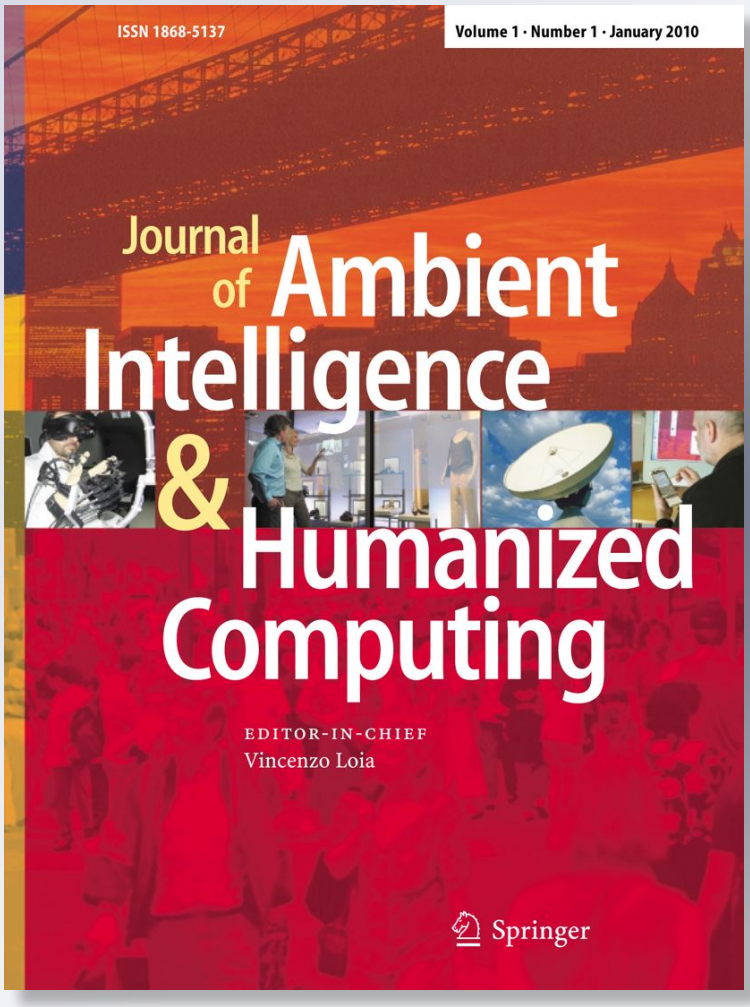
Emotional facial sensing and multimodal fusion in a continuous 2D affective space

Eva Cerezo, Isabelle Hupont, Sandra Baldassarri & Sergio Ballano

Journal of Ambient Intelligence and Humanized Computing

ISSN 1868-5137
Volume 3
Number 1

J Ambient Intell Human Comput (2012)
3:31-46
DOI 10.1007/s12652-011-0087-6



Springer

J Ambient Intell Human Comput (2012) 3:31–46
DOI 10.1007/s12652-011-0087-6

ORIGINAL RESEARCH

Emotional facial sensing and multimodal fusion in a continuous 2D affective space

Eva Cerezo · Isabelle Hupont · Sandra Baldassarri · Sergio Ballano

Received: 3 February 2011 / Accepted: 24 September 2011 / Published online: 30 October 2011

Abstract This paper deals with two main research focuses on Affective Computing: facial emotion recognition and multimodal fusion of affective information coming from different channels. The facial sensing system developed implements an emotional classification mechanism that combines, in a novel and robust manner, the five most commonly used classifiers in the field of affect sensing, obtaining at the output an associated weight of the facial expression to each of the six Ekman's universal emotional categories plus the neutral. The system is able to analyze any subject, male or female, of any age, and ethnicity and has been validated by means of statistical evaluation strategies, such as cross-validation, classification accuracy ratios and confusion matrices. The categorical facial sensing system has been subsequently expanded to a continuous 2D affective space which has made it also possible to face the problem of multimodal human affect recognition. A novel fusion methodology able to fuse any number of affective modules, with very different timescales and output labels, is proposed. It relies on the 2D Whissell affective space and is able to output a continuous emotional path characterizing the user's affective progress over time. A Kalman filtering technique controls this path in real-time to ensure temporal consistency and robustness to the system. Moreover, the methodology is adaptive to eventual temporal changes in the reliability of the different inputs' quality. The potential of the multimodal fusion methodology is demonstrated by fusing dynamic affective information extracted from different channels (video, typed-in text and emoticons) of an Instant Messaging tool.

Keywords Affective Computing · Kansei (sense/emotion) engineering · Human factors · Facial expression analysis · Multimodal fusion

E. Cerezo (✉) · S. Baldassarri
Dpto. Informática e Ingeniería de Sistemas, Universidad de Zaragoza, C/Maria de Luna 1, 50018 Zaragoza, Spain
e-mail: ecerezo@unizar.es

I. Hupont · S. Ballano
Instituto Tecnológico de Aragón (ITA), P.T. Walqa - Edificio I+D+i, Ctra. Zaragoza, N-330a, Km 566, 22197 Cuarte (Huesca), Spain

1 Introduction

Human computer intelligent interaction is an emerging field aimed at providing natural ways for humans to use computers as aids. It is argued that for a computer to be able to interact with humans it needs to have the communication skills of humans. One of these skills is the affective aspect of communication (Boukricha et al. 2007). For this reason, affect sensing is becoming an indispensable part of advanced human–computer interfaces. This paper deals with two main research focuses on Affective Computing: emotion recognition from the user's facial expressions and multimodal fusion of affective information extracted from different human communicative channels. A review of the state of the art in both issues follows.

1.1 Emotional facial recognition

The most expressive manner humans display emotions is through facial expressions. Facial expression is the most powerful, natural and direct way used by humans to communicate and understand each other's affective state and intentions (Keltner and Ekman 2000). 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 (HCI) designs.

The most long-standing way that facial affect has been described by psychologists is in terms of discrete categories, an approach that is rooted in the language of daily life. Facial expressions are often evaluated by classifying face images into the six universal emotions proposed by Ekman (1999) which include "happiness", "sadness", "fear", "anger", "disgust" and "surprise". Examples of studies using this categorization are those of Hammal et al. (2005) and Littlewort et al. (2006). The labeling scheme based on category is very intuitive and thus matches peoples' experience. This *categorical* approach, where emotions are a mere list of labels, fails however to describe the wide range of emotions that occur in daily communication settings and intrinsically ignore the intensity of an emotion. In this case, a small variation on face due to emotion may still be regarded as "neutral" face. There are a few tentative efforts to detect non-basic affective states from deliberately displayed facial expressions, including "fatigue" (Ji et al. 2006), and mental states such as "agreeing", "concentrating", "interested", "thinking", "confused", and "frustrated" (Kapoor et al. 2007; Yeasin et al. 2006). In any case, *categorical* approach presents a discrete list of emotions with no real link between them. It does not represent a *dimensional* space and has no algebra: every emotion must be studied and recognized independently.

