APÉNDICE A

Publicaciones Relacionadas

A continuación se incluyen íntegramente dos publicaciones que recogen el trabajo llevado a cabo en el análisis facial de vídeos y en la fusión multimodal en este Trabajo Fin de Máster.

Continuous Facial Affect Recognition from Videos

Sergio Ballano[1], Isabelle Hupont[1], Eva Cerezo[2] and Sandra Baldassarri[2]

[1] Aragon Institute of Technology, Department of R&D and Technology Services,
Zaragoza. 5018, María de Luna 7-8, Spain
[2] University of Zaragoza, Computer Science and Systems Engineering Department,
Zaragoza. 50018, María de Luna 3, Spain
{sballano, ihupont}@ita.es, {ecerezo, sandra}@unizar.es

Abstract. The interpretation of user facial expressions is a very useful method for emotional sensing and constitutes an indispensable part of affective HCI designs. This paper proposes an effective system for continuous facial affect recognition from videos. The system operates in a continuous 2D emotional space, characterized by evaluation and activation factors, enabling a wide range of intermediary affective states to be worked with. It makes use, for each video frame, of a classification method able to output the exact location (2D point coordinates) of the still facial image in that space. It also exploits the Kalman filtering technique to control the 2D point movement along the affective space over time and to improve the robustness of the method by predicting its future locations in cases of temporal facial occlusions or inaccurate tracking. The system has been tuned with an extensive universal database and preliminary evaluation results are very encouraging.

Keywords: Affective computing, facial expression analysis.

1 Introduction

Facial expressions are the most powerful, natural and direct way used by humans to communicate affective states. Thus, the interpretation of facial expressions is the most common method used for emotional detection and forms an indispensable part of affective Human Computer Interface designs.

Facial expressions are often evaluated by classifying still face images into one of the six universal "basic" emotions or categories proposed by Ekman [1] which include "happiness", "sadness", "fear", "anger", "disgust" and "surprise" [2-4]. There are a few tentative efforts to detect non-basic affective states, such as "fatigue", "interested", "thinking", "confused" or "frustrated" [5, 6]. In any case, this categorical approach, where emotions are a mere list of labels, fails to describe the wide range of emotions that occur in daily communication settings and ignores the intensity of emotions.

To overcome the problems cited above, some researchers such as Whissell [7] and Plutchik [8], prefer to view affective states not independent but rather related to one another in a systematic manner. They consider emotions as a continuous 2D space whose dimensions are evaluation and activation. The evaluation dimension measures how a human feels, from positive to negative. The activation dimension measures whether humans are more or less likely to take some action under the emotional state, from active to passive.

For many years, a lot of effort was dedicated to recognize facial expressions in still images. Given that humans inherently display facial emotions following a continuous temporal pattern [9], more recently attention has been shifted towards sensing facial affect from video sequences. The study of facial expressions' dynamics reinforces the limitations of categorical approach, since it represents a discrete list of emotions with no real link between them and has no algebra: every emotion must be studied and recognized independently. Dimensional approach is much more able to deal with variations in emotional states over time, since in such cases changing from one universal emotion label to another would not make much sense in real life scenarios.

Continuous dimensional annotation is best suited if the annotated sample has more than one emotional apex or blended emotions [10] but this can be very time consuming, with a very poor inter-annotator agreement and can make difficult the posterior evaluation when working with long videos. An intermediate approach is to annotate only certain moments in time (key-frames). The key-frames, selected by the user which expressed the emotions, will usually correspond to the onset, apex and offset of an emotion and any other moment that the user may find interesting especially in case of blended emotions which don't pass through the neutral state.

This paper proposes a method for continuous facial affect recognition from video. The system operates in a 2D emotional space, characterized by evaluation and activation factors. It combines a classification method able to output, frame per frame, the exact location (2D point coordinates) of the shown facial image and a Kalman filtering technique that controls the 2D point movement over time through an "emotional kinematics" model. In that way, the system works with a wide range of intermediary affective states and is able to define a continuous emotional path that characterizes the affective video sequence. The system is capable of analyzing any subject, male or female of any age and ethnicity, and has been validated considering human assessment.

The structure of the paper is the following: Section 2 describes the method for facial images classification in a continuous 2D affective space. In Section 3 the step from still images to video sequences through the "emotional kinematics" model is explained in detail and Section 4 comprises the conclusions and future work.

2 A Novel Method for Facial Images Classification in a Continuous 2D Affective Space

This section describes a novel method for sensing emotions from still facial images in a continuous 2D affective space. The facial images classification method starts with a classification mechanism in discrete emotional categories that intelligently combines different classifiers simultaneously to obtain a confidence value to each Ekman universal emotional category (Section 2.1). Then, this output is subsequently expanded in order to be able to work in a continuous emotional space and thus to consider intermediate emotional states (Section 2.2).

2.1 Classifiers Combination for Discrete Emotional Classification

The starting point of the system is the method for facial emotional classification presented in authors' previous work [4]. The inputs to this method are the variations with respect to the "neutral" face of the set of facial distances and angles shown in Fig. 1. In that way, the face is modeled in a computationally simple way without losing relevant information about the facial expression. The facial points that allow to calculate the facial distances and angles are obtained thanks to faceAPI [11], a commercial real-time facial feature tracking program.

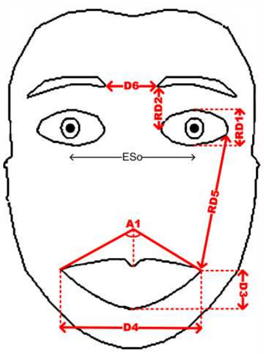
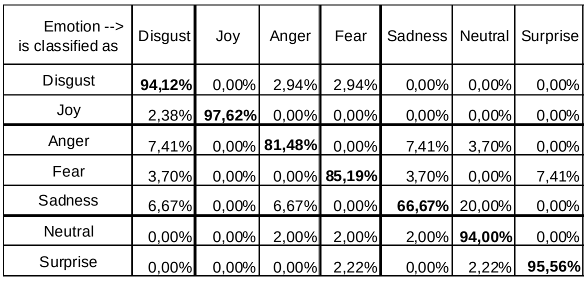


Fig. 1. System's facial inputs.

This initial method combines through a majority voting strategy [4] the five most commonly used classifiers in the literature (Multilayer Perceptron, RIPPER, SVM, Naïve Bayes and C4.5) to finally assign at its output a confidence value CV(E_i) of the facial expression to each of Ekman's six emotions plus "neutral". It has been well-tuned and tested with a total of 1500 static frames selected from the apex of the video sequences from the well-known FG-NET [12] and MMI [13] facial expression databases. Therefore, it has been validated with a large database of individuals of all races, ages and genders.

Table 1 shows the confusion matrix obtained when applying the initial discrete facial emotional classification method to the 1500 selected static frames. As can be observed, the success rates for the "neutral", "joy", "disgust", "surprise", "disgust" and "fear" are very high (81.48%-97.62%). The lowest result is for "sadness", which is confused with the "neutral" emotion on 20% of occasions, due to the similarity of their facial expressions.

Table 1. Confusion matrix obtained after applying the discrete emotional classification method to the 1500 selected static frames.



Emotion --> is classified as	Disgust	Joy	Anger	Fear	Sadness	Neutral	Surprise
Disgust	**94,12%**	0,00%	2,94%	2,94%	0,00%	0,00%	0,00%
Joy	2,38%	**97,62%**	0,00%	0,00%	0,00%	0,00%	0,00%
Anger	7,41%	0,00%	**81,48%**	0,00%	7,41%	3,70%	0,00%
Fear	3,70%	0,00%	0,00%	**85,19%**	3,70%	0,00%	7,41%
Sadness	6,67%	0,00%	6,67%	0,00%	**66,67%**	20,00%	0,00%
Neutral	0,00%	0,00%	2,00%	2,00%	2,00%	**94,00%**	0,00%
Surprise	0,00%	0,00%	0,00%	2,22%	0,00%	2,22%	**95,56%**

In his work, Plutchik [8] assigned “emotional orientation” values to a series of affect words. For example, two similar terms (like “joyful” and “cheerful”) have very close emotional orientation values while two antonymous words (like “joyful” and “sad”) have very distant values, in which case Plutchik speaks of “emotional incompatibility”. According to Plutchik’s findings, the obtained results can be considered positive as emotions with distant “emotional orientation” values (such as “disgust” and “joy” or “neutral” and “surprise”) are confused on less than 2.5% of occasions and incompatible emotions (such as “sadness” and “joy” or “fear” and “anger”) are never confused.

