# An Open Dataset for Impression Recognition from Multimodal Bodily Responses

Wang, Chen; Chanel, Guillaume

# An Open Dataset for Impression Recognition from Multimodal Bodily Responses

1st Chen Wang
*Computer Science Department)*
*University of Geneva*
Geneva, Switzerland
chen.wang@unige.ch

2nd Guillaume Chanel
*Computer Science Department*
*University of Geneva*
Geneva, Switzerland
guillaume.chanel@unige.ch

***Abstract*—We present a dataset (IMPRESSION) for multimodal recognition of impressions on individuals and dyads. Compared to other databases, we did not only elicit impression using video stimuli, but also recorded natural impression formation of strangers meeting for the first time through video call. The database allows machine learning studies on impression recognition, using multimodal signals of individuals in relation to their emotion expressivity, and with respect to the interlocutor's reactions. The experiment setup was arranged with 62 participants' synchronized recordings of face videos, audio signals, eye gaze data, and peripheral nervous system physiological signals (Electrocardiogram-ECG, Blood Volume Pulse-BVP and Galvanic Skin Response-GSR) using wearable sensors. Participants reported their formed impressions in the W & C dimensions in real-time. We present the database in detail as well as baseline methods and results for impression recognition in W & C.**

***Index Terms*—Open Database, Impressions, Warmth, Competence, LSTM**

## I. Introduction

In daily life we interact with various people ranging from complete strangers to intimate partners. When we meet strangers, the first moments are critical. According to J. Willis and T. Alexander [1], first impressions to unfamiliar faces could be formed with a limited exposure (as little as 100ms). The impressions that are formed during those first moments can have important consequences such as success at job interviews or a second date with a potential partner [2]. This is why people often attempt to control the impression they leave on others. This process is usually done by controlling one's own appearance (physical aspect, clothing style, etc.) and non-verbal behaviour [3]. An increasing interest has arisen in impression considering the people's bodily expressions and responses when interacting with stranger [4].

In this study, we distinguish the person emitting social signals (i.e. the emitter) from the receiver who will form an impression based on the interpretation of these signals. We then define *impression prediction* (yellow arrow shown in Fig.1) as the process of using the social signals (e.g. facial expressions, audio signals, gestures) of the emitter to predict the impression that the receiver will form of him/her. When we have already formed an impression of a stranger (e.g. love at first sight), this impression can be reflected through bodily responses such as physiological signals (e.g. heart rate) [5]–[7]. However, due to social context, we may regulate these response tendencies, for example, facial expressions. We define *impression detection* (blue arrow shown in Fig.1) as the recognition of receiver's formed impression using his/her own expressive signals. Advances on impression prediction have been boosted by the availability of annotated impression databases. These databases have used stimuli, such as images and short videos [4], [8]. They include information from different modalities (e.g. audio, facial expression). Available multimodal impression databases mainly studied impression expressions of participants in individual. However, in real life, impression formation is mostly associated with social contexts (e.g. meeting a new colleague and blind dating). In such contexts, the individual expressions from emitter is not the only factor for impression formation. It also depends on the implicit and explicit interactions that can occur between the emitter and receiver. Additionally, the studies on impression mainly target personality traits instead of studying dimensional impression in warmth and competence. Therefore, current databases have ignored the receiver's responses and lack of continuous impression annotations for the study of impression recognition.

Our main contribution in the field is an open dataset named IMPRESSION (http://doi.org/10.26037/yareta:7bm3myp5tveybcgmubfpdx6ske) for multimodal research of impression recognition on individuals and dyads. The databset targets impressions in the Warmth and Competence (W & C) dimensions instead of personality traits. It contains multimodal recordings including videos, audios, eye movements and physiological signals. The impressions reported in the IMPRESSION dataset were triggered by stimuli, as well as formed naturally by face-to-face video calls. Our second contribution is that we provide a baseline model for impression detection from receivers and impression prediction from emitters.

## II. Related Work

### A. Impression Formation and Recognition

Goffman et al. [9] defined impression formation as the process of information perception, organization and integration in order to form coherent impressions of others (e.g., in terms of personality and interpersonal attitudes). When people are

978-1-6654-0019-0/21/$31.00 

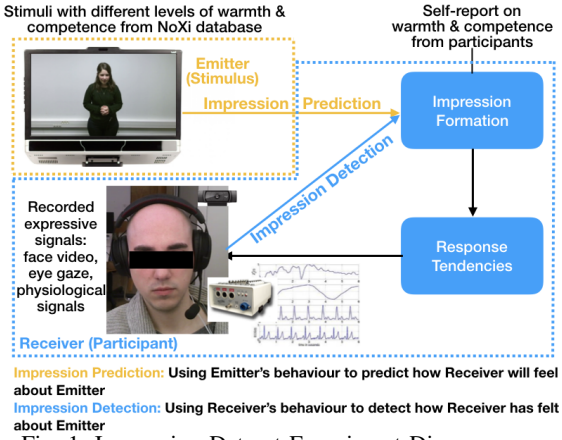

Fig. 1: Impression Dataset Experiment Diagram

aware of these mechanisms and attempt to control the impression that others form of us, it is called impression management [9]. It mainly concerns the control of appearance (e.g. haircut, dressing style). People also attempt to control their bodily responses, however it may be difficult to manage all the social cues that are exhibited during the interaction. Many researchers attempted to find what components influence social cognition when forming impressions of others. Two dimensions are often proposed: warmth and competence (W & C) [5], [10]. Warmth reflects the intentions of others and includes traits like kindness, trustworthiness, sociability. Competence reflects the ability of the other to enact his/her intentions and includes traits like intelligence, dominance and efficacy. In addition, W & C impressions elicit consequent emotional (admiration, contempt, envy, and pity) [5] and behavioral responses (active and passive, facilitative and harmful) [10].