To overcome the problems cited above, some researchers, such as Whissell (1989) and Plutchik (1980), prefer to view affective states not independent of one another 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. Besides *categorical* approach, *dimensional* approach is attractive because it provides a way of describing a wide range of emotional states and measuring the intensity of emotion. It is much more able to deal with non-discrete emotions and 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. However, in comparison with category-based description of affect, very few works have chosen a dimensional description level, and the few that do are more related to the design of synthetic faces (Stoiber et al. 2009), data processing (Du et al. 2007) or psychological studies (Gosselin and Schyns 2001) than to emotion recognition. Moreover, in existing affective recognition works the problem is simplified to a two-class (positive vs. negative and active vs. passive) (Fragopanagos and Taylor 2005) or a four class (quadrants of 2D space) classification (Garidakis et al. 2006), thereby losing the descriptive potential of 2D space. Apart from seeking effective features that reflect affective factors, the main difficulty comes from the labeling of ground-truth data since there is no available public facial expression database that provides emotional annotations in terms of evaluation and activation dimensions.

Independently of the description level chosen to classify emotions (*categorical* or *dimensional*), a classification mechanism must be established to categorize the facial posture shown in terms of the defined description level. In the literature, the facial expression analyzers that obtain the best success rates for emotional classification make use of neural networks, rule-based expert systems, Support Vector Machines or Bayesian nets based classifiers. In (Zeng et al. 2009b), an excellent state-of-the art summary is given of the various methods recently used in facial expression emotional recognition. However, the majority of those studies confine themselves to select only one type of classifier for emotional detection, or at the most compare different classifiers and then use that which provides the best results (Littlewort et al. 2006).

In this paper, an effective system for sensing facial emotions in a continuous 2D affective space is described. Its inputs are a set of carefully selected facial distances and angles that modelize the face in a simple way but without losing relevant facial expression information. The system starts with a classification method in discrete categories that is subsequently expanded in order to be able to work in a continuous emotional space and thus to consider intermediate emotional states. As regards the classification mechanism itself, the system intelligently combines the outputs of different classifiers simultaneously. In this way, the overall risk of making a poor selection with a given classifier for a given facial input is considerably reduced. The system is capable of analyzing any subject, male or female of any age and ethnicity, and has been validated considering human assessment.

1.2 Multimodal affect fusion

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 (Gilroy et al. 2009; Zeng et al. 2009a, b; Kapoor et al. 2007) 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 (Gilroy et al. 2009). Multimodal fusion of different affective channels is still in its initial stage and far from being

solved (Gunes et al. 2008). 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 (Kapoor et al. 2007; Shan et al. 2007; Pun et al. 2006), 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 modelled 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 (Kuncheva 2004) and the fact that time and feature dependence are abstracted. Various, mainly bimodal, decision-level fusion methods have been proposed in the literature (Zeng et al. 2007; Gunes and Piccardi 2007; Pal et al. 2006), but optimal fusion designs are still undefined. Most available multimodal recognizers have designed ad hoc 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. For example in (Wöllmer et al. 2009) 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.

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, compared to category-based description of affect, very few works have chosen a dimensional description level. 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 (Douglas-Cowie et al. 2007; Grimm et al. 2008), 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 labelling variation between different human raters is reported when working with the dimensional approach (Fragopanagos and Taylor 2005). 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.

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 (IM) tool able to feed three 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. 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: Sect. 2 describes the categorical facial classification method. In Sect. 3 the step from the discrete perspective to the continuous emotional space is explained in detail and Sect. 4 proposes the novel fusion methodology, which is put into practice in the application presented in Sect. 5. Section 6 comprises the conclusions and future work.

2 A novel method for facial discrete emotional classification

In this section, an effective method is presented for the automatic classification of facial expressions into discrete emotional categories. The method is able to classify the user's emotion in terms of the six Ekman's universal emotions (plus "neutral"), giving a confidence value to each emotional category. Section 2.1 explains the selection and extraction process of the features serving as inputs to the system. Section 2.2 describes the criteria taken into account when selecting the various classifiers and how they are combined. Finally, the obtained results are presented in Sect. 2.3.

2.1 Selection and extraction of facial inputs

Facial Action Coding System (FACS) (Ekman et al. 2002) was developed by Ekman and Friesen to code facial expressions in which the individual muscular movements in the face are described by Action Units (AUs). This work inspired many researchers to analyze facial expressions by means of image and video processing, where by tracking of facial features and measuring a set of facial distances and angles, they attempt to classify different facial expressions. In particular, existing works demonstrate that high emotional classification accuracy can be obtained by analyzing a small set of facial distances and angles. Examples are the work of Soyel and Demirel (2007) that studies six 3D facial distances; the method proposed by Hammal et al. (2005) that analyzes a set of five 2D facial distances; or the approach of Chang et al. (2009), that measures 12 feature distances.

Following that methodology, the initial inputs of our classifiers were established in a set of distances and angles obtained from 20 characteristic facial points (Hupont et al. 2008). In fact, the inputs are the variations of these angles and distances with respect to the "neutral" face. The chosen set of initial inputs compiles the distances and angles that have been proved to provide the best classification performance in existing works of the literature, such as the aforementioned. The points are obtained thanks to faceAPI, a commercial real-time facial feature tracking program that provides cartesian facial 3D coordinates. It is able to track up to ±90° of head rotation and is robust to occlusions, lighting conditions, presence of beard, glasses, etc. The initial set of parameters tested is shown in Fig. 1. In order to make the distance values consistent (independently of the scale of the image, the distance to the camera, etc.) and independent of the expression, all the distances are normalized with respect to the distance between the eyes. The choice of angles provides a size invariant classification and saves the effort of normalization.

In order to determine the goodness and usefulness of the parameters, a study of the correlation between them was carried out using the data (distance and angle values) obtained from a set of training images. For this purpose, two different facial emotion databases were used: the FGNET database (Wallhoff 2006) that provides spontaneous (non-acted) video sequences of 19 different young Caucasian people, and the MMI facial expression database (Pantic et al. 2005) that holds 1,280 acted videos of 43 different subjects from different races (Caucasian, Asian, South American and Arabic) and ages ranging from 19 to 62. Both databases show Ekman's six universal emotions plus the "neutral" one and provide expert annotations about the emotional apex frame of the video sequences. A new database has been built for this work with a total of 1,500 static frames selected from the apex of the video sequences from the FG-NET and MMI databases. It has been used as a training set in the correlation study and in the tuning of the classifiers.