2.2 Emotional Mapping to a 2D Continuous Affective Space

To enrich the emotional output information from the system in terms of intermediate emotions, one of the most influential evaluation-activation 2D models has been used: that proposed by Whissell. In her study, Whissell assigns a pair of values <evaluation, activation> to each of the approximately 9000 selected affective words that make up her “Dictionary of Affect in Language” [7]. Figure 2 shows the position of some of these words in the evaluation-activation space. The next step is to build an emotional mapping so that an expressional face image can be represented as a point on this plane whose coordinates (x,y) characterize the emotion property of that face.

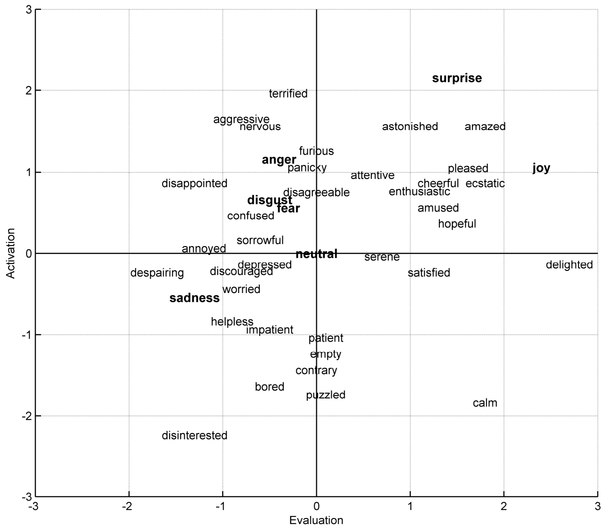


Fig. 2. Simplified Whissell's evaluation-activation space.

It can be seen that the words corresponding to each of Ekman's six emotions have a specific location (x_i, y_i) in the Whissell space (in bold in Fig. 2). Thanks to this, the output of the classifiers (confidence value of the facial expression to each emotional category) can be mapped onto the space. This emotional mapping is carried out considering each of Ekman's six basic emotions plus "neutral" as weighted points in the evaluation-activation space. The weights are assigned depending on the confidence value CV(E_i) obtained for each emotion. The final coordinates (x,y) of a given image are calculated as the centre of mass of the seven weighted points following equation (1) (see Fig. 3). In this way, the output of the system is enriched with a larger number of intermediate emotional states.

$$x = \frac{\sum_{i=1}^{7} x_i \, CV(E_i)}{\sum_{i=1}^{7} CV(E_i)} \quad and \quad y = \frac{\sum_{i=1}^{7} y_i \, CV(E_i)}{\sum_{i=1}^{7} CV(E_i)} \qquad \textbf{(1)}$$

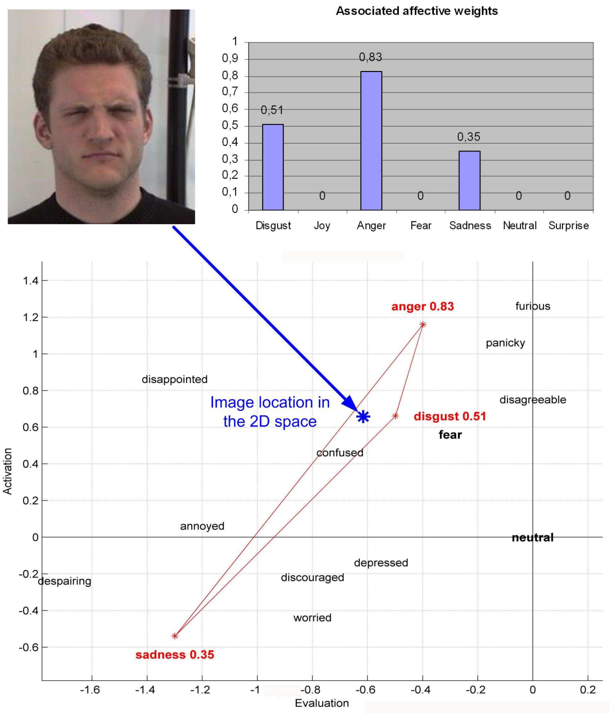


Fig. 3. Diagram for obtaining the location of a facial image in the 2D emotional space. A graphic illustration of the 2D emotional mapping process is included as an example.

Fig. 4 shows several images of the database with their nearest label in the Whissell space after applying the proposed emotional mapping.

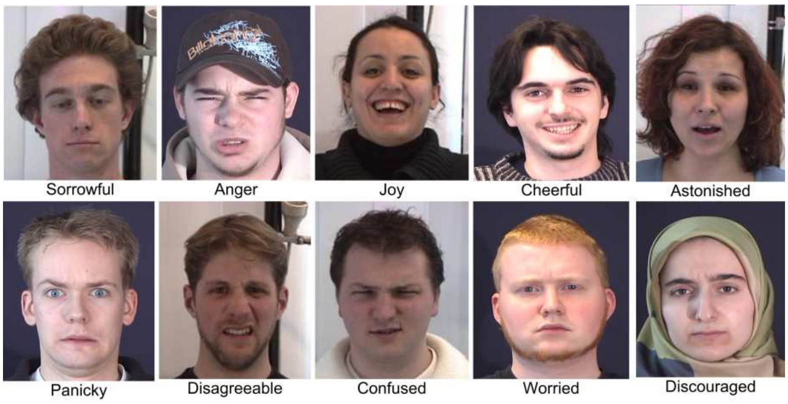


Fig. 4. Example of images from the database with their nearest label in the Whissell space after applying the 2D emotional mapping.

3 From Still Images to Video Sequences through 2D Emotional Kinematics Modeling

As pointed out in the introduction, humans inherently display facial emotions following a continuous temporal pattern. With this starting postulate and thanks to the use of the 2-dimensional description of affect, which supports continuous emotional input, an emotional facial video sequence can be viewed as a point (corresponding to the location of a particular affective state in time *t*) moving through this space over time. In that way, the different positions taken by the point (one per frame) and its velocity over time can be related mathematically and modeled, finally obtaining an "emotional path" in the 2D space that reflects intuitively the emotional progress of the user throughout the video. In Section 3.1, a Kalman filtering technique is proposed to model the "emotional kinematics" of that point when moving along the Whissell space and thus enable to both smooth its trajectory and improve the robustness of the method by predicting its future locations (e.g. in cases of temporal facial occlusions or inaccurate tracking). Section 3.2 presents the results obtained when applying the emotional kinematics model to different complex video sequences.

3.1 Modeling Emotional Kinematics with a Simple Kalman Filter

For real-time "emotional kinematics" control, the well-known Kalman filter is exploited [14]. Kalman filters are widely used in the literature for estimation problems ranging from target tracking to function approximation. Their purpose is to estimate a system's state by combining an inexact (noisy) forecast with an inexact measurement of that state, so that the most weight is given to the value with the least uncertainty at each time *t*.

Analogously to classical mechanics, the "emotional kinematics" of the point in the Whissell space (x-position, y-position, x-velocity and y-velocity) are modeled as the system's state in the Kalman framework at time t_k. The output of the 2D classification system described in Section 2 is modeled as the measurement of the system's state. In this way, the Kalman iterative estimation process -that follows the well-known recursive equations detailed in Kalman's work [14]- can be applied to the recorded user's emotional video sequence, so that each iteration corresponds to a new video frame (i.e. to a new sample of the computed emotional path). For the algorithm initialization at t_0, the predicted initial condition is set equal to the measured initial state and the 2D point is assumed to have null velocity.

One of the main advantages of using Kalman filter for the 2D point emotional trajectory modeling is that it can be used to tolerate small occlusions or inaccurate tracking. As pointed out in Section 2.1, the input facial feature points of the classification method are obtained thanks to the commercial facial tracker faceAPI [11]. In general, existing facial trackers do not perform the detection with high accuracy: most of them are limited in terms of occlusions, fast movements, large head rotations, lighting, beards, glasses, etc. Although faceAPI deals with these problems quiet robustly, on some occasions its performance is poor, especially when working in real-time. For that reason, its measurements include a confidence weighting, from 0 to 1, allowing the acceptability of the tracking quality to be determined. Thanks to it, when a low level of confidence is detected (lower than 0.5), the measurement will not be used and only the filter prediction will be taken as the 2D point position.