Studies addressing impression recognition in the W & C dimensions are scarce, though several researchers have investigated behaviors for the prediction of personality traits (e.g. [11], [12]). Escalante et al. [13] proposed a deep residual network, trained on a large dataset of short YouTube video clips, for predicting personality traits and whether persons are suitable for a job interview. In their work, the hireability was predicted as a function of personality traits using a linear regression model. Farnadi et al. [12] identified and highlighted audiovisual information for a deep residual network in the five dimensions of the Big-Five personality model. They used face representation and audio/video occlusion for predictions. Kaya et al. [14] applied an end-to-end system with audio, facial and scene features with late fusion to predict apparent personality traits. The performance of prediction models was evaluated with different metrics. For example, $R^2$ was used and results on trustworthiness and dominance were 0.57 and 0.46 respectively [11] while in the study of Escalante et al. [13], a relative mean absolute error below 0.09 was obtained on all five traits of the Big-Five personality model.

### *B. Open Impression Corpora*

In the last decade, several multimodal corpora on human interaction have been recorded and published [4], [15], [16]. However, most do not contain multimodal recordings or impression related annotations. There are a few open corpora with impression annotations such as Noxi [15], Youtuber personality [4], AMIGOS [8] and Mission Survior II [17]. The first one has impression annotations in the W & C dimensions while the latter ones are annotated with personality scores.

The Noxi dataset is a multi-lingual database of natural dyadic novice-expert interactions, featuring screen-mediated dyadic human interactions in the context of knowledge sharing. Noxi integrated mediated face-to-face interactions, i.e. participants interacted through a screen in different rooms. The expert participant was presumed to be knowledgeable about one or more topics that were of interest to the novice. The experts usually are those who talked more during the dyadic interaction. In total, 87 people (26 female and 61 male) were recorded during 84 dyadic interactions. The original Noxi dataset does not contain impression annotations. Biancardi, Cafaro and Pelachaud [18] analyzed the videos of the experts (i.e. emitters) and provided continuous impression annotations in W & C using the NOVA tool [19]. The impression annotations were given for the first 5 minutes of the Noxi expert-invoice interaction. To avoid the perception bias from languages, the W & C were annotated from non-verbal behaviour only by excluding the speech content. That is, impression annotations rely only on the visual modality, without considering speech content and prosody features. In total, 14 videos (lasting 70 minutes) from the Noxi dataset were annotated. The Noxi database does not include physiological signals, eye movement recordings and impression self-report on formed impressions of the receiver. The impression annotations are from external annotators and the number of annotators is very limited.

One multimodal database for personality research is the Mission Survival II corpus [17]. It is a annotated dataset with video and audio recordings of 4 participants without physiological recordings. The ASCERTAIN [20] corpus includes recordings of the EEG, ECG, GSR and facial video of 58 users, while viewing short movie clips for implicit personality and affect recognition. This database only includes participants in individual configuration. The YouTube personality dataset [4] includes 404 YouTubers, who explicitly present themselves in front of a camera talking about a certain topic. Amazon Mechanical Turk (AMT) was used for collecting Big-Five personality scores using the Ten-Item Personality Inventory (TIPI). This dataset does not contain any recording from annotators, thus it can only be used for predicting impressions (in personality) and not for detecting impressions.

In Table I, we summarize the characteristics of the reviewed databases and compare them to ours. To the best of our knowledge there is no impression database studying participants in dyadic interaction settings in the W & C dimensions. Most aforementioned datasets contain personality traits for classification and the audio-visual recordings only from the emitter or the receiver. The Noxi dataset consists of continuous impression annotations in W & C and concrete gestures, rest positions, smiling and head movements annotations. The multimodal recordings are from both emitters and

receivers. However, neither eye movements nor physiological signals were recorded. Besides, there are only 2 annotators for impression annotation which is a limitation. Although we could not use the Noxi dataset directly, the design of the Noxi recording is used as a reference.