A correlation-based feature selection technique (Hall 1998) was carried out in order to identify the most influential parameters in the variable to predict (emotion) as well as to detect redundant and/or irrelevant features. Subsets of parameters that are highly correlated with the class while having low intercorrelation are preferred. In that way, from the initial set of parameters only the most significant ones were selected to work with: RD1, RD2,

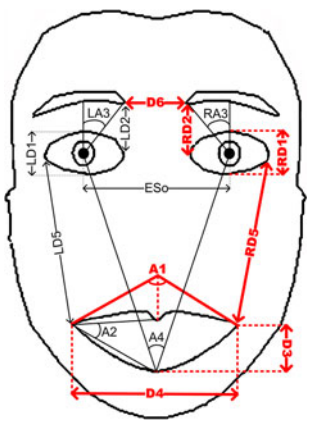


Fig. 1 Facial parameters tested (in *bold*, the final selected parameters)

RD5, D3, D4, D6 and A1 (marked in bold in Fig. 1). This reduces the number of irrelevant, redundant and noisy inputs in the model and thus computational time, without losing relevant facial information.

2.2 Classifiers selection and novel combination

In order to select the best classifiers, the Waikado Environment for Knowledge Analysis (Weka) tool was used (Witten and Frank 2005). It provides a collection of machine learning algorithms for data mining tasks. From this collection, five classifiers were selected after tuning and benchmarking: RIPPER, Multilayer Perceptron, SVM, Naive Bayes and C4.5. The selection was based on their widespread use as well as on the individual performance of their Weka implementation.

A tenfold cross-validation test over the 1,500 training images has been performed for each selected classifier. The success rates obtained for each classifier and each emotion are shown in the first five rows of Table 1. As can be observed, each classifier is very reliable for detecting certain specific emotions but not so much for others. For example, the C4.5 is excellent at identifying "joy" (92.90% correct) but is only able to correctly detect "fear" on 59.30% of occasions, whereas Naive Bayes is way above the other classifiers for "fear" (85.20%), but is below the others in detecting "joy" (85.70%) or "surprise" (71.10%). Therefore, an intelligent combination of the five classifiers in such a way that the strong and weak points of each are taken into account appears as a good solution for developing a method with a high success rate.

The classifier combination chosen follows a weighted majority voting strategy. The voted weights are assigned depending on the performance of each classifier for each emotion. From each classifier, a confusion matrix formed by elements $P_{jk}(E_i)$, corresponding to the probability of having emotion i knowing that classifier j has detected emotion k, is obtained. The probability assigned to each emotion $P(E_i)$ is calculated as:

$$P(E_i) = \frac{P_{1k'}(E_i) + P_{2k''}(E_i) + \cdots + P_{5k^v}(E_i)}{5} \quad (1)$$

where, k', $k''\ldots k^v$ are the emotions detected by classifiers 1, 2,…5, respectively.

1. Firstly, the confidence value $CV(E_i)$ is obtained by normalizing each $P(E_i)$ to a 0 through 1 scale:

$$CV(E_i) = \frac{P(E_i) - \min\{P(E_i)\}}{\max\{P(E_i)\} - \min\{P(E_i)\}} \quad (2)$$

   where,

   - $\min\{P(E_i)\}$ is the greatest $P(E_i)$ that can be obtained by combining the different $P_{jk}(E_i)$ verifying that $k \neq i$ for every classifier j. In other words, it is the highest probability that a given emotion can reach without ever being selected by any classifier.
   - $\max\{P(E_i)\}$ is that obtained when combining the $P_{jk}(E_i)$ verifying that $k = i$ for every classifier j. In other words, it is the probability that obtains a given emotion when selected by all the classifiers unanimously.

2. Secondly, a rule is established over the obtained confidence values in order to detect and eliminate emotional incompatibilities. The rule is based on the work of Plutchik (1980), who 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". The rule to apply is the following: if emotional incompatibility is detected, i.e. two non-null incompatible emotions exist simultaneously, that chosen will be the one with the closer emotional orientation to the rest of the non-null detected emotions. For example, if "joy", "sadness" and "disgust" coexist, "joy" is assigned zero since "disgust" and "sadness" are emotionally closer according to Plutchik.

2.3 Results

The results obtained when applying the strategy explained in the previous section to combine the scores of the five

Table 1 Success rates obtained with a tenfold cross-validation test over the 1,500 training images for each individual classifier and emotion (first five rows) and when combining the five classifiers (sixth row in bold)

	Disgust (%)	Joy (%)	Anger (%)	Fear (%)	Sadness (%)	Neutral (%)	Surprise (%)
RIPPER	50.00	85.70	66.70	48.10	26.70	80.00	80.00
SVM	76.50	92.90	55.60	59.30	40.00	84.00	82.20
C4.5	58.80	92.90	66.70	59.30	30.00	70.00	73.30
Naive Bayes	76.50	85.70	63.00	85.20	33.00	86.00	71.10
Multilayer Perceptron	64.70	92.90	70.40	63.00	43.30	86.00	77.80
Combination of classifiers	**94.12**	**97.62**	**81.48**	**85.19**	**66.67**	**94.00**	**95.56**

classifiers with a tenfold cross-validation test are shown in sixth row of Table 1. As can be observed, the success rates for the "neutral", "joy", "disgust", "surprise", "disgust" and "fear" emotions are very high (81.48–97.62%). The lowest result of our classification is for "sadness", which is confused with the "neutral" emotion on 20% of occasions, due to the similarity of their facial expressions. Confusion between this pair of emotions occurs frequently in the literature and for this reason many works do not consider "sadness". Nevertheless, the 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. Table 2 shows the confusion matrix obtained after the combination of the five classifiers.

3 From a discrete perspective to a 2D continuous affective space

As discussed in the introduction, the use of a discrete set of emotions (labels) for emotional classification has important limitations. To avoid these limitations and enrich the emotional output information from the system in terms of intermediate emotions, use has been made of one of the most influential evaluation–activation 2D models in the field of psychology: that proposed by Whissell (1989). The methodology, that will be explained in Sect. 3.1, starts from the confidence values associated to each of Ekman's emotions obtained by the discrete emotional classification, and calculates the (x,y) coordinates in the Whissell space of the analyzed facial expression (see Fig. 2). The results of the 2D emotional mapping are analyzed in detail taking human assessment into account in Sects. 3.2 (with database images) and 3.3 (with images obtained in uncontrolled environments).

3.1 Emotional mapping to a continuous affective space

In her study, Whissell assigns a pair of values ⟨evaluation, activation⟩ to each of the approximately 9,000 carefully selected affective words that make up her "Dictionary of Affect in Language" (Whissell 1989). Figure 3 shows the position of some of these words in the evaluation–activation space. The idea 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.

It can be seen that the emotion-related words corresponding to each one of Ekman's six emotions have a specific location (x_i, y_i) in the Whissell space (in bold in Fig. 3). Thanks to this, the output information of the classifiers (confidence value of the facial expression to each emotional category) can be mapped in the space. This emotional mapping is carried out considering each of Ekman's six basic emotions plus "neutral" as 2D 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 (x,y) coordinates of a given image are calculated as the centre of mass of the seven weighted points in the Whissell space following:

$$x = \frac{\sum_{i=1}^{7} x_i CV(E_i)}{\sum_{i=1}^{7} CV(E_i)} \quad \text{and} \quad y = \frac{\sum_{i=1}^{7} y_i CV(E_i)}{\sum_{i=1}^{7} CV(E_i)} \tag{3}$$

In this way the output of the system is enriched with a larger number of intermediate emotional states.

