Continuous Prediction of Spontaneous Affect from Multiple Cues and Modalities in Valence-Arousal Space

Mihalis A. Nicolaou, *Student Member*, *IEEE*, Hatice Gunes, *Member*, *IEEE*, and Maja Pantic, *Senior Member*, *IEEE*

Abstract—Past research in analysis of human affect has focused on recognition of prototypic expressions of six basic emotions based on posed data acquired in laboratory settings. Recently, there has been a shift toward subtle, continuous, and context-specific interpretations of affective displays recorded in naturalistic and real-world settings, and toward multimodal analysis and recognition of human affect. Converging with this shift, this paper presents, to the best of our knowledge, the first approach in the literature that: 1) fuses facial expression, shoulder gesture, and audio cues for dimensional and continuous prediction of emotions in valence and arousal space, 2) compares the performance of two state-of-the-art machine learning techniques applied to the target problem, the bidirectional Long Short-Term Memory neural networks (BLSTM-NNs), and Support Vector Machines for Regression (SVR), and 3) proposes an output-associative fusion framework that incorporates correlations and covariances between the emotion dimensions. Evaluation of the proposed approach has been done using the spontaneous SAL data from four subjects and subject-dependent leave-one-sequence-out cross validation. The experimental results obtained show that: 1) on average, BLSTM-NNs outperform SVR due to their ability to learn past and future context, 2) the proposed output-associative fusion framework outperforms feature-level and model-level fusion by modeling and learning correlations and patterns between the valence and arousal dimensions, and 3) the proposed system is well able to reproduce the valence and arousal ground truth obtained from human coders.

Index Terms—Dimensional affect recognition, continuous affect prediction, valence and arousal dimensions, facial expressions, shoulder gestures, emotional acoustic signals, multicue and multimodal fusion, output-associative fusion.

1 Introduction

Most of the past research on automatic affect sensing and recognition has focused on recognition of facial and vocal affect in terms of basic emotions, and then based on data that have been posed on demand or acquired in laboratory settings [29], [67], [49]. Additionally, each modality has been considered in isolation. However, a number of researchers have shown that in everyday interactions, people exhibit nonbasic, subtle, and rather complex mental/affective states like thinking, embarrassment, or depression [4]. Such subtle and complex affective states can be expressed via dozens (or hundreds) of anatomically possible facial expressions or bodily gestures. Accordingly, a single label (or any small number of discrete classes) may not reflect the complexity of the affective state conveyed by such a rich source of information [54]. Hence, a number of researchers advocate the use of dimensional description of human affect, where an affective state is characterized in terms of a number of latent dimensions (e.g., [54], [56], [55]).

• *The authors are with the Department of Computing, Imperial College, 180 Queen's Gate, London SW7 2AZ, UK.*
E-mail: {mihalis, h.gunes, m.pantic}@imperial.ac.uk.

Manuscript received 1 Oct. 2010; revised 31 Jan. 2011; accepted 22 Feb. 2011; published online 16 Mar. 2011.
Recommended for acceptance by A.A. Salah, T. Gevers, and A. Vinciarelli.
For information on obtaining reprints of this article, please send e-mail to: taffc@computer.org, and reference IEEECS Log Number TAFFCSI-2010-10-0080.
Digital Object Identifier no. 10.1109/T-AFFC.2011.9.

It is not surprising, therefore, that researchers in automatic affect sensing and recognition have recently started exploring how to model, analyze, and interpret the subtlety, complexity, and continuity of affective behavior in terms of latent dimensions, rather than in terms of a small number of discrete emotion categories [27].

The work introduced here converges with this recent shift in affect recognition, from recognizing posed expressions in terms of discrete and basic emotion categories to the recognition of spontaneous expressions in terms of dimensional and continuous descriptions. It contributes to the *affect sensing and recognition research field* as follows:

- It presents the first approach in the literature toward automatic, *dimensional*, and *continuous* affect prediction in terms of arousal (A) and valence (V) based on facial expression, shoulder gesture, and audio cues.
- It proposes an *output-associative prediction framework* that incorporates correlations between the emotion dimensions and demonstrates significantly improved prediction performance.
- It presents a comparison of two state-of-the-art machine learning techniques, namely, the bidirectional Long Short-Term Memory neural networks (BLSTM-NNs) and Support Vector Machines for Regression (SVR), for continuous affect prediction.
- It proposes a set of evaluation metrics and demonstrates their usefulness to dimensional and continuous affect prediction.

The paper is organized as follows: Section 2 describes theories of emotion and perception of emotions from visual and audio modalities. Section 3 summarizes the related work in the field of automatic dimensional affect analysis. Section 4 describes the overall methodology employed. Section 5 presents the naturalistic database used in the experimental studies and describes the preprocessing of the data. Section 6 explains the audio and visual feature extraction and tracking. Section 7 describes the learning techniques and the evaluation measures employed for continuous emotion prediction, and introduces the output-associative fusion framework. Section 8 discusses the experimental results. Section 9 concludes the paper.

2 Background

2.1 Theories of Emotion

The description of affect has been a long standing problem in the area of psychology. Three major approaches can be distinguished [24]: 1) the categorical approach, 2) the dimensional approach, and 3) the appraisal-based approach.

According to the categorical approach, there exist a small number of emotions that are basic, hard-wired in our brain, and recognized universally. Ekman conducted various experiments and concluded that six basic emotions can be recognized universally, namely, happiness, sadness, surprise, fear, anger, and disgust [19].

According to the dimensional approach, affective states are not independent from one another; rather, they are related to one another in a systematic manner. In this approach, the majority of affect variability is covered by two dimensions: valence and arousal [44], [54]. The valence dimension (V) refers to how positive or negative the emotion is, and ranges from unpleasant feelings to pleasant feelings of happiness. The arousal dimension (A) refers to how excited or apathetic the emotion is, and it ranges from sleepiness or boredom to frantic excitement. Psychological evidence suggests that these two dimensions are interrelated [48], [1], [39], [41]. More specifically, there exist repeating configurations and interdependencies within the values that describe each dimension.

In the categorical approach, where each affective display is classified into a single category, complex mental/affective state or blended emotions may be too difficult to handle [66]. Instead, in the dimensional approach, emotion transitions can be easily captured, and observers can indicate their impression of moderate (less intense) and authentic emotional expressions on several continuous scales. Hence, dimensional modeling of emotions has proven to be useful in several domains (e.g., affective content analysis [65]).

It should be possible to describe affect in a continuous manner in terms of any relevant dimension (or axes). However, for practical reasons, we opted for the dimensions of arousal and valence in a continuous scale due to their widespread use in psychology and behavioral science.

For further details on different approaches to modeling human emotions and their relative advantages and disadvantages, the reader is referred to [56] and [24].

2.2 Perception of Emotions from Audio and Visual Cues

The prosodic features which seem to be reliable indicators of the basic emotions are the continuous acoustic measures, particularly pitch-related measures (range, mean, median, and variability), intensity, and duration. For a comprehensive summary of acoustic cues related to vocal expressions of basic emotions, readers are referred to [14]. There have also been a number of works focusing on how to map audio expression to dimensional models. Cowie et al. used valence-activation space, which is similar to the V-A space, to model and assess emotions from speech [14], [13]. Scherer and colleagues have also proposed how to judge emotion effects on vocal expression, using the appraisal-based theory [56], [24].

The most widely known and used visual signals for automatic affect sensing and recognition are facial action units (e.g., pulling eyebrows up) and facial expressions (e.g., producing a smile). More recently, researchers have also started exploring how bodily postures (e.g., backward head bend and arms raised forward and upward) and bodily gestures (e.g., head nod) communicate affective information. Dimensional models are considered important in these tasks as a single label may not reflect the complexity of the affective state conveyed by a facial expression, body posture, or gesture. Ekman and Friesen [20] considered expressing discrete emotion categories via face, and communicating dimensions of affect via body as more plausible. A number of researchers have investigated how to map various visual signals onto emotion dimensions. For instance, Russell [54] mapped the facial expressions to various positions on the 2D plane of arousal valence (e.g., joy is mapped on the high arousal—positive valence quadrant), while Cowie et al. [15] investigated the emotional and communicative significance of head nods and shakes in terms of arousal and valence dimensions, together with dimensional representation of solidarity, antagonism, and agreement.

Ambady and Rosenthal reported that human judgments of behaviors that were based jointly on face and body cues were 35 percent more accurate than those based on the face cues alone [2]. In general, however, body and hand gestures are much more varied than face gestures. There is an unlimited vocabulary of body postures and gestures with combinations of movements of various body parts. Unlike facial expressions, communication of emotions by bodily movement and expressions is still a relatively unexplored and unresolved area in psychology, and further research is needed in order to obtain a better insight on how they contribute to the perception and recognition of affect dimensions or various affective states.

In this work, we chose to focus on acoustic cues, facial expressions, and shoulder gestures (and their fusion) since they have been reported as informative cues for a number of spontaneous human nonverbal behavior analysis tasks (e.g., automatic recognition of posed versus spontaneous smiles [52]).

3 Related Work

Affect sensing is now a well-established field, and there is an enormous amount of literature available on different aspects

of affect sensing. As it is virtually impossible to include all of these works, we only introduce the most relevant literature on dimensional affect sensing and recognition. Affect recognition using multiple cues and modalities, and its shift from the lab to the real-world settings, are reviewed and discussed in detail in [29]. An exhaustive survey of past efforts in audiovisual affect sensing and recognition (e.g., facial action unit recognition, posed versus spontaneous expression recognition, etc.), together with various visual, audio, and audiovisual databases, is presented in [67]. For a recent survey of affect detection models, methods, and their applications, reviewed in an interdisciplinary perspective, the reader is referred to [8].