## III. Proposed IMPRESSION Dataset

### A. Experiment Setting

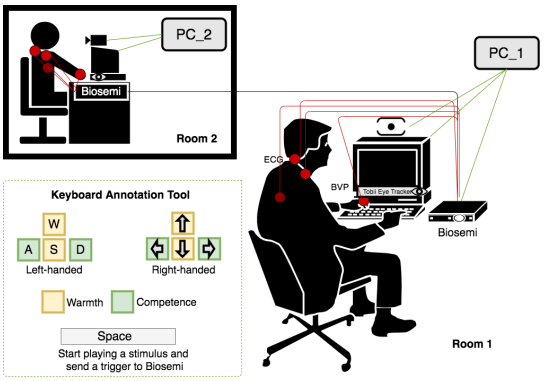

Fig. 2: Data Acquisition Architecture

In order to explore eye movements and physiological modality, as well as the relation between emotion expressivity and impressions, we designed an experiment for collecting impression data. The previous studies of impression/emotion data collection provided interesting results which suggest the adequateness of video stimuli to elicit emotions/impressions on participants [8], [21]. Based on the findings of annotation processing [12], [22], [23], a set of reliable annotations could be extracted from over 6 annotators. The AVEC challenge used data from 27 participants to test different machine learning structures and emotion changes [24]. We took these numbers as reference for designing the experiment and recruiting participants.

The IMPRESSION dataset was collected using 2 standard webcams, 2 Tobii eye trackers and a Biosemi (master-slave amplifiers) with a set of physiological sensors, annotated with real-time self-reported continuous W & C annotations. We recruited participants with English proficiency level over B2 in the Common European Framework of Reference, to guarantee that they were able to understand and follow experiment instructions. Participants with a history of epilepsy were excluded. The experiment took approximately 75 minutes and proceeded as follows:

- Before the experiment: Participants read and sign the consent form. Participants fill the Berkeley Expressivity Questionnaire (BEQ) [25] and demographic information before they come to the lab.
- Preparation: We give instructions about the experiment to participants, let them practise with the annotation tool and attach the electrode sensors for collecting physiological signals.
- Session 1: Participants watch Noxi stimuli and annotate the stimuli with respect to W & C (i.e. stimuli as emitter and participant as receiver).
- Session 2: Video call with another participant (stranger). The second participant is in another room (Room 2 in Fig. 2). Participants annotate their impressions of the interlocutor in W & C during the interaction and answer a questionnaire after the call. That is participants are both emitters and receivers.
- Clear up: Remove the sensors and help the participant to clean up and give the participant their compensation.

The whole experimental design was approved by the ethic committee of the University. Session 1 is under a controlled setting, as the W & C of the stimuli were evaluated and selected. Session 2 is a natural human-human interaction between 2 strangers. Participants were given time to get familiar with the physiological sensors, eye tracker as well as the annotation tool. Keyboard was used for annotation. The arrows keys for right-handed participants and ADWS keys for left-handed participants (shown in Fig.2). The pressed time stamp and the pressed key were recorded and a trigger was sent to Biosemi. Participants were well informed to the meaning of W & C. To help participants better annotate their impressions, a paper copy of W/C traits and corresponding annotating keyboard keys was provided to them. Once the participant was familiarized with the experiment, the researcher started the experiment and left the participant alone in the laboratory to watch and annotate the stimuli in Session 1, or talk with the other participant in Session 2.

In Session 1, participants watched stimuli (13 short video clips) from the Noxi database [15] (in SectionII-B) with physiological sensors (ECG, BVP and GSR sensors) attached on their skin. Only the Noxi experts' videos were used as stumuli for Session 1, as shown in the yellow box in Fig.1 (up left corner). While watching the stimuli, the participants reported their formed impression in W & C continuously by pressing the keyboard: up and down arrows (W and S keys) for warmth; left and right arrows (A and D keys) for competence. Taking warmth for example, participants pressed the up arrow key when they felt warmth was increasing. The more times they pressed the up arrow key, the stronger the increase was. If they felt warmth decreased, they could press the down arrow key accordingly. In order to remove bias between stimuli, a break was taken from 5s up to 30s. That is, participant could start the next stimulus autonomously after 5s by pressing the spacebar key when they felt ready. Once the session started, the upper body video (Logitech webcam C525 & C920, sample rate 30 fps), the eye movements (Tobii TX300 for Room 1 & T120 for Room 2, recording at 300Hz and 120Hz respectively) and physiological signals (using a Biosemi Active II amplifier, sample rate 512 Hz) of participants were recorded.

Session 2 recorded the same set of signals as Session 1. Participant 1 in Room 1 called Participant 2 in Room 2 through a Skype video call. Once the other participant's face appeared on the screen, the spacebar key was pressed and they could start their conversation. They could talk about whatever topics they preferred (i.e. open choices to the participants) and the conversation lasted at least 10 minutes. The participants were asked to annotate each other's W & C during the communi-

| Dataset | Participants | Recorded Modality | Annotation |
|---|---|---|---|
| Youtube Impression [4] | 442 YouTube vlogs | YouTube vlogs and a collection of personality for each vlogger | Big Five and hirablity using AMT |
| Noxi [15] | 84 individuals | Audio, video and depth recording | External Warmth/Competence/Gestures annotations |
| AMIGOS [8] | 40 individuals | Audio, Visual, Depth, EEG, GSR and ECG | Big-Five personality traits and PANAS. Valence, arousal, dominance, liking, familiarity and basic emotions. |
| ASCERTAIN [20] | 58 Individuals | EEG, ECG, GSR and Visual | Big-Five personality traits, self-assessment of valence and arousal |
| Mission Survival II [17] | 16 individuals (4-people group) | Audio-visual | Personality states by the Ten Item Personality Inventory |
| **IMPRESSION dataset** | 31 dyads | Audio-visual, eye movement, physiological signals (BVP,ECG and GSR) | Continuous self-reported W & C and BEQ scores |

TABLE I: Multimodal databases for impression and emotion recognition

cation by pressing the keyboard. After the conversation, the participants filled a questionnaire about each topics they had discussed, and rated their own competence on each topic with a 7-point Likert-scale. The participants did not know each other before the experiment. We prevented the participants from meeting by setting different experiment time (usually a 5-min gap) and locations.