3.2 Experimental Results

In order to demonstrate the potential of the proposed "emotional kinematics" model, it has been tested with a set of emotionally complex video sequences, recorded in a natural (unsupervised) setting. These videos are complex owing to three main factors:

- An average user's home setup was used. A VGA resolution webcam placed above the screen is used, with no special illumination, causing shadows to appear in some cases. In addition, the user placement, not covering the entire scene, reduces the actual resolution of the facial image.
- Different emotions are displayed contiguously, instead of the usual neutral→emotional-apex→neutral pattern exhibited in the databases, so emotions such as surprise and joy can be expressed without neutral periods between them.
- Some facial occlusions occur due to the user covering his/her face or looking away during a short period of time, causing the tracking program to lose the facial features. In these cases, only the prediction from the Kalman filter is used, demonstrating the potential of the "emotional kinematics" filtering technique.

15 videos from three different users were tested, ranging from 20 to 70 seconds from which a total of 127 key-frames were extracted to evaluate different key-points of the emotional path. The key-frames were selected by the user who recorded the video, looking for each of the emotional apex and neutral points.

These key-points were annotated in the Whissell space thanks to 18 volunteers. The collected evaluation data have been used to define a region where each image is considered to be correctly located. The algorithm used to compute the shape of the region is based on Minimum Volume Ellipsoids (MVE) and follows the algorithm described by Kumar and Yildrim [15]. MVE looks for the ellipsoid with the smallest volume that covers a set of data points. The obtained MVEs are used for evaluating results at four different levels:

1. Ellipse criteria. If the point detected by the system is inside the ellipse, it is considered a success; otherwise it is a failure.

2. Quadrant criteria. The output is considered to be correctly located if it is in the same quadrant of the Whissell space as the ellipse centre.

3. Evaluation axis criteria. The system output is a success if situated in the same semi-axis (positive or negative) of the evaluation axis as the ellipse centre. This information is especially useful for extracting the positive or negative polarity of the shown facial expression.

4. Activation axis criteria. The same criteria projected to the activation axis. This information is relevant for measuring whether the user is more or less likely to take an action under the emotional state.

The results obtained following the different evaluation strategies are presented in Table 2. As can be seen, the success rate is 61.90% in the most restrictive case, i.e. with ellipse criteria. It rises to 84.92% when considering the activation axis criteria.

Table 2. Results obtained in an uncontrolled environment.

	Ellipse criteria	Quadrant criteria	Evaluation axis criteria	Activation axis criteria
Success Rate	61.90%	74.60%	79.37%	84.92%

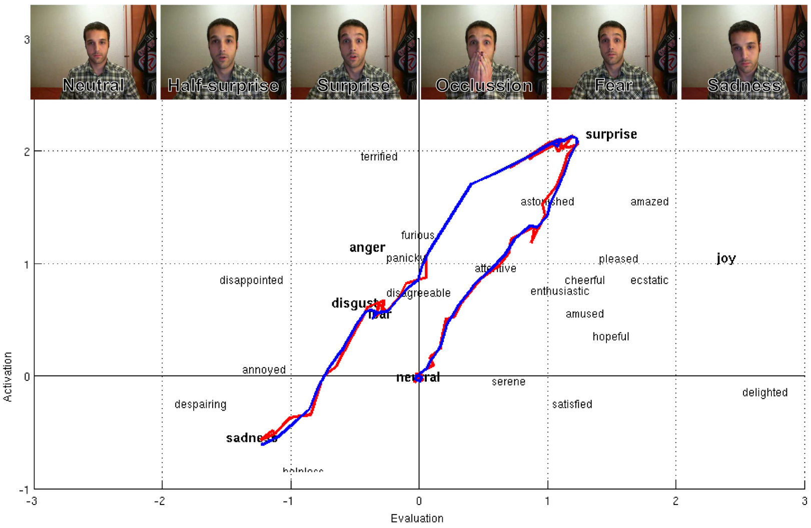


Fig. 5. “Emotional kinematics” model response during the different affective phases of the video and the occlusion period. In dashed red, emotional trajectory without Kalman filtering; In solid blue, reconstructed emotional trajectory using Kalman filter.

4 Conclusions and Future Work

This paper describes an effective system for continuous facial affect recognition from videos. The inputs are a set of facial parameters (angles and distances between facial points) that enable the face to be modeled in a computationally simple way without losing relevant information about the facial expression. The system makes use, frame per frame, of a classification method able to output the exact location (2D point coordinates) of a still facial image in the Whissell evaluation-activation space. The temporal consistency and robustness (to occlusions or inaccurate tracking) of the recognized affective sequence is ensured by a Kalman filtering technique that, through an “emotional kinematics” model, controls the 2D point trajectory when moving along the Whissell space.

The main distinguishing feature of our work compared to others is that the output does not simply provide a classification in terms of a set of emotionally discrete labels, but goes further by extending the emotional information over an infinite range of intermediate emotions and by allowing a continuous dynamic emotional trajectory to be detected from complex affective video sequences. Another noteworthy feature of the work is that it has been tuned with an extensive database of 1500 images showing individuals of different races and gender, giving universal results with very promising levels of correctness.

References

1. Keltner, D., Ekman, P.: Facial Expression Of Emotion. Handbook of emotions. pp. 236-249. New York: Guilford Publications, Inc. (2000).
2. Hammal, Z., Couvreur, L., Caplier, A., Rombaut, M.: Facial expression classification: An approach based on the fusion of facial deformations using the transferable belief model. International Journal of Approximate Reasoning. 46, 542-567 (2007).
3. Soyel, H., Demirel, H.: Facial Expression Recognition Using 3D Facial Feature Distances. Image Analysis and Recognition. pp. 831-838. Springer Berlin / Heidelberg (2007).
4. Hupont, I., Cerezo, E., Baldassarri, S.: Sensing facial emotions in a continuous 2D affective space. Presented at the Systems Man and Cybernetics (SMC) , Istanbul Octubre 10 (2010).
5. Kapoor, A., Burleson, W., Picard, R.W.: Automatic prediction of frustration. Int. J. Hum.-Comput. Stud. 65, 724-736 (2007).
6. Yeasin, M., Bullot, B., Sharma, R.: Recognition of facial expressions and measurement of levels of interest from video, (2006).
7. Whissell, C.M.: The Dictionary of Affect in Language, Emotion: Theory, Research and Experience. New York Academic (1989).
8. Plutchik, R.: Emotion: a Psychoevolutionary Synthesis. Harper & Row (1980).
9. Petridis, S., Gunes, H., Kaltwang, S., Pantic, M.: Static vs. dynamic modeling of human nonverbal behavior from multiple cues and modalities. Proceedings of the 2009 international conference on Multimodal interfaces. 23-30 (2009).
10. Ellen Douglas-Cowie, Laurence Devillers, Jean-Claude Martin, Roddy Cowie, Suzie Savvidou, Sarkis Abrilian, Cate Cox: Multimodal Databases of Everyday Emotion: Facing up to Complexity. Ninth European Conference on Speech Communication and Technology. pp. 813-816. , Lisbon, Portugal (2005).
11. Face API technical specifications brochure, http://www.seeingmachines.com/pdfs/brochures/faceAPI-Brochure.pdf.
12. Facial Expressions and Emotion Database. Technische Universität München, http://cotesys.mmk.e-technik.tu-muenchen.de/isg/content/feed-database.
13. Pantic, M., Valstar, M., Rademaker, R., Maat, L.: Web-based database for facial expression analysis. Proc. Int’l Conf. Muntimedia and Expo. 317-321 (2005).
14. Kalman, R.: A New Approach to Linear Filtering and Prediction Problems. Transactions of the ASME – Journal of Basic Engineering. 35-45 (1960).
15. Kumar, P., Yildirim, E.A.: Minimum-Volume Enclosing Ellipsoids and Core Sets. Journal of Optimization Theory and applications. 126, 1-21 (2005).