When it comes to automatic dimensional affect recognition, the most commonly employed strategy is to simplify the problem of classifying the six basic emotions to a three-class valence-related classification problem: positive, neutral, and negative emotion classification (e.g., [66]). A similar simplification is to reduce the dimensional emotion classification problem to a two-class problem (positive versus negative or active versus passive classification) or a four-class problem (classification into the quadrants of 2D V-A space, e.g., [10], [22], [23], [32], [64]). For instance, Wagner et al. [62] analyzes four emotions, each belonging to one quadrant of the V-A emotion space: high arousal positive valence (joy), high arousal negative valence (anger), low arousal positive valence (relief), and low arousal negative valence (sadness).

Systems that target automatic dimensional affect recognition, considering that the emotions are represented along a continuum, generally tend to quantize the continuous range into certain levels. Kleinsmith and Bianchi-Berthouze [37] discriminate between high-low, high-neutral, and low-neutral affective dimensions, while Wöllmer et al. [42] use the Sensitive Artificial Listener database (SAL-DB) and quantize the V-A into four or seven levels and use Conditional Random Fields (CRFs) to predict the quantized labels.

Methods for discriminating between more coarse categories, such as low, medium, and high [38], excited negative, excited positive, and calm neutral [11], positive versus negative [45], and active versus passive [10] have also been proposed. Of these, Caridakis et al. [10] use the SAL database, similar to our work presented in this paper, and combine information from audio (acoustic cues) and visual (Facial Animation Parameters used in animating MPEG-4 models) modalities. Nicolaou et al. focus on audiovisual classification of spontaneous affect into negative or positive emotion categories, and utilize 2 and 3-chain coupled Hidden Markov Models and likelihood space classification to fuse multiple cues and modalities [45]. Kanluan et al. [34] combine facial expression and audio cues exploiting SVR and late fusion, using weighted linear combinations and discretized annotations (on a 5-point scale, for each dimension).

As far as actual continuous dimensional affect prediction (without quantization) is concerned, four attempts have been proposed so far, three of which deal exclusively with speech (i.e., [64], [42], [26]). The work presented in [64] utilizes a hierarchical dynamic Bayesian network combined with BLSTM-NN performing regression and quantizing the results into four quadrants (after training). The work by Wöllmer et al. uses Long Short-Term Memory neural networks and Support Vector Machines for Regression [42]. Grimm and Kroschel use SVRs and compare their performance to that of the distance-based fuzzy k-Nearest Neighbor and rule-based fuzzy-logic estimators [26]. The work of Gunes and Pantic [28] focuses on dimensional prediction of emotions from spontaneous conversational head gestures by mapping the amount and direction of head motion, and occurrences of head nods and shakes into arousal, expectation, intensity, power, and valence level of the observed subject using SVRs.

For comparison purposes, in Table 1, we briefly summarize the automated systems that attempt to model and recognize affect in continuous dimensional space using multiple cues and modalities, together with the work presented in this paper. We also include the works proposed in [45], [34], and [9] as they are relevant for the current study (although classification has been reported for a discretized rather than a continuous dimensional affect space). Works using dimensions other than valence and arousal have been also included. Subsequently, Table 2 presents utilized classification methodology and the performance attained by the methods listed in Table 1. This overview is intended to be illustrative rather than exhaustive, and for systems most relevant to our work. For systems that deal with dimensional affect recognition from a single modality or cue, the reader is referred to [27] and [67].

As can be seen from Tables 1 and 2, the surveyed systems use different training and testing sets (which differ in the way emotions are elicited and annotated), they differ in the underlying model of emotions (i.e., target emotional categories) as well as in the employed modality or combination of modalities and the applied evaluation method. All of these make it difficult to quantitatively and objectively evaluate the accuracy of the V-A modeling and the effectiveness of the developed systems.

Compared to the works introduced in Tables 1 and 2, and surveyed in [27], the methodology introduced in this paper 1) presents the first approach toward automatic, dimensional, and continuous affect prediction based on facial expression, shoulder gesture, and audio cues, and 2) proposes a framework that integrates temporal correlations between continuous dimensional outputs (valence and arousal) to improve regression predictions. Our motivation for the latter is twofold. First, there is strong (theoretical and experimental) psychological evidence reporting that the valence and arousal dimensions are intercorrelated (i.e., repeated configurations do manifest between the dimensions) [48], [1], [39], [41]. Despite this fact, automatic modeling of these correlations has not been attempted yet. Second, there is a growing interest in the pattern recognition field in modeling not only the input but also the output covariances (e.g., [7], [63], [6]).

Additionally, as pointed out in [35], (dis)agreement between human annotators affects the performance of the automated systems. The system should ideally take into account the interobserver (dis)agreement level and correlate this to the level of (dis)agreement attained between the

TABLE 1
Overview of the Systems for Dimensional Affect Recognition from Multiple Modalities in Terms of Modality/Cue, Database Employed, Number of Samples Used, Features Extracted, and Dimensions Recognized

System	modality/cue	database	# of samples	features	dimensions
Caridakis et al. [9]	audiovisual	SAL, 4 subjects	not reported	various visual and acoustic features	neutral & 4 V-A quadrants (quantized)
Kanluan et al. [34]	audiovisual	VAM corpus recorded from the German TV talk show Vera am Mittag, 20 subjects	234 sentences &1600 images	prosodic & spectral features, 2-dimensional Discrete Cosine Transform applied to blocks of a predefined size in facial images	valence, activation & dominance (5-level annotation, mapped to continuous)
Forbes-Riley & Litman [21]	audio and text	student emotions from tutorial spoken dialogs	not reported	acoustic and prosodic, text-based, and contextual features	negative, neutral & positive
Karpouzis et al. [35]	audiovisual	SAL, 4 subjects	76 Passages, 1600 tunes	various visual & acoustic features	negative vs. positive, active vs. passive
Kim [36]	speech & physiological signals	data recorded using a version of the quiz Who wants to be a millionaire?, 3 subjects	343 samples	EMG, SC, ECG, BVP, Temp, RSP & acoustic features	4 A-V quadrants (quantized)
Nicolaou et al. [45]	audiovisual (facial expression, shoulder, audio cues)	SAL, 4 subjects	30,000 visual, 60,000 audio samples	trackings of 20 facial feature points, 5 shoulder points for video; MFCC and prosody features for audio	negative vs. positive valence (quantized)
This work	audiovisual (facial expression, shoulder, audio cues)	SAL, 4 subjects	30,000 visual, 60,000 audio samples	trackings of 20 facial feature points, 5 shoulder points for video; MFCC and prosody features for audio	valence and arousal (continuous)

This work is also included for comparison.

ground truth and the results provided by the system. To address the aforementioned issue, this work introduces novel evaluation measures and demonstrates their usefulness to dimensional and continuous affect prediction.

4 Outline of the Proposed Methodology

The methodology proposed in this paper consists of preprocessing, segmentation, feature extraction, and prediction components, and is illustrated in Fig. 1.

The first two stages, that of preprocessing and segmentation, depend mostly on the set of annotations provided with the SAL database (in terms of valence and arousal dimensions). We introduce various procedures to 1) obtain the ground truth corresponding to each frame by maximizing intercoder agreement, and 2) to determine the audiovisual segments that capture the transition *from* one emotional state *to* another (and back). Essentially, these procedures automatically segment spontaneous multimodal data in terms of negative and positive audiovisual segments that contain an offset before and after (i.e., the baseline) the displayed expression (Section 5.3).

During the feature extraction stage, the presegmented audiovisual segments from the SAL database are used. For

TABLE 2
The Utilized Classification Methodology and the Performance Attained by the Methods Listed in Table 1

System	Classification	Explicit fusion	Results
Caridakis et al. [9]	a feed-forward back-propagation network	not reported	reduced MSE for every tune
Kanluan et al. [34]	Support Vector Regression for 3 continuous dimensions	model-level fusion by a weighted linear combination	average estimation error of the fused result was 17.6% and 12.7% below the individual error of the acoustic and visual modalities, the correlation between the prediction and and ground truth was increased by 12.3% and 9.0%
Forbes-Riley & Litman [21]	AdaBoost to boost a decision tree algorithm	not reported	84.75% recognition accuracy for a 3 class problem
Karpouzis et al. [35]	a Recurrent Network that outputs one of the 4 classes	not reported	67% recognition accuracy with vision, 73% with prosody, 82% after fusion (whether on unseen data is not specified)
Kim [36]	modality-specific LDA-based classification	hybrid fusion by integrating results from feature- and model-level fusion	51% for bio-signals, 54% for speech, 55% for feature fusion, 52% for decision fusion, 54% for hybrid fusion, subject independent validation
Nicolaou et al. [45]	HMM and Likelihood Space via SVM	model-level fusion and likelihood space fusion	over 10-fold cross validation, best mono-cue result is 91.76% from facial expressions, best fusion result is 94% by fusing facial expressions, shoulder and audio cues
This work	SVR and BLSTM-NNs	feature-level, decision-level, and output-associative fusion	over leave-one-sequence-out cross validation, best result is attained by fusion of face, shoulder and audio cues, RMSE=0.15 and COR=0.796 for valence and RMSE=0.21 and COR=0.642 for arousal.

