Sentic Blending: Scalable Multimodal Fusion for the Continuous Interpretation of Semantics and Sentics

Erik Cambria, *Member, IEEE,* Newton Howard, *Member, IEEE,*
Jane Hsu, *Member, IEEE,* and Amir Hussain, *Senior Member, IEEE*

Abstract—The capability of interpreting the conceptual and affective information associated with natural language through different modalities is a key issue for the enhancement of human-agent interaction. The proposed methodology, termed sentic blending, enables the continuous interpretation of semantics and sentics (i.e., the conceptual and affective information associated with natural language) based on the integration of an affective common-sense knowledge base with any multimodal signal-processing module. In this work, in particular, sentic blending is interfaced with a facial emotional classifier and an opinion mining engine. One of the main distinguishing features of the proposed technique is that it does not simply perform cognitive and affective classification in terms of discrete labels, but it operates in a multidimensional space that enables the generation of a continuous stream characterising user's semantic and sentic progress over time, despite the outputs of the unimodal categorical modules have very different time-scales and output labels.

Keywords—Multimodal fusion; SenticNet; Facial expression analysis; Affective common-sense; Emotion recognition.

1 Introduction

Multimodality is a key aspect when trying to achieve natural interaction in machines. People, in fact, communicate not only through dialogues, but also through many other channels, e.g., facial expressions, gestures, eye contact, posture, and voice tone. Besides semantics, hence, machines also need to be able to recognise, interpret, and process sentics, that is, emotional information. In human cognition, thinking and feeling are mutually present: emotions are often the product of our thoughts, as well as our reflections are often the product of our affective states. Emotions are intrinsically part of our mental activity and play a key role in communication and decision-making processes.

Emotion is a chain of events made up of feedback loops. Feelings and behaviour can affect cognition, just as cognition can influence feeling. Emotion, cognition, and action interact in feedback loops and emotion can be viewed in a structural model tied to adaptation [1]. There is actually no fundamental opposition between emotion and reason. In fact, it may be argued that reason consists of basing choices on the perspectives of emotions at some later time. Reason dictates not giving in to one's impulses because doing so may cause greater suffering later [2]. Reason does not necessarily imply exertion of the voluntary capacities to suppress emotion. It does not necessarily involve depriving certain aspects of reality of their emotive powers. On the contrary, our voluntary capacities allow us to draw more of reality into the sphere of emotion. They allow one's emotions to be elicited not merely by the proximal, or the perceptual, or that which directly interferes with one's actions, but by that which, in fact, touches on one's concerns, whether proximal or distal, whether occurring now or in the future, whether interfering with one's own life or that of others.

Cognitive functions serve emotions and biological needs. Information from the environment is evaluated in terms of its ability to satisfy or frustrate needs. What is particularly significant is that each new cognitive experience that is biologically important is connected with an emotional reaction such as fear, pleasure, pain, disgust, or depression [3]. Emotions, in fact, are special states shaped by natural selection to adjust various aspects of our organism in order to make it better face particular situations, e.g., anger evolved for reaction, fear evolved for protection, and affection evolved for reproduction.

For these reasons, an important strand of emotion-related research in human-computer interaction is the simulation of emotional expressions made by embodied computer agents. Although several studies prove that multisensory fusion (e.g., audio, visual, physiological responses) improves the robustness and accuracy of machine analysis of semantics and sentics [4], [5], most cognitive and affective recognition works still focus on

- *E. Cambria is with Temasek Laboratories, National University of Singapore, 5A Engineering Drive 1, Singapore 117411. E-mail: cambria@nus.edu.sg*
- *N. Howard is with the Media Laboratory, MIT, 20 Ames Street, Cambridge, Massachusetts 02139-4307, USA E-mail: nhmit@mit.edu*
- *J. Hsu is with the Department of Computer Science and Information Engineering, National Taiwan University, Taipei 106, Taiwan E-mail: yjhsu@csie.ntu.edu.tw*
- *A. Hussain is with the Dept. of Computing Science and Mathematics, University of Stirling, Cottrell Building, FK9 4LA, UK E-mail: ahu@cs.stir.ac.uk*

978-1-4673-5923-8/13/$31.00 ©2013 IEEE

increasing the success rates in sensing semantics and sentics from a single channel, rather than merging complementary information across different channels [6]. The multimodal fusion of multiple channels is far from being solved [7] and represents an active and open research issue. In this context, sentic blending proposes a methodology for integrating and analysing conceptual and affective information coming from different modules. In this work, in particular, it is shown how a scalable semantics and sentics fusion technique allows to integrate different multimodal signals coming from a natural language text analyser and a facial emotional classifier, by presenting user cognitive and affective information in the form of a continuous multidimensional stream that shows user semantic and sentic evolution along the user-agent interaction process.

The structure of the paper is the following: Section 2 discusses the challenges of cognitive and affective multimodal fusion; Section 3 details how a multidimensional space is used for implementing the multimodal fusion methodology between different channels; Section 4 presents a proof of concept (PoC) and an evaluation of the proposed methodology; Section 5, finally, sets out conclusions and a description of future work.

2 Related Work

Cognitive and affective information fusion is essential to achieve successful user-agent interaction, but there are several challenges that make it a particularly difficult task. A major challenge is the definition of a reliable strategy to fuse the cognitive and affective information coming from different sources with very different time scales, metric levels, and temporal structures. Existing fusion strategies follow two main streams: feature-level fusion and decision-level fusion. Feature-level fusion combines the 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 [5], [8], [9], 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 synchronise the different inputs and retrain the whole classification system.

