unevenly sampled annotations from 7 annotations per minute to 52 annotations per minute. Our annotations are 3D arrays with sequential but not consecutive timestamps (frame numbers) and corresponding annotation values for warmth and competence. Once the participant was familiarized with the experiment, the researcher left the participant alone in the laboratory to watch and annotate the stimuli. In order to remove bias between stimuli, a break was taken from 5s up to 30s to assure that the participant's formed impression returned to neutral. Participants could start the next stimulus autonomously after 5s by pressing the space key when they felt ready. When the participant started a stimulus, a trigger was sent to all modality recordings for synchronization. All the stimuli took overall 25 minutes per participant. In total, we obtained $62 \times 25 = 1625$ minutes of multimodal recordings and warmth/competence annotations with mean = 0.36/0.31, std = 0.25/0.27. There experiments were conducted over a period of 4 months. Due to ethic reasons, the database is not publicly available at the current stage. But we may be able to publish the extracted features when we get approvals from participants and the ethics committee.

IV. Method

A. Signal Processing

The recorded modalities have various sampling frequency ranging from 512 Hz for physiological recordings and 30 Hz for the video. To be able to apply early fusion for our regression models, we firstly synchronized the multimodal recordings with the impression annotations using the recorded trigger. Then we resampled the original recordings or extracted features to get the same length of data. In this paper, we resampled each modality as well as annotations to 30 Hz for simplification.

1) Annotation Processing: To homogenize sampling frequencies we used the face video frame rate as a standard and applied 1D polynomial interpolation on warmth and competence annotations respectively to achieve the same sampling rate. Firstly we segmented annotations based on the time interval of consecutive annotations. With the segments of annotations that time intervals were smaller than 5 seconds, we were able to calculate the interpolation polynomial coefficients a_k. Then we interpolated this series with $p(x) = a_n x^n + a_{n-1}x^{n-1} + ... + a_2x^2 + a_1x + a_0$, where p(x) is the new annotation, and x is the interpolated timestamp (frame). If the time interval is larger than 5 seconds, we assumed that there was no change in warmth or competence during this time interval and interpolated as 0 (neutral). We chose 5 seconds as a boundary to make sure there is no uncovered delay of impression impulse. It takes around 100 ms to form an impression [36] and generally less than 300 ms to react to press a key [41]. Taking into account that the impression may lead to emotions, we added up the time of emotion formation ranging from 0.5 to 4 seconds [39]. Thus in total 5 seconds should be able to cover the possible annotation responses.

After the interpolation, we followed [29] and applied a 10 seconds sliding window with overlap (one frame shift per time) to smooth warmth and competence annotations. Then we applied standard scaler on the smoothed annotations. The standard score of an annotation x is calculated as: $z = (x - u)/s$, where u is the mean of the smoothed annotations and s is the standard deviation of the smoothed annotations. Standardization is a common requirement for many machine learning estimators. After applying the standard scaler, the scaled warmth and competence annotations were used as the ground truth for training and testing the impression detection models.

2) Feature Extraction: Features were extracted from three modalities of our database: facial video, eye gaze and physiological signals (ECG and GSR signals). As suggested in [18], different modality may require different time length for crafting features. According to [24], temporal models prefer short-time features since the model can capture the temporal context and short-time features contain more details than long-time features. Thus for facial features and eye movements we extracted frame-based features. For the physiological modality, we extracted both long-time features (e.g. mean heart rate over 1 minute) and frame-based features (e.g. heart rate variability). According to [2], facial expression can be deconstructed into specific action units (AU). For the facial modality, we extracted AUs from participants video on each frame using OpenFace [28], an open source tool. We had 17 AUs intensity (from 0 to 5) and 18 AUs presence (0 or 1) features. We avoid using the geometric facial features since they are linked directly with stereotypes judgment [11]. For the facial modality, we did not apply resampling. For eye movements, the 2D gaze location on the display, the 3D locations of the left and right eyes, and the gaze duration are recorded by eye tracker. All the 9 features from eye movements are down sampled (120 Hz or 300 Hz) to the video frame rate (30 Hz). To process physiological signals, we used the TEAP toolbox [30]. We filtered out the noise with a median filter and then extracted Skin Conductance Response (SCR) from the GSR signal, heart rate (HR), heart rate variability (HRV) from the ECG signal as frame-time features. HR multi-scale entropy, mean heart rate and standard deviation over 1 minute were extracted as long-time features. We resampled the extracted features to 30 Hz instead of resampling the raw signals directly to conserve more information. All the extracted multimodal features were smoothed using the same sliding window as for annotations to get the same sample sizes. Afterwards, we standardized the feature matrix by removing the mean and scaling to unit variance.