Scalable Multimodal Fusion for Continuous Affect Sensing

Isabelle Hupont and Sergio Ballano
Department of R&D and Technology Services
Aragon Institute of Technology
Zaragoza, Spain
Email: {ihupont,sballano}@ita.es

Sandra Baldassarri, *Senior Member, IEEE*, and Eva Cerezo
Computer Science and Systems Engineering Department
University of Zaragoza
Zaragoza, Spain
Email: {sandra,ecerezo}@unizar.es

***Abstract*—The success of affective interfaces lies in the fusion of emotional information coming from different modalities. This paper proposes a scalable methodology for fusing multiple affect sensing modules, allowing the subsequent addition of new modules without having to retrain the existing ones. It relies on a 2-dimensional affective model and is able to output a continuous emotional path characterizing the user's affective progress over time.**

***Index Terms*—Affective Computing, facial expression analysis, multimodal recognition, sentiment analysis.**

I. Introduction

Emotions are a fundamental component of human experience, cognition, perception, learning and communication. For this reason, affect sensing is becoming an increasingly popular research field for the enhancement of human-computer interaction (HCI).

Natural human-human affective interaction is inherently multimodal: people communicate emotions through multiple channels such as facial expressions, gestures, dialogues, etc. Although several studies prove that multisensory fusion (e.g. audio, visual, physiological responses...) improves the robustness and accuracy of machine analysis of human emotion [1]–[3], most emotional recognition works still focus on increasing the success rates in sensing emotions from a single channel rather than merging complementary information across channels [1]. Multimodal fusion of different affective channels is still in its initial stage and far from being solved [4]. There are several problems that make it an especially difficult task.

One of these problems is the definition of a reliable strategy to fuse the affective information coming from different sources with very different time scales, metric levels and temporal structures. Existing fusion strategies follow three main streams: feature-level fusion, decision-level fusion and hybrid fusion. Feature-level fusion combines the data (features) extracted from each channel in a joint vector before classification. Although several works have reported good performances when fusing different modalities at a feature-level [3], [5], [6] this strategy becomes more challenging as the number of input features increases and they are of very different natures (different timing, metrics, etc.). Adding new modalities implies a big effort to synchronize the different inputs and retrain the whole classification system. To overcome these difficulties, most researchers choose decision-level fusion, in which the inputs coming from each modality are modeled and classified independently, and these unimodal recognition results are integrated at the end of the process by the use of suitable criteria (expert rules, simple operators such as majority vote, sum, product, adaptation of weights, etc.).

Many studies have demonstrated the advantage of decision-level fusion over feature-level fusion, due to the uncorrelated errors from different classifiers [7] and the fact that time and feature dependence are abstracted. Various -mainly bimodal-decision-level fusion methods have been proposed in the literature [8]–[10], but optimal fusion designs are still undefined. Most available multimodal recognizers have designed adhoc solutions for fusing information coming from a set of given modalities but cannot accept new modalities without re-defining and/or re-training the whole system. Moreover, in general they are not adaptive to the input quality and therefore do not consider eventual changes in the reliability of the different information channels. Decision-level methods allow the integration of different algorithms without knowing their inner workings, which can be common when one or more of them are based on commercial software.

The hybrid methods try to combine the flexibility of the decision-level methods, by maintaining different classifiers for each modality, while using part of the information from every sensor in each modality, and taking advantage, when there is a statistical dependence between modalities, as in feature-level methods. For example in [11] a Multidimensional Dynamic Time Warping algorithm is used to improve speech recognition by fusing the audio channel with mouth gestures from a video channel. The common drawbacks of these methods with feature-level ones is the need to retrain the whole system when adding a new channel.

Another key factor that directly affects multimodal fusion is related to the chosen output description level of affect. Affect is often classified into one of the six universal "basic" emotions or categories proposed by Ekman [12] which include "happiness", "sadness", "fear", "anger", "disgust" and "surprise". There have been a few tentative efforts to detect non-basic affective states, such as "fatigue", "anxiety", "confused" or "frustrated" [3], [13]. In any case, this categorical approach, where emotions are a mere list of labels, fails to describe the wide range and intensities of emotions that occur in daily communication settings. Especially in the case of emotion transitions that do not go through "neutral" state, for example a transition from "surprise" to "angry" can not be represented with only two labels. To overcome these problems, some researchers such as Whissell [14] and Plutchik [15] prefer to view affective states not independently but rather related to one another in a systematic manner. They consider emotions as a continuous 2D space whose dimensions are evaluation and activation. The evaluation dimension measures how a human feels, from positive to negative. The activation dimension measures whether humans are more or less likely to take some action under the emotional state, from active to passive. Unlike the categorical approach, the dimensional approach describes an infinite number of affective states and intensities.

The multimodal fusion problem reinforces the limitations of categorical descriptions of affect. Discrete emotional labels have no real link between them and, at the fusion stage, every studied emotion must be recognized independently. The dimensional approach is best suited to deal with variations in emotional states over time. It provides an algebra and allows the emotional inputs coming from different modalities to be related mathematically. This is especially useful when integrating modules with different time-scales.

However, very few works have chosen a dimensional description level, and the few that do are more related to the design of synthetic faces [16], affective video content annotation [17] or psychological studies [18] than to recognition of emotions. This is mainly due to the current lack of (both unimodal and multimodal) databases annotated in terms of evaluation-activation dimensions. Some interesting dimensional databases are publically available [19], [20], but, in comparison to categorical ones, they are limited in terms of number of modalities (in general, they explore audio and/or video channels exclusively), annotators, subjects, samples, etc. Moreover, manual dimensional annotation of ground truth is very time consuming and unreliable, since a large labeling variation between different human raters is reported when working with the dimensional approach [21]. For these reasons, although working at the dimensional level would be more appropriate to face the problem of multimodal fusion, for training and validation of the individual modules to be fused using databases with categorical annotations is more reliable. In this way the introduction of noise into the training (due to scarce or poor data) and consequently the building of systems that are not very robust can be avoided.

Furthermore, it can not be forgotten that human emotions are usually continous and smooth over time. For a person is rare to go from "angry" to "happy" without slowly passing trough a number of intermediate states over time. This behaviour is modelled thanks to the "emotional kinematics" concept and the use of kalman filter. This way, the presented method will have a continous emotional output not only in the 2D space but also in time.

This paper proposes an original and scalable methodology for fusing multiple affect recognition modules. In order to let the modules be defined in a robust and reliable way by means of existing categorical databases, each module is assumed to classify in terms of its own list of emotional labels. Whatever these labels are, the method is able to map each module's output to a continuous evaluation-activation space, fuse the different sources of affective information over time through mathematical formulation and obtain a 2D dynamic emotional path representing the user's affective progress as final output. To show the potential of the proposed methodology, we applied it to an Instant Messaging tool able to feed 3 different affect recognition modules that sense emotions by analyzing user's facial expressions, typed-in text and "emoticons", respectively. Thanks to the scalability of the method, the IM tool would be easily improved by adding new modules, such as voice emotion recognition, without having to retrain the whole system each time a new module is added. This article aims to be a first step towards bringing a new perspective to the open issue of emotional multimodal fusion and to open the door to further discussion.

The structure of the paper is the following. Section II details the proposed multimodal affective fusion methodology. In section III this methodology is put into practice using the Instant Messaging tool. Finally, Section IV sets out our conclusions and a description of future work.

II. A Scalable Multimodal Fusion Methodology for Continuous Affect Sensing

This section details a general methodology for fusing multiple affective recognition modules and obtaining, as an output, a global 2D dynamic emotional path in the evaluation-activation space. It is assumed that every module i to be fused outputs a list -of one or more- discrete emotional labels characterizing the affective stimulus recognized at a given time t_{0i}. The possible output labels can be different for each module i. In this way, the modules' performances are maximized since unimodal databases annotated in categorical terms are -to date- more complete and reliable than dimensional and/or multimodal ones, allowing the individual modules to be better trained and validated. The proposed methodology is sufficiently scalable to add new modules coming from new channels without having to retrain the whole system. Fig. 1 shows the general fusion scheme that will be explained step by step in sections II-A, II-C and II-C.

A. Emotional Mapping to a Continuous 2D Affective Space

The first step of the methodology is to build an emotional mapping so that the output of each module i at a given time t_{0i} can be represented as a two-dimensional coordinates vector $p_i(t_{0i}) = [x_i(t_{0i}); y_i(t_{0i})]$ on the evaluation-activation space that characterizes the affective properties extracted from that module.