To overcome such difficulties, most researchers opt for a decision-level fusion, in which the inputs coming from each modality are modelled and classified independently, and their relative unimodal recognition results are integrated at the end of the process by the use of suitable criteria, e.g., expert rules and simple operators such as majority vote, sum, product, and adaptation of weights. Many studies have demonstrated the advantage of decision-level fusion over feature-level fusion, due to the uncorrelated errors from different classifiers [10] and the fact that time and feature dependence are abstracted. Various (mainly bimodal) decision-level fusion methods have been proposed in the literature [11], [12], [13], but optimal fusion designs are still undefined.

Most available multimodal recognisers 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 input quality changes and, hence, do not consider eventual adjustments in the reliability of the different information channels. In summary, there is not a general consensus when defining multimodal fusion strategies.

Another key challenge that directly affects multimodal fusion is related to the chosen output description level of semantics and sentics. Cognitive and affective information, in fact, is often classified into a limited set of labels, e.g., Ekman's six universal emotions [14], which often fail to describe the wide range and intensities of semantics and sentics that occur in daily communication settings. To overcome such a hurdle, Whissell [15] and Plutchik [1] proposed to view affective information 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 former measures how a human feels, from positive to negative. The latter 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 can describe an infinite number of cognitive and affective states and intensities, and is best suited to deal with variations of these over time. It provides an algebra and allows inputs coming from different modalities to be related mathematically, which makes it especially useful when integrating modules with different time-scales. However, compared to category-based description of semantics and sentics, very few works have chosen a dimensional description level and, in the few that have, the 2D space is discretised to a two-class (positive vs negative and active vs passive) [16] or a four class (space quadrants) classification [17], thereby losing the descriptive potential of the bidimensional space. 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 publicly available [18], [19] but, in comparison to categorical ones, they are limited in terms of number of modalities (in general, they explore only audio and/or video channels), 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 [20].

For these reasons, although working at dimensional level would be more appropriate to face the problem of multimodal fusion, training and validation of individual modules is performed using databases with categorical annotations. In order to exploit both categorical and dimensional classification techniques, the proposed multimodal fusion methodology exploits a multidimensional

space, built by means of sentic computing [21], which enables analogical reasoning between affective common-sense concepts.

3 A Methodology for Fusing Semantics and Sentics

In cognitive science, the term 'blending' refers to a general theory of cognition describing the way humans process and rationalise information through a set of mental operations [22]. The theory explains the process by which humans assign meaning to incoming information from sensory input, integrate it, and learn and gain knowledge.

In the same wake of conceptual blending, sentic blending proposes a methodology for fusing multiple unimodal signals in order to obtain a global multidimensional dynamic stream that shows how semantics and sentics evolve over time. In order to let the unimodal signals be defined in a robust and reliable way by means of existing categorical databases, each signal-processing module is assumed to classify in terms of its own list of cognitive and affective labels. Irrespective of the labels used, sentic blending maps each modules output to AffectiveSpace [21], a multidimensional vector space of affective common-sense knowledge, fuses the different sources of conceptual and emotional information over time through mathematical formulation, and obtains a dynamic multidimensional stream representing the users cognitive and affective progress as final output. The proposed methodology is sufficiently scalable to add new modules coming from new channels without having to retrain the whole system.

The first step of sentic blending consists in building a mapping such that the output of each module i at a given time t_{0i} can be represented as a 100-dimensional coordinate vector $s_i(t_{0i}) = [x_{0i}(t_{0i}), x_{1i}(t_{0i}), ... x_{99i}(t_{0i})]$ in AffectiveSpace. The majority of categorical modules described in the literature provide as output at the time t_{0i} (corresponding to the detection of the cognitive or affective stimulus) a list of labels with some associated weights. Irrespective of the categorisation used, each label has a specific location, i.e., an associated 100D point in AffectiveSpace. The components $< x_{0i}(t_{0i}), x_{1i}(t_{0i}), ... x_{99i}(t_{0i}) >$ of the coordinates vector $s_i(t_{0i})$ are then calculated as the barycenter of those weighted points. Hence, the users cognitive and affective progress can be viewed as a point (corresponding to the location of a particular state in time t) moving through this space over time.

The second step of sentic blending aims to compute the ensemble stream by fusing the different $s_i(t_{0i})$ vectors obtained from each modality over time. The main difficulty in achieving multimodal fusion is related to the fact that t_{0i} stimulus arrival times may be known a-priori or not, and may be very different for each module. To overcome such a hurdle, the following equation is proposed to calculate the ensemble stream $s(t) = [x_0(t), x_1(t), ... x_{99}(t)]$ at any arbitrary time t:

$$s(t) = \frac{\sum_{i=1}^{N} \gamma_i(t) s_i(t_{0i})}{\sum_{i=1}^{N} \gamma_i(t)} \quad (1)$$

where N is the number of fused modalities, t_{0i} is the arrival time of the last stimulus detected by module i, and $\gamma_i(t)$ 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 response is the sum of each modalitys contribution $s_i(t_{0i})$ modulated by the $\gamma_i(t)$ coefficients over time. Therefore, the definition of $\gamma_i(t)$ is especially important given that it governs the temporal behaviour of the fusion. Because human responses are analogous to systems with additive responses with decay, in which, in the absence of input, the response decays back to a baseline, $\gamma_i(t)$ weights are defined as:

$$\gamma_i(t) = \begin{cases} b_i c_i(t_{0i}) e^{-d_i(t-t_0)} & \text{if greater than } \epsilon \\ 0 & \text{elsewhere} \end{cases} \quad (2)$$

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 a 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 $\gamma_i(t)$ 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 ensemble stream that characterises the users cognitive and affective progress over time can be computed by calculating successive $s(t)$ values with any desired time between samples $\triangle t$. In other words, the ensemble stream is progressively built by adding $s(t_k)$ samples to its trajectory, where $t_k = k\triangle t$ (with k integer). However, there are two main issues in the stream calculation process:

1) If the contribution of every fused module is null at a given sample time, i.e., every $\gamma_i(t)$ is null at that time, the denominator in (1) is zero and the stream 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 jumps in AffectiveSpace can appear if semantic conflicts arise (e.g., if the distance between two close coordinates vectors $s_i(t_{0i})$ is long).