3.2 Evaluation of results taking human assessment into account

The method described in the previous section has been put into practice with the outputs of the classification system when applied to the database facial expressions images. In Fig. 4 the general location of all classified images is plotted (markers size is proportional to the percentage of images

Table 2 Confusion matrix obtained combining the five classifiers

	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**

Confusion matrix obtained combining the five classifiers (in bold, succes rates of each emotion coming from the sixth row in Table 1)

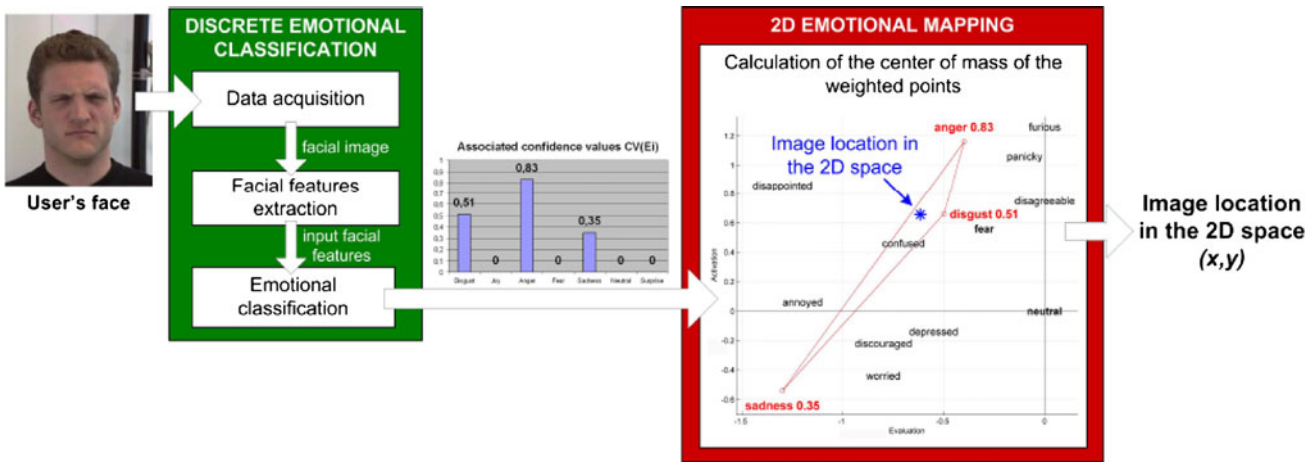


Fig. 2 Overall block diagram for obtaining the location of a facial image in the 2D emotional space. A graphic illustration of the 2D emotional mapping process is included as an example

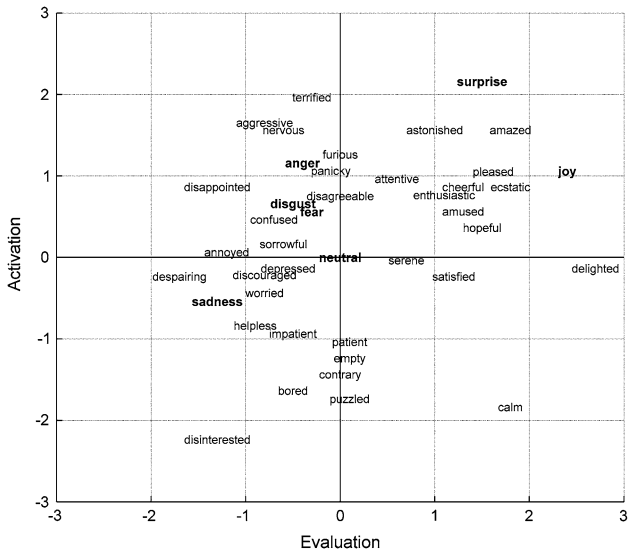


Fig. 3 Simplified Whissell's evaluation–activation space

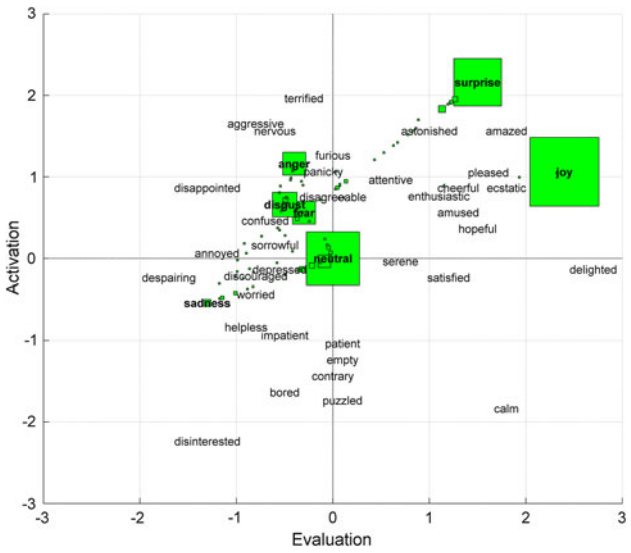


Fig. 4 Location of the different images of the database in the Whissell space, according to the method explained in Sect. 3.1 (marker size is proportional to the percentage of images situated at the same location)

situated at the same location). Figure 5 shows several images with their nearest label in the Whissell space.

The database used in this work provides images labeled with one of the six Ekman universal emotions plus "neutral", but there is no a priori known information about their location in the Whissell 2D space. In order to evaluate the system results, there is a need to establish the region in the Whissell space where each image can be considered to be correctly located. For this purpose, a total of 43 persons participated in one or more evaluation sessions (50 images per session). In the sessions they were told to locate a set of images of the database in the Whissell space (as shown in Fig. 3, with some reference labels). As result, each one of the frames was located in terms of evaluation–activation by 16 different persons.

The collected evaluation data have been used to define an ellipsoidal 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). MVE looks for the ellipsoid with the smallest volume that covers a set of data points. Although there are several ways to compute the shape of a set of data points (e.g. using a convex hull, rectangle, etc.), the MVE was chosen because of the fact that real-world data often exhibits a mixture of Gaussian distributions, which have equi-density contours in the shape of ellipsoids. First, the collected data are filtered in order to remove outliers: a

Fig. 5 Examples of images from the database with their nearest label in the Whissell space, according to the method described in Sect. 3.1

point is considered an outlier if its coordinate values (in both dimensions) are greater than the mean plus three times the standard deviation. Then, the MVE is calculated following the algorithm described by Kumar and Yildirim (2005). The MVEs obtained are used for evaluating results at four different levels:

1. *Ellipse criteria.* If the point detected by the system (2D coordinates in the Whissell space) is inside the defined 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.