To achieve this mapping, one of the most influential evaluation-activation 2D models is used: the Whissell space. In her study, Whissell assigns a pair of values $\langle evaluation; activation \rangle$ to each of the approximately 9000 affective words that make up her "Dictionary of Affect in

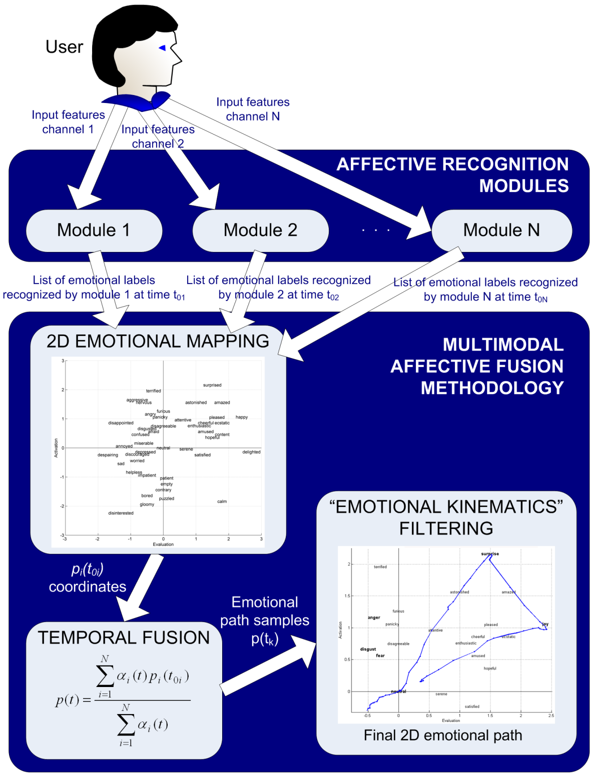

User
Input features channel 1
Input features channel 2
Input features channel N
AFFECTIVE RECOGNITION MODULES
Module 1
Module 2
. . .
Module N
List of emotional labels recognized by module 1 at time t_{01}
List of emotional labels recognized by module 2 at time t_{02}
List of emotional labels recognized by module N at time t_{0N}
2D EMOTIONAL MAPPING
MULTIMODAL AFFECTIVE FUSION METHODOLOGY
$p_i(t_{0i})$ coordinates
TEMPORAL FUSION
$p(t) = \frac{\sum_{i=1}^{N} \alpha_i(t) p_i(t_{0i})}{\sum_{i=1}^{N} \alpha_i(t)}$
Emotional path samples p(t_k)
"EMOTIONAL KINEMATICS" FILTERING
Final 2D emotional path

Fig. 1. Continuous multimodal affective fusion methodology.

Language" [14]. Fig. 2 shows the position of some of these words in the evaluation-activation space. The majority of categorical modules described in the literature provide as output at the time t_{0i} (corresponding to the detection of the affective stimulus) a list of emotional labels with some associated weights. Whatever the labels used, each one has a specific location, i.e. an associated 2D point, in the Whissell space. The components $\langle x_i(t_{0i}); y_i(t_{0i})\rangle$ of the coordinates vector $p_i(t_{0i})$ are then calculated as the barycenter of those weighted points.

B. Temporal Fusion of Individual Modules: Obtaining a Continuous 2D Emotional Path

Humans inherently display emotions following a continuous temporal pattern [22]. With this starting postulate, and thanks to the use of evaluation-activation space, the user's emotional progress can be viewed as a point (corresponding to the location of a particular affective state in time t) moving through this space over time. The second step of the methodology aims to compute this emotional path by fusing the different $p_i(t_{0i})$ vectors obtained from each modality over time.

The main difficulty to achieve multimodal fusion is related to the fact that t_{0i} affective stimulus arrival times may be known a-priori or not, and may be very different for each module. To overcome this problem, the following equation is proposed to calculate the overall affective response $p(t) = [x(t); y(t)]$ at any arbitrary time t:

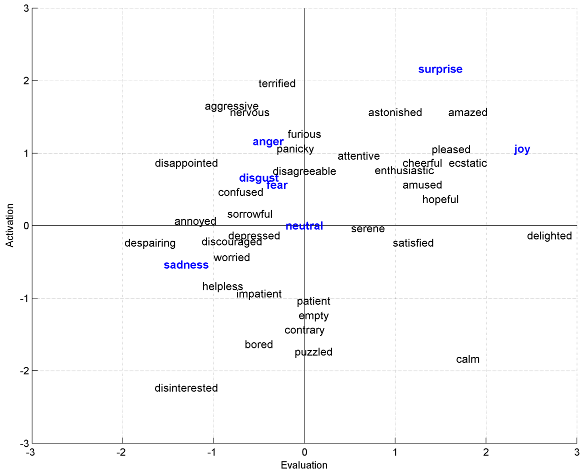


Fig. 2. Simplified Whissell's evaluation-activation space.

$$p(t) = \frac{\sum_{i=1}^{N} \alpha_i(t) p_i(t_{0i})}{\sum_{i=1}^{N} \alpha_i(t)} \tag{1}$$

where N is the number of fused modalities, t_{0i} is the arrival time of the last affective stimulus detected by module i and $\alpha_i(t)$ are the 0 to 1 weights (or confidences) that can be assigned to each modality i at a given arbitrary time t.

In this way, the overall fused affective response is the sum of each modality's contribution $p_i(t_{0i})$ modulated by the $\alpha_i(t)$ coefficents over time. Therefore, the definition of $\alpha_i(t)$ is especially important given that it governs the temporal behaviour of the fusion. As suggested by Picard [23], human affective responses are analogous to systems with additive responses with decay where, in the absence of input, the response decays back to a baseline. Following this analogy, the $\alpha_i(t)$ weights are defined as:

$$\alpha_i(t) = \begin{cases} b_i c_i(t_{0i}) e^{-d_i(t-t_0)} & t > \varepsilon \\ 0 & t \leq \varepsilon \end{cases} \tag{2}$$

where:

- b_i is the general confidence that can be given to module i (e.g. the general recognition success rate of the module).
- $c_i(t_{0i})$ is the temporal confidence that can be assigned to the last output of module i due to external factors (i.e. not classification issues themselves). For instance, due to sensor errors if dealing with physiological signals, or due to facial tracking problems if studying facial expressions (such as occlusions, lighting conditions, etc.).
- d_i is the rate of decay (in s^{-1}) that indicates how quickly an emotional stimulus decreases over time for module i.
- ε is the threshold below which the contribution of a module is assumed to disappear. Since exponential functions tend to zero at infinity but never completely disappear, ε

indicates the $\alpha_i(t)$ value below which the contribution of a module is small enough to be considered non-existent.

By defining the aforementioned parameters for each module i and applying (1) and (2), the emotional path that characterizes the user's affective progress over time can be computed by calculating successive $p(t)$ values with any desired time between samples Δt. In other words, the emotional path is progressively built by adding $p(t_k)$ samples to its trajectory, where $t_k = k\Delta t$ (with k integer).

C. "Emotional Kinematics" Path Filtering

Two main problems threaten the emotional path calculation process:

1) If the contribution of every fused module is null at a given sample time, i.e. every $\alpha_i(t)$ is null at that time, the denominator in (1) is zero and the emotional path sample cannot be computed. Examples of cases in which the contribution of a module is null could be the failure of the connection of a sensor of physiological signals, the appearance of an occlusion in the facial/postural tracking system, or simply when the module is not reactivated before its response decays completely.
2) Large "emotional jumps" in the Whissell space can appear if emotional conflicts arise (e.g. if the distance between two close coordinates vectors $p_i(t_{0i})$ is long).

To solve both problems, a Kalman filtering technique is applied to the computed emotional path. By definition, Kalman filters estimate a system's state by combining an inexact (noisy) forecast with an inexact measurement of that state, so that the biggest weight is given to the value with the least uncertainty at each time t. In this way, on the one hand, the Kalman filter serves to smooth the emotional path's trajectory and thus prevent large "emotional jumps". On the other hand, situations in which the sum of $\alpha_i(t)$ is null are prevented by letting the filter prediction output be taken as the 2D point position for those samples.

In an analogy to classical mechanics, the "emotional kinematics" of the 2D point moving through the Whissell space (position and velocity) are modelled as the system's state X_k in the Kalman framework, i.e. $X_k = [x, y, v_x, v_y]_k^T$ representing x-position, y-position, x-velocity and y-velocity at time t_k. The successive emotional path samples $p(t_k)$ are modelled as the measurement of the system's state. The two well-known main equations involved in the Kalman filtering technique are defined in the following way:

1) Process equation:

$$X_{k+1} = F_{k+1;k} X_k + w_k$$

$$\begin{bmatrix} x \\ y \\ v_x \\ v_y \end{bmatrix}_{k+1} = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ v_x \\ v_y \end{bmatrix}_{k} + w_k$$

where $F_{k+1;k}$ is the transition matrix taking the state X_k from time k to time $k+1$ (i.e. from one emotional path sample to the next). The process noise w_k is assumed to be additive, white, Gaussian and with zero mean. As suggested in the literature [24], its covariance matrix Q_k is defined as:

$$Q_k = \sigma^2 \begin{bmatrix} \frac{1}{3} & 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{3} & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & 1 & 0 \\ 0 & \frac{1}{2} & 0 & 1 \end{bmatrix}$$

where σ^2 is the intensity of a white continuous-time Gaussian noise process modeling the 2D point acceleration (which has not been considered as an element in the system's state).

