Crawling to Improve Multimodal Emotion Detection

Diego R. Cueva[1], Rafael A. M. Gonçalves[1], Fábio Cozman[1],
Marcos R. Pereira-Barretto[1]

[1]Departamento de Engenharia Mecatrônica e Sistema Mecânicos
Escola Politécnica da Universidade de São Paulo (EPUSP)
Av. Prof. Melo Moraes 2231 – São Paulo – SP - Brazil
marcos.barretto@poli.usp.br

Abstract. This paper demonstrates multimodal fusion of emotion sensory data in realistic scenarios of relatively long human-machine interactions. Fusion, combining voice and facial expressions, has been enhanced with semantic information retrieved from Internet social networks, resulting in more accurate determination of the conveyed emotion.

Keywords: Affective Computing, Emotion Detection, Artificial Intelligence, Web Crawling, Sensor Fusion

1 Introduction

On the ongoing evolution of computing, the development of friendly user experiences has presented restrained innovation since the dawn of the first fully graphical applications in the 70's. Although graphical improvements of such systems have been of great relevance, the lack of a paradigm change is hindering richer interaction between humans and machines. Without such a change, humans will still have to learn and adapt to the manner of operation of each machine, always strictly bounded by use cases established by the developer.

To emulate the human capacity of contextualization of a conversation and the flexible reaction to its semantic meaning would be the most natural way of interaction: if computers had the ability to process cues from the several linguistic and non-linguistic signs that permeate human behavior (like facial expressions, tone of voice and the affective context), they would be capable of comprehending in a more appropriate way the needs and issues of the user, bringing us closer to a genuinely human centered interaction.

Driven by these issues, several studies aim to develop computational algorithms to detect emotions, whether extracted from facial or vocal features. In general, these studies support themselves on the reputable and widely used facial classification developed by Ekman and Friesen [1], which distinguishes six basic blocks of emotional expressions: happiness, sadness, disgust, fear, anger and surprise.

In the last decade, with the vast expansion of opinion-oriented Internet databases (blogs, forums and social networks), some studies have also attempted to observe

emotions in basic semantic elements of conversation, comparing them with common sense available on the web.

However, although the field in question is rapidly developing, the existing techniques for dealing with emotions are still limited. The difficulty in acquiring robustness makes these systems unreliable and often unfeasible in applications of high complexity. Furthermore, the unimodal approach (i.e. the observation of only one source of emotion) ignores the intrinsic relationships between these various affective inputs. Their relevance is studied in detail in [2].

Having such issues as motivation, this paper investigates the fusion of this diverse set of emotional components – or "sensors" – to determine whether the emotion shown by the interlocutor can by more precisely detected in a multimodal solution. Two classifiers are used for the evaluation of weights and relationships between the various inputs provided, which are then compared to unimodal solutions.

The paper presents some of the relevant work in this field in section 2, describes the tools used in section 3 and presents experiments in Sections 4 through 5.

2 Related Work

The formal modeling and understanding of human behavior is a widely discussed subject in psychology and neuroscience. Among the lines of thought, the Appraisal Theories, as discussed by Roseman [3] and Schorr [4], provide a model that explains the behavior differences of each unique individual at the same time that specifies aspects that are common to all. They describe the processes underlining emotion elicitation, which are the same for everybody, although their development is unique to the experience path of each individual. The quantification of these elicitations is also a vital topic, discussed by Sander [5].

Some of the first successful algorithms for emotion recognition on faces were described in [6]. In the decade that followed - with a substantial increase in computing power - approaches based on three-dimensional models of the face and optical flow in real time have become more common, improving over previous methods.

On the analysis of emotion in speech, one can cite recent advances in [7]. Some papers classify speech emotion through the valence in tone of the conversation (positive, neutral or negative), while others attempt to get more well-defined behaviors; a comparison may be found in [8].

On the use of the Internet as a database for emotional comprehension of words, some recent studies, like the one from [9], are starting to present encouraging results.

The multimodal analysis, in turn, still has a small and recent number of studies and publications. Besides [2], Campanella and Belin [10] perform an up-to-date discussion of cognitive studies, supporting the correlation between voice and facial expressions in the construct of emotion. In [11], data fusion of voice, facial and corporal expressions resulted in more than 10% of improvement over the unimodal approach, while [12] investigated the level at which the fusion should occur, using face and voice as inputs.

3 Tools

For facial expression analysis, the commercial software eMotion [13], developed at the University of Amsterdam, has been incorporated into the study. This choice was based on the good performance of the fitting algorithm for three-dimensional meshes and the ability to assess all the emotional states which are evaluated in this article.