In order to address both issues, a Kalman filtering technique is applied to the computed stream. By definition, Kalman filters estimate a systems 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 ensemble streams trajectory and thus prevent large jumps. On the other hand, situations in which the sum of $\gamma_i(t)$ is null are prevented by letting the filter prediction output be taken as the 100D point position for those samples. In an analogy to classical mechanics, the 'sentic kinematics' of the 100D point moving through AffectiveSpace are modelled as the systems state X_k in the Kalman framework, i.e., $X_k = [x_0, x_1, ... x_{99}, v_{x_0}, v_{x_1}, ... v_{x_{99}}]_k^T$ representing position and velocity in 100 dimensions at time t_k. The following stream samples $s(t_k)$ are modelled as the measurement of the systems state in the following way:

$$X_{k+1} = F\, X_k + w_k = \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_{99} \\ v_{x_0} \\ v_{x_1} \\ \vdots \\ v_{x_{99}} \end{bmatrix}_{k+1} =$$

$$= \begin{bmatrix} 1 & 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 & 0 & 0 & \cdots & 1 \\ 0 & 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 1 \end{bmatrix} \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_{99} \\ v_{x_0} \\ v_{x_1} \\ \vdots \\ v_{x_{99}} \end{bmatrix}_k + w_k$$

where F is the transition matrix taking the state X_k from time k to time $k+1$ (i.e., from one stream sample to the next in the 100-dimensional vector space), and w_k is the process noise (which is usually assumed to be additive, white, Gaussian, and with zero mean). The measurement equation, hence, is defined as:

$$Y_k = H\, X_k + z_k = \begin{bmatrix} x_{0m} \\ x_{1m} \\ \vdots \\ x_{99m} \end{bmatrix}_k =$$

$$= \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 & 0 & 0 & \cdots & 0 \end{bmatrix} \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_{99} \\ v_{x_0} \\ v_{x_1} \\ \vdots \\ v_{x_{99}} \end{bmatrix}_k + z_k$$

where Y_k is the measurable at time k and H is the measurement matrix. The measurement noise z_k is also assumed to be additive, white, Gaussian, and with zero mean (but uncorrelated with the process noise w_k). Once the process and measurement equations are defined, the Kalman iterative estimation process can be applied to the stream, so that each iteration corresponds to a new sample.

4 PoC: Fusing Natural Language and Facial Expressions

As a PoC, sentic blending is employed for fusing natural language text and facial expressions within the design of an embodied conversational agent (ECA) based on a multimodal animation engine [23]. ECAs are graphical interfaces capable of using verbal and non-verbal modes of communication to interact with users in computer-based environments. These agents are sometimes just as an animated talking face, may be displaying simple facial expressions and, when using speech synthesis, with some kind of lip synchronisation, and sometimes they have sophisticated 3D graphical representation, with complex body movements and facial expressions.

Besides expressing emotions, ECAs should also be capable of understanding users' emotions and reacting accordingly. Recent research focuses on the psychological impact of affective agents endowed with the ability to behave empathically with the user [24], [25], [26]. The findings demonstrate that bringing about empathic agents is important in human-computer interaction. Moreover, addressing user's emotions significantly enhances the believability and lifelikeness of virtual humans. Nevertheless, to date, there are not many examples of agents that can sense in a completely automatic and natural (both verbal and non-verbal) way human emotion and respond realistically.

The architecture of the proposed ECA consists of four main modules: Perception, Cognitive, Deliberative/Generative, and Motor module (Fig. 1). The Perception module simply consists of the hardware necessary to gather the multimodal information from the user, i.e., keyboard, microphone, and webcam. The Cognitive module aims to infer the user's cognitive and affective state from the different inputs and integrate it. The Deliberative/Generative module is in charge of processing the extracted semantics and sentics to manage the virtual agent's decisions and reactions, which are finally generated by the Motor module.

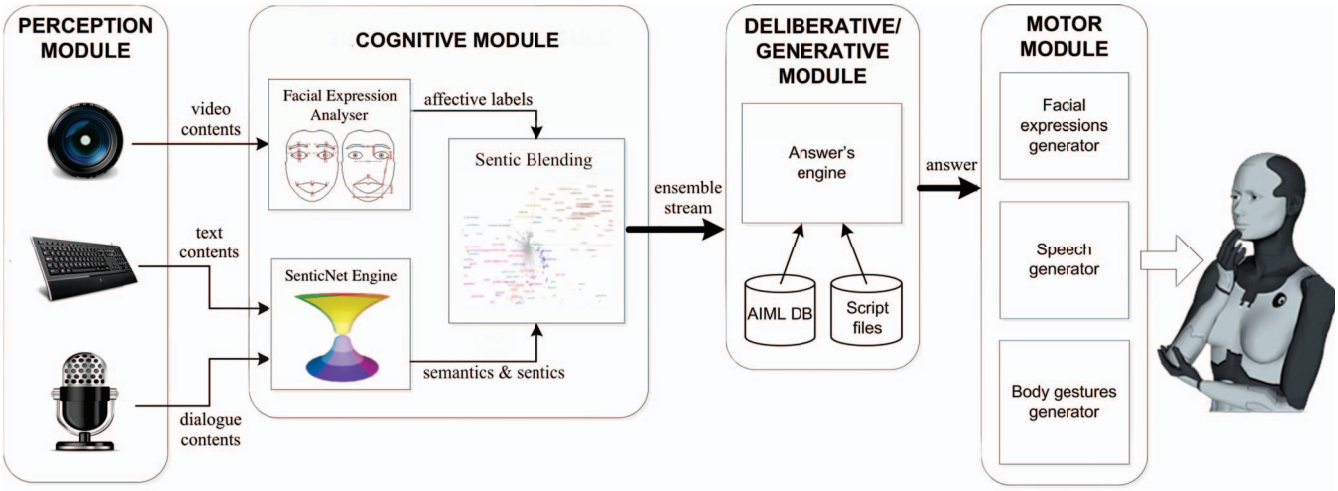

Fig. 1. ECA's architecture. The system mainly consists of two modules for managing the avatar's inputs and outputs, and two modules for performing affective common-sense reasoning.