2) Measurement equation:

$$Y_k = H_k X_k + z_k$$

$$\begin{bmatrix} x_m \\ y_m \end{bmatrix}_k = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}_k \begin{bmatrix} x \\ y \\ v_x \\ v_y \end{bmatrix}_k + z_k$$

where Y_k is the measurable at time k and H_k is the measurement matrix. The measurement noise z_k is assumed to be additive, white, Gaussian, with zero mean and uncorrelated with the process noise w_k. Its covariance matrix R_k is the identity matrix:

$$R_k = \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix}$$

so that it is assumed that the x and y measurements contain independent errors with $\lambda\ units^2$ variance.

Once the process and measurement equations are defined, the Kalman iterative estimation process can be applied to the emotional path, so that each iteration corresponds to a new sample.

III. Multimodal Fusion Application to Instant Messaging

Instant Messaging (IM) is a widely used form of real-time text-based communication between people using computers or other devices. Advanced IM software clients also include enhanced modes of communication, such as live voice or video calling. As users typically experience problems in accurately expressing their emotions in IM text conversations (e.g. statements intended to be ironic may be taken seriously, or humorous remarks may not be interpreted exactly as intended), popular IM programs have resorted to providing mechanisms referred to as "smileys" or "emoticons" seeking to overcome the IM systems' lack of expressiveness.

This section aims to show the potential of the multimodal affective fusion methodology presented in section II through the use of an Instant Messaging tool that combines different communication modalities (text, video and "emoticons"), each one with very different time scales. Section III-A describes the IM tool. In Section III-B the modules that extract emotional information from each modality are presented. Section III-C explains how the methodology has been tuned to achieve multimodal affective fusion. Finally, section III-D presents the experimental results obtained when applying the fusion methodology to an IM emotional conversation.

A. Instant Messaging Tool Description

Although any publicly available IM tool could be used (e.g. Skype [25] or Yahoo! Messenger [26]), a simple ad-hoc IM tool has been designed. It allows two persons to communicate via text, live video and "emoticons". Fig. 3 shows a snapshot of the tool during a conversation. The tool enables access in real-time to the following:

1) The introduced text contents, when the user presses the "enter" key (i.e. sends the text contents to his/her interlocutor).
2) The inserted "emoticons", when the user presses the "enter" key.
3) Each recorded remote user video frame (with a video rate f=25fps).

This information will serve as input to the three different affect recognition modules presented in section III-B.

B. Fusion Modalities

Three different modules are used to extract emotional information from the IM tool. Each one explores a different IM tool modality (text, "emoticons" or video) and makes use of a different set of output emotional categories:

1) **Module 1: text analysis module.** To extract affective cues from user's typed-in text, the "Sentic Computing" sentiment analysis paradigm presented in the authors' previous work [27], [28] is exploited. By using Artificial Intelligence and Semantic Web techniques, this module is able to process natural language texts to extract a "sentic vector" containing a list of up to 24 emotional labels. "Sentic Computing" enables the analysis of documents not only on the page or paragraph-level but even on the sentence level (i.e. IM dialogues level), obtaining a very high precision (73%) and significantly good recall and F-measure rates (65% and 68% respectively) at the output.
2) **Module 2: "emoticon" module.** "Emoticons" are direct affective information from the user. For this reason, this module simply outputs the list of emotional labels associated to the inserted "emoticons". Fig. 4 shows the 16 available "emoticons" and their corresponding labels, designed to be a good representation of each affective state [29]. Although the use of emoticons could be seen as a form of self-report and therefore making irrelevant the rest of modules, not all people use emoticons in the same way nor with the same frequency. There are differences in use, for example, depending on the user's gender [30], [31] and the cultural differences impose the level of contextual information required for communication [32]. Even more, a user could not be willing to directly express his/her emotional state. For these reasons emoticons can not be the only emotional sensor, but when used, they provide a reliable information, helping to solve complex emotional misunderstandings for example when sarcasm is present.
3) **Module 3: facial expression analysis module.** This module, also presented elsewhere by the authors [33], studies each frame of the recorded video sequence to automatically classify the user's facial expression in terms of Ekman's six universal emotions (plus the neutral one), giving a membership confidence value to each output emotional category. The classification mechanism inputs are a set of facial distances and angles between feature points of the face (eyebrows, mouth and eyes) extracted thanks to a real-time facial feature tracking program. The module is capable of analysing any subject, male or female, of any age and ethnicity with an average success rate of 87%.

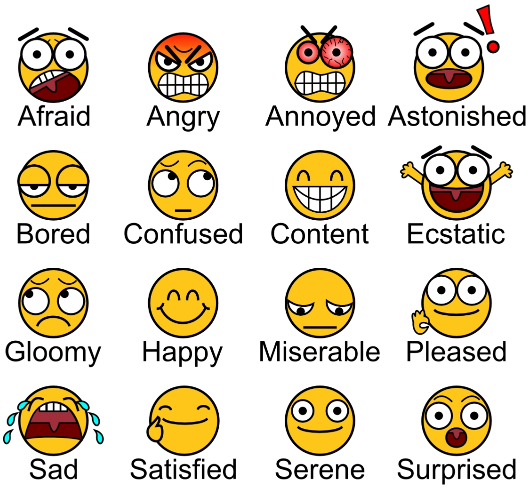


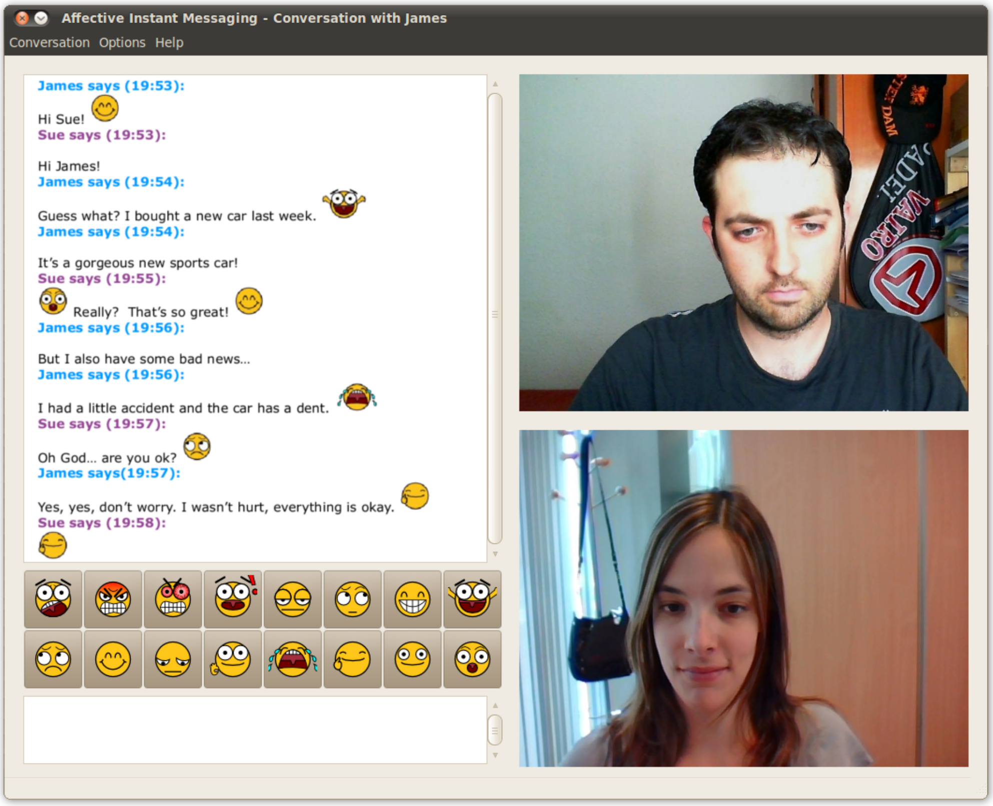
Fig. 4. "Emoticons" designed for the Instant Messaging tool and their corresponding emotional labels.

