Affective State Level Recognition in Naturalistic Facial and Vocal Expressions

Hongying Meng, *Member, IEEE,* and Nadia Bianchi-Berthouze, *Member, IEEE*

***Abstract*—Naturalistic affective expressions change at a rate much slower than the typical rate at which video or audio is recorded. This increases the probability that consecutive recorded instants of expressions represent the same affective content. In this paper, we exploit such a relationship to improve the recognition performance of continuous naturalistic affective expressions. Using datasets of naturalistic affective expressions (AVEC 2011 audio and video dataset, PAINFUL video dataset) continuously labeled over time and over different dimensions, we analyze the transitions between levels of those dimensions (e.g., transitions in pain intensity level). We use an information theory approach to show that the transitions occur very slowly and hence suggest modeling them as first-order Markov models. The dimension levels are considered to be the hidden states in the Hidden Markov Model (HMM) framework. Their discrete transition and emission matrices are trained by using the labels provided with the training set. The recognition problem is converted into a best path-finding problem to obtain the best hidden states sequence in HMMs. This is a key difference from previous use of HMMs as classifiers. Modeling of the transitions between dimension levels is integrated in a multistage approach, where the first level performs a mapping between the affective expression features and a soft decision value (e.g., an affective dimension level), and further classification stages are modeled as HMMs that refine that mapping by taking into account the temporal relationships between the output decision labels. The experimental results for each of the unimodal datasets show overall performance to be significantly above that of a standard classification system that does not take into account temporal relationships. In particular, the results on the AVEC 2011 audio dataset outperform all other systems presented at the international competition.**

***Index Terms*—Affective computing, continuous emotion recognition, dimensional model of affect, HMM, machine learning, naturalistic affective expressions.**

I. Introduction

In the affective computing field [1], various studies have been carried out to create systems that can recognize the affective states of their user by analyzing their vocal [2], [3], facial [4]–[6], body expressions [7]–[10], touch [11] and even their physiological changes [12]. Most of that work has been done on acted or stereotypical expressions. More recently, there has been a shift toward using naturalistic expressions to create systems that can interact with people in their everyday life (e.g., [11], [13]–[20]), as well as to provide automatic measures for user experience studies [21]–[26] or even platforms for research in affective computing [27].

Naturalistic expressions present a big challenge to the research community because they are less stereotypical and not always fully-fledged expressions. Furthermore, the dynamic of these expressions is more complex, leading to a larger variability in the way affect is expressed. Finally, naturalistic expressions tend to change more slowly than acted expressions. A few datasets of naturalistic expressions have been made available recently, with the aim to support the creation of emotion recognition systems that are usable in a real context.

An example of such a dataset is the FAU dataset [28] containing naturalistic vocal expressions of children playing with Sony's pet Robot Aibo. The FAU dataset was manually segmented and annotated at word level with 11 discrete emotion categories by different human labelers. This dataset was used in various academic events [16], [29], challenging the research community to improve the automatic recognition of continuous naturalistic vocal expressions.

More recently, the research community has been moving toward datasets that are labeled using a dimensional, rather than a categorical approach. An affective dimensional approach allows a more complete description of the emotional state [30]. An example of such a dataset is the AffectME naturalistic corpus [17] containing a large set of naturalistic affective postures of whole body game players. The affective postures of players playing different sport games were manually extracted and annotated accordingly not only to four discrete emotional states but also to four affective dimensions.

An even more interesting case is provided by the combination of both the continuous labeling along the temporal dimension and the labeling along a set of affective dimensions. Among the datasets featuring this type of labeling process, the audio and video dataset used in the AVEC 2011 challenge [31] provides a unique dataset of naturalistic vocal and facial expressions to help investigate the modeling of the transition between levels of affective dimensions over time. Another interesting dataset is the UNBC-McMaster Shoulder Pain Expression Archive Database (also called PAINFUL dataset) [32] containing videos (but no audio) of naturalistic expressions of acute pain. The facial expressions in the videos are labeled

Manuscript received June 9, 2012; revised October 25, 2012 February 21, 2013 accepted September 27, 2012 February 7, 2013, March 11, 2013. Date of publication April 23, 2013; date of current version February 12, 2014. This work was supported in part by the EPSRC under Grant EP/G043507/1: Pain rehabilitation: E/Motion-based automated coaching. The work of H. Meng was supported in part by the Award of the Brunel Research Initiative and Enterprise Fund. This paper was recommended by Associate Editor B. Wolfgang.

H. Meng is with the School of Engineering and Design, Brunel University, Uxbridge, Middlesex UB8 3PH, U.K. (e-mail: hongying.meng@brunel.ac.uk).

N. Bianchi-Berthouze is with UCLIC, University College London, London WC1E 6BT, U.K. (e-mail: n.berthouze@ucl.ac.uk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCYB.2013.2253768

by using the Prkachin and Solomon pain intensity metric [33] based on the activation of core facial action units. All these datasets provide a more natural scenario of everyday life where expressions change continuously.

In this paper, we tackle the problem of affective dimension level recognition by extending the methodology presented in our AVEC 2011 audio subchallenge workshop paper [34], which we showed to outperform the Latent-Dynamic Conditional Random Fields method proposed in [35]. In this methodology, the temporal relationships between consecutive levels of a given affective dimension are analysed and modeled by using a Markov model based approach [36]. The affective dimension level recognition problem is solved through a multistage automatic pattern recognition system where the temporal relationships are modeled through the Hidden Markov Model (HMM) framework. Here, we refine the multi-stage classification architecture and test its performance over four affective dimensions (arousal, valence, expectation, dominance) not only on the audio but also on the video data of the AVEC 2011 dataset. We further test it on the pain intensity dimension by using the PAINFUL dataset in order to get a better understanding of the pros and cons of the approach. In particular, we expand the analysis to understand how the duration of affective expressions affects its performance. Our experimental results show its effectiveness and generalization capabilities as well as highlight some insights on the importance of each stage.

The remainder of the paper is structured as follows. Section II provides a review on related works. Section III analyzes and models the labeling of the naturalistic expression datasets. Sections IV presents the multistage pattern recognition system and its variations. Sections V and VI present the experimental results. The paper ends with a discussion on the approach and some conclusions.

II. Related work

As promising results have been obtained in emotion recognition on acted expressions, it is now necessary to move toward modeling naturalistic expressions [37], [38], [10]. In particular, an important challenge is to create systems that can continuously (i.e., over time) monitor and classify affective expressions into either discrete affective states or continuous affective dimensions.

Various continuous and dimensional emotion recognition systems have been built using machine learning techniques, such as support vector machines (SVM) [31], [39]. The typical approach is to model each unit of expression (e.g., a video frame, a word) independently and to make it a standard classification problem at frame or word level. The results have been very encouraging [31], [37], [39]. Another interesting approach uses the temporal relationship between different concurrent information to provide a better classification over levels of affective dimensions. Eyben *et al.* [40] proposed a string-based prediction model and multimodel fusion of verbal and nonverbal behavioral events for the automatic prediction of human affect in a dimensional space. Recently, Nicolaou *et al.* [41] described a dimensional and continuous prediction method for emotions from naturalistic facial expressions that augments the traditional output-associative relevance vector machine (RVM) regression framework by learning nonlinear input and output dependencies inherent to the affective data.

Long short-term memory (LSTM) recurrent neural networks have also been successfully used for modeling the relationship between observations [42]–[45]. Wöllmer *et al.* [42] first proposed a method based on LSTM recurrent neural networks for continuous emotion recognition that included modeling of long-range dependencies between observations. This method outperformed techniques, such as support vector regression (SVR). Eyben *et al.* [43] used it for audiovisual classification of vocal outbursts in human conversation and the results showed significant improvements over a static approach based on SVM. Nicolaou *et al.* [44] also used LSTM networks to outperform SVR due to their ability to learn past and future contexts. Wöllmer *et al.* [45] used bidirectional long short-term memory (BLSTM) networks to exploit long-range contextual information for modeling the evolution of emotions within a conversation.

HMMs are another method typically used to model processes characterized by temporal relationships and they have been applied for facial expression recognition [46], affective vocal expression recognition [47], [48] and audiovisual affect recognition [49], [50], [51]. Nwe *et al.* [47] used a four-state fully connected HMM to recognize six archetypical emotions from speech, obtaining recognition performance comparable to subjective observers' ratings. Lee *et al.* [52] showed that HMMs produce better results when the unit of recognition is not the entire emotional expression (i.e., from the onset of the expression to its end) but the subunits that compose it as the expression develops and ends (phonemes in their case). HMMs have also been combined into multilevel systems with other machine learning algorithms, such as k-nearest neighbor (k-NN) (e.g., [53]).

Like HMMs, dynamic Bayesian network based methods [54], [55] have been used for facial expression recognition and semantic relationship modeling between facial affect behaviors [56]. They have been used to describe this temporal relationship in the affective states [57] and action units [58] in a probabilistic framework. Latent-dynamic conditional random fields [35] were used to represent indirectly the extrinsic dynamic between emotional labels. The temporal relationships between labels were computed on the basis of the structure of the expressions (e.g., relationship between AUs). More recently, Hammal and Kunz [59] proposed a system for dynamic recognition of spontaneous and nonprototypic pain expressions based on the transferable belief model from video sequences.

All these works show that learning techniques that exploit the relationship between consecutive observations outperformed approaches based on local information only. In particular, LSTM appears to provide higher performance than approaches based on local information, especially when the contextual information is very important but the exact nature of the relationship is not known *a priori* [43]. Similarly, in the HMM and dynamic Bayesian-based methods presented above, the temporal relationship was represented by the transition probabilities between hidden states. The main shortcoming

of these approaches is that the hidden states are unknown and need to be estimated based on assumed probability distributions of the data. Although optimization methods, such as the expectation maximization algorithm, could be used, the estimation is not always accurate because the data might violate the assumptions.