The Cognitive module, in particular, is in charge of extracting cognitive and affective information from the textual, vocal, and video inputs and integrating them. It consists of three main parts: a facial expression analyser [23], for extracting affective information from video; the SenticNet engine [27], for inferring semantics and sentics associated with typed-in text and speech-to-text converted contents; and sentic blending, for integrating the outputs coming from the two previous modules.

4.1 Facial Expression Analyser

The facial expression analyser achieves an automatic classification of the shown facial expressions into discrete emotional categories. It is able to classify the user's emotion in terms of Ekman's six universal emotions (*fear*, *sadness*, *joy*, *disgust*, *surprise*, and *anger*) [14] plus *neutral*, giving a membership confidence value to each emotional category. The face modelling selected as input for the facial expression analyser follows a feature-based approach: the inputs are a set of facial distances and angles calculated from feature points of the mouth, eyebrows, and eyes.

The system intelligently combines the outputs of five different classifiers simultaneously. In this way, the overall risk of making a poor selection with a given classifier for a given input is reduced. The classifier combination chosen follows a weighted majority voting strategy, where the voted weights are assigned depending on the performance of each classifier for each emotion. In order to select the best classifiers to combine, the Waikado Environment for Knowledge Analysis (Weka) tool was used [28]. This provides a collection of machine learning algorithms for data mining tasks. From this collection, five classifiers were selected after tuning: RIPPER, MLP, SVM, NB, and C4.5. The selection was based on their widespread use as well as on the individual performance of their Weka implementation.

To train the classifiers and evaluate the performance of the system, two different facial emotion databases were used: the FGNET database [29], which provides video sequences of 19 different Caucasian people, and the MMI Facial Expression Database [30], which holds 1,280 videos of 43 different subjects from different races (Caucasian, Asian, and Arabic). Both databases are classified according to Ekman's six universal emotions plus *neutral*. A new database has been built for testing this work with a total of 1,500 static frames carefully selected from the apex of the video sequences from the FG-NET and MMI databases.

The results obtained when applying the above-mentioned strategy for combining the scores of the five classifiers are shown in the form of confusion matrix in Table 1 (results have been obtained with a 10-fold cross-validation test over the 1500 database images). As it can be observed, the success rates for *neutral*, *joy*, and *surprise* are very high (84.44%–95.23%). However, the system tends to confuse *disgust* with *fear*, *anger* with *disgust*, and *fear* with *surprise*; therefore, the performance for those emotions is slightly worse. The lowest result of the classification is for *sadness*: it is confused with *neutral* on 67.80% of occasions, due to the similarity of the facial expressions. Confusion between these pairs of emotions occurs frequently in the literature and for this reason many classification works do not consider some of them.

Nevertheless, the results can be considered positive as two incompatible emotions (such as *sadness* and *joy* or *fear* and *anger*) are confused on less than 0.2% of occasions. Another relevant aspect to be taken into account when evaluating the results is human opinion. The labels provided in the database for training classifiers correspond to the real emotions felt by users although they do not necessarily have to coincide with the perceptions other human beings may have about the facial expressions shown.

Undertaking this kind of study is very important when dealing with human-computer interaction, since the system is proved to work in a similar way to the human brain. In order to take into account the human factor in the evaluation of the results, 60 persons were told to classify the 1,500 images of the database in terms of emotions. As a result, each one of the frames was classified by 10 different people in 5 sessions of 50 images. The Kappa statistic obtained from raters annotations is equal to 0.74 (calculated following the formula proposed in [31]), which indicates an adequate inter-rater agreement in the emotional images annotation.

With this information, the evaluation of the results was repeated: the recognition was marked as good if the decision was consistent with that reached by the majority of the human assessors. The results (confusion matrix) of considering users' assessment are shown in Table 2. As it can be seen, the success ratios have considerably increased. Therefore, it can be concluded that the confusions of the algorithms go in the same direction as those of the users, which means that the adopted classification strategy is consistent with human classification.

4.2 SenticNet Engine

In order to effectively mine semantics and sentics from text, it is necessary to bridge the gap between unstructured natural language data and structured machine-processable data. To this end, an intelligent software engine based on SenticNet has been proposed [32] that aims to extract the conceptual and affective information associated with natural language text. Such an engine consists of four main components: a pre-processing module, which performs a first skim of the input text; a semantic parser, whose aim is to extract concepts from natural language data; the IsaCore [33] module, for inferring the semantics associated with the given concepts; and the SenticNet API[1], for the extraction of sentics.

The pre-processing module firstly exploits linguistic dictionaries to interpret all the affective valence indicators usually contained in text, e.g., special punctuation, complete upper-case words, cross-linguistic onomatopoeias, exclamation words, degree adverbs, and emoticons. Secondly, the module detects negation and spreads it in a way that it can be accordingly associated to concepts during the parsing phase. Handling negation is an important concern in sentiment-related analysis, as it can reverse the meaning of a statement. Such a task, however, is not trivial as not all appearances of explicit negation terms reverse the polarity of the enclosing sentence and that negation can often be expressed in rather subtle ways [34]. Lastly, the module converts text to lower-case and, after lemmatising it, splits it into single clauses according to grammatical conjunctions and punctuation.