For speech processing, Emo-Voice [14], developed by the Institute of Computer Sciences, University of Augsburg, was used. Emo-Voice is available as an open source package and allows for customized training of classifiers for detection of emotions in speech.

Finally, for the analysis of the emotional connotation of the speech, a dedicated tool was developed by the authors. Named "emoCrawler", the tool will have its relevance evaluated in the experiments presented on this article. emoCrwaler collects the verbs, adjectives and nouns in the discourse and uses them in queries on social networks, evaluating the emotional reaction presented in the results, as described in the next section.

3.1 The emoCrawler

As it was developed solely in the context of this paper, we discuss emoCrawler here. emoCrawler is a modular application currently in development, having as goal the evaluation of emotional contexts without the requirement of complex syntactic analysis. The program searches through the databases of online social networks, processing large volumes of results and analyzing the relative frequencies of emotional expressions. Thus, the program includes a dictionary inspired by the proposals of Goleman (apud [15]) and Laros [16], which maps different expressions to the emotions explored in this article.

The words expressed by the user which are considered significant are extracted from the discourse and are looked up in the social database. The messages related to that query are then indexed and checked against the existence of the emotional expressions contained in the dictionary. The algorithm works through a rule set, aiming at eliminating false positives. As an example, it considers the presence of negation elements in the treatment of queries.

For instance, let's suppose the user inputs the phrase "I'm sorry that she died". Since both keywords "sorry" and "died" do not define emotions, the software uses them in queries in the chosen social network. The resulting database is a set of messages that contain not only these words, but also the emotion words associated with them, which are then processed by the software. The application data flow is shown on Figure 1.

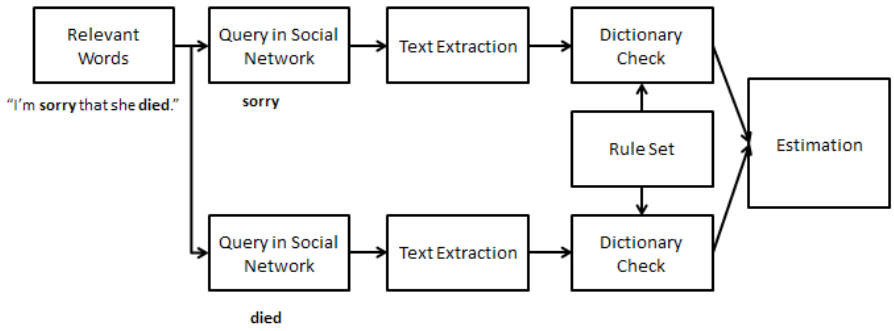

Fig. 1. Data flow example in emoCrawler

Given the modular context of emoCrawler, several social networks may be examined. For the present work, tough, only a sample from the microblog Twitter is considered [17]. The choice of Twitter as a data source is related to the conciseness and emotiveness of its posts. Users are frequently interested in expressing their opinion on a specific topic, but at the same time they have to do this in a summarized manner, without elaborate sentences – given the 140 character limitation.

The system must also consider the temporal factor in its analysis. The search for the emotional response connected to an element may present different results in different periods; the current popular opinion might diverge from the past. That being said, given the non-temporal context of the phrases used in the experiment, the mode of operation applied in this article queries only the current common sense regarding the elements of emotion.

4 Methodology

Having as objective the multimodal analysis, comprehending facial expression, voice and speech context, the *corpus eNTERFACE'05 Audio-Visual Emotion Database* [18] was selected. The database is a collection of videos comprised of scenes in which individuals are invited to express an emotional sentence in a way they feel is natural (Figure 2).

For the execution of experiments, three sets of samples that could be promptly classified by human observers were selected. The first set was reserved for the training of Emo-Voice, the second was used in the training of fusion classifiers and the third was selected for testing. For Emo-Voice, a SVM (Support Vector Machine) classifier was trained from samples of the corpus (20 for each emotion), allowing the algorithm to adapt to the characteristics of the recordings and language present in the database.

Even though there was a preliminary exclusion of unacceptable videos, many samples in non-ideal conditions were intentionally left in the experiment. Uneven

lighting, audio noise and rapid movement of head and torso are some of the elements that were purposely left.

Fig. 2. Illustrative expressions presented in the corpus

Neural networks were used for the data fusion, as they are a common choice in situations where the input sources contain noise and in which the degree of relative reliability between them is unknown. Furthermore, these networks are also able to assess the importance of results provided by a single source, what has been used to evaluate the contribution of emoCrawler.