In this paper, we aim to overcome this problem by modeling the transitions (over time) between affective dimension levels as first order Markov models. The temporal sequences of affective dimension levels (i.e., labels) can be defined as the hidden states sequences in the HMM framework. Then the probabilities of these hidden states and their state transitions can be accurately computed from the labels of the training set. In contrast to how HMMs are used in the works above, in our approach the hidden states during the training process are known, as they correspond to the sequence of affective labels in the training set. This approach transforms the dimension level classification problem into a best path-finding optimization problem in the HMM framework. Through a multistage classification approach, the output of a first-stage classification is used as observation sequences for a second-stage classification, modeled as a HMM-based framework. A third classification stage, a decision fusion tool, is then used to boost overall performance. Indeed, it has been shown that multiclassifier systems can outperform traditional approaches while simultaneously reducing computational requirements (see [2] for a review).

In the following section, we describe the AVEC 2011 and PAINFUL datasets along with the features used for the modeling and labeling processes. Then, using an information theory approach, we confirm the suitability of using a first-order Markov model approach to model the transition between the levels of each affective dimension.

III. Datasets and Affective Transition Modeling

A. Data and Feature

1) *AVEC 2011 Dataset: Conversational Context:* The AVEC 2011 challenge dataset is part of the SEMAINE corpus [60], which consists of a large number of emotionally-colored interactions between human participants and an emotionally-stereotyped character. The videos were recorded by using five high-resolution (780×580 pixels), high frame-rate cameras (50 fps), and four microphones. The AVEC 2011 dataset was created from the first 140 operator–user interactions, which constitutes the sensitive artificial listener (Solid-SAL) partition of the SEMAINE corpus. The Solid-SAL partition consists of a human participant interacting with another person, who plays the role of the emotionally stereotyped character. The SEMAINE database is fully described in [60]. The AVEC 2011 challenge dataset consists of 31 videos used for training, 32 videos for development, and another 31 videos for testing. Fig. 1 shows examples of video frames. There are 20 participants in the whole data, and approximately 15 000 frames (50 min) per video.

In this paper, we use the video and audio features provided by the AVEC 2011 challenge database [31] briefly described here.

Fig. 1. Video frames from the AVEC 2011 dataset [60].

The unit of classification for the video dataset is a video frame. For each video frame, a set of features is computed. They consist of information describing the position and the pose of the face and of the eyes, and a locally dense appearance description. The OpenCV implementation of the Viola and Jones face detector [61] was employed to identify the face position and its features. The dense local appearance descriptor used is the uniform local binary pattern (LBP) [62], [63]. By employing uniform LBPs, instead of full LBPs, and by aggregating the LBP operator responses in histograms taken over regions of the face, the dimensionality of the features is kept relatively low (59 dimensions per region). For this reason, the registered face region is divided into 10×10 blocks, resulting in a feature vector with 5900 components.

The unit of classification for the audio dataset is a user-uttered word. The feature vector for each word consists of 1941 components, composed of 25 energy- and spectral-related low-level descriptors (LLD)×42 functionals, six voicing-related LLD×32 functionals, 25 delta coefficients of the energy/spectral LLD×23 functionals, six delta coefficients of the voicing-related LLD×19 functionals, and ten voiced/unvoiced durational features. The set of LLD covers a standard range of commonly-used features in audio signal analysis and emotion recognition. The functional set is carefully reduced to avoid LLD/functional combinations that produce values that are constant, contain very little information, and/or high amounts of noise. A detailed description of the features can be found in [31].

2) *The PAINFUL Dataset: Facial Expressions of Pain:* The second dataset used is the PAINFUL dataset, i.e., part of the UNBC-McMaster Shoulder Pain Expression Archive Database [32]. This dataset contains only videos of patients experiencing shoulder pain. The videos were collected while patients were performing a series of active and passive range-of-motion tests with either their affected limb or the unaffected one. The dataset contains 200 video sequences containing spontaneous facial expressions (no audio is provided). The unit of classification is a video frame. 48 398 frames of the dataset were coded by experts using the facial action coding system (FACS) [64]. Examples of frames from the PAINFUL dataset are shown in Fig. 2. There are 25 participants in this dataset.

As with the previous dataset, we used the features provided with the dataset. There are 66-point Active Appearance Model (AAM) [65] landmarks for each frame of the videos. The AAMs were used to track the face and extract its visual features. For each point, the horizontal and vertical coordinates

Fig. 2. Video frames from the PAINFUL dataset.

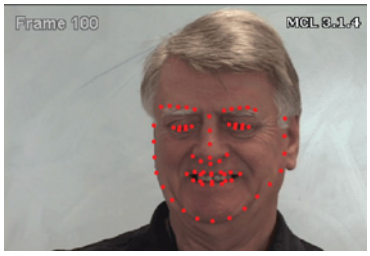


Fig. 3. Locations of 66 AAM points on the face were used as feature vector in the PAINFUL dataset.

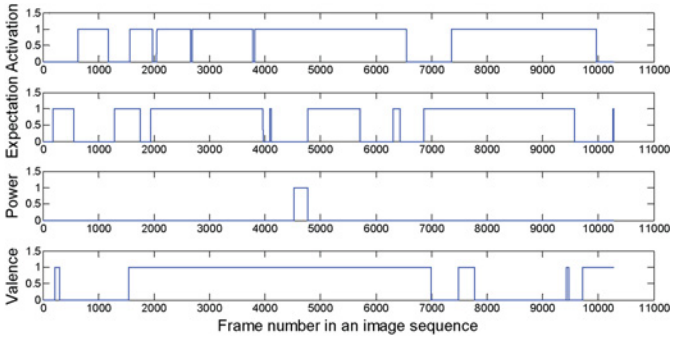


Fig. 4. Continuous labeling of the four affective dimensions (activation, expectation, power, and valence) in a sample of the AVEC 2011 video dataset.

are provided, forming a vector of 132 components for each frame. Fig. 3 shows an example of these points on a video frame. Detailed information on how these landmarks were generated can be found in [32].

B. Data Labels

1) *AVEC 2011 Video Labeling:* The video frames were continuously labeled over time by at least two raters according to four affective dimensions: activation, expectation, power, and valence. These dimensions are well established in the psychological literature and appear to account for most of the variability between everyday emotion categories [66]. The valence dimension indicates the overall positive or negative feeling of an individual toward the object at the focus of his/her emotional state. Activation indicates the individual's global level of dynamism or lethargy. The power dimension subsumes two related concepts—power and control. It relates to social experience of dominance and is also characterized by vocal and action tendency responses. The expectation dimension also subsumes various concepts, such as expecting, anticipating, and being taken by surprise.

The original continuous label traces were binned in temporal units of a duration corresponding to a single video frame for the facial expressions. The levels considered in the AVEC 2011 database were binarized by testing each value against the mean, i.e., a video frame is labeled as 1 (high) if the frame's original label (affective dimension level) is above the mean level for that dimension, 0 (low) otherwise. Fig. 4 shows an example of binarized labeling of a video recording along the four affective dimensions.

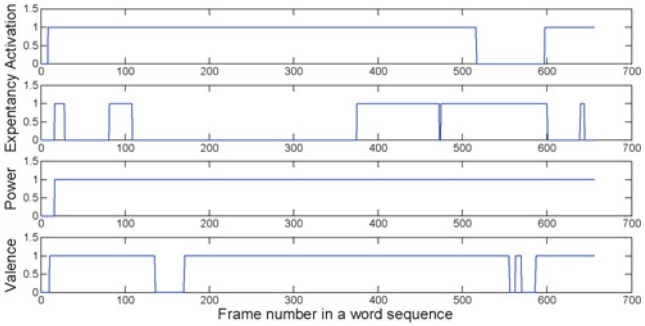


Fig. 5. Continuous labeling for the four affective dimensions (activation, expectation, power, and valence) in a sample of the AVEC 2011 audio dataset. Expectation tends to show shorter sequences, i.e., faster changes.

TABLE I

ENTROPY RATE FOR THE LABEL SEQUENCES IN THE THREE DATASETS

Dataset	Activation	Expectation	Power	Valence
AVEC 2011 Video	0.007	0.011	0.010	0.006
AVEC 2011 Audio	0.087	0.123	0.124	0.074
PAINFUL	0.043			

2) *AVEC 2011 Audio Labeling:* The AVEC 2011 audio dataset was labeled using a similar process. However, the traces were binned over the duration of the words (unit of classification for the AVEC 2011 audio dataset) uttered by the recorded participant, resulting in a single binary label per word. The word timings were obtained by applying forced alignment on the manual transcripts of the interactions. An example of four-dimensional labels for a word sequence in the training set is shown in Fig. 5.

3) *PAINFUL Dataset Labeling:* Each video frame contains action units (AUs) coded by certified FACS coders. Self-report and observer measures at the sequence level were taken as well. 1738 frames selected from one affected-side trial and one unaffected-side trial of 20 participants were randomly sampled and independently coded for interobserver agreement assessment. The Prkachin and Solomon pain intensity (PSPI) [67] metric was used to classify the level of pain (PSPI–FACS pain scale), as it is currently the only metric that can define pain intensity on a frame-by-frame basis. The pain intensity is computed as the sum of intensities of brow lowering, orbital tightening, levator contraction, and eye closure [32].

A PSPI value greater than 0 indicates a certain level of pain, with a maximum value of 15 in the dataset. Fig. 6 shows the pain level labels of the frames of a sample video recording. In this paper, following the study in [32], the pain levels were binarized, i.e., any PSPI higher than 0 was set to 1 to denote the presence of pain.

C. Label Analysis and Modeling

To provide a measure of the relationship between affective labels of consecutive units of classification (words or video frames), we computed the rate entropy of binary label sequences associated with each dataset. Given an affective or

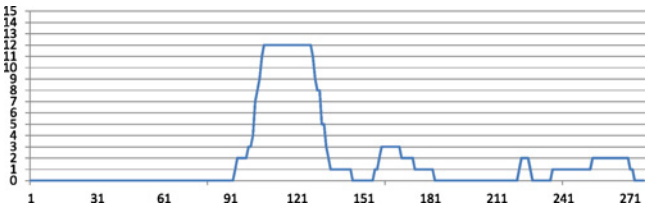


Fig. 6. Pain intensity level for each frame of a video clip.

