Multiple users' emotion recognition: Improving performance by joint modeling of affective reactions


DOI: 10.1109/ACII.2017.8273584
Multiple users’ emotion recognition: improving performance by joint modeling of affective reactions

Guillaume Chanel¹, Sunny Avry²,³, Gaëlle Molinarí²,³, Mireille Bétrancourt³, Thierry Pun¹

¹ Dpt. of Computer Science – Faculty of Sciences
University of Geneva
University of Geneva, Switzerland
Guillaume.Chanel@unige.ch

² Distance learning university Switzerland (UniDistance)

³ TECFA
Faculty of Psychology
University of Geneva, Switzerland

Abstract—This paper studies emotion recognition in the context of collaboration. When people are interacting with each other they tend to reach a similar emotional state through mechanisms like empathy and emotion contagion. We thus investigated if participants’ emotions could be determined from the affective reactions and behaviors of their partner. Two types of emotional expressions were studied: physiological reactions and speech.

Results show that emotions could be recognized with similar performance when employing affective features from the self or the partner. In addition, performance was improved when combining self and partner information. The results demonstrate that in social situations an emotion recognition model should include information about partners.

1. Introduction

The developments of affective and physiological computing allow machines to gain insight into the emotions, cognitive states and inner experiences of their users [1], [2]. However, modern interfaces not only serve human-computer interactions, but are also increasingly used to communicate with acquaintances. Hence contemporary computers need to develop methods to measure and quantify social interactions as done in the field of social signal processing [3]. Despite of this trend only a few studies analyzed emotional reactions, particularly physiological, in a social context (see section 2).

In the context of social interactions, emotions do not only develop in one’s mind but rather unfold according to the emotional expressions of surrounding others [4]. Similarly, the behaviors, expressions and physiology of interacting people are known to be in synchrony during interactions [5], [6]. This is partly due to processes such as emotional contagion and empathy [7] which drive people to share feelings with others and to answer compassionately to their distress. Furthermore, emotions are intrinsically social and can influence the behaviors and feelings of others [8]. People collaborating together and having the same objectives can feel similar emotions because they appraised a joint situation in the same way.

The current work focuses on the development of methods to assess emotions in a social and collaborative context. Based on the fact that emotions develop according to the behaviors and emotions of others, we state the following hypothesis:

HI: The emotions of a given person can be better predicted when including emotional expressions of a collaborative partner.

This hypothesis was tested by collecting emotional expressions (physiological reactions and speech) of people collaborating on a computer mediated task [9]. As can be seen in Fig 1, we trained models to recognize the emotions of each participant either using: (i) the expressions of this participant (called self model in Fig. 1), (ii) the expressions of the partner (called partner model), (iii) the combination of partner’s and participant’s expressions (called joint model). To validate our hypothesis, we compared the performance of these different approaches.

Figure 1. Emotion recognition frameworks. Traditional models predict self-reports based on self features. The proposed method includes partner affective features in a joint model to predict emotions.

2. Background

The detection of an individual’s emotions can be achieved by several means. For instance through the inference of facial expressions and vocal prosody [1], [10] but also by using physiological signals [11]. Results are now clearly converging, demonstrating that the multimodal fusion of several emotional cues is necessary for accurate emotion recognition [12].
However, several of the attempts to recognize emotions have been done using acted expressions or emotions elicited in the laboratory by controlled stimuli. This limits the potential transfer of these technologies to field applications.

In order to reach applications there has been a shift in the last decade toward the assessment of emotions in natural and realistic situations [13], [14]. This shift has demonstrated that most algorithms suffer from a drop of performance as compared to the recognition of controlled and acted emotions. This could be attributed to the fact that natural emotions are often more subtle, to the natural context which induces noise in the recorded signals and to the skewed distribution of class labels (i.e. some emotion is less frequently felt than others).

Note that this class imbalance should be accounted for at the evaluation stage by using appropriate performance measures such as average recall, F1 or Cohen kappa scores [14]. For instance, average recall between 62% and 66% was obtained to detect 2 emotional classes from speech in realistic situation [10]. For comparison on acted data results have been reported in the range of 60% to 82% [15]. A second important point for applications is to create participant independent models: models that do not need to be trained for each user. Again, this second challenge impacted the performance of the trained classifiers. As an example, Alzoubi et al. [14] showed that for physiological signals, and in a naturalistic context, participant dependent models could distinguish non-emotional vs. emotional moment with a kappa score around 0.21, while this score would drop to 0.01 when building participant independent models.