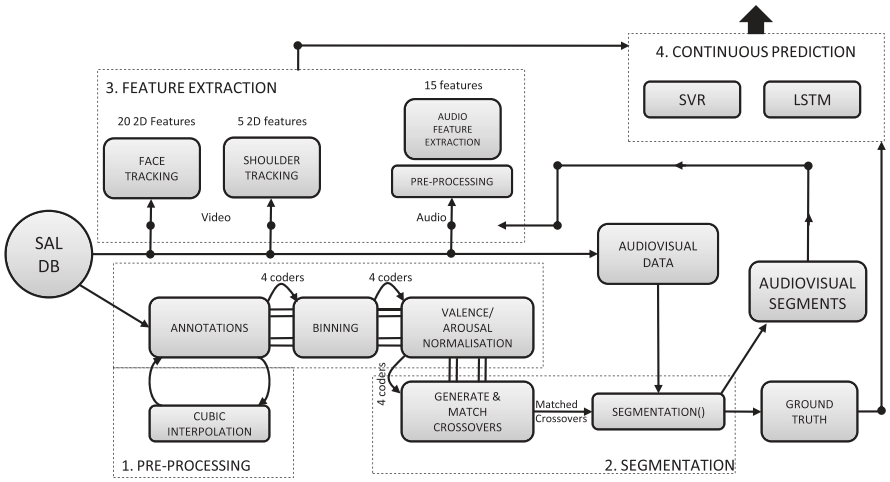


Fig. 1. Methodology employed: Preprocessing, segmentation, feature extraction, and prediction.

the audio modality, the Mel-frequency Cepstrum Coefficients (MFCC) [33], as well as prosody features, such as pitch and energy features are extracted. To capture the facial and shoulder motion displayed during a spontaneous expression, we use the Patras-Pantic particle filtering tracking scheme [50] and the standard Auxiliary Particle Filtering (APF) technique [53], respectively.

The final stage, that is based on all of the aforementioned steps, consists of affect prediction, multicue and multimodal fusion, and evaluation. SVRs and BLSTM-NNs are used for single-cue affect prediction. Due to their superior performance, BLSTM-NNs are further used for feature and model-level fusion of multiple cues and modalities. An output-associative fusion framework that employs a first layer of BLSTM-NNs for predicting V-A values from the original input features and a second layer of BLSTM-NN using these predictions jointly as intermediate features to learn the V-A interdependencies (correlations) are introduced next. Performance comparison shows that the proposed output-associative fusion framework provides a significantly improved prediction accuracy compared to feature-level and model-level fusion via BLSTM-NNs.

5 Data Set and Preprocessing

5.1 Data Set

We use the Sensitive Artificial Listener Database [17] that contains spontaneous data capturing the audiovisual interaction between a human and an operator undertaking the role of an avatar with four personalities: Poppy (happy), Obadiah (gloomy), Spike (angry), and Prudence (pragmatic).

The audiovisual sequences have been recorded at a video rate of 25 fps (352 $\times$ 288 pixels) and at an audio rate of 16 kHz. The recordings were made in a lab setting, using one camera, a uniform background, and constant lighting conditions. The SAL data have been annotated by a set of coders who provided continuous annotations with respect to valence and arousal dimensions using the FeelTrace annotation tool [13]. Feeltrace allows coders to watch the audiovisual recordings and move their cursor, within the 2D emotion space (valence and arousal) confined to $[-1, 1]$, to rate their impression about the emotional state of the subject. Although there are approximately 10 hours of footage available in the SAL database, V-A annotations have only been obtained for two female and two male subjects. We used this portion for our experiments.

5.2 Challenges

Using spontaneous and naturalistic data that have been manually annotated along a continuum presents us with a set of challenges which essentially motivate the adopted methodology.

The first issue is known as *reliability of ground truth*. In other words, achieving agreement among the coders (or observers) that provide annotations in a dimensional space is very challenging [27]. In order to make use of the manual annotations for automatic recognition, most researchers take the mean of the observers ratings, or assess the annotations manually. In Section 5.3, we describe the process of producing the ground truth with respect to the coders' annotations in order to maximize the intercoder agreement.

The second issue is known as *the baseline problem*. This is also known as the concept of having "a condition to compare against" in order for the automatic recognizer to successfully learn the recognition problem at hand [27]. For instance, in the context of acted (posed) facial expression recognition, the subjects are instructed to express a certain emotional state starting (and ending) with an expressionless face. Thus, posed affect data contain all the temporal transitions (neutral-onset-apex-offset-neutral) that provide a classifier with a sequence that begins and ends with an expressionless display: the baseline. Since such expressionless states are not guaranteed to be present in spontaneous data [27], [40], we use the transition *to* and *from* an emotional state (i.e., the frames where the emotional state changes) as the baseline to compare against.

The third issue refers to *unbalanced data*. In naturalistic settings, it is very difficult to elicit balanced amount of data for each emotion dimension. For instance, Caridakis et al. [9] reported that a bias toward quadrant 1 (positive arousal, positive valence) exists in the SAL database. Other researchers (e.g., [12]) handle the issue of unbalanced classes by imposing equal a priori probability. As classification results strongly depend on the a priori probabilities of

class appearance, we attempt to tackle this issue by automatically presegmenting the data at hand. More specifically, the segmentation stage consists of producing (approximately equal number of) negative and positive audiovisual segments with a temporal window that contains an offset before and after the displayed expression (i.e., the baseline).

5.3 Data Preprocessing and Segmentation

The data preprocessing and segmentation stage consists of 1) producing ground truth by maximizing intercoder agreement, 2) eliciting frames that capture the transition *to* and *from* an emotional state, and 3) automatic segmentation of spontaneous audiovisual data. A detailed description of these procedures is presented in [46].

In general, the V-A annotations of each coder are not in total agreement, mostly due to the variance in human observers' perception and interpretation of emotional expressions. Thus, in order to deem the annotations comparable, we normalized the data and provided some compensation for the synchronization issues. We experimented with various normalization techniques and opted for the one that minimized the intercoder mean squared error (MSE). To tackle the synchronization issues, we allow the time shifting of the annotations for each specific segment up to a threshold of 0.5 seconds given that this increases the agreement between coders.

In summary, achieving agreement from all participating coders is difficult and not always possible for each extracted segment. Thus, we use the intercoder correlation to obtain a measure of how similar one coder's annotations are to the rest. This is then used as a weight to determine the contribution of each coder to the ground truth.

More specifically, the averaged correlation cor'_{S,c_j} assigned to coder c_j is defined as follows:

$$cor'_{S,c_j} = \frac{1}{|S|-1} \sum_{i \in S, c_i \neq c_j} cor(c_i, c_j), \quad (1)$$

where S is the relevant session annotated by $|S|$ number of coders and each coder annotating S is defined as $c_i \in S$.

Typically, an automatically produced segment consists of a single interaction of the subject with the avatar (operator), starting with the final seconds of the avatar speaking, continuing with the subject responding (and thus reacting and expressing an emotional state), and concluding where the avatar starts responding. Given that in naturalistic data, emotional expressions are not generally preceded by neutral emotional states [27], [40], we considered this window to provide the best baseline possible. For more details, we refer the reader to [46]. It should be noted that this method is purely based on the annotations, unlike other methods which are based on features, e.g., turn-based segmentation based on voice activity detection [42].

6 Feature Extraction

In this section, we describe the audio and visual features that have been extracted using the automatically segmented audiovisual SAL data.

Fig. 2. Illustration of tracked points of (left) the face ($T_{f1} - T_{f20}$) and (right) the shoulders ($T_{s1} - T_{s5}$).

6.1 Audio Features

Our audio features include Mel-frequency Cepstrum Coefficients [33] and prosody features (the energy of the signal, the Root Mean Squared Energy, and the pitch obtained by using a Praat pitch estimator [51]). Mel-frequency Cepstrum (MFC) is a representation of the spectrum of an audio sample which is mapped onto the nonlinear mel scale of frequency to better approximate the human auditory system's response. The MFCC collectively make up the MFC for the specific audio segment.

We used six cepstrum coefficients, thus obtaining six MFCC and six MFCC-Delta features for each audio frame. We have essentially used the typical set of features used for automatic affect recognition [67], [52]. Along with pitch, energy, and RMS energy, we obtained a set of features with dimensionality $d = 15$ (per audio frame). Note that we used a 0.04 second window with a 50 percent overlap (i.e., first frame 0-0.04, second from 0.02-0.06, etc.) in order to obtain a double frame rate for audio (50 Hz) compared to that of video (25 fps). This is an effective and straightforward way to synchronize the audio and video streams.

6.2 Facial Expression Features

To capture the facial motion displayed during a spontaneous expression, we track 20 facial feature points (FFP), as illustrated in Fig. 2. These points are the corners of the eyebrows (four points), eyes (eight points), nose (three points), the mouth (four points), and the chin (one point). To track these facial points, we used the Patras-Pantic particle filtering tracking scheme [50]. Prior to tracking, initialization of the facial feature points has been done using the method introduced in [61]. For each video segment containing n frames, we obtain a set of n vectors containing 2D coordinates of the 20 points tracked in n frames ($T_f = \{T_{f1} \dots T_{f20}\}$ with dimensions $n * 20 * 2$).

6.3 Shoulder Features

The motion of the shoulders is captured by tracking two points on each shoulder and one stable point on the torso, usually just below the neck (see Fig. 2). The points to be tracked are initialized manually in the first frame. We then use the standard Auxiliary Particle Filtering [53] to track the shoulder points. This scheme is less complex and faster compared to the Patras-Pantic particle filtering tracking scheme; it does not require learning the model of prior probabilities of the relative positions of the shoulder points, while resulting in sufficiently high accuracy. The shoulder tracker results in a set of points $T_s = \{T_{s1} \dots T_{s5}\}$ with dimensions of $n * 5 * 2$.