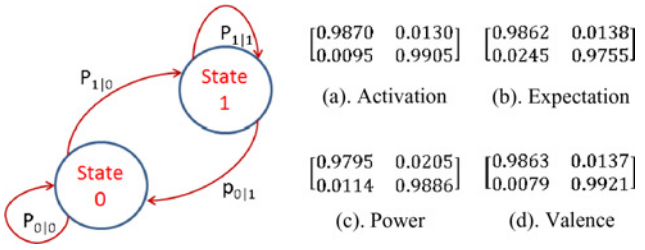


Fig. 7. Two hidden states in the HMMs and their transition matrices for four affective dimensions computed on a subset of the AVEC 2011 audio dataset. These two states are associated with the two levels of an affective dimension.

pain intensity dimension d, the label sequence associated with d is treated as a first-order Markov source, and its entropy is defined as

$$\textbf{Entropy}(d) = -\sum_i p_i \sum_j p_{j|i} \times \log_2 p_{j|i} \tag{1}$$

where i and j represent two possible levels of d; p_i is the probability that the level i occurs; and $p_{j|i}$ is the probability that level j occurs given the occurrence of level i as the previous level. For a sequence of binarized levels (0 or 1), the entropy for an affective dimension d is therefore given by

$$\begin{aligned}\textbf{Entropy}(d) = -[&p_0(p_{0|0}\log_2 p_{0|0} + p_{1|0}\log_2 p_{1|0}) \\ &+p_1(p_{0|1}\log_2 p_{0|1} + p_{1|1}\log_2 p_{1|1})].\end{aligned} \tag{2}$$

This implies that the assumption that the probability of occurrence of a certain affective dimension level depends only on the immediately preceding level. Table I shows very low entropy rate values (i.e., very low level of randomness) for each affective dimension, which is illustrated by Figs. 4, 5, and 6. Fig. 5 shows that, for all dimensions except expectation, the sequences of high levels are very long, i.e., the probability of two neighbor units having the same high level is very high. For all dimensions, the low levels are relatively long, which confirms that a first-order Markov model is an appropriate way to describe these types of label sequences. We thus suggest that, for naturalistic affective expressions, sequences of affective dimension levels have the Markov property.

Accordingly, an HMM framework can be used to model each affective dimension or the pain intensity dimension. We propose to design an HMM with two hidden states: 0 and 1. These two states are exactly associated with the two dimension levels. These hidden states capture the temporal structure of the data. $p_{0|0}$ and $p_{1|1}$ are the probabilities that the system remains in the current state and $p_{0|1}$ and $p_{1|0}$ are the transition probabilities between states. For each dimension, a typical transition matrix is represented in Fig. 7. The complete topology of the HMM is discussed in Section IV-C.

IV. A Multistage Classification System

A. System Overview

The aim of the classification system is to classify consecutive units (either words or video frame) of the input modality (audio or video) according to the levels of the affective dimension to be modeled. Each affective dimension of each modality is treated separately. Three variations of the architecture are proposed, which are shown in Fig. 8. Each of them performs an initial preprocessing of the input data and, through two or three classification stages, maps each unit of the input data into a dimension level value. For all three variations, feature extraction and dimension reduction processes in the preprocessing stage are applied to each unit of expression (e.g., frames for the video set or uttered words for the audio set).

In the first stage, the system classifies each unit by treating it independently of the other units. The output of each classification is a set of decision values indicating the likelihood that the classified unit expresses a particular affective dimension level (e.g., the probability to express a high rather than a low activation level). For simplicity, we call this set of values soft decision values. In the variation (a) shown in Fig. 8(a), this stage is performed by 1 classifier. However, to maximize the performance of this classification level, a number N of different classifiers could be used as illustrated in the first stage of Fig. 8 (b) and (c).

The three architecture variations differ mainly in their second stage. In the two-stage architecture version [Fig. 8(a)], the output of the first-stage classifier is used as input to its paired HMM in the second classification stage. The HMM reclassifies each input unit by taking into account not only the output of its paired first-stage classifier but also its classification of the previous units. In this way, each unit is classified on the basis of both its feature vector and its temporal relationship with previous units.

In the second variation [Fig. 8(b)], each of the N classifiers used in the first stage is paired with one HMM in the second stage. To optimize the final classification result, a third classification stage is added that uses a single HMM to combine the predicted labels from the N second-stage HMMs.

In the final variation [Fig. 8(c)], the second stage is skipped and the predicted labels from the first stage are combined and processed directly by a single HMM. The task of this third stage is, therefore, both to fuse the outputs of the first stage and to account for the relationship between consecutive units. The rationale behind this variation is that, when some of the classifiers in the first stage perform poorly, the HMMs in the second stage may be too sensitive. Skipping this stage thus reduces the weight of these first-stage classifiers.

B. First-Stage Classification

The first stage is a standard pattern recognition system in which every unit (e.g., frame or word) of the data is treated as an independent sample. The temporal relationship between these units is not taken into account. In this stage, any type of classifier can be used. The output of the classifier can be a real value, such as the posterior probabilities in naive Bayes classification, or the decision values in the SVM. In

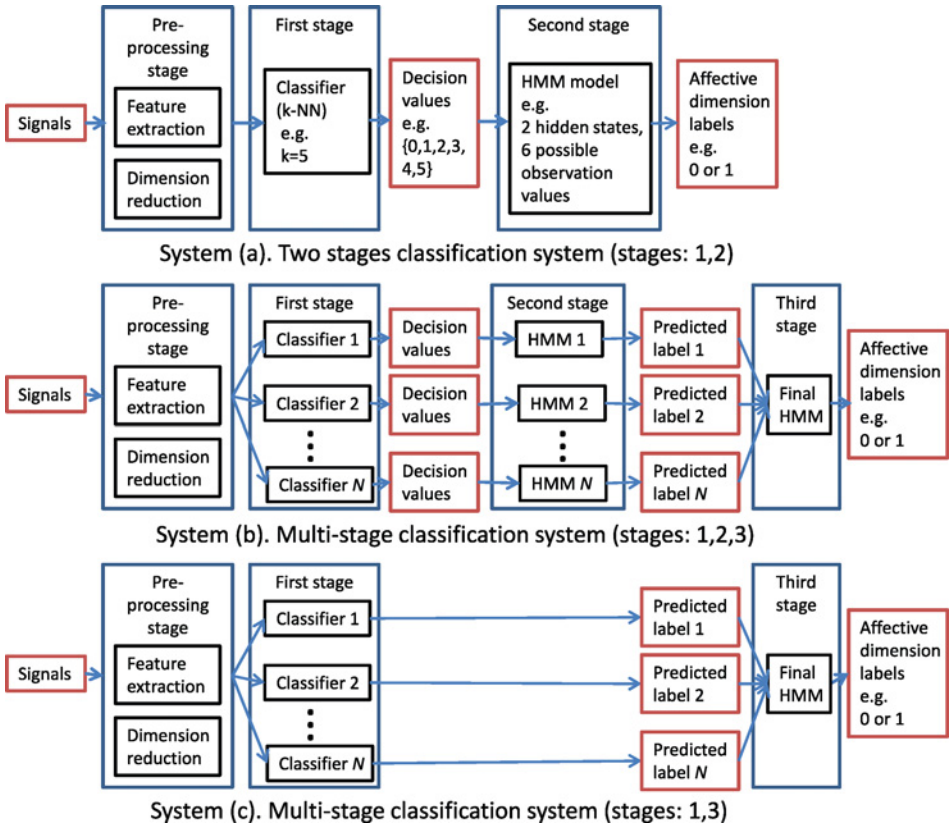


Fig. 8. Overview of the multi-stage automatic affective dimension level classification system. The system was tested under three different variants of the architecture: system (a), system (b) and system (c). After a common pre-processing stage, there are three possible classification stages of which the first can be modeled using any type of classifiers, whereas the last two are based on HMMs.

this paper, we propose to use the k-NN algorithm because it is simple and can conveniently output a very limited set of discrete values for each sequence of units. These outputs form the observed sequence to be input to the paired HMM in the second stage. In variations (b) and (c) of the system, the first stage uses N k-NN classifiers differing in their k value. This allows the system to take into consideration different degrees of variability in expressive cues (features) between levels of an affective dimension.

k-NN is a lazy learning method for classifying objects based on the closest training examples in the feature space. For a classification problem with binary label $\{0, 1\}$ and M training samples, the predicted label $\hat{y}$ of a test sample x can be decided by the majority of the labels in its k neighbors, namely

$$\hat{y} = \begin{cases} 1 \text{ if } \sum_{l=1}^{k} y_l > k/2, \\ 0 \text{ otherwise} \end{cases} \tag{3}$$

Since the predicted label is decided based on the value of $\sum_{l=1}^{k} y_l$, we can define a decision function for k-NN as the count of 0 neighbors as

$$\mathbf{Count}(x) = k - \sum_{l=1}^{k} y_l \tag{4}$$

This decision function will be the input for the HMMs in the subsequent stage of the system as the observed variables. As shown in Fig. 9(a), when k=5, the neighbors of the testing sample are five samples with labels 1 or 0. The predicted label of the testing sample, 0, is decided by the number of 0s in its neighbors' labels. It should be noted that there is no difference with counting the number of label 1 instead.