Session 1 occurred before Session 2 to avoid a potential effect of priming. In Session 2, one participant interacted with another participant (unacquainted) through a video call. This session could trigger intense impressions and consequent emotions that we could not control. If Session 2 was to occur before Session 1, this could lead to bias in reported impressions of the stimuli. In contrast, Session 1 was a controlled condition and the 13 stimuli in this session remained the same for all participants limiting the priming effect.

One ECG sensor was put on one side (left or right) of the clavicle, and the other ECG electrode on the lower rib of the opposite side of the previous sensor. BVP sensor and GSR sensors were attached on the hand that was not used for annotation to reduce the noise caused by movement. To be more specific, we attached the BVP probe on the index finger, while GSR electrodes to the middle and ring finger on the proximal part of the finger and avoided the joint. All the electrodes were placed on the palm side of the hand. The overall sensor placement is roughly presented by the red dots in Fig.2.

### B. Impression Stimuli

The impression stimuli mentioned in this section are for experiment Session 1 and considered as emitters. Participants who watched the stimuli and reported their formed impressions are receivers. The 13 stimuli used to evoke participants' impressions are from the Noxi database [15] as mentioned in Section II-B. All the stimuli were from 'expert' videos, since this role is more related to W & C expressions, and experts were those who talked more during the dyadic interactions. The stimuli used were cut based on the warmth (range[0,1]), competence (range[0,1]) and gesture annotations (e.g.iconic [18]). We firstly applied peak detection on the Noxi W & C annotations and selected the video clips that contain at least one change (peak) in warmth or competence. Then among the W & C changing clips, we chose the ones containing most gesture annotations. Examples of such gestures are shown in Fig.3 [18]. Each stimulus lasts 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). The 13 stimuli were displayed in a random sequence during the experiment.

Fig. 3: Annotated gesture examples in the stimuli [18]

### C. Collected Multimodal Data

In total we recorded multimodal data of 31 dyads (23 female and 39 male). All the participants were above 18 years old and there was no upper age limit. Participants were from different cultural backgrounds such as French, Chinese and Arabian. Among all the participants, 60 participants answered all the questionnaires and allowed the use of the questionnaire data. After excluding participants who did not complete the experiment or did not give consent for some recording modalities, we ended up with 27 dyads (20 female and 34 male). In total, we obtained 1350 minutes of multimodal recordings with W & C annotations for Session 1. In Session 2, we obtained 540 minutes multimodal and mutual recordings. In total, the experiment was conducted over a period of 4 months.

### D. Collected Data from Questionnaires

*Berkeley Expressivity Questionnaire*. In order to study the relation between impression recognition performance and emotion expressivity, all the participants filled the Berkeley expressivity questionnaire (BEQ) before the experiment. 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 [25]. 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. The impulse strength facet represents individual differences in the intensity of emotional response tendencies. This facet includes questions as "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. Likewise the positive expressivity concerns expressions of positive emotions. A 7-point Likert scale is used for the expressivity measurement. The BEQ score calculation is presented in [25]. 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 60 participants are shown in Table II. 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.

| 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 |

TABLE II: Participant BEQ Score

*Questionnaire on Discussed Topics*. At the end of Session 2, participants filled a questionnaire to specify the topics they discussed and reported how competent (knowledgeable) they were on each topic. They were allowed to add any topics that they talked about. In total, we got responses from 60 participants on this questionnaire. In Session 2, during a 10 minute conversation, one third of the dyads reached a maximum of 5 topics. Without any collaboration, the participants listed the topics that they were discussed after their conversation. As a result, the list of topics was not the same between the participants of each dyad. Out of the 30 dyads, 18 dyads reported the same on the first topic, 14 on the second topic, 3 on the third topic, and only 1 on the fourth and fifth topic. For the first reported topic, only 3 participants (out of 60) reported low competence (lower than 4). As shown in Table.III, the first reported topic has the highest competence value and lowest standard deviation. Overall, more than half of the participants talked about 4 topics during the 10 minutes interaction. The majority of participants first reported the topics that they are confident with and then the less competent topics.

| Topic | 1st | 2nd | 3rd | 4th | 5th |
|---|---|---|---|---|---|
| Mean | 5.8 | 5.35 | 5.44 | 5.21 | 5.35 |
| STD | 1.15 | 1.39 | 1.36 | 1.60 | 1.57 |

TABLE III: Reported 7-Likert Scale Competence Rating On Topics

## IV. Data Processing

With the collected raw data, we first synchronized the multimodal recordings and impression annotations using the recorded trigger. That is, based on the trigger timestamp, we cropped and grouped the recordings and annotations into the length of each corresponding stimulus for Session 1 and conversation length for Session 2. In order to keep the participant's identity anonymous, we are unable to share the raw recording data with participant's face and voice. Thus we share the extracted features from face and audio modality. For the other modalities (eye movements and physiological signals) which do not reveal participant's personal information, we share the synchronized cropped data.