The SAL database consists of challenging data with sudden body movements and out-of-plane head rotations. As the focus of this paper is on dimensional and continuous affect prediction, we would like to minimize the effect of imperfect and noisy point tracking on the automatic prediction. Therefore, both facial point tracking and shoulder point tracking have been done in a semi-automatic manner (with manual correction when tracking is imperfect).

7 Dimensional Affect Prediction

7.1 Bidirectional Long Short-Term Memory Neural Networks

The traditional Recurrent Neural Networks (RNN) are unable to learn temporal dependencies longer than a few time steps due to the vanishing gradient problem [31]. LSTM Neural Networks (LSTM-NNs) were introduced by Graves and Schmidhuber [25] to overcome this issue. Analysis of the error flow [30] has shown that the backpropagated error in RNNs either grows or decays exponentially. LSTMs introduce recurrently connected memory blocks instead of traditional neural network nodes, which contain memory cells and a set of multiplicative gates. The gates essentially allow the network to learn when to maintain, replace, or reset the state of each cell. As a result, the network can learn when to store or relate to context information over long periods of time, while the application of nonlinear functions (similar to transfer functions in traditional NN) enables learning *nonlinear dependencies*.

Traditional RNNs process input in a temporal order, thus learning characteristics of the input by relating only to past context. Bidirectional RNNs (BRNNs) [58], [3] instead modify the learning procedure to overcome the latter issue of the past and future context: They present each of the training sequences in a forward and a backward order (to two different recurrent networks, respectively, which are connected to a common output layer). In this way, the BRNN is aware of both future and past events in relation to the current time step. The concept is directly expanded for LSTMs, referred to as Bidirectional Long Short-Term Memory neural networks.

BLSTM-NNs have been shown to outperform unidirectional LSTM-NN for speech processing (e.g., [25]) and have been used for many learning tasks. They have been successfully applied to continuous emotion recognition from speech (e.g., [42], [64]), proving that modeling the sequential inputs and long range temporal dependencies appears to be beneficial for the task of automatic emotion prediction.

To the best of our knowledge, to date, BLSTM-NNs have only been used for affect prediction from the audio modality (e.g., [42]). No effort has been reported so far on using BLSTM-NNs for prediction of affect from visual modality or multiple cues and modalities.

7.2 Support Vector Regression

SVM for regression [18] is one of the most dominant kernel methods in machine learning. A nonlinear function is learned by the model in a mapped feature space, induced by the kernel used. An important advantage of SVMs is the convex optimization function employed which guarantees that the optimal solution is found. The goal is to optimize the generalization bounds for regression by a loss function which is used to weight the actual error of the point with respect to the distance from the correct prediction.

Various loss functions could be used to this aim (e.g., quadratic loss function, Laplacian loss function, and ϵ-insensitive loss function). The ϵ-insensitive loss function, introduced by Vapnik, is an approximation of the Huber loss function and enables a more reliable generalization bound [16]. This is due to the fact that unlike the Huber and quadratic loss functions (where all the data will be support vectors), the support vectors can be sparse with the ϵ-insensitive loss function. Sparse data representations have been shown to reduce the generalization error [60] (see [57, chapter 3.3] for details).

In this work, we employ ϵ-insensitive regression that is based on the idea that all points that fall within the ϵ-band have a zero cost. The ones outside the band have a cost assigned which is relative to their distance measured by the variables.

We choose to use SVRs in our experiments due to the fact that they are commonly employed in works reporting on continuous affect prediction (e.g., [42], [26], [34]).

7.3 Evaluation Metrics

Finding optimal evaluation metrics for dimensional and continuous emotion prediction and recognition remains an open research issue [27]. The mean squared error is the most commonly used evaluation measure by related work in the literature (e.g., [42], [26], [34]), while correlation coefficient (COR) is also employed by several studies (e.g., [26], [34]).

MSE evaluates the prediction by taking into account the squared error of the prediction from the ground truth. Let $\hat{\theta}$ be the prediction and θ be the ground truth. MSE is then defined as

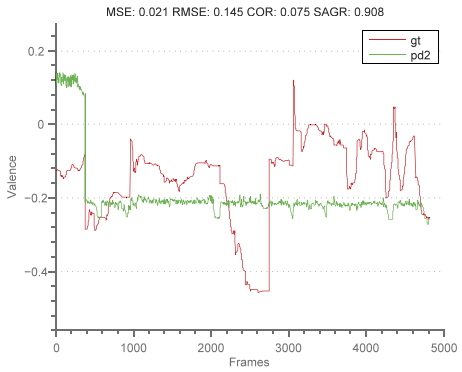
$$MSE = \frac{1}{n}\sum_{f=1}^{n}(\hat{\theta}(f) - \theta(f))^2 = \sigma_{\hat{\theta}}^2 + E([\hat{\theta} - \theta])^2. \quad (2)$$

As can be seen from the equation, MSE is the sum of the variance and the squared bias of the predictor, where E is the expected value operator. Therefore, the MSE provides an evaluation of the predictor based on its variance and bias. This also applies for other MSE-based metrics, such as the root mean squared error (RMSE), defined as

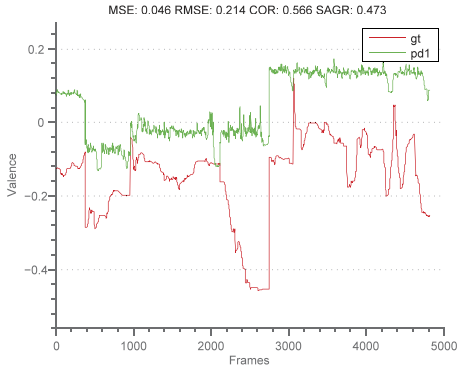
$$\mathrm{RMSE} = \sqrt{\mathrm{MSE}}.$$

In this work, we use the RMSE since it is measured in the same units as our actual data (as opposed to the squared units measuring MSE). MSE-based evaluation has been criticized for heavily weighting outliers [5]. Most importantly, it is unable to provide any structural information regarding how θ and $\hat{\theta}$ change together, i.e., the covariance of these values. The correlation coefficient that we employ for evaluating the prediction and ground truth compensates for the latter and is defined as follows:

$$COR(\hat{\theta}, \theta) = \frac{COV\{\hat{\theta}, \theta\}}{\sigma_{\hat{\theta}}\sigma_{\theta}} = \frac{E[(\hat{\theta} - \mu_{\hat{\theta}})(\theta - \mu_{\theta})]}{\sigma_{\hat{\theta}}\sigma_{\theta}}, \quad (3)$$



(a)



(b)

Fig. 3. Illustration of how MSE-based (both MSE and RMSE), COR, and SAGR evaluation metrics provide different results for two different predictions on the same sequence (gt: ground truth, pd: prediction).

where σ stands for the standard deviation, COV stands for the covariance, while μ symbolizes the mean (expected value).

COR provides an evaluation of the *linear* relationship between the prediction and the ground truth, and subsequently, an evaluation of whether the model has managed to capture linear structural patterns inhibited in the data at hand. As for the covariance calculation, since the means are subtracted from the values in question, it is independent of the bias (and differs from the MSE-based evaluation).

In addition to the two aforementioned metrics, we propose the use of another metric which can be seen as *emotion-prediction-specific*. Our aim is to obtain an agreement level of the prediction with the ground truth by assessing the valence dimension, as being positive $(+)$ or negative $(-)$, and the arousal dimension, as being active $(+)$ or passive $(-)$. Based on this heuristic, we define a sign agreement metric (SAGR) as follows:

$$SAGR = \frac{1}{n}\sum_{f=1}^{n}\delta_{(sign(\hat{\theta}(f)),sign(\theta(f)))}, \tag{4}$$

where δ is the Kronecker delta function, defined as

$$\delta_{(a,b)} = \begin{cases} 1, & a = b, \\ 0, & a \neq b. \end{cases} \tag{5}$$

As a proof of concept, we provide two cases from our experiments that demonstrate how each evaluation metric contributes to the evaluation of the prediction with respect to the ground truth. In Fig. 3, we present two suboptimal predictions from audio cues, for the valence dimension, using two BLSTM-NNs with different topologies. Notice how each metric informs us of a specific aspect of the prediction. The MSE of Fig. 3a is smaller than Fig. 3b, demonstrating that the first case is numerically closer to the ground truth than the second case. Despite this fact, the first prediction does not seem to follow the ground truth structurally; it rather fluctuates around the mean of the prediction (generating a low bias). This is confirmed by observing COR, which is significantly higher for the second prediction case (0.566 versus 0.075). Finally, SAGR demonstrates that the first prediction case is in high agreement with the ground truth in terms of classifying the emotional states as negative or positive. In summary, we conclude that a high COR accompanied by a large MSE is undesirable, as well as a high SAGR accompanied by a large MSE (such observations also apply to the RMSE metric).

Our empirical evaluations show that there is an inherent trade-off involved in the optimization of these metrics. By using all three metrics simultaneously, we attain a more detailed and complete evaluation of predictor versus ground truth, i.e., 1) a variance-and-bias-based evaluation with MSE (how much prediction and ground-truth values vary), 2) a structure-based evaluation with COR (how closely the prediction follows the structure of the ground truth), and 3) emotion-prediction-specific evaluation with SAGR (how much prediction and ground truth agree on the positive versus negative, and active versus passive aspect of the exhibited expression).