A natural way to collect data containing realistic emotional expressions is to record social situations. Since emotions are intrinsically social they are present in most interactions between people or when people perform a common task. For instance, the common physiological reactions of several people watching a movie was employed to detect moments of emotional importance in [16]. Similarly, the level of synchrony measured between the physiological signals of several spectators was shown to be predictive of aesthetic and emotional highlights [17]. Measuring speech during an interaction can also be used to detect moments of conflict [15], while speech turn taking, a joint measure of speech behaviors, has been shown to be an indicator of leadership [18]. Hence, the combination of several people signals has been used to assess emotional processes (e.g. conflict) at the level of the group.

In a social interaction, it is also possible to combine the information of several group members in order to recognize the emotions of a specific member. In [19] this was achieved by modeling the dynamic of a dyadic interaction using Hidden Markov Models (HMMs) and Long Short Term Memory networks (LSTMs) on facial expressions and speech. Although their models were context sensitive they only accounted for temporal relationships in the features without considering from whom the features were computed. In addition, the joint model was not compared to a self model. A similar HMM approach was proposed in [20]. By comparing to a self-model, their joint model showed improvement only when partners had more reliable modalities. Finally, Yang and Narayanan [21] focused on the analysis of mutual influence between partners. They extracted features from two modalities (speech and hand movements) representing the amount of influence between a participant and their partner. It was shown that models obtained from those features were more accurate than self models and joint models obtained by feature concatenation. All the joint models reported above ([19]–[21]) were trained based on data collected in acted (i.e. non-natural) dyadic interactions, and none were relying on physiological activity.

In this paper, we want to analyze the emotional, both physiological and vocal, reactions of two participants collaborating together through computer-mediated interactions. We are thus in a natural context where we address the problem of participant independent models. The main novelty of this paper is that we aim at recognizing emotions of individuals not only based on their own emotional (physiological and vocal) reactions but also from their partners’ reactions. To our knowledge this is the first attempt to combine multi-user information, particularly physiological reactions, for the recognition of individual emotions in a natural context.

### 3. Methods

#### 3.1. EATMINT database

The EATMINT database\(^1\) contains multimodal recordings of dyads collaborating on the design of a slogan against violence at school [9]. During their collaboration, participants’ behaviors, facial expressions, speech, eye-movements and physiological signals were recorded. The participants were free to display any social and emotional behavior. Both participants’ recordings are synchronized which allows combining information from the two participants easily. Since it also includes emotional annotations this database was an ideal choice to answer our research question. This paper reports on the use of physiological signals and speech to assess emotional moments in the interaction. More information on the database can be obtained from [9].

At the end of the collaboration, the participants visualized a video of the interaction to annotate 20 emotional and 20 non-emotional moments. They were asked to report on their own emotion. The video consisted of a facial view of themselves as well as a sound track of what they said. Hence the participants made the annotations only on the basis of their own expressions without seeing and hearing the expressions of their partner. They could freely browse the video to make their annotations. Each annotation was made at a specific time point without specifying any duration. When registering an emotional moment the participants were asked to report which specific emotion they felt using a modified version of the Geneva Emotion Wheel [22].

---

\(^1\) https://eatmint.unige.ch
3.2. Feature extraction

In order to compare the performance of features computed from the signals of the participant who reported an emotion to the performance of features computed from the signals of the partner, two feature vectors are computed for each emotional annotation (Fig. 1). Features computed from the participant (resp. partner) are called self features (resp. partner features) in the rest of this document.

Features were extracted for each annotation and for the two modalities: physiological signals and speech. A window was centered on each annotation and emotional features (either self features or partner features) were computed from that window. The features of all modalities were concatenated into a single features vector for each window. The current work focuses on simple features (e.g. mean, variance,) and features computed from existing frameworks (e.g. openSMILE [23]). This choice was done to demonstrate that our hypothesis can be validated in the most basic data processing scenario.

1) Physiological signals

The following physiological signals were recorded using a Biosemi Active II device with a sampling frequency of 512 Hz: an electrocardiogram (ECG), skin conductance (SC), abdomen expansion, blood volume pulse (BVP) and skin temperature. All those signals were shown to be predictive of emotions[11]. Physiological signals were processed on a window of 30 seconds (i.e. 15 seconds before and after the annotation). This duration was chosen based on the reaction time of the slowest signal (i.e. 4-10 seconds for SC) and the duration needed to accurately compute some features (e.g. respiration main frequency). Finally the first seconds of the collaboration were used as a baseline to normalize some features as described below for each of them.