C. Multimodal Fusion Methodology Tuning

This section describes, step by step, how the multimodal fusion methodology is tuned to fuse the 3 different affect recognition modules in an optimal way.

1) Step 1: Emotional Mapping to the Whissell space: Every output label extracted by the text analysis module, the "emoticon" module and the facial expression analyzer has a specific location in the Whissell space. Thanks to this, the first step of the fusion methodology (section II-A) can be applied and vectors $p_i(t_{oi})$ can be obtained each time a given module i outputs affective information at time t_{oi} (with i comprised between 1 and 3).

2) Step 2: Temporal Fusion of Individual Modalities: It is interesting to notice that vectors $p_i(t_{0i})$ coming from the text analysis and "emoticons" modules can arrive at any time t_{0i}, unknown a-priori. However, the facial expression module outputs its $p_3(t03)$ vectors with a known frequency, determined by the video frame rate f. For this reason, and given that the facial expression module is the fastest acquisition module, the emotional path's time between samples is assigned to $\Delta t = \frac{1}{f}$.


Affective Instant Messaging - Conversation with James
Conversation Options Help

James says (19:53):
Hi Sue!
Sue says (19:53):
Hi James!
James says (19:54):
Guess what? I bought a new car last week.
James says (19:54):
It's a gorgeous new sports car!
Sue says (19:55):
Really? That's so great!
James says (19:56):
But I also have some bad news...
James says (19:56):
I had a little accident and the car has a dent.
Sue says (19:57):
Oh God... are you ok?
James says(19:57):
Yes, yes, don't worry. I wasn't hurt, everything is okay.
Sue says (19:58):

Fig. 3. Snapshot of the Instant Messaging tool during a conversation.

The next step towards achieving the temporal fusion of the different modules (section II-B) is assigning a value to the parameters that define the $\alpha i(t)$ weights, namely b_i, $c_i(t_{0i})$, d and ε. Table 1 summarizes the values assigned to each parameter for each modality and the reasons for their choice. It should be noted that it is especially difficult to determine the value of the different di given that there are no works in the literature providing data for this parameter. Therefore it has been decided to establish the values empirically. Once the parameters are assigned, the emotional path calculation process can be started following (1) and (2).

3) Step 3: "Emotional Kinematics" Filtering: Finally, the "emotional kinematics" filtering technique (section II-C) is iteratively applied in real-time each time a new sample is added to the computed emotional path. As in most of the works that make use of Kalman filtering, parameters σ and λ are established empirically. An optimal response has been achieved for $\sigma = 0.5\ units/s^2$ and $\lambda = 0.5\ units^2$.

# Module	**1**	**2**	**3**
Modality	text	"emoticons"	video
Total number of possible output labels	24 weighted emotional labels	16 emotional labels	6 Ekman's universal labels (plus "neutral") + confidence valude to each output label
General confidence b_i	$b_1 = 0.65$ The general confidence is assigned the value of the module's recall rate.	$b_2 = 1$ The maximum general condicence value is assigned since emoticons are the direct expression of user's affective state.	$b_3 = 0.87$ The general conficence is assigned the value of the module's general success rate
Temporal confidence $c_i(t_{0i})$	$c_1(t_{01}) = c_2(t_{02}) = 1$ The temporal confidence is assigned constant value 1 since the modules do not depend on external factors.		$c_3(t_{03})$ is assigned to the tracking quality confidence weighting, from 0 to 1, provided by the facial feature tracking program for each analyzed video frame.
Decay value d_i	$d_1 = d_2 = 0.035s^{-1}$ Value established empirically.		Irrelevant since the emotional path sample rate is equal to the video frame rate.
Threshold value ϵ	$\epsilon = 0.1$ Value established empirically.		

TABLE I
TEMPORAL FUSION PARAMETERS

D. Experimental Results

In order to demonstrate the potential of the presented fusion methodology, it has been applied to the Instant Messaging conversation shown in Fig. 3 (James' side). This conversation is emotionally complex owing to the fact that contrasting emotions are displayed contiguously (at first, James is excited and happy about having bought a wonderful new car and shortly afterwards becomes sad when telling Sue he has dented it).

Fig. 5 shows the emotional paths obtained when applying the methodology to each individual module separately (i.e. the

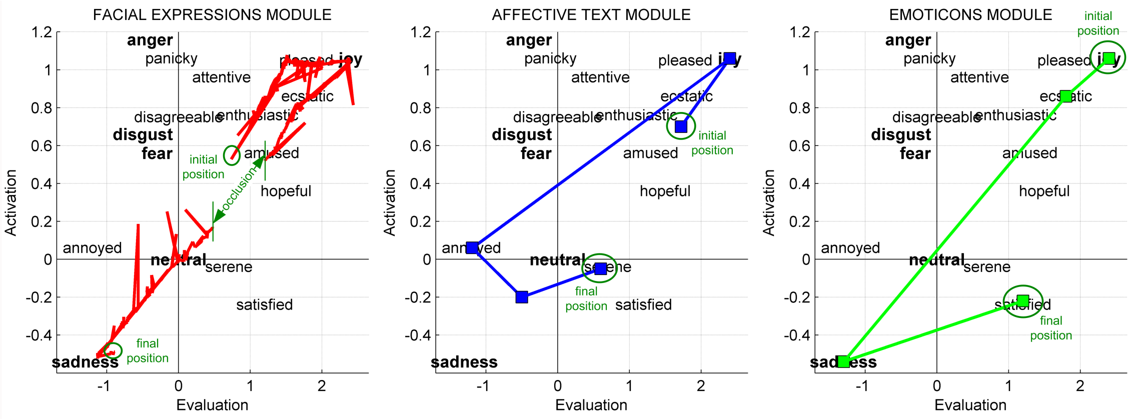
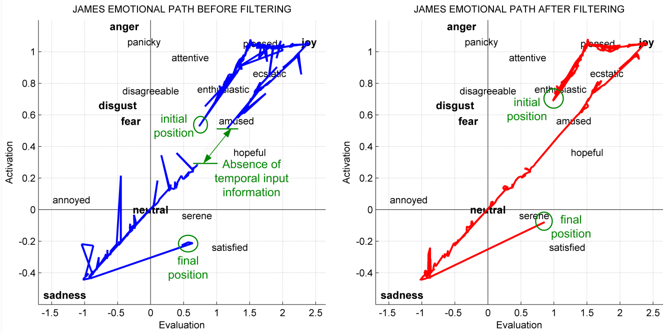


Fig. 5. Emotional paths obtained when applying the methodology to each individual module separately without "emotional kinematics" filtering. Square markers inidicate the arrival time of an emotional stimulus (not shown for facial expression module for figure clarity reasons).



(a) (b)

Fig. 6. Continuous emotional path obtained when applying the multimodal fusion methodology to James' Instant Messaging conversation shown in Fig. 3, without using "emotional kinematics" filtering (a), and using "emotional kinematics" filtering (b).

modules are not fused, only the contribution of one module is considered) without using "emotional kinematics" filtering. At first sight, the timing differences between modalities are striking: the facial expressions module's input stimuli are much more numerous than those of the text and "emoticons", making the latter's emotional paths look more linear. Another noteworthy aspect is that the facial expression module's emotional path calculation is interrupted during several seconds (14s approximately) due to the appearance of a short facial occlusion during the user's emotional display, causing the tracking program to temporarily lose the facial features.

Fig. 6 presents the continuous emotional path obtained when applying the methodology to fuse the 3 modules, both without (a) and with (b) the "emotional kinematics" filtering step. As can be seen, the complexity of the user's affective progress is shown in a simple and efficient way. Different modalities complement each other to obtain a more reliable result. Although the interruption period of the emotional path calculation is considerably reduced with respect to the facial expressions module's individual case (from 14s to 6s approximately), it still exists since both the text and "emoticons" modules' decay process reaches the threshold ε before the end of the facial occlusion, causing the $\alpha_1(t)$ and $\alpha_2(t)$ weights to be null. Thanks to the use of the "emotional kinematics" filtering technique, the path is smoothed and the aforementioned temporal input information absence is solved by letting the filter prediction output be taken as the 2D point position for those samples.