7.4 Single-Cue Prediction

The first step in our experiments consists of prediction based on single cues. Let $\mathcal{D} = \{V, A\}$ represent the set of dimensions, $\mathcal{C}$ the set of cues consisting of the facial expressions, shoulder movement, and audio cues. Given a set of input features $\mathbf{x_c} = [\mathbf{x_{1_c}}, \ldots, \mathbf{x_{n_c}}]$, where n is the training sequence length and $c \in \mathcal{C}$, we train a machine learning technique f_d in order to predict the relevant dimension output, $\mathbf{y_d} = [y_1, \ldots, y_n]$, $d \in \mathcal{D}$:

$$f_d : \mathbf{x} \mapsto y_d. \tag{6}$$

This step provides us with a set of predictions for each machine learning technique and each relevant dimension employed.

7.5 Feature-Level Fusion

Feature-level fusion is obtained by concatenating all of the features from multiple cues into one feature vector which is then fed into a machine learning technique.

In our case, the audio stream has a double frame rate with respect to the video stream (50 Hz versus 25 fps). When fusing audio and visual features (shoulder or facial expression cues) at the feature level, each video feature vector is repeated twice, and the ground truth for the audio cues is then used for training and evaluation. This practice is in accordance with similar works in the field that focus on human behavior understanding from audio-visual data (e.g., [52]).

7.6 Model-Level Fusion

In the decision-level data fusion, the input coming from each modality and cue is modeled independently, and these

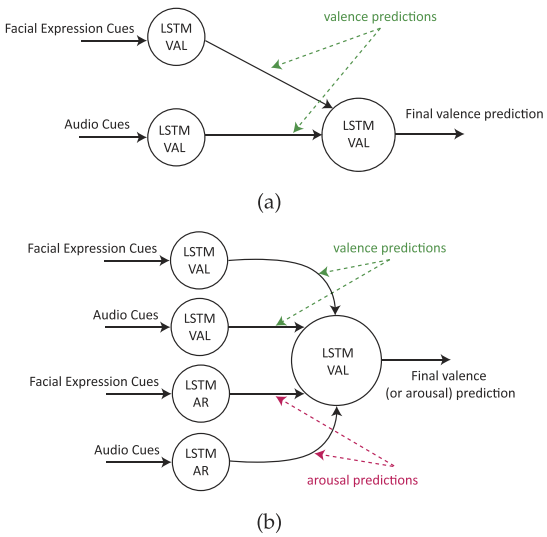


Fig. 4. Illustration of (a) model-level fusion and (b) output-associative fusion using facial expression and audio cues. Model-level fusion combines valence predictions from facial expression and audio cues by using a third network for the final valence prediction. Output-associative fusion combines both valence and arousal values predicted from facial expression and audio cues, again by using a third network which outputs the final prediction.

single-cue and single-modal recognition results are combined in the end. Since humans display multicue and multimodal expressions in a complementary and redundant manner, the assumption of conditional independence between modalities and cues in decision-level fusion can result in loss of information (i.e., mutual correlation between the modalities). Therefore, we opt for model-level fusion of the continuous predictions as this has the potential of capturing correlations and structures embedded in the continuous output of the regressors (from different sets of cues). This is illustrated in Fig. 4a.

More specifically, during model-level fusion, a function learns to map predictions to a dimension d from the set of cues as follows:

$$f_{mlf} : f_d(\mathbf{x_1}) \times \cdots \times f_d(\mathbf{x_m}) \mapsto y_d, \tag{7}$$

where m is the total number of fused cues.

7.7 Output-Associative Fusion

In the previous sections, we have treated the prediction of valence or arousal as a 1D regression problem. However, as mentioned in Section 2, psychological evidence shows that valence and arousal dimensions are correlated [48], [1], [65].

In order to exploit these correlations and patterns, we propose a framework capable of learning the dependencies that exist among the predicted dimensional values. We use BLSTM-NN as the basis for this framework as they appear to outperform SVR in the prediction task at hand (see Section 8). Given the setting described in Section 7.4, this framework learns to map the outputs of the intermediate predictors (each BLSTM-NN as formulated in (6)) onto a higher (and final) level of prediction by incorporating cross-dimensional (output) dependencies (see Fig. 4b). This method, which we call *output-associative fusion*, can be represented by a function f_{oaf}:

$$f_{oaf} : f_{Ar}(\mathbf{x_1}) \times f_{Val}(\mathbf{x_1}) \times \cdots \times f_{Ar}(\mathbf{x_m}) \times f_{Val}(\mathbf{x_m}) \mapsto y_d, \tag{8}$$

where m is again the total number of fused cues.

As a result, the final output, taking advantage of the temporal and bidirectional characteristics of the regressors (BLSTM-NNs), depends not only on the entire sequence of input features $\mathbf{x_i}$ but also on the entire sequence of intermediate output predictions $\mathbf{f_d}$ of both dimensions (see Fig. 4b).

7.8 Experimental Setup

Prior to experimentation, all features used for training the machine learning techniques have been normalized to the range of $[-1, 1]$, except for the audio ones, which have been found to perform better with z-normalization (i.e., normalizing to $\text{mean} = 0$ and standard $\text{deviation} = 1$).

For validation purposes, we use a subset of the SAL-DB that consists of 134 audiovisual segments (a total of 30,042 video frames) obtained by the automatic segmentation procedure (see [46]). We employ subject-dependent leave-one-out-validation evaluation as most of the works in the field report only on subject-dependent dimensional emotion recognition when the number of subjects and data are limited (e.g., [42]).

For automatic dimensional affect prediction, we employ two state-of-the-art machine learning techniques: Support Vector Machines for Regression and bidirectional Long Short-Term Memory Neural Networks. Experimenting with SVR and BLSTM-NN requires that various parameters within these learning methods are configured and the interaction effect between various parameters is investigated. For SVR, we experiment with Radial Basis Function (RBF) kernels ($e^{(-\gamma\|\mathbf{x}-\mathbf{x}'\|^2)}$) as the results outperformed our initial polynomial kernel experiments. To this aim, kernel specific parameters, such as the γ RBF kernel parameter (which determines how closely the distribution of the data is followed) and the polynomial kernel degree, as well as generic parameters, including the outlier cost C, the tolerance of termination, and the ϵ of the loss function, need to be optimized. We perform a grid search (using the training set) and select the best performing set of parameters to be used.

The respective parameter optimization for BLSTM-NNs refers to mainly determining the topology of the network along with the number of epochs, momentum, and learning rate. Our networks typically have one hidden layer and a learning rate of 10^{-4}. The momentum is varied in the range of $[0.5, 0.9]$. All of these parameters can be determined by optimizing on the given training set (e.g., by keeping a validation set aside) while avoiding overfitting.

8 Experimental Results

8.1 Single-Cue Prediction

To evaluate the performance of the employed learning techniques for continuous affect prediction, we first experiment with single cues. Table 3 presents the results of applying BLSTM-NN and SVR (with radial basis function kernels) for the prediction of valence and arousal dimensions.

We initiate our analysis with the valence dimension. From both BLSTM-NNs and SVR, it is obvious that the

TABLE 3
Single-Cue Prediction Results for Valence and Arousal Dimensions (F: Facial Expressions, S: Shoulder Cues, A: Audio)

		BLSTM-NN			SVR		
		RMSE	COR	SAGR	RMSE	COR	SAGR
Valence	F	**0.17**	**0.712**	**0.841**	0.21	0.551	0.740
	S	**0.21**	**0.592**	**0.781**	0.25	0.389	0.718
	A	**0.22**	**0.444**	**0.648**	0.25	0.146	0.538
Arousal	F	**0.25**	**0.493**	**0.681**	0.27	0.418	0.700
	S	0.29	**0.411**	**0.687**	**0.27**	0.388	0.667
	A	**0.24**	**0.586**	**0.764**	0.26	0.419	0.716

visual cues appear more informative than audio cues. Facial expression cues provide the highest correlation with the ground truth $(\mathrm{COR} = 0.71)$ compared to shoulder cues $(\mathrm{COR} = 0.59)$ and audio cues $(\mathrm{COR} = 0.44)$. This fact is also confirmed by the RMSE and SAGR values. Facial expression cues provide the highest SAGR (0.84) indicating that the predictor was accurate in predicting an emotional state as positive or negative for 84 percent of the frames.

Works on automatic affect recognition from audio have reported that arousal can be much better predicted than valence using audio cues [26], [59]. Our results are in agreement with such findings for prediction of the arousal dimension audio cues appear to be superior to visual cues. More specifically, audio cues (using BLSTM-NNs) provide $\mathrm{COR} = 0.59$, $\mathrm{RMSE} = 0.24$, and $\mathrm{AGR} = 0.76$. The facial expression cues provide the second best results with $\mathrm{COR} = 0.49$, while the shoulder cues are deemed less informative for arousal prediction. These findings are also confirmed by the SVR results.

From Table 3, we also obtain a comparison of the performance of the employed learning techniques. We clearly observe that BLSTM-NNs outperform SVRs. In particular, COR and SAGR metrics provide better results for BLSTM-NNs (for all cues and all dimensions). The RMSE metric also confirms these findings except for the prediction of arousal from shoulder cues. Overall, we conclude that capturing temporal correlations and *remembering* the temporally distant events (or storing them in memory) is of utmost importance for continuous affect prediction.