### A. Annotation Processing

The values for W & C are represented by a stepwise continuous ground truth (see Fig.4). When reporting W & C, the participants were allowed to press as many times as they wished on the keyboard. Consequently, there was no value range limitation. The annotation process involves interpolation, normalisation, denoising and extracting the inter-coder agreed annotation for impression prediction.

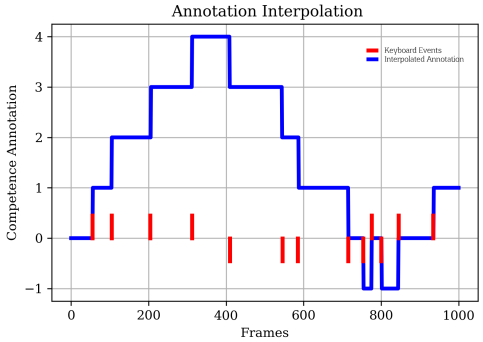


Fig. 4: Stepwise Annotation Interpolation

*Interpolation*. In order to deal with the uneven sampling, similar to other works reporting on data annotated in continuous dimensional spaces (e.g. [26]), we interpolated the impression annotations. As mentioned in Section III-A, participants pressed the keyboard when they felt the changes in W & C, shown as the red line in Fig.4. If no keyboard events were detected, we assume that the W & C remains the same level. A regularly sampled time series was created from the unevenly sampled annotations by computing the cumulative sum of the annotations as can be seen in Fig.4. It preserves the most information of the original data. An interpolation example is shown in Fig.4.

*Normalization*. The W & C annotation for participants are not in total agreement, mostly due to the variance in human participants' perception and interpretation of formed impressions. Thus, in order to make the annotations comparable, data normalisation is necessary. Similar procedures have been adopted by other works such as in [23], [26]. In this part, we opted for normalisation which removes the median and scales the data according to the quantile range [27]. This helps to avoid propagating noise in cases where there are a few number of very large marginal outliers.

*Denoising*. After the normalization, we followed the method proposed by Thammasan et al. [28] and applied a 10 seconds sliding window with overlap (one frame shift per time) to smooth W & C annotations. The smoothed W & C annotations, namely $Annotation_R$ were used as the ground truth for training and testing the impression detection models. For each participant (receiver), there is 1 set corresponding $Annotation_R$ to detect.

*Generalized Annotation*. In Session 1, 1 emitter (stimulus) leaved different impressions ($Annotation_R$) on each participant (in total 54). In order to have one-to-one relation as impression detection, we computed a generalized impression from 54 $Annotation_R$ for each emitter, namely $Annotation_E$. We followed the agglomeration method proposed in [23], which has been proven to work efficiently on continuous dimensional emotion recognition. Since the annotations have already been normalised and denoised, the outliers have been removed. Thus the extracted annotation minimizes the sum of squared differences within all annotators. These processed annotations (i.e.$Annotation_E$) were used as ground

truth for training and testing the impression prediction model.

### B. Feature Extraction

In this section, we present features extracted from the receiver and the emitter during the interaction. Due to our experiment setting, the modalities recorded in Session 1 for stimuli (i.e. emitters) and participants (i.e. receivers) were not the same. For the stimuli, only the audio and visual modalities were available. For participants, features were extracted from three modalities of our database: facial video, eye movement and physiological signals (BVP, ECG and GSR signals). The feature presentations for each modality of participants and stimuli are listed in Table.IV. The extracted features are

| Modality/Agent | Receiver | Emitter |
|---|---|---|
| Visual | AU presence & intensity | AU presence & intensity |
| Eye Movements | 3D locations of each eye, 2D eye gaze location on the screen, gaze duration | 3D eye gaze direction, 2D eye gaze direction |
| Physiological | the spectral entropy and mean frequency of SCR, HR, HRV, HR multi-scale entropy, mean HR & STD | N/A |
| Audio | N/A | PCM RMS energy, MFCC, PCM ZCR, voiceProb |

TABLE IV: Extracted Features from each modality for the receiver and the emitter respectively

widely used in dimensional emotion recognition and impression recognition [29], [30]. Kevin Brady et al. [29] suggested that different modality may require different time length for crafting features. According to Chen et al and Ringeval et al. [24], [30], temporal models prefer frame-based features since the model can capture the temporal context and short-time features contain more details than long-time features. 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 [7], facial expression can be deconstructed into specific action units (AU). For the facial modality, we extracted AUs from both participant's and stimuli videos on each frame using OpenFace [31], an open source tool. We could extract the intensity (values from 0 to 5) of 17 AUs and the presence (values of 0 or 1) of 18 AUs. We avoid using geometric facial features since they are linked directly with stereotypes judgment [6] while we are interested in more interactional features. For the facial modality, we did not apply any resampling.