1. http://sentic.net/api

The semantic parser deconstructs text into concepts using a lexicon based on sequences of lexemes that represent multiple-word concepts extracted from SenticNet and IsaCore. These n-grams are not used blindly as fixed word patterns but exploited as reference for the module, in order to extract multiple-word concepts from information-rich sentences. So, differently from other shallow parsers, the module can recognise complex concepts also when irregular verbs are used or when these are interspersed with adjective and adverbs, e.g., the concept 'buy christmas present' in the sentence "I bought a lot of very nice Christmas presents". The semantic parser, additionally, provides, for each retrieved concept, its relative frequency, valence, and status, i.e., the concept's occurrence in the text, its positive or negative connotation, and the degree of intensity with which the concept is expressed, respectively. For each clause, the module outputs a small bag of concepts (SBoC), which is later on analysed separately by the IsaCore module and the SenticNet API to infer the conceptual and affective information associated with the input text, respectively.

Once natural language text is deconstructed into concepts, these are given as input to both the IsaCore module and the SenticNet API. While the former exploits the graph representation of the common and common-sense knowledge base to detect semantics, the latter exploits the SenticNet API to infer sentics. In particular, the IsaCore module applies spectral association for assigning activation to key nodes of the semantic network, which are used as seeds or centroids for classification. Such seeds can simply be the concepts corresponding to the class labels of interest plus their available synonyms and antonyms, if any. Seeds can also be found by applying CF-IOF [35] on a training corpus, in order to perform a classification that is more relevant to the data under analysis. After seed concepts are identified, the module spreads their values across the IsaCore graph.

This operation, an approximation of many steps of spreading activation, transfers the most activation to concepts that are connected to the seed concepts by short paths in affective common-sense knowledge. Therefore, the concepts of each SBoC provided by the semantic parser are projected on the matrix resulting from spectral association in order to calculate their semantic relatedness to each seed concept and, hence, their degree of belonging to each different class. Such classification measure is directly proportional to the degree of connectivity between the nodes representing the retrieved concepts and the seed concepts in the IsaCore graph.

The concepts retrieved by the semantic parser are also given as input to the SenticNet API, which, in turn, exploits an ensemble of graph-mining and dimensionality-reduction techniques [21] to infer the affective information associated with them. In particular, the SenticNet API provides concept polarity and affective labels in terms of Pleasantness, Attention, Sensitivity, and Aptitude, the dimensions of the Hourglass of Emotion categorisation model [36].

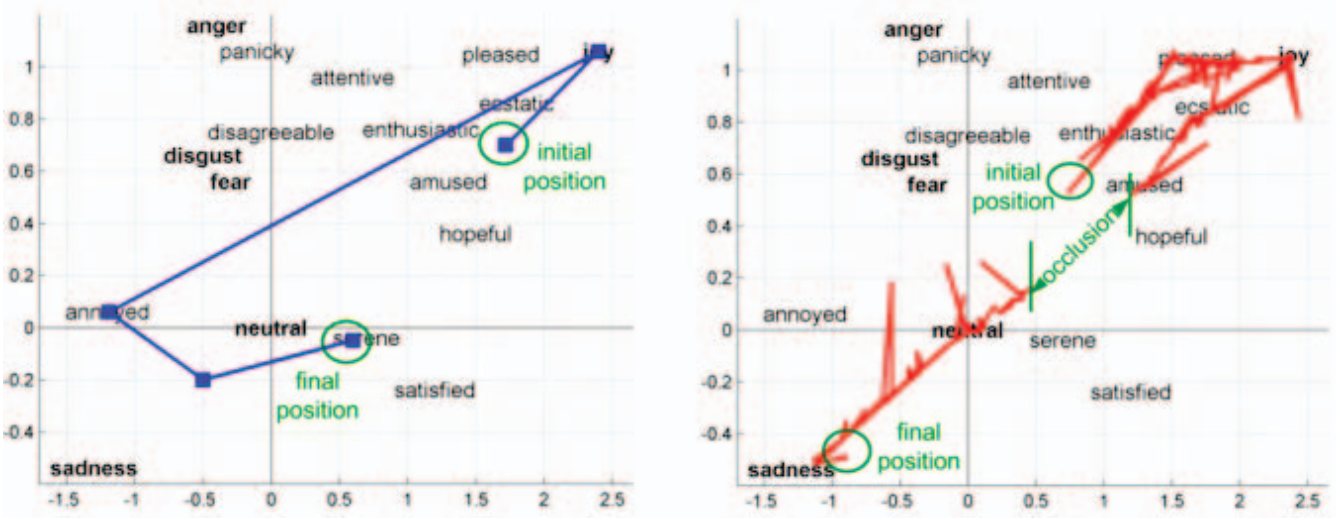

Fig. 2. Ensemble streams obtained when applying sentic blending to the SenticNet engine (left) and the facial expression analyser (right), without 'sentic kinematics' filtering.

As well as in the IsaCore module, the categorisation does not consist in simply labelling each concept, but also in assigning a confidence score to each emotional label, which is directly proportional to the degree of belonging to a specific affective cluster.

4.3 Sentic Blending Module

The sentic blending module aims to fuse the affective classification performed by the facial expression analyser with the labels produced by the SenticNet engine in real-time. In particular, the module considers a projection of AffectiveSpace on its first two eigenmoods, that is, its first two bases e_0 and e_1. The most significant eigenmood, e_0, represents concepts with positive affective valence. That is, the larger a concept's component in the e_0 direction is, the more affectively positive it is likely to be. Concepts with negative e_0 components, then, are likely to have negative affective valence.