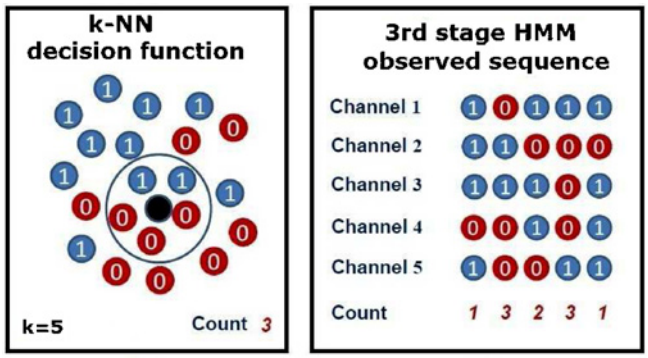
C. Second-Stage Classification

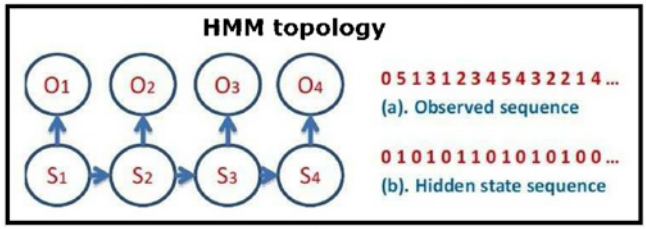
Each classifier of the first stage is paired with an HMM in this second stage. For each HMM, the observed sequence is based on the decision values output by its paired first-stage classifier. The decision values can be continuous values or discrete values depending on the classifier used. For continuous decision values, Gaussian mixture models can be used to estimate their probability distribution. For discrete decision values, a discrete probability distribution can be used for probability matrices estimation.

With k-NN in the first stage, a fully connected discrete HMM [Fig. 9(b)] with two hidden states (S_i is 0 or 1) can be built based on the decision function (observation O_i) from its paired k-NN classifier. The transition and emission probability matrices of the HMM can be estimated from the labels of the training set directly, see Figs. 9(c) and 7 for examples of emission and transition matrices when k=5 and the decision values are within $\{0, 1, 2, 3, 4, 5\}$.

D. Third-Stage Classification

In this decision fusion stage, the Markov property of temporal relationships in the sequences is further taken into account through the use of another HMM that fuses the outputs from the preceding stage. Two variations are proposed. In the variation presented in Fig. 8(b), the third-stage HMM fuses the





Discrete emission matrix

Observed		0	1	2	3	4	5
Hidden states	0	0.0	0.0748	0.2582	0.3523	0.2486	0.066
	1	0.1039	0.3061	0.3506	0.1974	0.0420	0.0

(c)

Fig. 9. Markov model modeling. (a) k-NN decision function for k-NN with k=5. The black point is a testing sample and the other points are training samples. The training sample labels are 1 or 0. The number of label 0 in the testing sample's five-neighborhood is the decision function for input of the later HMM. (b) Example of HMM used in the second stage. The observed values O_i are obtained from the decision values from its paired first-stage classifier. This example shows a possible observed sequence outputted by a k-NN when k=5. (c) An example of discrete emission matrix obtained from a k-NN with k=5 for activation in the audio training set. (d) Input to the third stage (decision fusion stage) from the multiple channels (i.e., five channels in this example). For each input unit to be classified, the output of each channel is either 1 or 0. The number (count) of labels 0 output by the N channels becomes the observed value for the third-stage HMM for an input unit.

output produced by the N k-NN-HMM pairs. In the variation shown in Fig. 8(c), the first stage feeds directly into the third-stage HMM. Hence, this HMM fuses the output produced by the N k-NNs. For simplicity, we will use the term channel to refer to either a k-NN-HMM pair feeding into the third stage or a single k-NN feeding into the third stage. There are N such channels.

The observed sequence for the third-stage HMM is based on the output of the N channels. For each input unit (i.e., video frame or uttered word), each channel produces a 0 or 1 output [Fig. 9(d)]. The number of 0's, predicted for an input unit by the N channels, is used as observed value for the third-stage HMM for that unit. For N=5 channels, the observed sequence for the third-stage HMM would therefore be a sequence of values between 0 and 5. It should be noted here that the topology of the channels and the type of classifiers need not be restricted to those used in our implementation. As discussed in the final section of this paper, more complex structures could be explored.

E. HMM Implementation

Each HMM is implemented as a fully-connected network with two hidden states. The training of each HMM in both the second and the third stages is based on the original training set. The state transition matrix (Fig. 7) is directly estimated from the labels associated with the training set. The state emission matrix is also estimated from the discrete probability distribution of the training labels in the training dataset. For the HMM testing, the classification problem is converted into a best path-finding problem for the decision value sequence. The Viterbi algorithm [68] is used to produce the best match label sequence. In our experiments, a topology with two hidden states and one-dimensional integer observation variables was sufficient. A more complex version of HMM could be used when the decision values are vectors.

V. Experimental Results

Independent experiments were carried out for each modality (audio or video) and for each affective dimension. Different evaluation methods were used for the three datasets (AVEC 2011 video, AVEC 2011 audio, PAINFUL dataset) according to the amount of data available. The methods used are detailed in dedicated subsections below. Performance was measured by weighted accuarcy (WA) and unweighted accuracy (UA) at unit of classification level (i.e., frame for video and word for audio). For binary classification, WA and UA are defined by equation 5 based on the number of true positive (tp), false positive (fp), true negative (tn), and false negative (fn) obtained.

$$\begin{aligned} \mathbf{WA} &= tp + tn/(tp + tn + fp + fn) \\ \mathbf{UA} &= (tp/(tp + fn) + tn/(fp + tn))/2 \end{aligned} \tag{5}$$

To evaluate the advantages of a three-stage architecture over a one-stage classifier, the one-stage classifier and the three variations of the proposed architecture were trained and tested. Different trials were run for each type of architecture. In the case of system (a), N=1, i.e., the first stage was formed by one k-NN and the second stage by one HMM. At each trial, a different value of k was used, varying between 2 and 20. Twenty was set as the maximum value because when k was increased, the performance of the classifier appeared to converge very quickly (Figs. 10, 11, 13). Furthermore, high values of k can result in over-smoothing of the boundary between classes.

In the case of systems (b) and (c), N k-NN classifiers with incremental values of k were used for the first stage. For example, the first trial used 19 k-NNs with k varying between 2 and 20 (indicated as [2:20]). At each trial, the value of N was decreased and the k-NN with the smaller k value was removed. For example, the second trial used only 18 k-NNs with k varying between 3 and 20 (indicated as[3:20]). Nineteen trials were run in total with the 19th trial using only one k-NN with k=20. The rationale was to evaluate if, by decreasing the number of k-NNs but keeping higher values of k, the performance of the architecture would increase. For k=20 only, one k-NN was used. This is because, by looking at the results obtained with system (a), the results appear to increase generally or stabilize

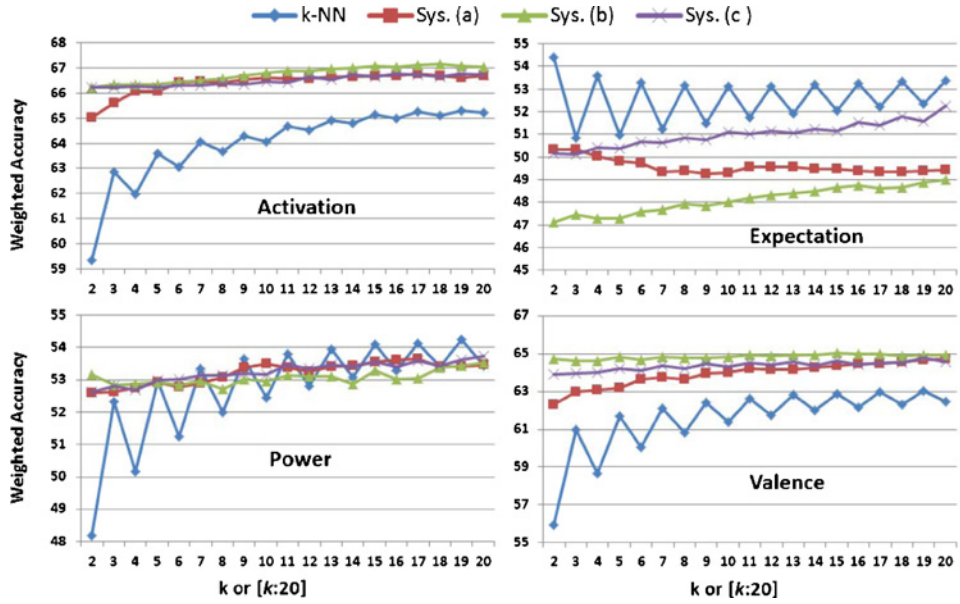


Fig. 10. AVEC 2011 video development dataset: Weighted accuracy comparison between the results of the one-stage classifier (k-NN in the graph) and the two-stage system (a), and between the results of the multi-stage systems (b) and (c). It should be noted that for the one-stage classifier (k-NN) and system (a), the x-axis indicates the value of k. For systems (b) and (c), the x-axis indicates the minimum value of k in the [k:20] range used for the multiple k-NNs in the first stage (i.e, k=[2:20],[3:20],...,[19:20],[20]).

as k increases. An optimization of the k value is beyond the scope of this paper. The results for k=20 are used to compare the gain of using systems (b) and (c) compared to using system (a) or a single-stage architecture (i.e., without modeling the temporal relationship between consecutive labels). In the following sections, we present the results for each dataset modality separately before providing a thorough comparison of the architecture variations in Section VI.

A. AVEC 2011: Video Development Dataset

In this dataset, each frame is represented by a vector of 5900 features. To reduce the dimensionality of the feature vector and to make calculations computationally tractable, PCA was used and only 1000 principal components (accounting for 72% of the variance) were selected for the classification stage. Thirty-one video sequences were used for the training phase (AVEC 2011 training dataset: number of frames = 501277) while 32 video sequences were used for the initial testing phase (AVEC 2011 development dataset: number of frames = 449074). Due to computer memory limitations, only 3% of the training frames were randomly selected for the training process.

Performance was measured by weighted accuracy at frame level, i.e., the percentage of correctly classified video frames, as shown in Fig. 10. Performance obtained with the multistage architectures show improvements with respect to the one-stage architecture for affective dimensions activation and valence. Only a very small improvement is observed for power, while lower performance is obtained for expectation.

In Table II, we compare the best results on the development set with the AVEC 2011 baseline results [31], with the associated confusion matrices shown in Table III. Our results show a clear improvement on the baseline rates for activation and valence. There was a slight drop in performance for power and expectation. Possible reasons for the reduction in performance will be discussed in Section VI.