IV. Conclusions and Future Work

This paper describes an original and scalable methodology for fusing multiple affective recognition modules. This methodology is able to fuse any number of unimodal categorical modules, with very different time-scales and output labels. This is possible thanks to the use of a 2-dimensional evaluation-activation description of affect that provides the system with mathematical capabilities to deal with temporal emotional issues. The key step from a discrete perspective of affect to a continuous emotional space is achieved by using the Whissell dictionary, that allows the mapping of any emotional label to a 2D point in the activation-evaluation space. The decision-level fusion allows the use of different recognition modules which could be integrated in our application as plugins, developed independently from different researchers and easily tuned together to maximize the recognition rate. The proposed methodology outputs a 2D emotional path that represents in a novel and efficient way the user's detected emotional progress over time. A Kalman filtering technique controls the emotional path in real-time through an "emotional kinematics" model to ensure temporal consistency and robustness. The methodology has been put into practice in the context of Instant Messaging by fusing 3 different affect sensing modalities (text, facial expressions and "emoticons"). The first experimental results are promising and the potential of the proposed methodology has been demonstrated. This work brings a new perspective and invites further discussion on the still open issue of multimodal affective fusion.

In general, evaluation issues are largely solved for categorical affect recognition approaches. Unimodal categorical modules can be exhaustively evaluated thanks to the use of large well-annotated databases and well-known measures and methodologies (such as percentage of correctly classified instances, cross-validation, etc.). The evaluation of the performance of dimensional approaches is, however, an open and difficult issue to be solved. In the future, our work is expected to focus in depth on evaluation issues applicable to dimensional approaches and multimodality. The proposed fusion methodology will be explored in different application contexts, with different numbers and natures of modalities to be fused.

Acknowledgment

This work has been partly financed by the Spanish DGICYT contract number TIN2007-63025, by the Government of Aragon: IAF N2008/0574 and CyT N2008/0486 agreements.

References

[1] S. Gilroy, M. Cavazza, M. Niranen, E. Andr, T. Vogt, J. Urbain, M. Benayoun, H. Seichter, and M. Billinghurst, "PAD-based multimodal affective fusion," in *Proceedings of the Conference on Affective Computing and Intelligent Interaction*, 2009, pp. 1–8.

[2] Z. Zeng, M. Pantic, and T. S. Huang, "Emotion recognition based on multimodal information," in *Affective Information Processing*, J. Tao and T. Tan, Eds. Springer London, 2009, pp. 241–265.

[3] A. Kapoor, W. Burleson, and R. Picard, "Automatic prediction of frustration," *International Journal of Human-Computer Studies*, vol. 65, no. 8, pp. 724–736, 2007.

[4] H. Gunes, M. Piccardi, and M. Pantic, "From the lab to the real world: Affect recognition using multiple cues and modalities," *Affective Computing: Focus on Emotion Expression, Synthesis, and Recognition*, pp. 185–218, 2008.

[5] C. Shan, S. Gong, and P. McOwan, "Beyond facial expressions: Learning human emotion from body gestures," in *Proceedings of the British Machine Vision Conference*, 2007.

[6] T. Pun, T. Alecu, G. Chanel, J. Kronegg, and S. Voloshynovskiy, "Brain-computer interaction research at the computer vision and multimedia laboratory, University of Geneva," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 14, no. 2, pp. 210–213, 2006.

[7] L. Kuncheva, *Combining pattern classifiers: methods and algorithms*. Wiley-Interscience, 2004.

[8] Z. Zeng, J. Tu, M. Liu, T. Huang, B. Pianfetti, D. Roth, and S. Levinson, "Audio-visual affect recognition," *IEEE Transactions on Multimedia*, vol. 9, no. 2, pp. 424–428, 2007.

[9] H. Gunes and M. Piccardi, "Bi-modal emotion recognition from expressive face and body gestures," *Journal of Network and Computer Applications*, vol. 30, no. 4, pp. 1334–1345, 2007.

[10] P. Pal, A. Iyer, and R. Yantorno, "Emotion detection from infant facial expressions and cries," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 2, 2006, pp. 721–724.

[11] M. Wöllmer, M. Al-Hames, F. Eyben, B. Schuller, and G. Rigoll, "A multidimensional dynamic time warping algorithm for efficient multimodal fusion of asynchronous data streams," *Neurocomputing*, vol. 73, no. 1-3, pp. 366–380, 2009.

[12] P. Ekman, T. Dalgleish, and M. Power, *Handbook of Cognition and Emotion*. Wiley Online Library, 1999.

[13] G. Castellano, L. Kessous, and G. Caridakis, "Multimodal emotion recognition from expressive faces, body gestures and speech," in *Proceedings of the Doctoral Consortium of the 2nd International Conference on Affective Computing and Intelligent Interaction*, 2007.

[14] C. Whissell, *The Dictionary of Affect in Language, Emotion: Theory, Research and Experience*. Academic, 1989, vol. 4.

[15] R. Plutchik, *Emotion: a Psychoevolutionary Synthesis*. Harper & Row, 1980.

[16] F. Gosselin and P. Schyns, "Bubbles: A technique to reveal the use of information in recognition tasks," *Vision Research*, vol. 41, no. 17, pp. 2261–2271, 2001.

[17] A. Hanjalic and L. Xu, "Affective video content representation and modeling," *Multimedia, IEEE Transactions on*, vol. 7, no. 1, pp. 143–154, 2005.

[18] N. Stoiber, R. Seguier, and G. Breton, "Automatic design of a control interface for a synthetic face," in *Proceedings of the 13th international conference on Intelligent user interfaces*. ACM, 2009, pp. 207–216.

[19] E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. McRorie, J. Martin, L. Devillers, S. Abrilian, A. Batliner *et al.*, "The HUMAINE database: Addressing the collection and annotation of naturalistic and induced emotional data," in *Proceedings of the 2nd International Conference on Affective Computing and Intelligent Interaction*, 2007, pp. 488–500.

[20] M. Grimm, K. Kroschel, and S. Narayanan, "The Vera am Mittag German audio-visual emotional speech database," in *Proceedings of the IEEE International Conference on Multimedia and Expo*, 2008, pp. 865–868.

[21] N. Fragopanagos and J. Taylor, "Emotion recognition in human-computer interaction," *Neural Networks*, vol. 18, no. 4, pp. 389–405, 2005.

[22] S. Petridis, H. Gunes, S. Kaltwang, and M. Pantic, "Static vs. dynamic modeling of human nonverbal behavior from multiple cues and modalities," in *Proceedings of the International Conference on Multimodal Interfaces*, 2009, pp. 23–30.

[23] R. Picard, *Affective Computing*. The MIT Press, 1997.

[24] D. Morrell and W. Stirling, "An extended set-valued kalman filter," in *Proceedings of ISIPTA*, 2003, pp. 396–407.

[25] "Skype official website," Accessed September 2010, http://www.skype.com/.

[26] "Yahoo! messenger official website," Accessed September 2010, http://messenger.yahoo.com/.

[27] E. Cambria, A. Hussain, C. Havasi, and C. Eckl, "Sentic computing: Exploitation of common sense for the development of emotion-sensitive systems," *Development of Multimodal Interfaces: Active Listening and Synchrony*, vol. 5967, pp. 153–161, 2010.

[28] ——, "AffectiveSpace: Blending Common Sense and Affective Knowledge to Perform Emotive Reasoning," *WOMSA at CAEPIA, Seville*, 2009.

[29] J. Sánchez, N. Hernández, J. Penagos, and Y. Ostróvskaya, "Conveying mood and emotion in instant messaging by using a two-dimensional model for affective states," in *Proceedings of VII Brazilian symposium on Human factors in computing systems*. ACM, 2006, pp. 66–72.

[30] A. Wolf, "Emotional expression online: Gender differences in emoticon use," *CyberPsychology & Behavior*, vol. 3, no. 5, pp. 827–833, 2000.

[31] A. Kring and A. Gordon, "Sex differences in emotion: Expression, experience, and physiology," *Journal of Persnaliw and Saial Psychology*, vol. 1, no. 74, pp. 3–686, 1998.

[32] S. Kayan, S. Fussell, and L. Setlock, "Cultural differences in the use of instant messaging in Asia and North America," in *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work*. ACM, 2006, pp. 525–528.

[33] I. Hupont, S. Baldassarri, R. Del-Hoyo, and E. Cerezo, "Effective emotional classification combining facial classifiers and user assessment," *Articulated Motion and Deformable Objects*, vol. 5098, pp. 431–440, 2008.