The multimodal fusion of this PoC takes place in two dimensions. Firstly, every output label inferred by the SenticNet engine and the facial expression analyser is projected into the 2D representation of AffectiveSpace. Therefore, $s_i(t_{oi})$ can be obtained each time a given module i outputs cognitive and affective information at time t_{oi} (with i comprised between 1 and 2). It is interesting to notice that vectors $s_1(t_{01})$ coming from the SenticNet engine can arrive at any time t_{01}, unknown a-priori. However, the facial expression module outputs its $s_2(t_{02})$ 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 ensemble streams time between samples is assigned to $\triangle t = \frac{1}{f}$. The next step towards achieving the temporal fusion of the different modules is assigning a value to the parameters that define the $\gamma_i(t)$ weights, namely b_i, $c_i(t_{0i})$, d and ϵ. 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, such values were established empirically for this PoC. Once the parameters are assigned, the ensemble stream calculation process can be started following (1) and (2).

Finally, the 'sentic kinematics' filtering technique is iteratively applied in real-time each time a new sample is added to the computed stream. In order to demonstrate the potential of the presented fusion methodology, sentic blending has been applied to a conversation in which the text typed by a user and his facial capture video have been analysed and fused. The user narrates an emotional incident: at first, he is excited and happy about having bought a wonderful new car but, shortly afterwards, he becomes sad when telling he has dented it (see Table 3).

Fig. 2 shows the ensemble streams 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 'sentic kinematics' filtering. At first sight, the timing differences between modalities are striking: the facial expressions modules input stimuli are much more numerous than those of the SenticNet engine, making the latters emotional paths look more linear. Another noteworthy aspect is that the facial expression modules stream calculation is interrupted during several seconds (14s approximately) due to the appearance of a short facial occlusion, causing the tracking program to temporarily lose the facial features. Fig. 3 presents the continuous ensemble stream obtained when applying the methodology to fuse both modules, both without (left) and with (right) the 'sentic kinematics' filtering step. As can be seen, the complexity of the users cognitive and affective progress is shown in a simple and efficient way.

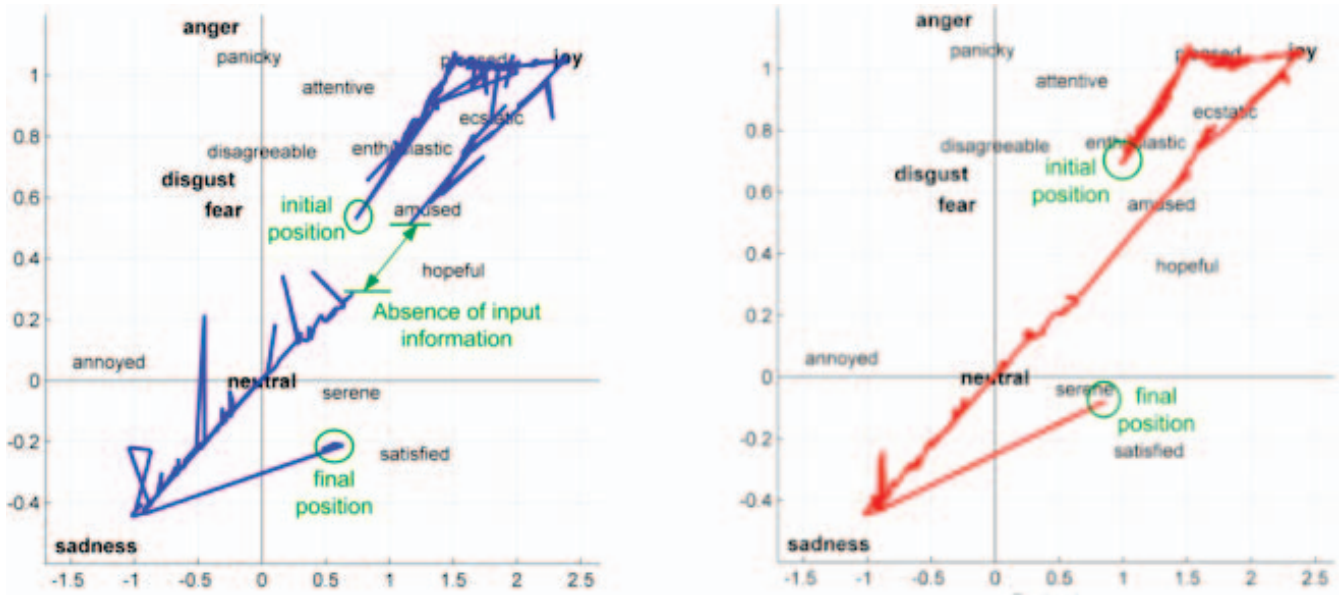

Fig. 3. Ensemble stream obtained when applying sentic blending to the proposed conversation, with (right) and without (left) using 'sentic kinematics' filtering.

Different modalities complement each other to obtain a more reliable result. Although the interruption period of the ensemble stream calculation is considerably reduced with respect to the facial expressions modules individual case (from 14s to 6s approximately), it still exists since the text modules decay process reaches the threshold ϵ before the end of the facial occlusion, causing the $\gamma_1(t)$ and $\gamma_2(t)$ weights to be null. Thanks to the use of the 'sentic kinematics' filtering technique, the ensemble stream is smoothed and the temporal input information absence is solved by letting the filter prediction output be taken as the 2D point position for those samples.