A subset of emotions classified by Ekman as “the big six” was used: happiness, sadness, fear, anger. This study left "surprise" out, because it is often understood as an expression without valence, not actually conveying an emotional state (i.e. surprise may be tied to any state). During the study, “disgust” was also removed from the set, for reasons discussed below.

Thus, the neural network has inputs originated from three different sources, one for each emotion from each source. The classifier also contains an output for every emotion, in order to perform the reverse path of nominal attribution, i.e., converting to "names of emotions". Even though some studies describe continuous transitions for some emotions, it would be impossible to map the output to a continuous scale in this particular case, as the presented emotions are not sortable in intensity or valence. Figure 3 illustrates the process.

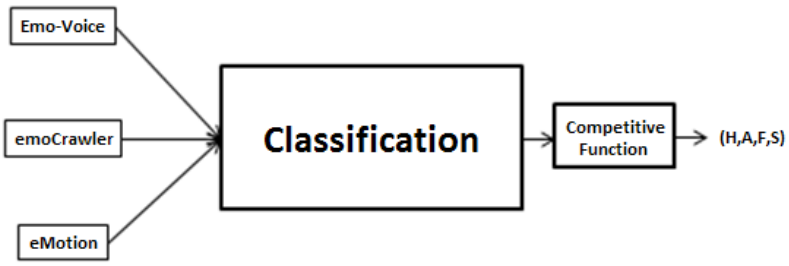

Fig. 3. Sensor Fusion

For the execution of experiments, two approaches were considered in regard to network topology and training algorithm: a feed-forward back propagation neural network (FFBPNN) and a probabilistic neural network (PNN). FFBPNN is a common solution for classification problems, while PNN, in turn, presents itself as a relevant alternative, given its training time orders of magnitude faster than the FFBPNN.

5 Experiments and Results

The first stage of evaluation consisted in the analysis of results coming directly from the sensors, i.e. the unimodal analysis. Table 1 shows the results of such tests for a heterogeneous set of samples, indicating the percentage of correct results by the tools.

Table 1. Rate of emotion detection (percentage of correct guesses) in unimodal analysis – heterogeneous sample set

Emotion	Face	Voice
Happiness	12.5%	12.5%
Anger	88.9%	11.1%
Disgust	0%	0%
Fear	50,0%	50,0%
Sadness	50,0%	100,0%

As it can be seen from Table 1, both sensors, emoVoice and eMotion, had a very bad performance in detecting "disgust", possibly due to the amateur profile of the actors. Therefore, this emotion has been removed from the classifier training and test sets, to avoid contamination issues in the classification of the other four emotions.

In a second step, data from the remaining four emotions was separated into training and test sets for the neural networks, with equal number of samples for each emotion. FFBPNN learning consisted in the application of the Robust Backpropagation algorithm to the training set. The choice of the amount of hidden layer nodes was performed by studying the rates of convergence of the network.

The probabilistic network adjustment, in turn, relies heavily on the smoothing parameter, a positive scalar related to the distance between training vectors. The methodology for choosing such parameter was to start it with a high value (generalist) and to perform successive reduction steps to ensure full compliance of training data in simulation, resulting in a spreading factor of 0.17 as the best match.

Table 2 shows, after the supervised learning, the results for the test set, comparing the unimodal assessment of emotions in face and voice with the data coming from multimodal fusion. Since emoCrawler is inherently associated with the classifier, it was not evaluated in a unimodal approach.

Table 2. Comparison of individual measurements against multimodal fusion: rate of detection for each method

Emotion	Voice	Face	FFBP Fusion (face/voice/semantics)	PNN Fusion (face/voice/semantics)
Happiness	20%	0%	60%	60%
Anger	100%	0%	100%	100%
Fear	40%	20%	80%	60%
Sadness	100%	60%	60%	60%
Average Rate	**65%**	**20%**	**75%**	**70%**

To assess the influence of emoCrawler, a bimodal FFPBNN was trained, using only face and voice information. Table 3 presents the results when submitted to the test set.

Table 3. Evaluation of emoCrawler's efficiency over the FFBPNN test group: rate of detection in each case

Emotion	emoCrawler disabled	emoCrawler enabled
Happiness	20%	60%
Anger	60%	100%
Fear	20%	80%
Sadness	100%	60%
Average Rate	**50%**	**75%**

It's possible to observe the superior multimodal behavior when semantic comprehension elements are present, particularly in emotions in which the face and voice systems had poor performance. Despite the coherent results, there's a decrease in accuracy in the specific case of sadness, likely a consequence of intense noise in emoCrawler's training dataset, which caused confusion in the learning process.