B. Multitask Temporal Model

Long-short term memory (LSTM) neural networks [32] are temporal models for sequence prediction. They have been proved to perform reliably on affective detection tasks [23], [24]. Multitask learning framework has been shown working efficiently targeting correlated tasks [33] (i.e. multi-label data classification/regression). As mentioned in Section II-A, some researchers believe that warmth and competence are correlated with each other [11], [13]. Therefore, we adopted a multitask LSTM model to detect warmth and competence simultaneously on our multimodal data.

Besides an input layer and an output layer, a hidden LSTM layer was followed by a hidden state layer in our sequence model. We use the truncated back propagation through time with max step of 100 timesteps to train our LSTM networks. As mentioned before, the time delay of forming an impression of the stimuli, reacting to the impression and pressing the annotation key generally takes less than 3 seconds. That means the possible time gap in annotations and facial expressions lies within $3s \times 30fps = 90frames$. To simplify the data processing, we used the sequences of 100 timesteps. Adam optimizer is applied and the learning rate is initialized from 0.01 and reduced as half every 10 epochs. We trained at most 50 epochs and applied early stopping to avoid overfitting with patience equal to 5 epochs. The output layer was set to 2 dimensions for multitask learning and 1 for single task. Mean squared error(MSE) was used as the loss function.

For comparison purposes, we also performed impression detection with another regression model: XGBoost [31]. XGBoost is an ensemble learning method which is widely adopted for regression. In boosting, the trees are built sequentially such that each subsequent tree aims to reduce the errors of the previous tree. Each tree learns from its predecessors and updates the residual errors. Hence, the tree that locates next in the sequence will learn from an updated version of the residuals. We used the python XGBoost [31] library with the same early stopping setting as for LSTM.

Besides, we investigated modeling impression with different modalities including facial modality only, eye gaze only, physiological modality only and the fused combination of them. In this work, only early fusion method is considered with all the features re-sampled to the video frame rate (30 fps) and concatenated together. Other fusion methods and model architectures will be considered in future work.

V. Results and Discussion

A. Evaluation Protocol

For the evaluation, we used a leave-one-participant out cross-validation scheme. We divided the data set into three partitions: 1 participant was left out for testing, the remaining data was randomly divided into two parts: 80 percent for a training set and 20 percent for a validation set. We applied cross validation (rotate the left-out testing participant) to estimate the model performance of all the participants.

The Concordance Correlation Coefficient (CCC) is used as the performance metric for impression detection. To be able to compare with the impression prediction work [14], [38], R square and mean absolute error (MAE) are calculated as well. The warmth and competence values predicted by regression models are compared with processed annotation in Section IV-A1. To test the relation between expressivity and model prediction performance, the Spearman correlation is applied.

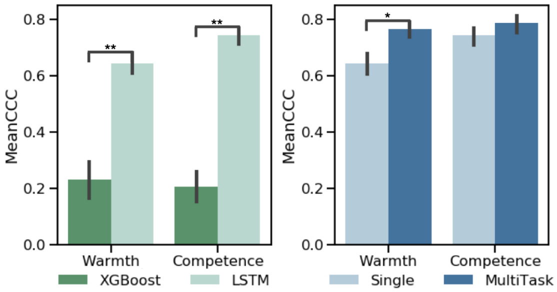

Fig. 2. Model Performance. (left) XGBoost vs. LSTM. (right) single task LSTM vs. multitask LSTM.

B. Results of Impression Detection

In order to find out how to detect formed impression based on one's multimodal recordings, we tested different models (XGBoost vs. LSTM), different learning frameworks (single task vs. multitask), and different modalities (features from a single modality vs. fused features from multiple modalities).

1) Comparison between models: In this session, we present the model prediction results from XGboost [31] and LSTM [32] to detect impressions in warmth and competence, in single task and multitask framework. We implemented the temporal model with Tensorflow [34] and non-temporal model with python XGBoost [31] library using facial modality only.

To assess which method performs better, we applied a pairwise t-test on the performances with the method as the independent variable. As shown in Fig 2, the LSTM network outperforms XGBoost significantly for both warmth ($t = 17.36, p < 0.001$) and competence ($t = 6.47, p < 0.001$). In addition multitask learning achieves higher performance than single task model ($t = 2.79, p < 0.006$).