5 Conclusion and Future Work

Sentic blending is a scalable methodology for fusing multiple cognitive and affective recognition modules. This methodology is able to fuse any number of unimodal categorical modules, with very different timescales and output labels. This is possible thanks to the use of a multidimensional vector space that provides the system with mathematical capabilities to deal with temporal issues. The proposed methodology outputs a continuous multidimensional stream that represents in a novel and efficient way the users detected cognitive and affective progress over time. A Kalman filtering technique controls the ensemble stream in real-time through a 'sentic kinematics' model to ensure temporal consistency and robustness. The methodology has been shown effective to fuse two different modalities: natural language data and facial expressions. 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 semantic fusion. In general, evaluation issues are largely solved for categorical 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.

References

[1] R. Plutchik, "The nature of emotions," *American Scientist*, vol. 89, no. 4, pp. 344–350, 2001.
[2] N. Frijda, "The laws of emotions," *American Psychologist*, vol. 43, no. 5, 1988.
[3] U. Neisser, *Cognitive Psychology*. Appleton Century Crofts, 1967.
[4] Z. Zeng, M. Pantic, and T. Huang, "Emotion recognition based on multimodal information," *Affective Information Processing*, vol. 4, pp. 241–265, 2009.
[5] A. Kapoor, W. Burleson, and R. Picard, "Automatic prediction of frustration," *International Journal of Human-Computer Studies*, vol. 65, pp. 724–736, 2007.
[6] S. Gilroy, M. Cavazza, M. Niiranen, E. Andre, T. Vogt, J. Urbain, M. Benayoun, H. Seichter, and M. Billinghurst, "Pad-based multimodal affective fusion," in *ACII*, Amsterdam, 2009, pp. 1–8.
[7] 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.

[8] C. Shan, S. Gong, and P. McOwan, "Beyond facial expressions: Learning human emotion from body gestures," in *BMVC*, Warwick, 2007.
[9] T. Pun, T. Alecu, G. Chanel, J. Kronegg, and S. Voloshynovskiy, "Brain-computer interaction research at the computer vision and multimedia laboratory," *IEEE Trans. on Neural Systems and Rehabilitation Engineering*, vol. 14, no. 2, pp. 210–213, 2006.
[10] L. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*. Wiley & Sons, 2004.
[11] Z. Zeng, J. Tu, M. Liu, T. Huang, B. Pianfetti, D. Roth, and S. Levinson, "Audio-visual affect recognition," *IEEE Trans. Multimedia*, vol. 9, no. 2, pp. 424–428, 2007.
[12] H. Gunes and M. Piccardi, "Bi-modal emotion recognition from expressive face and body gestures," *Network and Computer Applications*, vol. 30, no. 4, pp. 1334–1345, 2007.
[13] P. Pal, A. Iyer, and R. Yantorno, "Emotion detection from infant facial expressions and cries," in *International Conference on Acoustics, Speech and Signal Processing*, Dallas, 2006.
[14] P. Ekman, T. Dalgleish, and M. Power, *Handbook of Cognition and Emotion*. Chichester: Wiley, 1999.
[15] C. Whissell, "The dictionary of affect in language," *Emotion: Theory, Research, and Experience*, vol. 4, pp. 113–131, 1989.
[16] K. Karpouzis, G. Caridakis, L. Kessous, N. Amir, A. Raouzaiou, L. Malatesta, and S. Kollias, "Modeling naturalistic affective states via facial, vocal and bodily expressions recognition," in *Lecture Notes in Artificial Intelligence*. Springer, 2007, vol. 4451, pp. 92–116.
[17] G. Caridakis, K. Karpouzis, and S. Kollias, "User and context adaptive neural networks for emotion recognition," *Neurocomputing*, vol. 71, pp. 2553–2562, 2008.
[18] E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, and M. McRorie, "The HUMAINE database: Addressing the needs of the affective computing community," in *nduced Emotional Data Affective Computing and Intelligent Interaction*, ser. Lecture Notes in Computer Science. Berlin Heidelberg: Springer, 2007, vol. 4738, pp. 488–500.
[19] M. Grimm, K. Kroschel, and S. Narayanan, "The vera am mittag german audio-visual emotional speech database," in *IEEE International Conference on Multimedia and Expo*, 2008, pp. 865–868.
[20] F. Fragopanagos and J. Taylor, "Emotion recognition in human-computer interaction," *Neural Networks*, vol. 18, pp. 389–405, 2005.
[21] E. Cambria and A. Hussain, *Sentic Computing: Techniques, Tools, and Applications*. Dordrecht, Netherlands: Springer, 2012.
[22] G. Fauconnier and M. Turner, *The Way We Think: Conceptual Blending and the Mind's Hidden Complexities*. Basic Books, 2003.
[23] E. Cambria, I. Hupont, A. Hussain, E. Cerezo, and S. Baldassarri, "Sentic avatar: Multimodal affective conversational agent with common sense," in *Toward Autonomous, Adaptive, and Context-Aware Multimodal Interfaces: Theoretical and Practical Issues*, ser. Lecture Notes in Computer Science, A. Esposito, A. Hussain, M. Faundez-Zanuy, R. Martone, and N. Melone, Eds. Berlin Heidelberg: Springer-Verlag, 2011, vol. 6456, pp. 82–96.
[24] K. Isbister, *Better game characters by design: A psychological approach*. Morgan Kaufmann, 2006.
[25] N. Yee, J. Bailenson, M. Urbanek, F. Chang, and D. Merget, "The unbearable likeness of being digital: The persistence of nonverbal social norms in online virtual environments," *CyberPsychology & Behavior*, vol. 10, no. 1, pp. 115–121, 2007.
[26] M. Ochs, C. Pelachaud, and D. Sadek, "An empathic virtual dialog agent to improve human-machine interaction," in *International Joint Conference on Autonomous Agents and Multiagent Systems*, 2008, pp. 89–96.
[27] E. Cambria, C. Havasi, and A. Hussain, "SenticNet 2: A semantic and affective resource for opinion mining and sentiment analysis," in *FLAIRS*, Marco Island, 2012, pp. 202–207.
[28] I. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*. San Francisco: Morgan Kaufmann, 2005.
[29] F. Wallhoff, "Facial expressions and emotion database," Technische Universitat Munchen, Tech. Rep., 2006.
[30] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," in *ICME*, Singapore, 2005.
[31] S. Siegel and N. Castellan, *Nonparametric Statistics for the Social Siences*. New York: McGraw-Hill, 1988.
[32] E. Cambria, D. Rajagopal, D. Olsher, and D. Das, "Big social data analysis," in *Big Data Computing*, R. Akerkar, Ed. Chapman and Hall/CRC, 2013, ch. 13.
[33] E. Cambria, Y. Song, H. Wang, and N. Howard, "Semantic multi-dimensional scaling for open-domain sentiment analysis," *IEEE Intelligent Systems*, doi: 10.1109/MIS.2012.118, 2013.
[34] E. Lapponi, J. Read, and L. Ovrelid, "Representing and resolving negation for sentiment analysis," in *ICDM SENTIRE*, Brussels, 2012, pp. 687–692.
[35] E. Cambria, T. Benson, C. Eckl, and A. Hussain, "Sentic PROMs: Application of sentic computing to the development of a novel unified framework for measuring health-care quality," *Expert Systems with Applications*, vol. 39, no. 12, pp. 10 533–10 543, 2012.
[36] E. Cambria, A. Livingstone, and A. Hussain, "The hourglass of emotions," in *Cognitive Behavioral Systems*, ser. Lecture Notes in Computer Science, A. Esposito, A. Vinciarelli, R. Hoffmann, and V. Muller, Eds. Berlin Heidelberg: Springer, 2012, vol. 7403, pp. 144–157.