8.2 Multicue and Multimodal Fusion

The experiments in the previous section have demonstrated that using BLSTM-NNs provide better results (for all cues and all dimensions) than using SVRs. Therefore, BLSTM-NNs are employed for feature-level and model-level fusion, as well as output-associative fusion (described in Section 7.7). Experimental results are presented in Table 4, along with the statistical significance test results. We performed statistical significance tests (t-test) using $alpha = 0.05$ (95 percent confidence interval). We performed t-tests to compare the RMSE results of the proposed output-associative fusion to that of the best of model-level or feature-level fusion result (for each cue combination). Table 4 shows the significant results marked with a †. Overall, the output-associative fusion appears to be *significantly* better than the other fusion methods, except for prediction of valence from face-shoulder and shoulder-audio cue combinations.

Looking at Table 4, feature-level fusion appears to be the worst performing fusion method for the task and data at hand. Although, in theory, the cross-cue temporal correlations can be exploited by feature-level fusion, this does not seem to be the case for the problem at hand. This is possibly due to the increased dimensionality of the feature vector along with synchronicity issues between the fused cues.

In general, model-level fusion provides better results than feature-level fusion. This can be justified by the fact that the BLSTM-NNs are able to learn temporal dependencies and structural characteristics manifesting in the continuous output of each cue. Model-level fusion appears to be much better for predicting the valence dimension rather than the arousal dimension. This is mainly due to the fact that the single-cue predictors for valence dimension perform better, thus containing more correct temporal dependencies and structural characteristics (while the weaker arousal predictors contain less of these dependencies). Both fusion techniques reconfirm that visual cues are more informative for valence dimension than audio cues. Finally, the fusion of all cues and modalities provides us with the best (most accurate) results.

Regarding the arousal dimension, we observe that the performance gap between model-level and feature-level fusion is smaller compared to that of valence dimension. For instance, for the fusion of face and shoulder cues, the feature-level fusion provided better COR and SAGR results (but a worse RMSE) than model-level fusion.

Facial expression and audio cues have been the best performing single cues for continuous emotion prediction

TABLE 4
Fusion Results for the Three Methods Employed

		output-associative			model-level			feature-level		
		RMSE	COR	SAGR	RMSE	COR	SAGR	RMSE	COR	SAGR
Valence	FS	0.15	**0.777**	**0.89**	0.16	0.774	**0.890**	0.19	0.676	0.845
	SA	**0.18**	**0.664**	0.825	0.19	0.653	**0.830**	0.21	0.583	0.733
	FA	**0.16**†	**0.760**	**0.892**	0.17	0.748	0.856	0.20	0.604	0.790
	FSA	**0.15**†	**0.796**	**0.907**	0.16	0.782	0.892	0.19	0.681	0.856
	coders	0.141	0.85	0.86	0.141	0.85	0.86	0.141	0.85	0.86
Arousal	FS	**0.24**†	**0.536**	**0.719**	0.25	0.479	0.666	0.27	0.508	0.731
	SA	**0.23**†	**0.602**	**0.763**	0.26	0.567	0.637	0.28	0.461	0.685
	FA	**0.22**†	**0.628**	0.800	0.23	0.605	0.800	0.24	0.589	0.763
	FSA	**0.21**†	**0.642**	**0.766**	0.22	0.639	0.763	0.26	0.500	0.700
	coders	0.145	0.87	0.84	0.145	0.87	0.84	0.145	0.87	0.84

The best results are obtained by employing output-associative fusion. Significant results are marked with a †. For comparison purposes, the average agreement level between human coders is also shown in terms of RMSE, COR, and SAGR metrics.

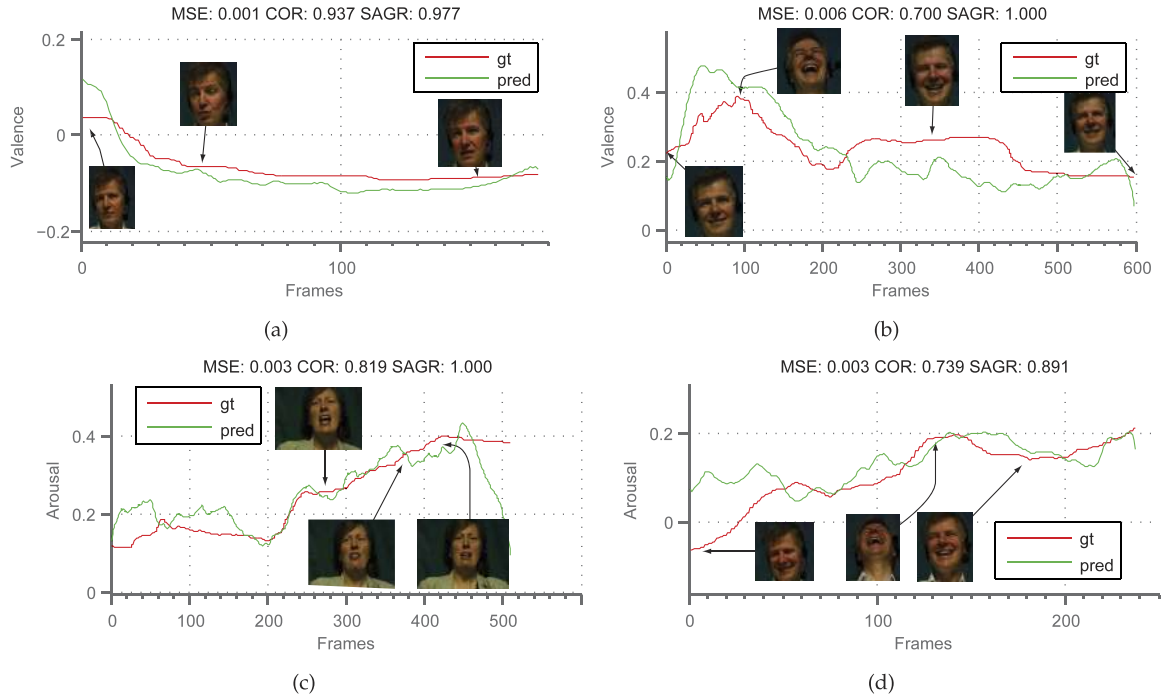


Fig. 5. Example valence (a), (b) and arousal (c), (d) predictions obtained by output-associative fusion (gt: ground truth, pd: prediction).

(see Section 8.1). Therefore, it is not surprising that fusion of these two cues provides the best feature-level fusion results. For model-level fusion instead, the best results are obtained by combining the predictions from all cues and modalities.

Finally, the proposed output-associative fusion provides the best results, outperforming both feature-level and model-level fusion. Similarly to the model-level fusion case, the best results (for both dimensions) are obtained when predictions from all cues and modalities are fused.

We denote that the performance increase of output-associative fusion is higher for the arousal dimension (compared to the valence dimension). This could be justified by the fact that the single-cue predictors for valence perform better than for arousal (Table 3) and thus more correct valence patterns are passed onto the output-associative fusion.

Table 4 also shows the average agreement level between human coders in terms of RMSE, COR, and SAGR metrics (calculated for each dimension separately). It is interesting to note that when predicting the valence dimension, the proposed output-associative fusion 1) appears to outperform the average human coder in terms of SAGR criterion, and 2) provides prediction results that are relatively close to human coders (in terms of RMSE and COR).

In Fig. 5, we illustrate a set of predictions obtained via output-associative fusion. As can be observed from the figure, the prediction results closely follow the structure and the values of the ground truth.

Overall, the temporal dynamics of spontaneous multimodal behavior (e.g., when a facial or a bodily expression starts, reaches an apex, and ends) have not received much attention in the affective and behavioral science research fields. More specifically, it is virtually unknown whether and how the temporal dynamics of various communicative cues are interrelated (e.g., whether a smile reaches its apex while the person is shrugging his shoulders). The facial, shoulder, and audio cues explored in this paper possibly have different temporal dynamics. Accordingly, the BLSTM-NNs are able to incorporate and model the temporal dynamics of each modality independently (and appropriately) in the output-associative and model-level fusion schemes. This may be one reason why output-associative and model-level fusion appear to perform better than feature-level fusion.

9 Conclusions

The affect sensing and recognition field has recently shifted its focus toward subtle, continuous, and context-specific interpretations of affective displays recorded in naturalistic and real-world settings and toward combining multiple modalities for automatic analysis and recognition. The work presented in this paper converges with this recent shift by

1. extracting audiovisual segments from databases annotated in dimensional affect space and automatically generating the ground truth,
2. fusing facial expressions, shoulder, and audio cues for dimensional and continuous prediction of emotions,
3. experimenting with state-of-the-art learning techniques such as BLSTM-NNs and SVRs, and
4. incorporating correlations between valence and arousal values via output-associative fusion to improve continuous prediction of emotions.

Based on the experimental results provided in Section 8, we are able to conclude the following:

- Arousal can be much better predicted than valence using audio cues. For valence dimension instead, visual cues (facial expressions and shoulder movements) appear to perform better. This has also been confirmed by other related work on dimensional

emotion recognition [42], [26], [59]. Whether such conclusions hold for different context and different data remain to be evaluated.

- Emotional expressions change over the course of time, and usually have start, peak, and end points (temporal dynamics). It appears that such temporal aspects (dynamics) are crucial in predicting both valence and arousal dimensions. A learning technique, such as the BLSTM-NNs, that can exploit these aspects, appears to outperform SVR (the static learning technique at hand).
- When working with temporal and structured emotion data, choosing predictors that are able to optimize not only the variance (of the predictor) and the bias (to the ground truth), but also the covariance of the prediction (with respect to the ground truth) is crucial for the prediction task at hand. Emotion-specific metrics (such as SAGR) that carry valuable information regarding the emotion-specific aspects of the prediction are also desirable.
- As confirmed by the psychological theory, valence and arousal are correlated. Such correlations appear to exist in our data where fusing predictions from both valence and arousal dimensions (via output-associative fusion) improves the results compared to using predictions from either valence or arousal dimension alone (both for feature-level and model-level fusion).
- In general, multimodal data appear to be more useful for predicting valence than for predicting arousal. While arousal is better predicted by using audio features alone, valence is better predicted by using multicue and multimodal data.