Our results show that both regression models are able to detect impressions in warmth and competence dimensions with facial modality, while the temporal model (LSTM) outperforms the non-temporal model (XGBoost). This could be explained by the fact that reporting felt impressions is a cognitive effort which requires a reaction time. This time lag between body response and annotations can be learned by the LSTM network while it is a more difficult task for XGBoost. Also according to [23], XGBoost is more suitable for data with a small number of variables, whereas LSTM is better for data with a large number of variables like our case. Besides, our results show that multitask learning improve the performance both on warmth and competence, which indicates the correlation between these 2 dimensions. Further studies could investigate the compensation halo effect of warmth and competence to improve the impress detection accuracy.

2) Comparison between modalities: Since the LSTM outperforms XGBoost and multitask learning outperforms single task learning, we tested the modalities' performance using the multitask LSTM model. We trained the LSTM with each unimodal data separately. In addition, we fused different

TABLE II
MEAN CCC BETWEEN MODALITIES ON WARMTH AND COMPETENCE

Modality	Facial	EyeGaze	Physio	Face&Eye	All
Warmth	0.767	0.751	0.401	0.806	0.824
Competence	0.781	0.737	0.212	0.816	0.833

modalities at early stage (i.e. features were concatenated in a unique feature vector) to explore various feature combinations. The mean performance of the multimodal fusion for each impression dimension is presented in Table II. All type of feature fusions we tested improve the performance compared to unimodal features. As shown in Table II, the multimodal fusion is beneficial for both warmth and competence detection. The performance of the physiological modality is lower than the other 2 unimodal performances with $t = 6.66, p < 0.05$ for facial modality and $t = 5.71, p < 0.05$ for eye gaze. However, fusing the physiological modality with the other two improved the model performance.

Different modalities contain complementary information about the formed impressions. For a single modality, there is no significant difference ($p > 0.5$) on warmth and competence detection from face and eye gaze. For physiological signals, the warmth detection outperforms competence detection ($t = 4.55, p < 0.06$). That may indicate that warmth have more influence on physiological reactions or that the timestep we set for LSTM was not big enough to capture the physiological changes caused by competence. Features from all modalities are manually selected. This may be the reason for the relatively poor performance of physiological signals as well. Since there is no existing impression detection work, we calculated the same evaluation metrics from impression prediction work to have a general view. The mean MAE of our multimodal multitask LSTM is 0.27 while in [38] is 0.09. In [38], it was found that some personality traits such as neuroticism are better predicted by verbal cues. In our experimental setting, audio data was recorded but basically there is no voice nor verbal content from the participants watching and annotating stimuli. In human-human and human-virtual agent/robot interactions, verbal data could be an important modality to detect formed impressions. For the R square, our model achieves mean R square value of 0.71 while in [14], the best performance is 0.57. In [14], impressions were predicted from images instead of videos. Thus, it lacked temporal information which could influence the impression prediction performance.

C. Model Performance and BEQ Score

To investigate how emotion expressivity influence the model performance, we calculated the Spearman correlation between the BEQ scores and the CCC of each participant. The result is presented in Table III. We can see that the impression detection accuracy on warmth ($modality_w$ in the table) and competence ($modality_c$) is positively correlated with both positive emotion expressivity (Positive in Table III) and negative emotion expressivity (Negative in Table III). We think this is due to the experimental setting consisting of leaving the participants alone watching the stimuli. Under our setting the participants response tendencies (the right corner box shown in Fig 1) may appear more similar with the recorded behaviors (the left corner box shown in Fig 1) since there is no social judgments on their responses. It also shows that the negative facet has the highest correlation with the warmth detection from physiological signals. All facets have higher correlation on warmth detection than on competence from physiological modality. This confirms with our modality performance that competence are more difficult to detect from this modality. For facial modality, both warmth and competence detection have higher correlation on impulse strength facet, and this facet influence more on warmth. Our findings to some extent confirm that impression is linked with emotion and behavior consequence [10], [11].

TABLE III
CORRELATION BETWEEN LSTM PERFORMANCE AND BEQ SCORES

Modality	$Face_w$	$Face_c$	Eye_w	Eye_c	Phy_w	Phy_c
Negative	0.12	0.17	0.10	0.09	0.21	0.08
Positive	0.38	0.13	0.20	0.12	0.14	0.24
Impulse	0.59*	0.45	0.16	0.20	0.15	0.12
Overall	0.26	0.39	0.11	0.17	0.11	0.09

* $p < 0.1$

VI. CONCLUSION AND FUTURE WORK

In this study, we explored the use of multitask LSTM to detect formed impressions. Different from the existing work, we studied how triggered impressions are expressed by participants, instead of focusing on how different behaviors are perceived and lead to a certain impression. We used a wide variety of features from facial expressions, eye movements and physiological signals. With a multitask LSTM, warmth and competence can be detected reliably with mean CCC equal to 0.82 and 0.83 respectively. We found that the higher the impulse expressivity of the subject the more accurately the impression can be detected. Overall our results demonstrate that it is possible to assess the impressions formed by people based on the analysis of their facial expressions, eye movements and physiological signals. Facial expressions and eye movements are the most reliable modalities for this purpose.