TABLE II
AVEC 2011 DEVELOPMENT DATASET: COMPARISON BETWEEN RECOGNITION RATES OF PROPOSED AND BASELINE [31] METHODS

AVEC 2011	Weighted Accuracy	Activation	Expectation	Power	Valence
Video	Baseline	60.2	58.3	56.0	63.6
	Multi-stage	67.2	54.4	53.7	65.0
Audio	Baseline	63.7	63.2	65.8	58.1
	Multi-stage	73.6	67.5	64.3	70.1

B. AVEC 2011: Audio Development Dataset

To reduce the dimensionality of the feature vector, PCA was used and only 100 principal components were selected as they accounted for 99% of the variance. Thirty-one audio sequences were used as the training set (AVEC 2011 audio training set) and 32 sequences of the AVEC 2011 audio development set were used for initial testing.

The weighted accuracy results for audio are shown in Fig. 11. Here, the multistage architectures show a clear improvement in the recognition rate with respect to the one-stage architecture for three of the dimensions, namely, activation, power, and valence, and only a marginal improvement for expectation as k increases. Comparisons were also made with the baseline weighted rates for the development dataset, as shown in Table II. Apart from a slight drop on dimension activation, our results clearly outperform the baseline rates. The associated confusion matrices are shown in Table III.

C. AVEC 2011: Challenge Test Dataset

Our approaches were also tested on the test datasets of the AVEC 2011 video and audio subchallenge, which contain 11 sample sequences each. The rates were computed by the AVEC 2011 organizers as the true labels of the test samples were not

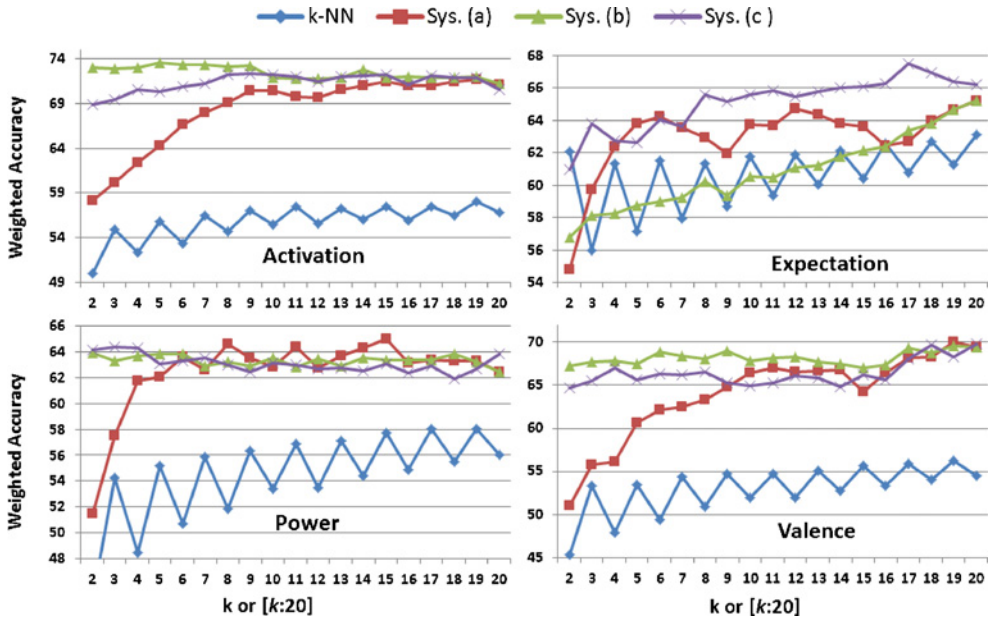


Fig. 11. AVEC 2011 audio development dataset: Weighted accuracy comparison between the results of the one-stage classifier (k-NN in the graph) and the two-stage system (a), and between the results of the multi-stage systems (b) and (c). It should be noted that for the one-stage classifier (k-NN) and system (a), the x-axis indicates the value of k. For systems (b) and (c), the x-axis indicates the minimum value of k in the [k:20] range used for the multiple k-NNs in the first stage (i.e, k=[2:20],[3:20],...,[19:20], [20]).

TABLE III

AVEC 2011 DEVELOPMENT DATASET: CONFUSION MATRICES FOR THE FOUR AFFECTIVE DIMENSIONS. THE ROWS ARE THE TRUE LABELS AND THE COLUMNS ARE THE PREDICTED LABELS

AVEC 2011	True\Predicted	Activation		Expectation		Power		Valence	
Development		1	0	1	0	1	0	1	0
Video	1	197923	50397	63360	114936	163327	99908	215120	69464
	0	96787	102878	89371	180318	107393	77357	87540	75861
Audio	1	7360	2114	740	4710	8884	2031	9367	1287
	0	2197	4629	586	10264	3782	1603	3591	2055

TABLE IV

AVEC 2011 TEST DATASET: COMPARISON BETWEEN CLASSIFICATION RATES FROM PROPOSED AND BASELINE [31] METHODS. WA = WEIGHTED ACCURACY, UA = UNWEIGHTED ACCURACY

AVEC 2011	Accuracy	Activation		Expectation		Power		Valence		Average	
Test	%	WA	UA	WA	UA	WA	UA	WA	UA	WA	UA
Audio	Baseline	55.0	57.0	52.9	54.5	28.0	49.1	44.3	47.2	45.1	51.9
	Multi-stage	64.3	66.2	57.0	58.6	41.3	54.4	50.5	51.4	53.3	57.7
Video	Baseline	42.4	52.5	53.6	49.3	36.4	37.0	52.5	51.2	46.2	47.5
	k-NN (k=9)	60.4	60.2	44.6	45.2	42.2	42.2	57.8	56.8	51.3	51.1
	System (a) (k=9)	64.5	64.3	44.7	45.8	39.3	39.0	60.8	60.1	52.3	52.3
	System (b) (k=9:20)	63.8	63.6	45.0	46.4	39.7	39.3	61.3	60.4	52.5	52.4
	System (c) (k=9:20)	63.2	63.0	47.4	46.7	39.9	39.6	59.9	58.9	52.6	52.1

available to us. Overall performance is shown in Table IV and was compared with the baseline performance provided in [31].

Table IV clearly shows our method to outperform the baseline rates for all affective dimensions in the AVEC 2011 audio subchallenge and all but the expectation dimension in the video subchallenge. For the audio data, the official overall average performance comparison on weighted and unweighted accuracy among all the participants of the audio subchallenge are shown in Fig. 12. The panels show that our method (denoted as UCL in the graphs [34]) performs similarly to that of the Uni-Ulm team [69] in weighted accuracy, but outperforms all teams in unweighted accuracy.

For the AVEC 2011 video subchallenge, results for four of our methods (i.e., k-NN with k=9, system (a) with k=9, system (b) and (c) with k=[9:20]) were higher than the baseline. Our accuracy (average: 52%) was bettered only by the top two performers in the competition,[1] with [35] (average: 61%) and [70] (average: 54%); although it should be noted that, whereas we limited ourselves to the features provided with the AVEC 2011 video dataset, the above works made use of optimized input features, and hence the results are not directly comparable.

D. PAINFUL Dataset

Since 83.6% of the frames corresponded to a no pain expression [15], this dataset (25 different subjects in 200 video

[1]http://sspnet.eu/avec2011/

TABLE V
PAINFUL DATASET: CONFUSION MATRIX FOR THE PAIN AND NO-PAIN CLASSES BY SYSTEM (C) WITH k=20

True\Predicted	1 (pain)	0 (no-pain)
1(pain)	130	8239
0(no-pain)	1195	38834

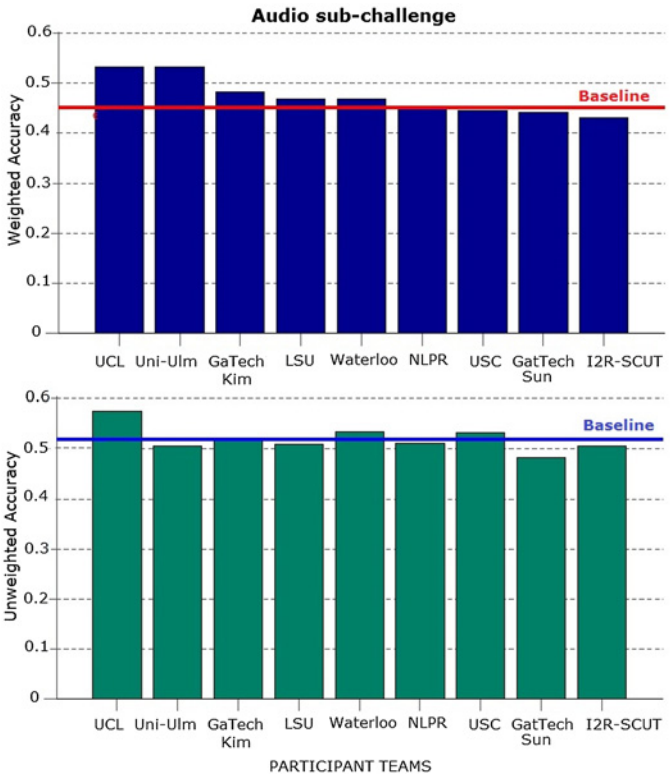


Fig. 12. Official results of weighted (top) and unweighted (bottom) accuracy among all participants of the AVEC 2011 audio subchallenge[1]. Our results [34] are denoted by the label UCL team. The image is courtesy of Michel Valstar.