Overall, we conclude that compared to an average human coder, the proposed system is well able to approximate the valence and arousal dimensions. More specifically, for valence dimension, our output-associative fusion framework approximates the intercoder RMSE (≈ 0.141) and intercoder correlation (0.84) by obtaining an $\mathrm{RMSE} = 0.15$ and $COR \approx 0.8$ (see Table 4). It also achieves a higher SAGR (≈ 0.91) than the intercoder SAGR (0.86).

As future work, the proposed methodology remains to be evaluated on extensive data sets (with a larger number of subjects) annotated using a richer emotional expression space with other continuous dimensions such as power, expectation, and intensity (e.g., the newly released Semaine Database [43]). Moreover, it is possible to exploit the correlations between valence and arousal dimensions inherent in naturalistic affective data utilizing other machine learning techniques. For instance, Nicolaou et al. [47] introduce an output-associative Relevance Vector Machine regression framework that augments the traditional Relevance Vector Machine regression by learning nonlinear input and output dependencies inherent in the affective data. We will focus on exploring such output-associative regression frameworks using unsegmented audiovisual sequences.

Acknowledgments

The research work presented in this paper has been funded by the European Research Council under the ERC Starting Grant agreement no. ERC-2007-StG-203143 (MAHNOB). The previous work of Hatice Gunes was funded by the European Community's Seventh Framework Programme [FP7/2007-2013] under the grant agreement no 211486 (SEMAINE).

References

[1] N. Alvarado, "Arousal and Valence in the Direct Scaling of Emotional Response to Film Clips," *Motivation and Emotion,* vol. 21, pp. 323-348, 1997.

[2] N. Ambady and R. Rosenthal, "Thin Slices of Expressive Behavior as Predictors of Interpersonal Consequences: A Meta-Analysis," *Psychological Bull.,* vol. 11, no. 2, pp. 256-274, 1992.

[3] P. Baldi, S. Brunak, P. Frasconi, G. Pollastri, and G. Soda, "Exploiting the Past and the Future in Protein Secondary Structure Prediction," *Bioinformatics,* vol. 15, pp. 937-946, 1999.

[4] S. Baron-Cohen and T.H.E. Tead, *Mind Reading: The Interactive Guide to Emotion.* Jessica Kingsley Publishers Ltd., 2003.

[5] S. Bermejo and J. Cabestany, "Oriented Principal Component Analysis for Large Margin Classifiers," *Neural Networks,* vol. 14, no. 10, pp. 1447-1461, 2001.

[6] L. Bo and C. Sminchisescu, "Twin Gaussian Processes for Structured Prediction," *Int'l J. Computer Vision,* vol. 87, pp. 28-52, 2010.

[7] L. Bo and C. Sminchisescu, "Structured Output-Associative Regression," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 2403-2410, 2009.

[8] R.A. Calvo and S. DMello, "Affect Detection: An Interdisciplinary Review of Models, Methods, and Their Applications," *IEEE Trans. Affective Computing,* vol. 1, no. 1, pp. 18-37, Jan.-June 2010.

[9] G. Caridakis, K. Karpouzis, and S. Kollias, "User and Context Adaptive Neural Networks for Emotion Recognition," *Neurocomputing,* vol. 71, nos. 13-15, pp. 2553-2562, 2008.

[10] G. Caridakis, L. Malatesta, L. Kessous, N. Amir, A. Raouzaiou, and K. Karpouzis, "Modeling Naturalistic Affective States via Facial and Vocal Expressions Recognition," *Proc. ACM Int'l Conf. Multimodal Interfaces,* pp. 146-154, 2006.

[11] G. Chanel, K. Ansari-Asl, and T. Pun, "Valence-Arousal Evaluation Using Physiological Signals in an Emotion Recall Paradigm," *Proc. IEEE Int'l Conf. Systems, Man and Cybernetics,* pp. 2662-2667, 2007.

[12] G. Chanel, J. Kronegg, D. Grandjean, and T. Pun, "Emotion Assessment: Arousal Evaluation Using EEG's and Peripheral Physiological Signals," *Proc. Multimedia Content Representation, Classification and Security,* pp. 530-537, 2006.

[13] R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, and M. Schroder, "Feeltrace: An Instrument for Recording Perceived Emotion in Real Time," *Proc. ISCA Workshop Speech and Emotion,* pp. 19-24, 2000.

[14] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J.G. Taylor, "Emotion Recognition in Human-Computer Interaction," *IEEE Signal Processing Magazine,* vol. 18, no. 1, pp. 32-80, Jan. 2001.

[15] R. Cowie, H. Gunes, G. McKeown, L. Vaclau-Schneider, J. Armstrong, and E. Douglas-Cowie, "The Emotional and Communicative Significance of Head Nods and Shakes in a Naturalistic Database," *Proc. LREC Int'l Workshop Emotion,* pp. 42-46, 2010.

[16] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods.* Cambridge Univ. Press, Mar. 2000.

[17] E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, L. Lowry, M. McRorie, L. Jean-Claude Martin, J.-C. Devillers, A. Abrilian, S. Batliner, A. Noam, and K. Karpouzis, "The Humaine Database: Addressing the Needs of the Affective Computing Community," *Proc. Second Int'l Conf. Affective Computing and Intelligent Interaction,* pp. 488-500, 2007.

[18] H. Drucker, C.J.C. Burges, L. Kaufman, A.J. Smola, and V. Vapnik, "Support Vector Regression Machines," *Advances in Neural Information Processing Systems,* pp. 155-161, MIT Press, 1996.

[19] P. Ekman, *Emotions in the Human Faces,* second ed. Cambridge Univ. Press, 1982.

[20] P. Ekman and W. Friesen, "Head and Body Cues in Gyrus and Inferior Medial Prefrontal Cortex in Social Perception," *Perceptual & Motor Skills,* vol. 24, pp. 711-724, 1967.

[21] K. Forbes-Riley and D. Litman, "Predicting Emotion in Spoken Dialogue from Multiple Knowledge Sources," *Proc. Human Language Technology Conf. North Am. Chapter of the Assoc. Computational Linguistics,* pp. 201-208, 2004.

[22] N. Fragopanagos and J.G. Taylor, "Emotion Recognition in Human-Computer Interaction," *Neural Networks,* vol. 18, no. 4, pp. 389-405, 2005.

[23] D. Glowinski, A. Camurri, G. Volpe, N. Dael, and K. Scherer, "Technique for Automatic Emotion Recognition by Body Gesture Analysis," *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition Workshops,* pp. 1-6, 2008.

[24] D. Grandjean, D. Sander, and K.R. Scherer, "Conscious Emotional Experience Emerges as a Function of Multilevel, Appraisal-Driven Response Synchronization," *Consciousness and Cognition,* vol. 17, pp. 484-495, 2008.

[25] A. Graves and J. Schmidhuber, "Framewise Phoneme Classification with Bidirectional LSTM and Other Neural Network Architectures," *Neural Networks,* vol. 18, pp. 602-610, 2005.

[26] M. Grimm and K. Kroschel, "Emotion Estimation in Speech Using a 3D Emotion Space Concept," *Proc. IEEE Automatic Speech Recognition and Understanding Workshop,* pp. 381-385, 2005.

[27] H. Gunes and M. Pantic, "Automatic, Dimensional and Continuous Emotion Recognition," *Int. J. Synthetic Emotions,* vol. 1, no. 1, pp. 68-99, 2010.

[28] H. Gunes and M. Pantic, "Dimensional Emotion Prediction from Spontaneous Head Gestures for Interaction with Sensitive Artificial Listeners," *Proc. Int'l Conf. Intelligent Virtual Agents,* pp. 371-377, 2010.

[29] 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, I-Tech Education and Publishing, 2008.

[30] S. Hochreiter, "Untersuchungen zu Dynamischen Neuronalen Netzen," diploma thesis, Institut Für Informatik, Lehrstuhl Prof. Brauer, Technische Universität München, 1991.

[31] S. Hochreiter, "The Vanishing Gradient Problem during Learning Recurrent Neural Nets and Problem Solutions," *Int'l J. Uncertainty, Fuzziness and Knowledge-Based Systems,* vol. 6, no. 2, pp. 107-116, 1998.

[32] S. Ioannou, A. Raouzaiou, V. Tzouvaras, T. Mailis, K. Karpouzis, and S. Kollias, "Emotion Recognition through Facial Expression Analysis Based on a Neurofuzzy Method," *J. Neural Networks,* vol. 18, no. 4, pp. 423-435, 2005.

[33] D. Jurafsky and J.H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition,* second ed. Prentice Hall, 2008.

[34] I. Kanluan, M. Grimm, and K. Kroschel, "Audio-Visual Emotion Recognition Using an Emotion Recognition Space Concept," *Proc. 16th European Signal Processing Conf.,* 2008.

[35] 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," *Lecture Notes in Artificial Intelligence,* vol. 4451, pp. 92-116, 2007.

[36] J. Kim, "Bimodal Emotion Recognition Using Speech and Physiological Changes," *Robust Speech Recognition and Understanding,* pp. 265-280, I-Tech Education and Publishing, 2007.