Erik Cambria received his BEng (2005) and MEng (2008) in Electronic Engineering from the University of Genova and his PhD (2012) in Computing Science and Mathematics from the University of Stirling. Today, he is a research scientist at the National University of Singapore. He is editorial board of Cognitive Computation and chair of several international conferences, e.g., Brain Inspired Cognitive Systems (BICS), symposia, e.g., Extreme Learning Machines (ELM), and workshops, e.g., ICDM SENTIRE.

Newton Howard is one of the directors of the Synthetic Intelligence Project and a resident scientist at the Massachusetts Institute of Technology. He received his Doctoral degree in Cognitive Informatics and Mathematics from La Sorbonne in . He is a national security advisor to several U.S. Government organisations and works with multi-disciplinary teams of physicists, chemists, biologists, brain scientists, computer scientists, and engineers to reach a deeper understanding of the brain.

Jane Hsu is Professor and Department Chair of Computer Science and Information Engineering at National Taiwan University. Her research interests include intelligent multi-agent systems, data mining, service oriented computing and web technology. She is on the editorial board of the International Journal of Service Oriented Computing and Applications. She is actively involved in many international conferences as organizer and program committee. She is a member of AAAI, IEEE, ACM, Phi Tau Phi, and TAAI.

Amir Hussain is a Professor of Computing Science at the University of Stirling, UK. He did his first degree and PhD in Electronic Engineering both from Strathclyde University in 1992 and 1997, respectively. He currently manages over a dozen interdisciplinary research projects funded by research councils, charities and industry, and has authored more than 200 papers in brain-inspired cognitive computing technology and applications. He is founding Editor-in-Chief of the Cognitive Computation journal (Springer).

	disgust	joy	anger	fear	sadness	neutral	surprise
disgust	**79.41%**	0%	2.39%	18.20%	0%	0%	0%
joy	4.77%	**95.23%**	0%	0%	0%	0%	0%
anger	19.20%	0%	**74.07%**	0%	3.75%	2.98%	0%
fear	9.05%	0%	0%	**62.96%**	8.53%	0%	19.46%
sadness	0.32%	0.20%	1.68%	0%	**30.00%**	67.80%	0%
neutral	0%	0%	1.00%	2.90%	4.10%	**92.00%**	0%
surprise	0%	0%	0%	11.23%	0%	4.33%	**84.44%**

TABLE 1
Confusion matrix obtained combining the five classifiers. Success rates for *neutral*, *joy*, and *surprise* are very high, but *disgust*, *anger*, and *fear* tend to be confused.

	disgust	joy	anger	fear	sadness	neutral	surprise
disgust	**84.24%**	0%	2.34%	13.42%	0%	0%	0%
joy	4.77%	**95.23%**	0%	0%	0%	0%	0%
anger	15.49%	0%	**77.78%**	0%	3.75%	2.98%	0%
fear	1.12%	0%	0%	**92.59%**	2.06%	0%	4.23%
sadness	0.32%	0.20%	1.68%	0%	**66.67%**	31.13%	0%
neutral	0%	0%	0%	0.88%	1.12%	**98.00%**	0%
surprise	0%	0%	0%	6.86%	0%	2.03%	**91.11%**

TABLE 2
Confusion matrix obtained after human assessment. Success ratios considerably increase, meaning that the adopted classification strategy is consistent with human classification.

# Module	1	2
Modality	text	video
Categorisation model	Hourglass of Emotions	Ekman's basic emotions
Total number of possible output labels	24	6
General confidence	$b_1 = 0.65$	$b_2 = 0.9461$
Temporal confidence	$c_1(t_{01}) = 1$	$c_2(t_{02})$ is assigned to the tracking quality confidence weighting, from 0 to 1, provided by the facial feature tracking program
Decay value	$d_1 = 0.035\ s^{-1}$	Irrelevant since the stream sample rate is equal to the video frame rate
Threshold value	$\epsilon = 0.1$	Irrelevant since the stream sample rate is equal to the video frame rate

TABLE 3
Temporal Fusion Parameters