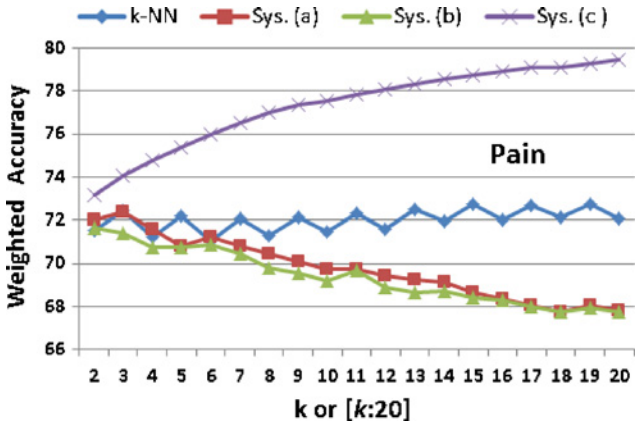


Fig. 13. PAINFUL dataset: Weighted accuracy comparison between the results of the one-stage classifier (k-NN in the graph) and the two-stage system (a), and between the results of the multi-stage systems (b) and (c). It should be noted that for the one-stage classifier (k-NN) and system (a), the x-axis indicates the value of k. For systems (b) and (c), the x-axis indicates the minimum value of k in the [k:20] range used for the multiple k-NNs in the first stage (i.e, k=[2:20],[3:20],...,[19:20], [20]).

clips) was tested only on binary labels, i.e., pain or no-pain. Person-independent tests were carried out using leaving one-person-out cross validation. Average performance over the 25 trials is shown in Fig. 13. Although systems (a) and (b) did not yield an improvement over the one-stage architecture, a very significant improvement was obtained with system (c). These results are comparable with those of [15], even if slightly lower. However, the two results cannot be directly compared, as the authors of [15] have preselected the input features. Here, we have taken a more general approach to focus on the ability of the architecture to learn the mappings.

Table V shows the confusion matrix for the results obtained with system (c) and k=20. It shows performance to be very good for the no-pain class but less convincing for the pain class. This may be due to the highly unbalanced training set. This imbalance may affect the performance of the k-NNs in the first-stage classification. It is possible that system (c) produced better results because the third stage carries out fusion and temporal modeling at the same time by favoring the first stage classifiers that produce better performance. However, given that in this case, the best k-NN classifiers may have overfitted the no-pain class, system (c) may have further enhanced that overfitting. In contrast, systems (a) and (b) may have detected the overfitting thanks to the intermediate level, however, failing to converge to a better solution.

It may be the case that we obtain better performance by simply balancing the data. However, this could reduce the ability to generalize to different non-pain expressions given that we may expect more variability in non-pain cases. Instead, it is possible that by using a more robust type of classifier capable of better generalization over small data sets (e.g., SVM, multitask learning [71]) or unbalanced datasets (e.g., weighted k-NN [72]) better performance would be achieved in the first stage, leading to a more effective action for each version of the multistage architecture in general.

VI. ARCHITECTURES COMPARISON

To evaluate if one architecture statistically performed better, the parametric repeated measure test with greenhouse correction was applied to the results presented in Figs. 10, 11 and 13. An effect of architecture type and interaction effect between architecture and affective dimensions was found for the AVEC 2011 video dataset (respectively, $F = 30.846$, $df = 1.261$, $p < 0.000$, $\eta_p^2 = 0.30$; $F = 92.274$, $df = 3.782$, $p < 0.000$, $\eta_p^2 = 0.794$) and the AVEC 2011 audio dataset (respectively, $F = 213.909$, $df = 1.729$, $p < 0.000$, $\eta_p^2 = 0.748$; $F = 225.474$, $df = 5.187$, $p < 0.000$, $\eta_p^2 = 0.904$). An effect of architecture type was also found for the PAINFUL dataset ($F = 103.817$, $df = 1.065$, $p < 0.000$, $\eta_p^2 = 0.852$).

These effects were further analyzed using a Bonferroni post hoc analysis as well as a nonparametric Wilcoxon signed rank test since many of the data deviated from normality. The results for each dataset are presented in Table VI, where the p-value reported is the most conservative of the two p-values obtained by each statistical test. Fig. 14 illustrates the average accuracy for the four architectures. The statistical analysis reveals that, for activation and valence, system (b) yielded

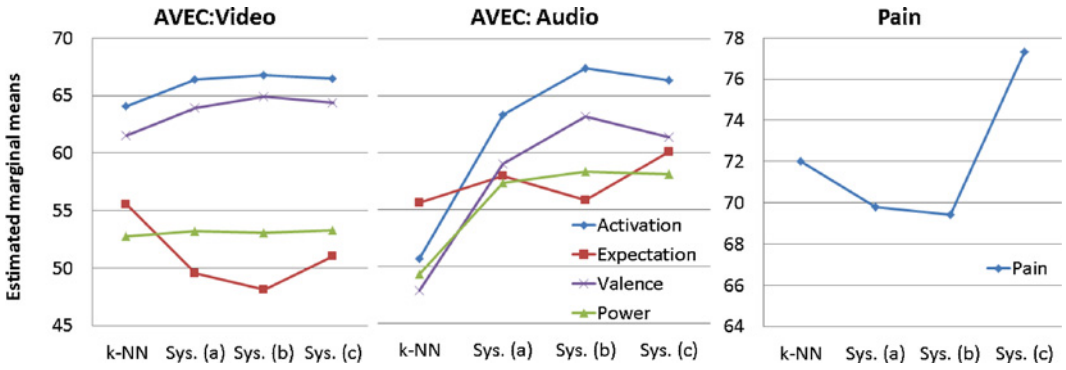


Fig. 14. Average weighted accuracy comparison between one-stage k-NN, two-stage system (a), multi-stage systems (b) and (c) over the different datasets.

TABLE VI

THE TABLE SHOWS THE BEST PERFORMING ARCHITECTURES FOR EACH AFFECTIVE DIMENSION AND DATASET. BOTH THE NONPARAMETRIC WILCOXON SIGNED RANK TEST AND THE PARAMETRIC REPEATED MEASURE TEST WITH BONFERRONI POST HOC CORRECTIONS CONFIRM THESE RESULTS (SEE p-VALUES)

Modality	Activation	Expectation	Power	Valence
AVEC 2011 Video	(b)>(c)>(a)>k-NN $p<.05$	k-NN>(c)>(a)>(b) $p<0.000$	(c)>(b) $p<.05$ No other diff.	(b)>(c)>(a)>k-NN $p<0.01$
AVEC 2011 Audio	(b)>(c)=(a)>k-NN $p<.05$	(c)>(a)>(b)=k-NN $p<.05$	(b),(c),(a)>k-NN $p<.05$	(b)>(c)=(a)>k-NN $p<.01$
Pain	(c)>k-NN>(a)>(b) $p<0.000$			

TABLE VII

DESCRIPTIONS OF SEQUENCES OF 1 AND SEQUENCES OF 0 FOR EACH DIMENSION AND DATASET. COMPARISON OF AVERAGE LENGTHS OF SEQUENCES OF 1 BY USING THE MANN–WHITNEY TEST

Dataset	**Affective Dimensions:** '1'-sequence count	**Average length:** '1'-sequences ('0'-sequences)	**Std length:** '1'-sequences ('0'-sequences)	**Mann-Whitney test** Length comparison
AVEC 2011 Video	Activation 299	1608.59 (1556.71)	2969.67 (4039.62)	Expectation<All (p-value<.001)
	Expectation 503	805.97 (1076.54)	1380.10 (1867.24)	Power<Valence (p-value<.05)
	Power 365	1422.70 (1171.80)	2550.50 (2121.65)	Activation<Valence (p-value<.01)
	Valence 274	2034.81 (1418.55)	3285.30 (3323.70)	
AVEC 2011 Audio	Activation 221	84.34 (76.57)	134.88 (177.29)	Expectation<All (p-value<.001)
	Expectation 341	39.20 (66.54)	61.84 (109.95)	Power<Valence (p-value<.001)
	Power 292	75.00 (48.70)	134.22 (81.98)	Activation<Valence (p-value<.05)
	Valence 204	105.92 (71.42)	166.33 (157.16)	
PAINFUL	Pain intensity 130	63.37 (303.85)	72.87 (482.49)	——

better performance than the other three variants. In the case of the pain intensity dimension for the PAINFUL dataset and expectation for the AVEC 2011 audio dataset, the best performance was obtained with the multistage system (c). However, in the case of expectation for the AVEC 2011 video dataset, the best accuracy was obtained with the one-stage architecture.

The only cases in which all multistage architectures performed worse than, or equal to, the one-stage architecture are expectation and power in the AVEC 2011 video dataset. A possible explanation for this lower performance is that, if in the initial stage some of the classifiers show lower performance, these inaccuracies are propagated to the higher stage, unless an optimization of the results from the first level is performed, as is done by multistage system (c). Another possible explanation for the lower performance for expectation could be that this type of expression is sudden (e.g., surprise) and lasts, on average, shorter than the others [73], thus making the assumption of Markov property slightly less suitable. This can be observed in Table VII, where the average lengths for the sequences of 1 (high level) for each dimension are reported. The Mann–Whitney test was used to compare the lengths of the sequences between affective dimensions (see last column of the table). The Mann–Whitney test shows that the expectation sequences are significantly shorter than those from other dimensions. A similar, but less significant, situation can be observed for the power dimension. This higher randomness is confirmed by a slightly higher entropy rate, as shown in Table I. It would be of interest to explore further this apparent difference in duration between affective dimensions and possibly exploit this information in the modeling process.

VII. DISCUSSION AND CONCLUSION

In this paper, a Markov model approach was proposed to model the transitions between levels of affective dimensions

or between pain intensity levels. This approach exploits the naturally slow changes of natural affective expressions. As discussed in Section II, our approach differs from other works that model the temporal relationship between observations, by associating the hidden states of the HMMs with levels of affective dimension labels. Therefore, the classification problem is converted into a best path-finding problem to obtain the best hidden state sequence in the HMM framework, hereby, using the Viterbi algorithm. Three variants of a multistage classification system were described and compared to a one-stage classifier. Upon testing of these four variants on the video and audio data of the AVEC 2011 challenge datasets, as well as the PAINFUL dataset, it was found that, on the whole, the multistage approach outperformed a one-stage k-NN classifier that does not consider the temporal information. However, the fact that accuracy gains were uneven between dimensions (e.g., no improvement in power and expectation for the video data) revealed that the rate at which expression transitions occur should be taken into account when deciding what architecture variant should be used. If the transitions are relatively frequent, as in the case of sudden and short duration emotional expressions, a one-stage architecture or system (c) appear most suitable; however, further work is needed to identify a suitable set of selection criteria. When compared to the baseline rates proposed for the AVEC 2011 dataset, a significant improvement in accuracy was observed. Results for the PAINFUL dataset were promising, with multistage architecture system (c) yielding a clear improvement over one-stage classifiers. However, results on the AVEC 2011 video test dataset, while higher than the baseline, suggest that there is much scope for improvement.