Modelling and detecting impressions are challenging tasks. Although our results show that impression can be detected through multimodal cues, open research questions remain in the modeling process. How to map the formed impressions with annotations, what signals contains more impression manifestations and how to process the impression annotations properly require more investigations. Feature extraction and selection are crucial in the process of detecting impressions. For this work, we used conventional handcrafted features from each modality. In the future work, automatic methods or even ad-hoc feature extraction methods will be an interesting direction to investigate to improve the performance.

References

[1] N. Ambady and J. J. Skowronski, First impressions. Guilford Press, 2008.

[2] P. Ekman and K. Dacher, ”Universal facial expressions of emotion.” Segerstrale U, P. Molnar P, eds. Nonverbal communication: Where nature meets culture (1997): 27-46.

[3] T. Alexander, and Jenny M. Porter. ”Misleading first impressions: Different for different facial images of the same person.” Psychological science 25.7 (2014): 1404-1417.

[4] A. P. Cludio, A. Gaspar, E. Lopes and M.B. Carmo ”Virtual characters with affective facial behavior.” 2014 International Conference on Computer Graphics Theory and Applications (GRAPP). IEEE, 2014.

[5] S. E. Asch, ”Forming impressions of personality.” The Journal of Abnormal and Social Psychology 41.3 (1946): 258.

[6] R.S. Rosenberg, C. Nelson, and P. Vivekananthan, ”A multidimensional approach to the structure of personality impressions.” Journal of personality and social psychology 9.4 (1968): 283.

[7] B. Wojciszke, R. Bazinska and M. Jaworski, ”On the dominance of moral categories in impression formation.” Personality and Social Psychology Bulletin 24.12 (1998): 1251-1263.

[8] G. Peeters, ”Evaluative meanings of adjectives in vitro and in context: Some theoretical implications and practical consequences of positive-negative asymmetry and behavioral-adaptive concepts of evaluation.” Psychologica Belgica 32.2 (1992): 211-231.

[9] A. E. Abele and B. Wojciszke. ”Agency and communion from the perspective of self versus others.” Journal of personality and social psychology 93.5 (2007): 751.

[10] A. J. Cuddy, P. Glick, and A. Beninger, ”The dynamics of warmth and competence judgments, and their outcomes in organizations.” Research in organizational behavior 31 (2011): 73-98.

[11] C. M. Judd, et al. ”Fundamental dimensions of social judgment: understanding the relations between judgments of competence and warmth.” Journal of personality and social psychology 89.6 (2005): 899.

[12] A. J. Cuddy, S. T. Fiske, and P. Glick, ”Warmth and competence as universal dimensions of social perception: The stereotype content model and the BIAS map.” Advances in experimental social psychology 40 (2008): 61-149.

[13] V.Y. Yzerbyt, K. Nicolas and M.J. Charles, ”Compensation versus halo: The unique relations between the fundamental dimensions of social judgment.” Personality and Social Psychology Bulletin 34.8 (2008): 1110-1123.

[14] M. McCurrie, et al. ”Predicting first impressions with deep learning.” 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017). IEEE, 2017.

[15] G. Farnadi, et al. ”A multivariate regression approach to personality impression recognition of vloggers.” Proceedings of the 2014 ACM Multi Media on Workshop on Computational Personality Recognition. ACM, 2014.

[16] A. Subramaniam, et al. ”Bi-modal first impressions recognition using temporally ordered deep audio and stochastic visual features.” European Conference on Computer Vision. Springer, Cham, 2016.

[17] J.P. Forgas, ”On mood and peculiar people: Affect and person typicality in impression formation.” Journal of Personality and Social Psychology 62.5 (1992): 863.

[18] B. Kevin, et al. Multi-modal audio, video and physiological sensor learning for continuous emotion prediction. In Proceedings of the 6th Inter- national Workshop on Audio/Visual Emotion Challenge, pages 97104. ACM, 2016.

[19] P. Filip, et al. Multimodal emotion recognition for avec 2016 challenge. In Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge, pages 7582. ACM, 2016.