ECG signals were processed according to the Pan–Tompkins algorithm [24] in order to obtain inter beat intervals (IBI). The mean and variance of IBI were then used as features to measure heart rate variability. The first 10 seconds of the collaboration were used to compute an average baseline IBI which was subtracted from the IBI computed annotations. BVP is also related to the cardiac activity and is a relative measure of blood pressure. To measure the global evolution of the blood pressure we computed the mean of the BVP for the window of interest relatively to the first sample of the collaboration (i.e. the first sample was subtracted from the window average). The variance of the BVP signal was used as a measure of local change.

Skin conductance was recorded by placing two electrodes on the third phalange of the middle and ring fingers. The signal was first filtered using a 0.5 seconds moving average filter. The SC mean was computed to obtain an estimation of the tonic response. This mean was normalized by subtracting the first SC sample at the start of the collaboration. Other features were computed from the derivative of the signal: the sum of the positive values and the percentage of the positive values. The number of peaks in the SC signal was also computed. These last three features quantify the amount and amplitude of phasic responses in SC.

Abdomen expansion was measured by tying a respiration belt around the abdomen of the participants. The respiration signal was smoothed and de-trended using a bandpass filter with cutoff frequencies of 0.1Hz and 1.1Hz. The power spectral density of the signal was estimated with the Welch method and the frequency with most power was registered as a feature. Standard deviation was used as a measure of respiration energy and the maximum respiration amplitude was computed as the subtraction of the maximum and minimum of the windowed signal.

Skin temperature was measured by placing a sensor on the third phalange of the little finger. The mean of the windowed signal was subtracted from the first sample of the collaboration to obtain a global change of temperature. The average derivative was computed to obtain the general trend of the temperature in the time window.

2) Speech

During the collaboration the participants could talk to each other using headset microphones. This signal was recorded all along the interaction with a sampling frequency of 22 KHz. Due to the low quality of the headset a background noise reduction procedure was applied to the signals. Using the sox command line utility a noise profile was obtained from free of speech recordings taken before the collaboration. This profile was then used to denoise the signal recorded during collaboration. Finally, the signals were scaled so that the maximum of the scaled signal corresponds to the 90 percentile of the original signal. Higher values were clipped to the maximum.

Prosodic, spectral and voice quality features were computed over a window of 30 seconds centered on each annotation. Features were extracted using the openSMILE software [23]. In order to obtain state of the art features we used the set of features computed for the Interspeech challenge 2009 [10]. This feature set contains 384 features which are frame based statistics of the zero-crossing rate, root mean square energy, harmonics to noise ratio, fundamental frequency and mel-frequency cepstral coefficients.

3.3. Classification and evaluation

The performance of several classifiers to discriminate emotional moments vs. non-emotional moments was evaluated. The specific type of emotions (e.g. anxious, relaxed) is not considered in this paper. Two types of classifiers were trained for this purpose: random forests [25] and support vector machines [26]. Random forest was chosen as a non-linear classifier which remains robust in high dimensional spaces and belongs to the family of tree classifiers. 2000 trees were trained in the forest to ensure statistical robustness of the results. Linear SVMs were

http://sox.sourceforge.net
selected as linear classifiers which are robust in high dimensional spaces. The C parameter of SVMs was empirically set to 1.

Classifiers were either trained on self features, partner features or the concatenation of these two feature vectors (Fig. 1). This allows comparing their performance on those three problems. The concatenation of the feature vectors corresponds to a feature level fusion strategy [27]. We were interested to check if different fusion strategy would lead to similar results. Hence we applied a decision level fusion by combining the output scores of classifiers trained on self features with those trained on partner features (Fig. 1). This was achieved only for the same type of classifiers. A sum fusion strategy [27] was applied by averaging the scores of the two classifiers and taking a decision based on the new value.

Due to a low number of samples per participant (maximum 20 for each class) we opted for participant independent models. A leave one participant-out procedure was used. Hence for each participant who reported their emotions a classification model was trained using features of all other participants; the targets were then estimated by applying the trained model on the data of the tested participant. Note that the trained models are always trained to recognize different participants’ labels than those presented in the test set which guarantees the independence of the training and test sets. The performance of each algorithm was evaluated by computing Cohen kappa score. This score ranges from -1 to 1. A random accuracy corresponds to a value of 0 and a perfect classification to a value of 1.