In our experiments, the architecture was tested over a reasonable but limited set of values for k (the number of neighbors) and N. Use of the system would likely require a thorough optimization of these parameters. It should be noted here that we tested for both even and odd values of k. As it could have been expected, the systems performed better with k an odd number (Figs. 10, 11, 13) since we have an even number of classes (i.e., always two classes). Hence, in further testing the choice of k should take into consideration the number of classes to be recognized.

In this paper, k-NN classifiers were used for the first-stage classification. It is possible that, by using more robust classifiers (e.g., SVM, Bayesian classifiers), higher performance could be achieved. This is particularly important for naturalistic datasets, where there is a high probability of having unbalanced classes. Transfer learning algorithms could be worth exploring in such a scenario, as they show higher performance in the case of small, sparse, and unbalanced datasets [71] [74]. For systems (a) and (b), the use of other types of classifiers may require some refinement. For example, as SVM classifiers return real-valued decision values, the observation sequences for the paired HMM in the second stage would also be real values. While the transition probabilities would not change, a Gaussian distribution would have to be considered to estimate the emission probabilities.

A number of possible extensions of this approach are worth mentioning. First, it would be of interest to investigate and characterize the extent to which a multistage approach improves on the first-stage classification when the latter is already performing well. This may indeed be the case since the multistage approach makes it possible to consider different temporal length relationships and operates as a fusion level. While classifiers in the first stage only varied according to their k value, more interesting combinations could be used. For example, variations could be based on different units of classification for each classifier in order to take into account local variations and to incorporate a broader range of features.

Second, an interesting extension would be to consider the second and third stages as means to model the fusion between different modalities, and their temporal relationship. In such a scenario, the first-stage classifiers could be dedicated to the local classification of units from different modalities. A critical issue would then be how to perform the alignment between modalities. A possibility could be to use different channels tuned to different temporal units of classification and let the system identify, through training, the best temporal alignment function. In addition, other variations of the HMM topology could be used to explore more complex temporal relationships between modalities.

Finally, in this paper, we only considered binary dimensions by a fully connected HMM with two hidden states. However, the approach can be applied to multilevel dimensions by simply using a Markov model with a number of states equivalent to the number of dimension levels to be recognized.

ACKNOWLEDGMENT

The authors would like to thank M. Valstar for computing the results on AVEC 2011 challenge test set.

REFERENCES

[1] R. W. Picard, *Affective Computing*. Cambridge, MA, USA: The MIT Press, 1997.

[2] M. E. Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognit.*, vol. 44, no. 3, pp. 572–587, 2011.

[3] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Commun.*, vol. 53, no. 9–10, pp. 1062–1087, 2011.

[4] M. Pantic and L. J. Rothkrantz, "Automatic analysis of facial expressions: The state of the art," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 12, pp. 1424–1445, Dec. 2000.

[5] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 1, pp. 39–58, Jan. 2009.

[6] H. Meng, B. Romera-Paredes, and N. Bianchi-Berthouze, "Emotion recognition by two view SVM 2K classifier on dynamic facial expression features," in *Proc. FG*, Mar. 2011, pp. 854–859.

[7] P. De Silva, A. Kleinsmith, and N. Bianchi-Berthouze, "Towards unsupervised detection of affective body posture nuances affective computing and intelligent interaction," in *Proc. LNCS*, vol. 3784. 2005, pp. 32–39.

[8] C. Shan, S. Gong, and P. W. McOwan, "Beyond facial expressions: Learning human emotion from body gestures," in *Proc. BMVC*, 2007.

[9] H. Gunes and M. Piccardi, "Automatic temporal segment detection and affect recognition from face and body display," *IEEE Trans. Syst. Man Cybern. B*, vol. 39, no. 1, pp. 64–84, Feb. 2009.

[10] A. Kleinsmith and N. Bianchi-Berthouze, "Affective body expression perception and recognition: A survey," *IEEE Trans. Affective Comput.*, vol. 99, 2012, prePrint.

[11] Y. Gao, N. Bianchi-Berthouze, and H. Meng, "What does touch tell us about emotions in touchscreen-based gameplay?" *ACM Trans. Comp. Human Interaction*, vol. 19, no. 4, article no. 31, Dec. 2012.

[12] R. L. Mandryk, K. M. Inkpen, and T. W. Calvert, "Using psychophysiological techniques to measure user experience with entertainment technologies," *Behaviour IT*, vol. 25, no. 2, pp. 141–158, 2006.
[13] L. Devillers, C. Vaudable, and C. Chastagnol, "Real-life emotion-related states detection in call centers: A cross-corpora study," in *Proc. INTERSPEECH*, 2010, pp. 2350–2353.
[14] M. Nicolaou, H. Gunes, and M. Pantic, "Designing frameworks for automatic affect prediction and classification in dimensional space," in *Proc. Workshop Gesture Recognit. (CVPR)*, 2011, pp. 20–26.
[15] P. Lucey, J. F. Cohn, I. Matthews, S. Lucey, S. Sridharan, J. Howlett, and K. M. Prkachin, "Automatically detecting pain in video through facial action units," *IEEE Trans. Syst. Man Cybern. B*, vol. 41, no. 3, pp. 664–674, Jun. 2011.
[16] B. Schuller, R. Müller, F. Eyben, J. Gast, B. Hörnler, M. Wöllmer, G. Rigoll, A. Höthker, and H. Konosu, "Being bored? recognising natural interest by extensive audiovisual integration for real-life application," *Image Vision Comput.*, vol. 27, no. 12, pp. 1760–1774, Nov. 2009.
[17] A. Kleinsmith, N. Bianchi-Berthouze, and A. Steed, "Automatic recognition of non-acted affective postures," *IEEE Trans. Syst. Man Cybern. B*, vol. 41, no. 4, pp. 1027–1038, Aug. 2011.
[18] J. Sanghvi, G. Castellano, I. Leite, A. Pereira, P. W. McOwan, and A. Paiva, "Automatic analysis of affective postures and body motion to detect engagement with a game companion," in *Proc. 6th Int. Conf. Human-Robot Interaction*, 2011, pp. 305–312.
[19] N. Savva, A. Scarinzi, and N. Bianchi-Berthouze, "Continuous recognition of player's affective body expression as dynamic quality of aesthetic experience," *IEEE Trans. Comput. Intell. AI Games*, vol. 4, no. 3, pp. 199–212, Jun. 2012.
[20] M. Thrasher, M. D. Van der Zwaag, N. Bianchi-Berthouze, and J. H. D. M. Westerink, "Mood recognition based on upper body posture and movement features," in *Proc. 4th Int. Conf. Affective Comput. Intell. Interaction*, 2011, pp. 377–386.
[21] N. Bianchi-Berthouze, "Understanding the role of body movement in player engagement," *Human-Computer Interaction*, vol. 28, no. 1, pp. 40–75, 2013.
[22] C. Jennett, A. L. Cox, P. Cairns, S. Dhoparee, A. Epps, T. Tijs, and A. Walton, "Measuring and defining the experience of immersion in games," *Int. J. Human-Comput. Stud.*, vol. 66, no. 9, pp. 641–661, Sep. 2008.
[23] N. Bianchi-Berthouze, P. Cairns, and A. L. Cox, "On posture as a modality for expressing and recognizing emotions," in *Proc. Emotion HCI*, 2008, pp. 74–80.
[24] G. N. Yannakakis, J. Hallam, and H. H. Lund, "Entertainment capture through heart rate activity in physical interactive playgrounds," *User Modeling User-Adapted Interaction*, vol. 18, no. 1-2, pp. 207–243, 2008.
[25] P. Cairns, A. Cox, N. Berthouze, S. Dhoparee, and C. Jennett, "Quantifying the experience of immersion in games," in *Proc. Cognitive Sci. Games Gameplay workshop Cognitive Sci.*, 2006.
[26] N. Bianchi-Berthouze, "Mining multimedia subjective feedback," *J. Intell. Inf. Syst.*, vol. 19, no. 1, pp. 43–59, Jul. 2002.
[27] M. Schröder, E. Bevacqua, R. Cowie, F. Eyben, H. Gunes, D. Heylen, M. Maat, G. Mckeown, S. Pammi, M. Pantic, C. Pelachaud, B. Schuller, E. Sevin, M. F. Valstar, and M. Wöllmer, "Building autonomous sensitive artificial listeners," *IEEE Trans. Affective Comput.*, vol. 3, no. 2, pp. 165–183, Sep. 2012.
[28] S. Steidl, A. Batliner, and B. Schuller, "The hinterland of emotions: Facing the open-microphone challenge," in *Proc. 3rd Int. Conf. Affective Comput. Intell. Interaction Workshops.*, 2009, pp. 1–8.
[29] A. Batliner, S. Steidl, B. Schuller, D. Seppi, K. Laskowski, T. Vogt, L. Devillers, L. Vidrascu, N. Amir, L. Kessous, and V. Aharonson, "Combining efforts for improving automatic classification of emotional user states," in *Proc. IS-LTC*, 2006, pp. 240–245.
[30] H. Gunes, B. Schuller, M. Pantic, and R. Cowie, "Emotion representation, analysis and synthesis in continuous space: A survey," in *Proc. FG*, 2011, pp. 827–834.
[31] B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie, and M. Pantic, "AVEC 2011–-The first international audio/visual emotion challenge," in *Proc. Int. Conf. Affective Comput. Intell. Interaction*, Oct. 2011, pp. 415–424.
[32] P. Lucey, J. F. Cohn, K. M. Prkachin, P. E. Solomon, and I. Matthews, "Painful data: The unbc-mcmaster shoulder pain expression archive database," in *Proc. FG*, 2011, pp. 57–64.
[33] K. Prkachin and P. Solomon, "The structure, reliability and validity of pain expression: Evidence from patients with shoulder pain," *Pain*, vol. 139, pp. 267–274, May 2008.
[34] H. Meng and N. Bianchi-Berthouze, "Naturalistic affective expression classification by a multi-stage approach based on hidden markov models," in *Proc. Int. Conf. Affective Comput. Intell. Interaction*, vol. 6975. 2011, pp. 378–387.
[35] G. A. Ramirez, T. Baltrušaitis, and L.-P. Morency, "Modeling latent discriminative dynamic of multi-dimensional affective signals," in *Proc. Int. Conf. Affective Comput. Intell. Interaction*, vol. 6975. Oct. 2011, pp. 396–406.
[36] B. S. Everitt, *The Cambridge Dictionary of Statistics*. Cambridge, London, U.K.: Cambridge Univ. Press, 2006.
[37] M. S. Bartlett, G. Littlewort, M. G. Frank, C. Lainscsek, I. R. Fasel, and J. R. Movellan, "Automatic recognition of facial actions in spontaneous expressions," *J. Multimedia*, vol. 1, no. 6, pp. 22–35, 2006.
[38] J. F. Cohn, L. I. Reed, Z. Ambadar, J. Xiao, T. Moriyama, and T. Moriyama, "Automatic analysis and recognition of brow actions and head motion in spontaneous facial behavior." in *Proc. SMC*, 2004, pp. 610–616.
[39] F. Schwenker, S. Scherer, Y. Magdi, and G. Palm, "The GMM-SVM supervector approach for the recognition of the emotional status from speech," in *Proc. ICANN*, vol. 5768. 2009, pp. 894–903.
[40] F. Eyben, M. Wöllmer, M. Valstar, H. Gunes, B. Schuller, and M. Pantic, "String-based audiovisual fusion of behavioural events for the assessment of dimensional affect," in *Proc. FG*, 2011, pp. 322–329.
[41] M. Nicolaou, H. Gunes, and M. Pantic, "Output-associative RVM regression for dimensional and continuous emotion prediction," in *Proc. FG*, 2011.
[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," in *Proc. INTERSPEECH*, 2008, pp. 597–600.
[43] F. Eyben, S. Petridis, B. Schuller, G. Tzimiropoulos, and S. Zafeiriou, "Audiovisual classification of vocal outbursts in human conversation using long-short-term memory networks," in *Proc. ICASSP*, May 2011, pp. 5844–5847.
[44] M. A. Nicolaou, H. Gunes, and M. Pantic, "Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space," *IEEE Trans. Affective Comput.*, vol. 2, no. 2, pp. 92–105, Apr.–Jun. 2011.
[45] M. Wöllmer, A. Metallinou, F. Eyben, B. Schuller, and S. S. Narayanan, "Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional lstm modeling," in *Proc. INTERSPEECH*, 2010, pp. 2362–2365.
[46] I. Cohen, N. Sebe, A. Garg, L. S. Chen, and T. S. Huang, "Facial expression recognition from video sequences: temporal and static modeling," *Comp. Vision Image Understanding*, vol. 91, no. 1-2, pp. 160–187, 2003.
[47] T. L. Nwe, S. W. Foo, and L. C. D. Silva, "Speech emotion recognition using hidden markov models," *Speech Commun.*, vol. 41, no. 4, pp. 603–623, 2003.
[48] A. Nogueiras, A. Moreno, A. Bonafonte, and J. B. Mariño, "Speech emotion recognition using hidden markov models," in *Proc. INTERSPEECH*, 2001, pp. 2679–2682.
[49] L. De Silva and P. C. Ng, "Bimodal emotion recognition," in *Proc. FG*, 2000, pp. 332–335.
[50] Z. Zeng, J. Tu, B. Pianfetti, M. Liu, T. Zhang, Z. Zhang, T. Huang, and S. Levinson, "Audio-visual affect recognition through multi-stream fused HMM for HCI," in *Proc. CVPR*, vol. 2. 2005, pp. 967–972.
[51] T. Kitazoe, S.-I. Kim, Y. Yoshitomi, and T. Ikeda, "Recognition of emotional states using voice, face image and thermal image of face," in *Proc. INTERSPEECH*, 2000, pp. 653–656.
[52] C. M. Lee, S. Yildirim, M. Bulut, A. Kazemzadeh, C. Busso, Z. Deng, S. Lee, and S. Narayanan, "Emotion recognition based on phoneme classes," in *Proc. ICSLP*, 2004, pp. 889–892.
[53] M. Yeasin, B. Bullot, and R. Sharma, "From facial expression to level of interest: A spatio-temporal approach," in *Proc. CVPR*, 2004, pp. 922–927.
[54] Y. Zhang and Q. Ji, "Active and dynamic information fusion for facial expression understanding from image sequences," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 5, pp. 699–714, May 2005.
[55] X. Li and Q. Ji, "Active affective state detection and user assistance with dynamic bayesian networks," *IEEE Trans. Syst. Man Cybern. A*, vol. 35, no. 1, pp. 93–105, Jan. 2005.
[56] Y. Tong, W. Liao, and Q. Ji, "Facial action unit recognition by exploiting their dynamic and semantic relationships," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 10, pp. 1683–1699, oct. 2007.