Table 3 Results obtained according to different evaluation criteria

	Ellipse criteria (%)	Quadrant criteria (%)	Evaluation axis criteria (%)	Activation axis criteria (%)
Success rate	73.73	87.45	94.12	92.94

The results obtained following the different evaluation strategies are presented in Table 3.

As can be seen, the success rate is 73.73% in the most restrictive case, i.e. when the output of the system is considered to be correctly located when inside the ellipse. It rises to 94.12% when considering the evaluation axis criteria. Objectively speaking, these results are very good, especially when, according to Bassili (1979), a trained observer can correctly classify facial emotions with an average of 87%. However, they are difficult to compare with other emotional classification studies that can be found in literature, given that either such studies do not recognize emotions in evaluation–activation terms, or they have not been tested under common experimental conditions (e.g. different databases or evaluation strategies are used).

3.3 Evaluation of real video sequences

In order to demonstrate the potential of the proposed classification method it has been tested with a set of emotionally complex video sequences, recorded in a natural (unsupervised) setting. These videos are complex owing to two 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.

Fifteen videos from three different users were tested (Fig. 6), ranging from 20 to 70 s from which a total of 127

Fig. 6 Examples of images from the video sequences taken in uncontrolled environments

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, as explained in previous subsection. The results obtained following the different evaluation strategies are presented in Table 4. 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.

4 Expansion to multimodality: a scalable methodology for sensing emotions from multiple channels

In this section the use of the bidimensional Whissell affective space is expanded to go beyond unimodal facial affect sensing, in order to define a general methodology for fusing the responses of multiple emotional recognition modules. The modules to be fused can be of very different natures, exploring different modalities, time scales, metric levels, etc. The proposed methodology is able to fuse the different sources of affective information over time and to obtain as final output a global 2D dynamic emotional path in the activation-evaluation space representing the user's affective progress. Moreover, it is scalable enough to add new modules coming from new channels without having to retrain the whole system.

Similar to the facial emotions recognition module presented in Sect. 2, every module i to be fused is assumed to output 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.

Figure 7 shows the general fusion scheme that will be explained step by step in next sections. Since the proposed methodology can combine any number of modules covering different modalities and its overall performance is highly dependent on the accuracy of the individual modules to fuse, in this section the methodology is presented

Table 4 Results obtained in an uncontrolled environment according to different evaluation criteria

	Ellipse criteria (%)	Quadrant criteria (%)	Evaluation axis criteria (%)	Activation axis criteria (%)
Success rate	61.90	74.60	79.37	84.92

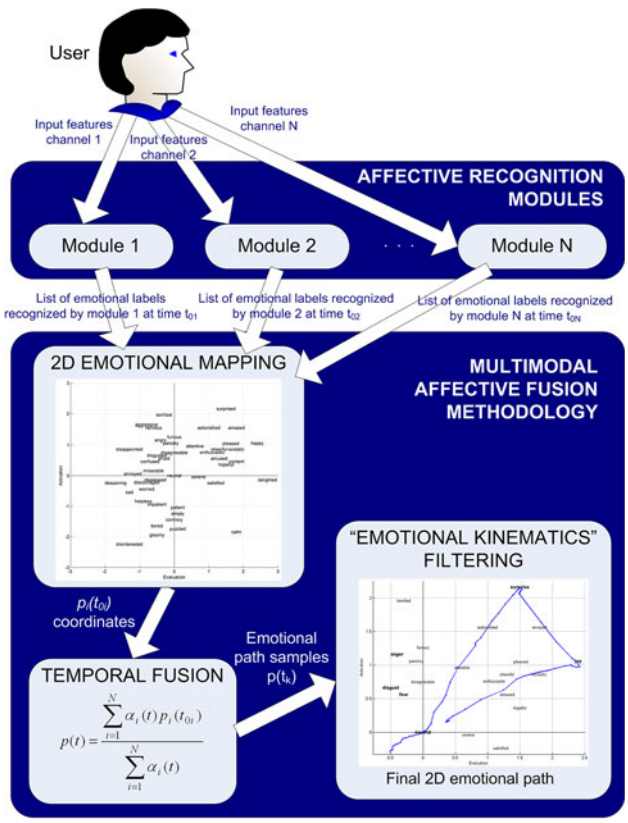


Fig. 7 Continuous multimodal affective fusion methodology

from a theoretical point of view exclusively. However, in order to show its potential and usefulness, it will be applied in Sect. 5 to a real IM interaction context by fusing the information coming from different affective channels.

4.1 Emotional mapping to a continuous 2D affective space

The first step of the proposed methodology expands the idea of mapping the output of the facial classification method to the evaluation–activation space to any categorical module i. The idea is to build an emotional mapping to the Whissell space so that the output of each module i at a given time t_{0i} can be represented as a two-dimensional coordinates point $p_i(t_{0i}) = (x_i(t_{0i}); y_i(t_{0i}))$ that characterizes the affective properties extracted from that module at time t_{0i}. The majority of the categorical modules described in the literature provide as output a list of emotional labels with some associated weights at the time t_{0i} corresponding to the detection of the affective stimulus. Since the Whissell dictionary (Whissell 1989) if composed of more than 9,000 affective words, whatever the labels used, each one has an associated 2D point in the Whissell space. Following the method explained in Sect. 3.1, the components $(x_i(t_{0i}); y_i(t_{0i}))$ of $p_i(t_{0i})$ are calculated.

4.2 Temporal fusion of individual modules: obtaining a continuous 2D emotional path

Humans inherently display emotions following a continuous temporal pattern (Petridis et al. 2009). 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:

$$p(t) = \frac{\sum_{i=1}^{N} \alpha_i(t) p_i(t_{0i})}{\sum_{i=1}^{N} \alpha_i(t)} \tag{4}$$

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 used affective response is the sum of each modality's contribution $p_i(t_{0i})$ modulated by the $\alpha_i(t)$ coefficients 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 (1997), 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:

$$\alpha_i(t) = \begin{cases} b_i c_i(t_{0i}) e^{-d_i(t-t_0)} & t > \varepsilon \\ 0 & t \leq \varepsilon \end{cases} \tag{5}$$

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 (4) and (5), 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).

4.3 "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:

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

$$\begin{bmatrix} x \\ y \\ v_x \\ v_y \end{bmatrix}_{k+1} = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ v_x \\ v_y \end{bmatrix}_k + w_k \quad (6)$$

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 (Morrell and Stirling 2003), its covariance matrix Q_k is defined as:

$$Q_k = \sigma^2 \begin{bmatrix} \frac{1}{3} & 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{3} & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & 1 & 0 \\ 0 & \frac{1}{2} & 0 & 1 \end{bmatrix} \quad (7)$$

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

$$Y_k = H_k X_k + z_k$$

$$\begin{bmatrix} x_m \\ y_m \end{bmatrix}_k = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}_k \begin{bmatrix} x \\ y \\ v_x \\ v_y \end{bmatrix}_k + z_k \quad (8)$$

where Y_k is the observable 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 wk. Its covariance matrix R_k is the identity matrix:

$$R_k = \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix} \quad (9)$$

so that it is assumed that the x and y measurements contain independent errors with λ 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.
The methodology presented in this section has been put into practice in the context of an IM tool that will be presented in next section.