6 Final remarks and future work

In realistic scenarios, emotion detection based on facial analysis provides somewhat poor results due to uneven lightning and body movement, but mostly because speech causes facial deformations that may be interpreted as different emotions. Voice-only emotion detection is strongly influenced by valence. A multimodal approach, as the one discussed in this work, can outperform a unimodal system, as shown by experimental results.

The addition of semantic information through emoCrawler, a tool built by the authors to search for emotional content of words on Internet social networks, was demonstrated to improve emotion detection over bi-modal (face+speech) systems.

As future work, the evaluation of different classifiers (particularly Bayesian Networks) and the comparative analysis in regard to the results of this paper is planned. It's also an upcoming goal to use multimodal fusion in situations where even longer (several minutes) human-machine interactions are present, perhaps considering windowing for a closer relationship with reality.

Acknowledgments. The authors would like to thank the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), the Department of Mechatronics Engineering at EPUSP and FAPESP (by support through process n. 2008/03995-5) for the collaboration and financial support in this research.

References

1. Ekman, P., Friesen, W. “Facial Action Coding System”, Consulting Psychologist Press (1977)
2. Scherer, K. Ellgring, H. “Multimodal Expression of Emotion: Affect Programs or Componential Appraisal Patterns?”. American Psychological Association (2007)
3. Roseman, I.J.; Smith, C.A. “Appraisal Theory – Overview, Assumptions, Varieties, Controversies”. In “Appraisal Processes in Emotion – Theory, Methods, Research” editado por Scherer, K; Schorr, ª; Johnstone, T. Oxford University Press, USA (2001)
4. Schorr, A. “Appraisal – The Evolution of an Idea”. In “Appraisal Processes in Emotion – Theory, Methods, Research” editado por Scherer, K; Schorr, ; Johnstone, T. Oxford University Press, UK (2001)
5. Sander, D.; Grandjean, D.; Scherer, K. R. “A systems approach to appraisal mechanisms in emotion” Neural Networks vol. 18 pgs. 317-352 (2005)
6. Bartlett M.S., Hager J.C., Ekman P., Sejnowski T.J.. “Measuring facial expressions by computer image analysis”. Department of Cognitive Science, University of California, San Diego, USA (1999)
7. Rachuri, K.K.; Musolesi, M.; Mascolo, C.; Rentfrow, P.; Longworth, C.; Aucinas, A. “EmotionSense: a mobile phone based adaptive platform for experimental social psychology research”. UbiComp ’10, Sep 26-Sep 29, 2010, Copenhagen, Denmark.
8. Vogt, T.; André, E."Comparing Feature Sets for Acted and Spontaneous Speech in View of Automatic Emotion Recognition," in IEEE International Conference on Multimedia & Expo (ICME 2005) (2005)
9. Ptaszynski, M., Dybala, P., Shi, W., Rzepka, R., Araki, K. “Towards Context Aware Emotional Intelligence in Machines: Computing Contextual Appropriateness of Affective States”. IJCAI'09 Proceedings of the 21st international jont conference on Artifical intelligence. (2009)
10. Campanella, S., Belin, P. “Integrating face and voice in person perception”. Trends in Cognitive Sciences, 11, 535−543. (2007)
11. Castellano, G., Kessous, L. Caridakis, G. “Multimodal emotion recognition from expressive faces, body gestures and speech”. In Fiorella de Rosis, Roddy Cowie (Ed.), Proc. of the Doctoral Consortium of 2nd International Conference on Affective Computing and Intelligent Interaction, Lisbon, September (2007)
12. Chetty, G. Wagner, M. “A Multilevel Fusion Approach for Audiovisual Emotion Recognition”. International Conference on Auditory-Visual Speech Processing (2008)
13. eMotion - Visual Recognition. <http://www.visual-recognition.nl> Accessed: March 2011.
14. T. Vogt, E. André and N. Bee, "EmoVoice - A framework for online recognition of emotions from voice," in Proceedings of Workshop on Perception and Interactive Technologies for Speech-Based Systems (2008)
15. Martinez-Miranda, J.; Aldea, A. “Emotions in human and artificial intelligence”. Computers in Human Behavior vol.21 pgs.323-341 (2005)
16. Laros, F.J.M.; Steenkamp, J.E.M. “Emotions in consumer behavior: a hierarchical approach”. Journal of Business Research vol.58 pgs.1437-1445 (2005)
17. Twitter – The best way to discover what’s new in your world. <http://www.twitter.com>. Accessed: March 2011.
18. Martin, O. Kotsia, I. Macq, B. Pitas, I. “The eNTERFACE’05 Audio-Visual Emotion Database”. Université Catholique de Louvain; Aristotle University of Thessaloniki (2005)