[37] A. Kleinsmith and N. Bianchi-Berthouze, "Recognizing Affective Dimensions from Body Posture," *Proc. Int'l Conf. Affective Computing and Intelligent Interaction,* pp. 48-58, 2007.

[38] D. Kulic and E.A. Croft, "Affective State Estimation for Human-Robot Interaction," *IEEE Trans. Robotics,* vol. 23, no. 5, pp. 991-1000, Oct. 2007.

[39] R. Lane and L. Nadel, *Cognitive Neuroscience of Emotion.* Oxford Univ. Press, 2000.

[40] R. Levenson, "Emotion and the Autonomic Nervous System: A Prospectus for Research on Autonomic Specificity," *Social Psychophysiology and Emotion: Theory and Clinical Applications,* pp. 17-42, John Wiley & Sons, 1988.

[41] P.A. Lewis, H.D. Critchley, P. Rotshtein, and R.J. Dolan, "Neural Correlates of Processing Valence and Arousal in Affective Words," *Cerebral Cortex,* vol. 17, no. 3, pp. 742-748, 2007.

[42] M. Wöllmer, F. Eyben, S. Reiter, B. Schuller, C. Cox, E. Douglas-Cowie, and R. Cowie, "Abandoning Emotion Classes—Towards Continuous Emotion Recognition with Modelling of Long-Range Dependencies," *Proc. Ninth Interspeech Conf.,* pp. 597-600, 2008.

[43] G. McKeown, M. Valstar, R. Cowie, and M. Pantic, "The Semaine Corpus of Emotionally Coloured Character Interactions," *Proc. IEEE Int'l Conf. Multimedia and Expo,* pp. 1079-1084, 2010.

[44] A. Mehrabian and J. Russell, *An Approach to Environmental Psychology.* MIT Press, 1974.

[45] M. Nicolaou, H. Gunes, and M. Pantic, "Audio-Visual Classification and Fusion of Spontaneous Affective Data in Likelihood Space," *Proc. IEEE Int'l Conf. Pattern Recognition,* pp. 3695-3699, 2010.

[46] M. Nicolaou, H. Gunes, and M. Pantic, "Automatic Segmentation of Spontaneous Data Using Dimensional Labels from Multiple Coders," *Proc. LREC Int'l Workshop Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality,* pp. 43-48, 2010.

[47] M. Nicolaou, H. Gunes, and M. Pantic, "Output-Associative RVM Regression for Dimensional and Continuous Emotion Prediction," *Proc. IEEE Int'l Conf. Automatic Face and Gesture Recognition,* 2011.

[48] A.M. Oliveira, M.P. Teixeira, I.B. Fonseca, and M. Oliveira, "Joint Model-Parameter Validation of Self-Estimates of Valence and Arousal: Probing a Differential-Weighting Model of Affective Intensity," *Proc. 22nd Ann. Meeting Int'l Soc. for Psychophysics,* pp. 245-250, 2006.

[49] M. Pantic and L. Rothkrantz, "Toward an Affect Sensitive Multimodal Human-Computer Interaction," *Proc. IEEE,* vol. 91, no. 9, pp. 1370-1390, Sept. 2003.

[50] I. Patras and M. Pantic, "Particle Filtering with Factorized Likelihoods for Tracking Facial Features," *Proc. Int'l Conf. Automatic Face and Gesture Recognition,* pp. 97-104, 2004.

[51] B. Paul, "Accurate Short-Term Analysis of the Fundamental Frequency and the Harmonics-to-Noise Ratio of a Sampled Sound," *Proc. Inst. of Phonetic Sciences,* pp. 97-110, 1993.

[52] S. Petridis, H. Gunes, S. Kaltwang, and M. Pantic, "Static versus Dynamic Modeling of Human Nonverbal Behavior from Multiple Cues and Modalities," *Proc. ACM Int'l Conf. Multimodal Interfaces,* pp. 23-30, 2009.

[53] M.K. Pitt and N. Shephard, "Filtering via Simulation: Auxiliary Particle Filters," *J. Am. Statistical Assoc.,* vol. 94, no. 446, pp. 590-616, 1999.

[54] J.A. Russell, "A Circumplex Model of Affect," *J. Personality and Social Psychology,* vol. 39, pp. 1161-1178, 1980.

[55] K. Scherer, A. Schorr, and T. Johnstone, *Appraisal Processes in Emotion: Theory, Methods, Research.* Oxford Univ. Press, 2001.

[56] K. Scherer, "Psychological Models of Emotion," *The Neuropsychology of Emotion,* pp. 137-162, Oxford Univ. Press, 2000.

[57] B. Schölkopf and A.J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond.* The MIT Press, 2001.

[58] M. Schuster and K.K. Paliwal, "Bidirectional Recurrent Neural Networks," *IEEE Trans. Signal Processing,* vol. 45, no. 11, pp. 2673-2681, Nov. 1997.

[59] K.P. Truong, D.A. Leeuwen van, M.A. Neerincx, and F.M. Jong de, "Arousal and Valence Prediction in Spontaneous Emotional Speech: Felt versus Perceived Emotion," *Proc. Ann. Conf. Int'l Speech Comm. Assoc.,* pp. 2027-2030, 2009.

[60] D. Vrakas and I.P. Vlahavas, *Artificial Intelligence for Advanced Problem Solving Techniques.* IGI Global Snippet, 2008.

[61] D. Vukadinovic and M. Pantic, "Fully Automatic Facial Feature Point Detection Using Gabor Feature Based Boosted Classifiers," *Proc. IEEE Int'l Conf. Systems, Man, and Cybernetics,* vol. 2, pp. 1692-1698, 2005.

[62] J. Wagner, J. Kim, and E. Andre, "From Physiological Signals to Emotions: Implementing and Comparing Selected Methods for Feature Extraction and Classification," *Proc. IEEE Int'l Conf. Multimedia and Expo,* pp. 940-943, 2005.

[63] J. Weston, O. Chapelle, A. Elisseeff, B. Schölkopf, and V. Vapnik, "Kernel Dependency Estimation," Technical Report 98, Aug. 2002.

[64] M. Wöllmer, B. Schuller, F. Eyben, and G. Rigoll, "Combining Long Short-Term Memory and Dynamic Bayesian Networks for Incremental Emotion-Sensitive Artificial Listening," *IEEE J. Selected Topics in Signal Processing,* vol. 4, no. 5, pp. 867-881, Oct. 2010.

[65] Y.-H. Yang, Y.-C. Lin, Y.-F. Su, and H.H. Chen, "Music Emotion Classification: A Regression Approach," *Proc. IEEE Int'l Conf. Multimedia and Expo,* pp. 208-211, 2007.

[66] C. Yu, P.M. Aoki, and A. Woodruff, "Detecting User Engagement in Everyday Conversations," *Proc. Eighth Int'l Conf. Spoken Language Processing,* 2004.

[67] Z. Zeng, M. Pantic, G. Roisman, and T. Huang, "A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 31, no. 1, pp. 39-58, Jan. 2009.

Mihalis A. Nicolaou received the BSc degree in informatics and telecommunications from the University of Athens, Greece, in 2008, and the MSc degree in advanced computing from Imperial College London, United Kingdom, in 2009. He is currently working toward the PhD degree in the Department of Computing at Imperial College as part of a European Union (EU-FP7) project called MAHNOB, aimed at multimodal analysis of human naturalistic nonverbal behavior. He has received various awards during his studies. His current research interests are in the areas of affective computing and pattern recognition, with a particular focus on continuous and dimensional emotion prediction. He is a student member of the IEEE.

Hatice Gunes received the PhD degree in computing sciences from the University of Technology Sydney (UTS), Australia, in September 2007. She is currently a postdoctoral research associate at Imperial College London, United Kingdom, and an honorary associate of UTS. At Imperial College, she has worked on SEMAINE, a European Union (EU-FP7) award-winning project that aims to build a multimodal dialog system which can interact with humans via a virtual character and react appropriately to the user's nonverbal behavior. She has (co)authored more than 35 technical papers in the areas of affective computing and multimodal human-computer interaction. She is a cochair of the EmoSPACE Workshop at IEEE FG 2011, a member of the Editorial Advisory Board for the Affective Computing and Interaction Book (IGI Global, 2011), and a reviewer for numerous journals and conferences in her areas of expertise. She was a recipient of both the Australian Government International Postgraduate Research Scholarship (IPRS) and the UTS Faculty of IT Research Training Stipend from 2004 to 2007. She is a member of the IEEE, the ACM, and the Australian Research Council Network in Human Communication Science.

Maja Pantic is a professor in affective and behavioral computing in the Department of Computing, Imperial College London, United Kingdom, and in the Department of Computer Science, University of Twente, The Netherlands. She is one of the world's leading experts in research on machine understanding of human behavior, including vision-based detection, tracking, and analysis of human behavioral cues like facial expressions and body gestures, and multimodal human affect/mental state understanding. She has published more than 100 technical papers in these areas of research. In 2008, for her research on Machine Analysis of Human Naturalistic Behavior (MAHNOB), she received European Research Council Starting Grant as one of 2 percent of the best young scientists in any research field in Europe. She is also a partner in several FP7 European projects, including the currently ongoing FP7 SSPNet NoE, for which she is the scientific coordinator. She currently serves as the editor-in-chief of *Image and Vision Computing Journal* and as an associate editor for the *IEEE Transactions on Systems, Man, and Cybernetics Part B*. She has also served as the general chair for several conferences and symposia including the IEEE FG 2008 and the IEEE ACII 2009. She is a senior member of the IEEE.