5 Multimodal fusion application to Instant Messaging

Instant Messaging 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 Sect. 4 through the use of an IM tool that combines different communication modalities (text, video and "emoticons"), each one with very different time scales. Section 5.1 describes the IM tool and presents the modules that extract emotional information from each modality. Section 5.2 explains how the methodology has been tuned to achieve multimodal affective fusion. Finally, Sect. 5.3 presents the experimental results obtained when applying the fusion methodology to an IM emotional conversation.

5.1 Instant Messaging tool and fusion modalities

Although any publicly available IM tool could be used a simple ad-hoc IM tool has been designed. It allows two persons to communicate via text, live video and "emoticons". Figure 8 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).

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 (Cambria et al. 2010) 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". Figure 9 shows the 16 available "emoticons" and their corresponding labels, designed to be a good representation of each affective state (Sánchez et al. 2006). 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 (Wolf 2000) and the cultural differences impose the level of contextual information required for communication (Kayan et al. 2006). 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 reliable information, helping to solve complex

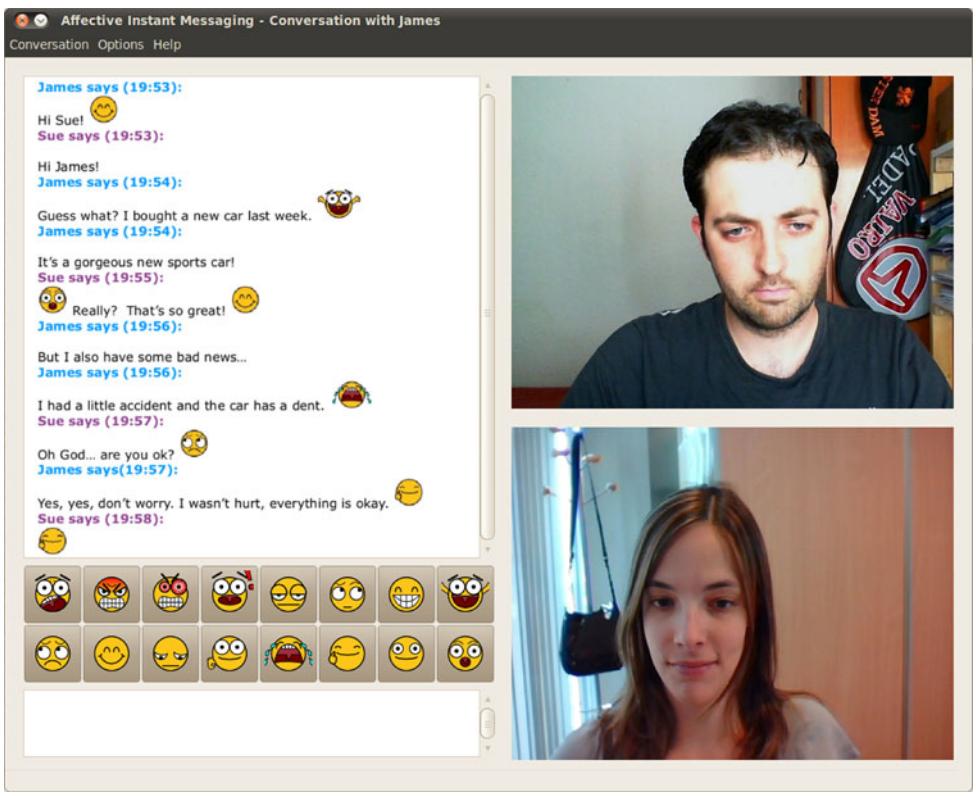


Fig. 8 Snapshot of the Instant Messaging tool during a conversation

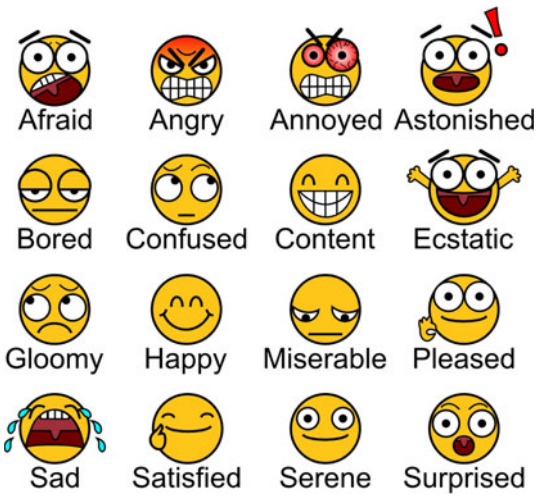


Fig. 9 "Emoticons" designed for the Instant Messaging tool and their corresponding emotional labels

emotional misunderstandings for example when sarcasm is present.

3. Module 3: facial expression analysis module, the one described in Sect. 2.

5.2 Multimodal fusion methodology tuning

This section describes, step by step, how the multimodal fusion methodology presented in Sect. 4 is tuned to fuse the three 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 (Sect. 4.1) 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 $p_3(t_{03})$ 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 $\Delta t = 1/f$. The next step towards achieving the temporal fusion of the different modules (Sect. 4.2) is assigning a value to the parameters that define the $\alpha_i(t)$ weights, namely b_i, $c_i(t_0i)$, d_i and ε. Table 5 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 d_i 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 (4) and (5).
3. Step 3: "Emotional Kinematics" filtering. Finally, the "emotional kinematics" filtering technique (Sect. 4.3) is iteratively applied in real-time each time a new sample is added to the computed emotional path. As in

Table 5 Temporal fusion parameters

# Module	1	2	3
Modality	Text	"emoticons"	Video
Total number of possible output labels	24 weighted emotional labels	16 emotional labels	six Ekman's universal labels (plus "neutral") + confidence value 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 confidence value is assigned since emoticons are the direct expression of user's affective state	$b_3 = 0.87$ the general confidence 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	$d_1 = d_2 = 0.035\ s^{-1}$ 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.		

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.

5.3 Experimental results

In order to demonstrate the potential of the presented fusion methodology, it has been applied to the IM conversation shown in Fig. 8 (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).

Figure 10 shows the emotional paths obtained when applying the methodology to each individual module separately (i.e. the 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 (14 s 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.

Figure 11 presents the continuous emotional path obtained when applying the methodology to fuse the three 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 14 to 6 s approximately), it still exists since both the text and "emoticons" modules' decay process reaches the threshold 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.