The eye-movements of the participants were recorded using a Tobii eye tracker however, this information was not available for the stimuli. To compensate we extracted the eye movements based on the stimuli face videos using OpenFace [31]. It includes 3D eye gaze direction vector in world coordinates of both eyes, and 2D eye gaze direction in radians. The features were not exactly the same with the eye tracker recordings. For participants, the 2D gaze location on the display, the 3D locations of the left and right eyes, and the gaze duration were recorded by the eye tracker. As Tobii TX300 and T120 have different sampling rates, the features recorded by both devices were down-sampled using decimate method [32] to match the video frame rate (30Hz).

Physiological signals were only recorded for the participants (receivers) to the experiment. We used the TEAP toolbox [33] to extract physiological features. We filtered out the noise with a median filter and then extracted the spectral entropy and mean frequency of Skin Conductance Response (SCR) from the GSR signal, heart rate (HR) and 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.

Audio features were extracted using Opensmile [34] for the stimuli only. We collected the audios from both Session 1 and Session 2. However, in Session 1, participants were focused on watching the stimuli and did not speak. Thus, we extracted audio features from the stimuli (emitters) but not from the participants (receivers) for Session 1. The features included the root-mean-square signal frame energy (pcm RMS), Mel-Frequency cepstral coefficients (MFCC), zero-crossing rate of time signal (pcm zcr) and voiceProb (voicing probability) shown in Table IV. The features were resampled to 30 Hz. All the extracted multimodal features were smoothed using the same sliding window as for the annotations to get the same sample sizes. Afterwards, we standardized the feature matrix by removing the mean and scaling to unit variance.

## V. Baseline Model and Results

### A. Baseline Model

Long-short term memory (LSTM) neural networks [35] are temporal models for sequence prediction. They have been proven to perform reliably on affective detection tasks [30]. Multitask learning framework has been shown to work efficiently by targeting correlated tasks [30] (i.e. multi-label data classification/regression).

The LSTM model structure is shown as Fig.5. We used truncated back propagation through time with a max step of 100 to train our LSTM networks. The window size of the LSTM was determined based on previous studies [1], [36]. Recommended by Mariooryad and Busso [36], the delay between the moment the impression was formed by the annotator and the concrete annotation made through annotation tool, could be addressed by shifting the annotations backwards by 2 seconds. According to [1], it takes around 100 ms to form an impression and generally less than 300 ms to react and press a key [36]. Thus, in total, 3 seconds should be able to cover the possible annotation responses. That means the possible time gap in annotations and facial expressions lies within $3s \times 30fps = 90frames$. To simplify the data processing, we used 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 over-fitting with patience equal to 5 epochs. The output layer was set to 2 dimensions for

multitask learning. Mean squared error(MSE) was used as the loss function.

To evaluate the influence of individual detection/prediction, we fed our LSTM model with corresponding multimodal features from one agent of the dyad (i.e. either stimulus or participant) (shown in Fig.5). For the evaluation, we used a leave-one-participant out cross-validation scheme for impression detection/prediction using Session 1 data. 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. Once 1 participant was left out, the rest of the data was mixed. The Concordance Correlation Coefficient (CCC) is used as the performance metric for impression recognition as it indicates local variations as well as the global trend of impression simultaneously [29], [30]. The reported CCC in this paper are the mean CCC of all tested participants.

To compute cross-corpus model performance, we trained our model with all participants from Session 1 and validate it with data from Session 2. As the social cognition workload for annotation in Session 2 is high, the annotations collected from Session 2 are much sparser than Session 1. Thus we filtered the model output based on participants' annotation timestamps and compared with the reported annotation as binary classification. F1 score is used to evaluate model performance.

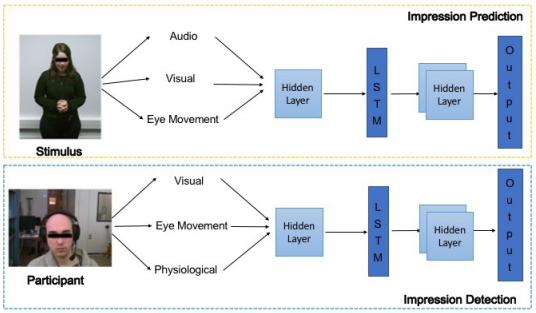


Fig. 5: Multitask LSTM for Impression Prediction & Detection

### B. Results and Discussion

We are not aware of studies for dimensional impression recognition, thus we compared the results with dimensional emotion recognition. The obtained results are better than the state of the art. For example, the highest CCC achieved for emotion recognition in [37] is 0.680 compared to 0.751 for warmth and 0.737 for competence. Comparing impression detection performance with prediction (Table V and Table VI), impression detection overall achieved better performance than impression prediction. Individual annotations ($Annotation_R$) were more detectable for impression detection while more difficult to predict by using emitter's features. Although all the impression detection outperformed impression prediction, this performance could be biased by the imbalanced training data size: 13 stimuli (emitters) for impression prediction and 54 participants (receivers) for detection.