4. Results and discussion

Fig. 2 presents the distribution of kappa scores obtained among participants for the proposed models. All emotion classification strategies obtained a kappa score oscillating around 0.1 with a wide range of performance since interquartile intervals are around 0.3. However, most boxes are situated above the 0 threshold indicating that the performance was higher than random for approximately 75% of the participants. The goal of this paper is not to outperform current algorithms for emotion recognition but to test if joint models can perform better than self models. Nevertheless, we compared the obtained performance with state of the art for natural and participant independent emotion recognition. This was done by looking at the self models. Concerning physiological signals, the median kappa score is of 0.08 for random forests classifiers and 0.19 for SVM. Although the performance remains modest it is higher than random. This shows that participant independent emotion assessment (emotion vs. no emotion) is feasible in natural interactions contrarily to what was concluded in [14] where a random score (kappa score = 0.01) was obtained on a neutral vs. emotion task. In addition, computing more complex physiological features could lead to a better performance. Concerning speech, the best median kappa score is of 0.07 which corresponds to an average recall of 53.3%. This is much lower than the recall obtained in [10] and close to random classification. However, this low performance could be explained by the fact that participants focused more on their facial expressions for emotional annotation than speech. Hence it is possible that the speech windows barely contain relevant emotional information.

The results presented in Fig. 2 show that kappa scores obtained by using partner features are in the same range than the ones obtained using self features. When looking at the median kappa score and the random forest classifier the performance is even improved from 0.08 to 0.1 for physiological signals and from 0.05 to 0.11 for speech. These results demonstrate that the emotions of a given participant can be detected as well from their affective reactions than from his/her partners’ reactions. This could be due to emotion alignment between the two partners who tend to reach emotional states at the same time.

Finally, the performance obtained using self features is compared to joint models (features fusion and decision fusion) in Fig. 3. For this purpose, we computed the pairwise differences between self models and joint models and applied a Wilcoxon test to check if the obtained distributions are significantly different from 0. For the random forest classifier results presented in Fig. 3 indicate that classification scores are significantly higher for joint models than for self models. The boxes being over 0 indicates that for 75% of the participants the performance was improved by using joint models. Concerning physiological signals there is an
improvement of 0.08 kappa for feature fusion ($T=249.5$, $p=0.001$) and 0.081 kappa for decision fusion ($T=273.5$, $p=0.004$). Concerning speech there is a median improvement of 0.012 kappa for feature fusion ($T=186$, $p=0.004$) and 0.081 kappa for decision fusion ($T=184.5$, $p=0.004$). Although the effects are lower for SVM similar results were obtained. Concerning physiological activity, a median kappa improvement of 0.042 for feature fusion ($T=361.5$, $p=0.05$) and 0.061 kappa for decision fusion ($T=353$, $p=0.04$) were observed. Only the decision fusion was effective for speech with a median improvement of 0.1 kappa ($T=184.5$, $p=0.004$). The weaker effects observed for SVM might be explained by its already good performance on self features. However, taken together these results demonstrate that joint models outperform self models for both fusion strategies and that combining partners’ affective reactions can be used to improve the emotion prediction of a group member.

5. Conclusion

This paper compares traditional emotion recognition models which are based on self-measures of emotions to joint models which include information about a collaborative partner. Overall, the results validated our hypothesis H1 that the emotions of a given person can be better predicted when including emotional expressions of a collaborative partner. The improvements were distributed around 0.1 kappa. This could be due to the fact that participants feel emotions in similar moments. Such emotional alignment can be due to similar appraisals of the situation by participants, to emotional contagion and empathy or to a group climate which develops among the participants (e.g. conflict, friendship, etc.). Further work would be needed to correlate model improvements with those different factors to check the precise reasons leading to classification improvements.

This work remains a proof of concept and several research directions could be followed to improve it. Firstly, the obtained results are modest, with most kappa scores in the range of 0 – 0.2. Since it is easier to improve low performance models the results presented in this study should be validated on higher performance models. This could be achieved by computing more complex and performant features. Of particular interest is the study of joint features, such as synchronization [17], which try to model the inter-dependencies existing between participants. The current work focused on the recognition of emotions vs. non emotional moments. Further work is needed to investigate if the results are still valid for other emotion recognition tasks. In addition, static models of emotions were employed in this study and the flow of emotions felt by the two partners was not considered. It is likely that a dynamic combination of partner’s emotions would allow to capture how partners responds emotionally to each other and consequently to improve the emotion recognition efficiency. Finally, the advantages of joint models for emotion assessment should be validated in several other contexts including interactions with more than two participants and other types of interactions such as negotiation.

Acknowledgments

This work is supported by the National Center of Competence in Research (NCCR) Affective sciences financed by the Swiss National Science Foundation (n° 51NF40_104897) and hosted by the University of Geneva.

References