Regarding performance all the processing can be achieved in real time when using a multi-core processor. As an example, in the case of a video of 10 s, the classification process takes 4.21 ms/frame, text processing 2.96 ms/sentence and the fusion process 2.84 ms/frame: these times result in a total processing time of 10.01 ms/frame. Tracking is performed with FaceAPI which works in two ways: off-line (more precise) or on-line (needing a Dual Core at least) giving processing times of less than 33 ms.

6 Conclusions and future work

This paper first presents a facial affect recognizer able to sense emotions from a user's captured static facial image. Its inputs are a set of facial parameters, angles and distances between characteristic points of the face, chosen so that the face is modeled in a simple way without losing relevant facial expression information. The system implements an emotional classification mechanism that

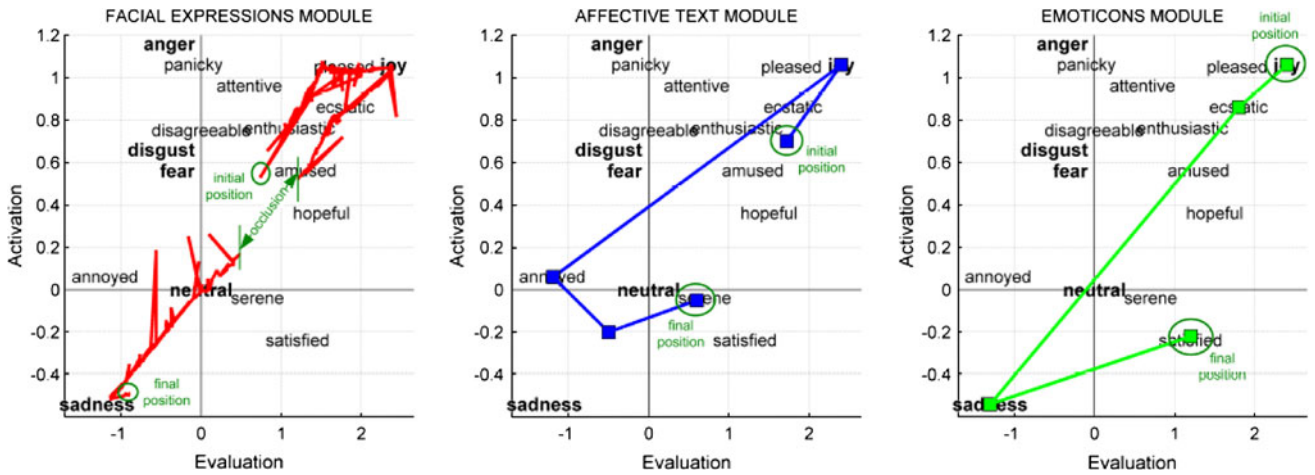


Fig. 10 Emotional paths obtained when applying the methodology to each individual module separately without "emotional kinematics" filtering. *Square* markers indicate the arrival time of an emotional stimulus (not shown for facial expression module for figure clarity reasons)

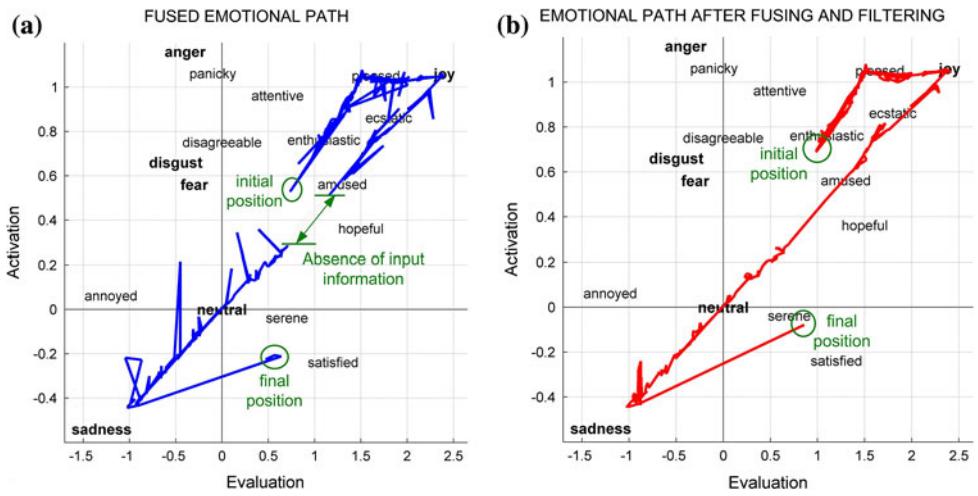


Fig. 11 Continuous emotional path obtained when applying the multimodal fusion methodology to James' Instant Messaging conversation shown in Fig. 6, without using "emotional kinematics" filtering (**a**), and using "emotional kinematics" filtering (**b**)

combines in a novel and robust manner the five most commonly used classifiers in the field of affect sensing, obtaining at the output an associated weight of the facial expression to each of the six Ekman's universal emotional categories plus neutral. It has been exhaustively validated by means of statistical evaluation strategies, such as cross-validation, classification accuracy ratios and confusion matrices and has been tested with an extensive database of 1,500 images showing individuals of different races and gender, giving universal results with very promising levels of correctness.

The expansion to dynamic and multimodal Affective Computing is achieved thanks to the use of a 2-dimensional description of affect that provides the system with mathematical capabilities to face temporal and multisensory emotional issues. A novel methodology is presented able to fuse any number of categorical modules, with very different time-scales and output labels. The proposed methodology outputs a 2D emotional path that represents the user's detected affective progress over time. A Kalman filtering technique controls this path in real-time to ensure temporal consistency and robustness to the system. Moreover, the methodology is adaptive to eventual temporal changes in the reliability of the different inputs' quality. The potential of the multimodal fusion methodology is demonstrated by fusing dynamic affective information extracted from the different channels of an IM tool. 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.

Acknowledgments The authors wish to thank Dr. Cynthia Whissell for her explanations and kindness, Dr. Hussain and E. Cambria for the text analyzing tool and all the participants in the evaluation sessions. This work has been partly financed by the University of Zaragoza through the AVIM project.