For cross corpus validation, we only test impression detection (features and annotations are both from the emitter). We trained the multitask LSTM model with all participants from Session 1, and then we used the data from Session 2 to check the model performance. The mean F1 scores for W & C are 0.692 and 0.607 respectively for all participants from Session 2. We computed the accuracy as well and it achieved 0.921 on warmth and 0.844 on competence. The low F1 scores obtained are due to cross-study analyses: our model is trained on a dataset and is validated on another with a different context. For cross-corpus dimensional emotion recognition (valence & arousal), the F1 score is around 0.63 and 0.51 in [38]. Our results are similar to the state of the art and show the reproducibility of the impression detection model.

| Modality | Facial | EyeGaze | Physio | Face&Eye | All |
|---|---|---|---|---|---|
| Warmth | 0.747 | 0.705 | 0.371 | 0.759 | 0.751 |
| Competence | 0.718 | 0.723 | 0.212 | 0.725 | 0.737 |

TABLE V: Mean CCC of Impression Detection on W & C(Reciver features and $Annotation_R$)

| Modality | Facial | Audio | Eyegaze | All |
|---|---|---|---|---|
| Warmth | 0.193 | 0.189 | 0.144 | 0.301* |
| Competence | 0.145 | 0.162 | ;0.155 | 0.212 |

* $p < 0.1$

TABLE VI: Mean CCC of Impression Prediction on W & C (Emitter features and $Annotation_E$)

## VI. Conclusion

Impression recognition plays a ubiquitous role in human interactions and is hence of interest to a range of disciplines including psychology, computer science, and neuroscience. In this work, we propose a multimodal dataset for impression recognition on individuals and dyads using bodily signals. Besides the database, we also present a baseline method to recognize impressions with a multitask LSTM model using features from the emitter or the receiver.

The database is unique with 2 sessions of recordings. It contains elicited impression using video stimuli, as well as recordings of natural impression formation of strangers meeting for the first time through video call. The database allows impression recognition with features from the emitter and/or the receiver in the W & C dimensions. The database contains multimodal recordings including frontal face videos, audio, eye movements, ECG, Galvanic GSR and BVP. With the baseline model, we found that the formed impression is more detectable from the receiver than predicting from the emitter. That indicates that person's bodily responses for certain formed impressions are similar. Meanwhile, how receivers interpreting and perceiving the behaviour from the emitter could vary enormously. The baseline model have shown a reliable cross-corpus performance.

## Acknowledgment

This study is supported by the Swiss National Science Foundation under Grant Number 2000221E-164326 and by ANR IMPRESSSIONS project number ANR-15-CE23-0023.

## REFERENCES


[1] J. Willis and A. Todorov, "First impressions: Making up your mind after a 100-ms exposure to a face," *Psychological science*, vol. 17, no. 7, pp. 592–598, 2006.

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

[3] A. J. Cuddy, P. Glick, and A. Beninger, "The dynamics of warmth and competence judgments, and their outcomes in organizations," *Research in organizational behavior*, vol. 31, pp. 73–98, 2011.

[4] J.-I. Biel and D. Gatica-Perez, "The youtube lens: Crowdsourced personality impressions and audiovisual analysis of vlogs," *Multimedia, IEEE Transactions on*, vol. 15, no. 1, pp. 41–55, 2013.

[5] 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*, vol. 40, pp. 61–149, 2008.

[6] C. M. Judd, L. James-Hawkins, V. Yzerbyt, and Y. Kashima, "Fundamental dimensions of social judgment: understanding the relations between judgments of competence and warmth." *Journal of personality and social psychology*, vol. 89, no. 6, p. 899, 2005.

[7] P. Ekman and D. Keltner, "Universal facial expressions of emotion," *Segerstrale U, P. Molnar P, eds. Nonverbal communication: Where nature meets culture*, pp. 27–46, 1997.

[8] J. A. M. Correa, M. K. Abadi, N. Sebe, and I. Patras, "Amigos: A dataset for affect, personality and mood research on individuals and groups," *IEEE Transactions on Affective Computing*, 2018.

[9] E. Goffman *et al.*, *The presentation of self in everyday life*. Harmondsworth, 1978.

[10] S. T. Fiske, A. J. Cuddy, and P. Glick, "Universal dimensions of social cognition: Warmth and competence," *Trends in cognitive sciences*, vol. 11, no. 2, pp. 77–83, 2007.

[11] M. McCurrie, F. Beletti, L. Parzianello, A. Westendorp, S. Anthony, and W. J. Scheirer, "Predicting first impressions with deep learning," in *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*. IEEE, 2017, pp. 518–525.

[12] G. Farnadi, S. Sushmita, G. Sitaraman, N. Ton, M. De Cock, and S. Davalos, "A multivariate regression approach to personality impression recognition of vloggers," in *Proceedings of the 2014 ACM Multi Media on Workshop on Computational Personality Recognition*. ACM, 2014, pp. 1–6.

[13] H. J. Escalante, H. Kaya, A. A. Salah, S. Escalera, Y. Gucluturk, U. Guclu, X. Baró, I. Guyon, J. J. Junior, M. Madadi *et al.*, "Explaining first impressions: Modeling, recognizing, and explaining apparent personality from videos," *arXiv preprint arXiv:1802.00745*, 2018.

[14] F. Gürpinar, H. Kaya, and A. A. Salah, "Multimodal fusion of audio, scene, and face features for first impression estimation," in *2016 23rd International conference on pattern recognition (ICPR)*. IEEE, 2016, pp. 43–48.