[20] C. Yunqiang, Z. Xiang Sean, and H. Thomas, 2001. One-class svm for learning in image retrieval. In Image Processing, 2001. Proceedings. 2001 International Conference on. Vol. 1. IEEE, 3437

[21] H. Zhaocheng, et al. An investigation of annotation delay compensation and output-associative fusion for multimodal continuous emotion prediction. In Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge, pages 4148. ACM, 2015.

[22] G. Hatice and S. Bjorn , 2013. Categorical and dimensional affect analysis in continuous input: current trends and future directions. Image and Vision Computing, 31, 2, 12013

[23] N. Jiquan, et al. 2011. Multimodal deep learning. In Proceedings of the 28th international conference on machine learning (ICML-11), 689696.

[24] C. Shizhe and Q. Jin. Multi-modal conditional attention fusion for dimensional emotion prediction. In Proceedings of the 2016 ACM on Multimedia Conference, pages 571575. ACM, 2016.

[25] C. Angelo, et al. ”The NoXi database: multimodal recordings of mediated novice-expert interactions.” Proceedings of the 19th ACM International Conference on Multimodal Interaction. ACM, 2017.

[26] B. Beatrice, C. Angelo and P. Catherine, ”Analyzing first impressions of warmth and competence from observable nonverbal cues in expert-novice interactions.” Proceedings of the 19th ACM International Conference on Multimodal Interaction. ACM, 2017.

[27] T. Baltruaitis, R. Peter and M. Louis-Philippe, ”Openface: an open source facial behavior analysis toolkit.” 2016 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2016.

[28] W. Chen, L. Phil, P. Thierry and C. Guillaume, ”Towards a Better Gold Standard: Denoising and Modelling Continuous Emotion Annotations Based on Feature Agglomeration and Outlier Regularisation.” Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop. ACM, 2018.

[29] N. Thammasan, K. Fukui, and M. Numao. 2016. An investigation of annotation smoothing for eeg-based continuous music-emotion recognition. In Systems, Man, and Cybernetics (SMC), 2016 IEEE International Conference on. IEEE, 003323003328.

[30] S. Mohammad, et al. ”Toolbox for Emotional feAture extraction from Physiological signals (TEAP).”Frontiers in ICT 4 (2017):1.

[31] C. Tianqi, and G. Carlos. ”Xgboost: A scalable tree boosting system.” Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. ACM, 2016.

[32] S. Hochreiter and J. Schmidhuber. ”Long short-term memory.” Neural computation 9.8 (1997): 1735-1780.

[33] C. Shizhe, et al. ”Multimodal multi-task learning for dimensional and continuous emotion recognition.” Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge. ACM, 2017.

[34] A. Martinn, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:1603.04467, 2016.

[35] J. J. Gross and O. P. John, 1995. Facets of emotional expressivity: Three self-report factors and their correlates. Personality and individual differences, 19(4), pp.555-568.

[36] J. Willis and T. Alexander, ”First impressions: Making up your mind after a 100-ms exposure to a face.” Psychological science 17.7 (2006): 592-598.

[37] W. Eric. ”How many seconds to a first impression?.” APS Observer 19.7 (2006).

[38] H. J. Escalante, et al. (2018). Explaining first impressions: modeling, recognizing, and explaining apparent personality from videos. arXiv preprint arXiv:1802.00745.

[39] L. Robert, ”Emotion and the autonomic nervous system: A prospectus for research on autonomic specificity.” Social psychophysiology: Theory and clinical applications (1988).

[40] C. Y. Olivola and T. Alexander, ”Fooled by first impressions? Reexamining the diagnostic value of appearance-based inferences.” Journal of Experimental Social Psychology 46.2 (2010): 315-324.

[41] S. J.Baker, J. P. Maurissen, and G. J. Chrzan, (1986). Simple reaction time and movement time in normal human volunteers: A long-term reliability study. Perceptual and motor skills, 63(2), 767-774.

[42] B. Beatrice, et al. ”A Computational Model for Managing Impressions of an Embodied Conversational Agent in Real-Time”,in press

[43] S. Muralidhar, M. Schmid Mast and D. Gatica-Perez, (2017, November). How may I help you? behavior and impressions in hospitality service encounters. In Proceedings of the 19th ACM International Conference on Multimodal Interaction (pp. 312-320). ACM.

[44] L. S.Nguyen, D.Frauendorfer, M. S.Mast and D.Gatica-Perez, (2014). Hire me: Computational inference of hirability in employment interviews based on nonverbal behavior. IEEE transactions on multimedia, 16(4), 1018-1031.