References

Bassili JN (1979) Emotion recognition: the role of facial movement and the relative importance of upper and lower areas of the face. J Pers Soc Psychol 37:2049–2058

Boukricha H, Becker C, Wachsmuth I (2007) Simulating empathy for the virtual human Max. In: Proceedings of 2nd International Workshop on Emotion and computing in conj. with the German Conference on Artificial Intelligence (KI2007), pp 22–27

Cambria E, Hussain A, Havasi C, Eckl C (2010) Sentic computing: exploitation of common sense for the development of emotion-sensitive systems. Development of Multimodal Interfaces: Active Listening and Synchrony 5967:153–161

Chang CY, Tsai JS, Wang CJ, Chung PC (2009) Emotion recognition with consideration of facial expression and physiological signals. In: Proceedings of the 6th Annual IEEE Conference on Computational intelligence in bioinformatics and computational biology, pp 278–283

Douglas-Cowie E, Cowie R, Sneddon I, Cox C, Lowry O, McRorie M, Martin J, Devillers L, Abrilian S, Batliner A et al (2007) 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, pp 488–500

Du Y, Bi W, Wang T, Zhang Y, Ai H (2007) Distributing expressional faces in 2-D emotional space. In: Proceedings of the 6th ACM International Conference on Image and video retrieval, pp 395–400

Ekman P (1999) In: Dalgleish T, Power M (eds) Handbook of cognition and emotion. Wiley, Chihester

Ekman P, Friesen WV, Hager JC (2002) Facial action coding system. Research Nexus eBook, Salt Lake City

Fragopanagos N, Taylor JG (2005) Emotion recognition in human–computer interaction. Neural Netw 18:389–405

Garidakis G, Malatesta L, Kessous L, Amir N, Paouzaiou A, Karpouzis K (2006) Modeling naturalistic affective states via facial and vocal expression recognition. Proc Int Conf on Multimodal Interfaces: 146–154

Gilroy S, Cavazza M, Niranen M, Andr E,Vogt T, Urbain J, Benayoun M, Seichter H, Billinghurst M (2009) PAD-based multimodal affective fusion. In: Proceedings of the Conference on Affective computing and intelligent interaction, pp 1–8

Gosselin F, Schyns PG (2001) Bubbles: a technique to reveal the use of information in recognition tasks. Vis Res 41:2261–2271

Grimm M, Kroschel K, Narayanan S (2008) The Vera am Mittag German audio-visual emotional speech database. In: Proceedings of the IEEE International Conference on multimedia and expo, pp 865–868

Gunes H, Piccardi M (2007) Bi-modal emotion recognition from expressive face and body gestures. J Netw Comp Appl 30(4):1334–1345

Gunes H, Piccardi M, Pantic M (2008) From the lab to the real world: affect recognition using multiple cues and modalities. In: Jimmy Or (ed) Affective computing, InTech, Vienna, pp 185–218

Hall MA (1998) Correlation-based feature selection for machine learning. Hamilton, New Zealand

Hammal Z, Caplier A, Rombaut M (2005) Belief theory applied to facial expressions classification. Pattern recognition and image analysis. Lect Notes Comput Sci 3687(2005):183–191

Hupont I, Baldassarri S, del Hoyo R, Cerezo E (2008) Effective emotional classification combining facial classifiers and user assesment. Lect Notes Comput Sci 5098:431–440

Ji Q, Lan P, Looney C (2006) A probabilistic framework for modeling and real-time monitoring human fatigue. IEEE Transac Syst, Man Cybernetics, Part A 36:862–875

Kapoor A, Burleson W, Picard RW (2007) Automatic prediction of frustration. Int J Human–Comp Studies 65:724–736

Kayan S, Fussell S, Setlock L (2006) Cultural differences in the use of Instant Messaging in Asia and North America. In: Proceedings of the 2006 Conference on Computer supported cooperative work, pp 525–528

Keltner D, Ekman P (2000) Facial expression of emotion. Handbook of emotions 2:236–249

Kumar P, Yildirim EA (2005) Minimum-volume enclosing ellipsoids and core sets. J Optimization Theory Appl 126:1–21

Kuncheva L (2004) Combining pattern classifiers: methods and algorithms. Wiley-Interscience, Hoboken

Littlewort G, Bartlett MS, Fasel I, Susskind J, Movellan J (2006) Dynamics of facial expression extracted automatically from video. Image Vis Comput 24:615–625

Morrell D, Stirling W (2003) An extended set-valued kalman filter. In: Proceedings of the 3rd International Symposium on Imprecise probabilities and their applications (ISIPTA'03), pp 396–407

Pal P, Iyer A, Yantorno R (2006) Emotion detection from infant facial expressions and cries. In: Proceedings of the IEEE International Conference on Acoustics, speech and signal processing 2, pp 721–724

Pantic M, Valstar M, Rademaker R, Maat L (2005) Web-based database for facial expression analysis. In: IEEE International Conference on Multimedia and Expo, pp 317–321

Petridis S, Gunes H, Kaltwang S, Pantic M (2009) Static vs. dynamic modeling of human nonverbal behavior from multiple cues and modalities. In: Proceedings of the International Conference on multimodal interfaces, pp 23–30

Picard R (1997) Affective Computing. The MIT Press

Plutchik R (1980) Emotion: a psychoevolutionary synthesis. Harper & Row, New York

Pun T, Alecu T, Chanel G, Kronegg J, Voloshynovskiy S (2006) Braincomputer interaction research at the computer vision and multimedia laboratory, University of Geneva. IEEE Transac Neural Syst Rehabilitat Eng 14(2):210–213

Sánchez J, Hernández N, Penagos J, Ostróvskaya Y (2006) 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, pp 66–72

Shan C, Gong S, McOwan P(2007) Beyond facial expressions: learning human emotion from body gestures. In: Proceedings of the British Machine Vision Conference

Soyel H, Demirel H (2007) Facial expression recognition using 3D facial feature distances. Lect Notes Comp Sci 4633:831–383

Stoiber N, Seguier R, Breton G (2009) Automatic design of a control interface for a synthetic face. In: Proceedings of the 13th International Conference on Intelligent user interfaces, 207–216

Wallhoff F (2006) Facial expressions and emotion database. Technische Universität München, 2006. Available: http://www.mmk.ei.tum.de/~waf/fgnet/feedtum.html

Whissell CM (1989) The dictionary of affect in language. emotion: theory, research and experience 4. The Measurement of Emotions. Academic Press, New York

Witten I, Frank E (2005) Data mining: practical machine learning tools and techniques, 2nd edn. Morgan Kaufmann, San Francisco

Wolf A (2000) Emotional expression online: gender differences in emoticon use. CyberPsychol Behav 3(5):827–833

Wöllmer M, Al-Hames M, Eyben F, Schuller B, Rigoll G (2009) A multidimensional Dynamic Time Warping algorithm for efficient multimodal fusion of asynchronous data streams. Neurocomputing 73(1–3):366–380

Yeasin M, Bullot B, Sharma R (2006) Recognition of facial expressions and measurement of levels of interest from video. IEEE Transac Multiméd 8:500–508

Zeng Z, Tu J, Liu M, Huang T, Pianfetti B, Roth D, Levinson S (2007) Audio-visual affect recognition. IEEE Transac Multiméd 9(2):424–428

Zeng Z, Pantic M, Huang TS (2009a) Emotion recognition based onmultimodal information. In: Tao J, Tan T (eds) Affective information processing. Springer, London, pp 241–265

Zeng Z, Pantic M, Roisman GI, Huang TS (2009b) A survey of affect recognition methods: audio, visual, and spontaneous expressions. IEEE Transac Pattern Anal Mach Intell 31(1):39–58