[15] A. Cafaro, J. Wagner, T. Baur, S. Dermouche, M. Torres Torres, C. Pelachaud, E. André, and M. Valstar, "The noxi database: multimodal recordings of mediated novice-expert interactions," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. ACM, 2017, pp. 350–359.

[16] F. Ringeval, F. Eyben, E. Kroupi, A. Yuce, J.-P. Thiran, T. Ebrahimi, D. Lalanne, and B. Schuller, "Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data," *Pattern Recognition Letters*, vol. 66, pp. 22–30, 2015.

[17] F. Pianesi, M. Zancanaro, B. Lepri, and A. Cappelletti, "A multimodal annotated corpus of consensus decision making meetings," *Language Resources and Evaluation*, vol. 41, no. 3-4, pp. 409–429, 2007.

[18] B. Biancardi, A. Cafaro, and C. Pelachaud, "Analyzing first impressions of warmth and competence from observable nonverbal cues in expert-novice interactions," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. ACM, 2017, pp. 341–349.

[19] T. Baur, G. Mehlmann, I. Damian, F. Lingenfelser, J. Wagner, B. Lugrin, E. André, and P. Gebhard, "Context-aware automated analysis and annotation of social human–agent interactions," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 5, no. 2, pp. 1–33, 2015.

[20] R. Subramanian, J. Wache, M. K. Abadi, R. L. Vieriu, S. Winkler, and N. Sebe, "Ascertain: Emotion and personality recognition using commercial sensors," *IEEE Transactions on Affective Computing*, vol. 9, no. 2, pp. 147–160, 2016.

[21] F. Ringeval, B. Schuller, M. Valstar, R. Cowie, H. Kaya, M. Schmitt, S. Amiriparian, N. Cummins, D. Lalanne, A. Michaud *et al.*, "Avec 2018 workshop and challenge: Bipolar disorder and cross-cultural affect recognition," in *Proceedings of the 2018 on audio/visual emotion challenge and workshop*, 2018, pp. 3–13.

[22] J. Junior, C. Jacques, Y. Güçlütürk, M. Pérez, U. Güçlü, C. Andujar, X. Baró, H. J. Escalante, I. Guyon, M. van Gerven *et al.*, "First impressions: A survey on computer vision-based apparent personality trait analysis," *arXiv preprint arXiv:1804.08046*, 2018.

[23] C. Wang, P. Lopes, T. Pun, and G. Chanel, "Towards a better gold standard: Denoising and modelling continuous emotion annotations based on feature agglomeration and outlier regularisation," in *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*, 2018, pp. 73–81.

[24] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the recola multimodal corpus of remote collaborative and affective interactions," in *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*. IEEE, 2013, pp. 1–8.

[25] J. J. Gross and O. P. John, "Facets of emotional expressivity: Three self-report factors and their correlates," *Personality and individual differences*, vol. 19, no. 4, pp. 555–568, 1995.

[26] M. Wöllmer, F. Eyben, S. Reiter, B. Schuller, C. Cox, E. Douglas-Cowie, and R. Cowie, "Abandoning emotion classes-towards continuous emotion recognition with modelling of long-range dependencies," in *Proc. 9th Interspeech 2008 incorp. 12th Australasian Int. Conf. on Speech Science and Technology SST 2008, Brisbane, Australia*, 2008, pp. 597–600.

[27] B. Booth, K. Mundnich, and S. S. Narayanan, "A novel method for human bias correction of continuous-time annotations," 2018.

[28] N. Thammasan, K.-i. Fukui, and M. Numao, "An investigation of annotation smoothing for eeg-based continuous music-emotion recognition," in *Systems, Man, and Cybernetics (SMC), 2016 IEEE International Conference on*. IEEE, 2016, pp. 003 323–003 328.

[29] K. Brady, Y. Gwon, P. Khorrami, E. Godoy, W. Campbell, C. Dagli, and T. S. Huang, "Multi-modal audio, video and physiological sensor learning for continuous emotion prediction," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2016, pp. 97–104.

[30] S. Chen, Q. Jin, J. Zhao, and S. Wang, "Multimodal multi-task learning for dimensional and continuous emotion recognition," in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*. ACM, 2017, pp. 19–26.

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

[32] O. Grewe, F. Nagel, R. Kopiez, and E. Altenmüller, "Emotions over time: synchronicity and development of subjective, physiological, and facial affective reactions to music." *Emotion*, vol. 7, no. 4, p. 774, 2007.

[33] M. Soleymani, F. Villaro-Dixon, T. Pun, and G. Chanel, "Toolbox for emotional feature extraction from physiological signals (teap)," *Frontiers in ICT*, vol. 4, p. 1, 2017.

[34] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 2010, pp. 1459–1462.

[35] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[36] S. Mariooryad and C. Busso, "Correcting time-continuous emotional labels by modeling the reaction lag of evaluators," *IEEE Transactions on Affective Computing*, vol. 6, no. 2, pp. 97–108, 2014.

[37] B. T. Atmaja and M. Akagi, "Multitask learning and multistage fusion for dimensional audiovisual emotion recognition," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 4482–4486.

[38] S. Rayatdoost and M. Soleymani, "Cross-corpus eeg-based emotion recognition," in *2018 IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2018, pp. 1–6.