[57] D. Jiang, Y. Cui, X. Zhang, P. Fan, I. Ganzalez, and H. Sahli, "Audio visual emotion recognition based on triple stream dynamic bayesian network models," in *Proc. Int. Conf. Affective Comput. Intelligent Interaction*, vol. 6974. 2011, pp. 609–618.
[58] A. M. Rahman, M. I. Tanveer, and M. Yeasin, "A spatio-temporal probabilistic framework for dividing and predicting facial action units," in *Proc. Int. Conf. Affective Comput. Intell. Interaction*, vol. 6975. 2011, pp. 598–607.
[59] Z. Hammal and M. Kunz, "Pain monitoring: A dynamic and context-sensitive system," *Pattern Recognit.*, vol. 45, no. 4, pp. 1265–1280, 2012.
[60] G. Mckeown, M. Valstar, R. Cowie, and M. Pantic, "The semaine corpus of emotionally colored character interactions," in *IEEE Int. Conf. Multimedia Expo*, 2010, pp. 1079–1084.
[61] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comp. Vision*, vol. 57, no. 2, pp. 137–154, 2004.
[62] T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern Recognit.*, vol. 29, no. 1, pp. 51–59, 1996.
[63] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.
[64] P. Ekman and W. V. Friesen, *Facial Action Coding System*. Mountain View, CA, USA: Consulting Psychologists Press, 1978.
[65] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active Appearance Models," *Eur. Conf. Comp. Vision* , vol. 2, pp. 484–498, 1998.
[66] J. R. J. Fontaine, K. R. Scherer, E. B. Roesch, and P. C. Ellsworth, "The world of emotions is not two-dimensional," *Psychological Sci.*, vol. 18, no. 12, pp. 1050–1057, 2007.
[67] K. M. Prkachin and P. E. Solomon, "The structure, reliability and validity of pain expression: Evidence from patients with shoulder pain," *PAIN*, vol. 139, no. 2, pp. 267–274, 2008.
[68] A. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Trans. Information Theory*, vol. 13, no. 2, pp. 260–269, Apr. 1967.
[69] M. Glodek, S. Tschechne, G. Layher, M. Schels, T. Brosch, S. Scherer, M. Kächele, M. Schmidt, H. Neumann, G. Palm, and F. Schwenker, "Multiple classifier systems for the classification of audio-visual emotional states," in *Proc. Int. Conf. Affective Comput. Intell. Interaction*, vol. 6975. 2011, pp. 359–368.
[70] A. Cruz, B. Bhanu, and S. Yang, "A psychologically-inspired match-score fusion model for video-based facial expression recognition," in *Proc. Int. Conf. Affective Comput. Intell. Interaction*, 2011, pp. 341–350.
[71] B. Romera-Paredes, A. Argyriou, N. Berthouze, and M. Pontil, "Exploiting unrelated tasks in multi-task learning," *J. Mach. Learn. Res. Proc. Track*, vol. 22, pp. 951–959, 2012.
[72] W. Liu and S. Chawla, "Class confidence weighted knn algorithms for imbalanced data sets," in *Proc. 15th Pacific-Asia Conf. Adv. Knowl. Discovery Data Mining*, 2011, pp. 345–356.
[73] S. D'Mello and A. Graesser, "The half-life of cognitive-affective states during complex learning," *Cognition Emotion*, vol. 25, no. 7, pp. 1299–1308, 2011.
[74] B. Romera-Paredes, H. Aung, M. Pontil, A. C. de C. Williams, P. Watson, and N. Bianchi-Berthouze, "Transfer learning to account for idiosyncrasy in face and body expressions," in *Proc. FG*, 2013.

Hongying Meng (M'10) received the Ph.D. degree in communication and electronic systems from Xi'an Jiaotong University, Xi'an, China, in 1998.

He is currently a Lecturer with the School of Engineering and Design, Brunel University, West London, U.K., and is an Honorary Research Associate at University College London, London, U.K. His current research interests include digital signal processing, machine learning, human–computer interaction, computer vision, image processing, and embedded systems with more than 60 publications and 600 citations in these areas.

Dr. Meng is a Technical Committee Member on computational intelligence of the IEEE Systems, Man, and Cybernetics Society and a Fellow of the Higher Education Academy in U.K.

Nadia Bianchi-Berthouze received the Laurea degree with honors in computer science in 1991 and the Ph.D. degree in science of biomedical images in 1996 from the University of Milano, Milano, Italy.

She is currently a Senior Lecturer at the UCL Interaction Centre, University College London, London, U.K. Her current research interests include studying body movement as a medium to automatically recognize and steering the quality of experience of humans interacting and engaging with/through whole-body technology.

Dr. Bianchi-Berthouze was awarded an EU FP6 International Marie Curie Reintegration Grant in 2006 to investigate the above issues in the clinical and entertainment contexts. She is currently Principal Investigator on an EPSRC-funded project on Pain rehabilitation: E/Motion-based automated coaching (EP/H007083/1, 2010–2014). She is a member of the Editorial Board of the *International Journal of Creative Interfaces* and *Computer Graphics Journal on Multimodal User Interfaces*.