Time-Delay Neural Network for Continuous Emotional Dimension Prediction from Facial Expression Sequences

Hongying Meng, *Member, IEEE,* Nadia Bianchi-Berthouze, *Member, IEEE,* Yangdong Deng, *Member, IEEE,* Jinkuang Cheng, John Cosmas, *Senior Member, IEEE*

***Abstract*—Automatic continuous affective state prediction from naturalistic facial expression is a very challenging research topic but very important in human-computer interaction. One of the main challenges is modeling the dynamics that characterize naturalistic expressions. In this paper, a novel two-stage automatic system is proposed to continuously predict affective dimension values from facial expression videos. In the first stage, traditional regression methods are used to classify each individual video frame, while in the second stage, a Time-Delay Neural Network (TDNN) is proposed to model the temporal relationships between consecutive predictions. The two-stage approach separates the emotional state dynamics modeling from an individual emotional state prediction step based on input features. In doing so, the temporal information used by the TDNN is not biased by the high variability between features of consecutive frames and allows the network to more easily exploit the slow changing dynamics between emotional states. The system was fully tested and evaluated on three different facial expression video datasets. Our experimental results demonstrate that the use of a two-stage approach combined with the TDNN to take into account previously classified frames significantly improves the overall performance of continuous emotional state estimation in naturalistic facial expressions. The proposed approach has won the affect recognition sub-challenge of the third international Audio/Visual Emotion Recognition Challenge (AVEC2013)[1].**

***Index Terms*—Affective computing, Neural networks, Emotion prediction, Emotion dimension, Facial expression.**

I. Introduction

EMOTIONAL expressions are very important in human communication. They mediate interaction between people, enrich and often clarify the meaning of words or sentences and help regulate tension. They also act as an important regulatory loop on oneself. Evidence has shown that, when portraying an expression through either our face or our body, our emotional state is also biased in the direction of the expressed emotion [1], [2]. As interactive technology becomes ubiquitous in our society and takes on social companionship and coaching roles in 'serious' tasks (e.g. education [3], physical rehabilitation [4]), it is critical that it is endowed with the capability to read people's emotional expressions in order to react or adapt appropriately. The work proposed in this paper aims to advance the state of the art in the recognition of continuous naturalistic affective expressions by taking into account their temporal dynamics.

[1]It should be noted that only the results of our algorithm were submitted to AVEC2013. Hence none of the work presented here has been previously published.

Since the emergence of the field of affective computing [5], much attention has been dedicated to creating systems that could recognize affective expressions. Work has focused on most modalities that people and animals consciously or unconsciously use to communicate or detect emotions: vocal (e.g. [6], [7]), facial (e.g. [8], [9], [10]), body expressions (e.g. [11]), touch behaviors (e.g. [12]), physiological (e.g. [13], [14]) and neurological activation patterns (e.g. [15], [16], [17]) or media-mediated expressions (e.g. [18]).

Initially, the work focused on acted or stereotypical expressions and on very controlled environmental conditions. The datasets created and used to develop such systems typically contained well-defined, separate acted expressions [19], [20], [21], [22]. Today, however, we are assisting to an increasing attempt to shift to expressions that reflect or are closer to those encountered in real-life situations. This shift is due in part to the successful results obtained on controlled and acted expressions, but also to the fact that sensing technology has entered our everyday life and is now embedded in many forms of technology (e.g. Google glasses, smart-phones). This, in turn, requires modeling the variability and richness found in everyday emotional expressions and also the fact that these are not pre-segmented but need to be continuously tracked over time.

Even if most of the work is still done in a controlled environment, applications are emerging built on real-life situations. For example, the work by [23], [24], [25], [26], [27] attempts to continuously monitor facial expressions and body movement to provide continuous and more objective measures of clinical conditions (e.g. depression, anxiety disorder, pain levels). Engagement in computer games is continuously achieved by reading the emotional state of the player [28], [14], [29] and consequently adapting the game according to the available cognitive resources or the type of experience the player is looking for in that moment [30]. There is also growing interest in stretching the challenge by considering non-controlled environments (e.g. changes in illumination, perspective, etc. [23]) since real-life applications need to work in such environments.

Another change that is occurring in the field of automatic emotion recognition is the shift in what needs to be modeled. Given the initial focus on stereotypical expressions, most of the work focused on modeling an emotional space consisting of discrete basic states such as *Anger*, *Disgust*, *Fear*, *Happiness*, *Sadness* and *Surprise* [31]. However, naturalistic

expressions present a bigger challenge to the research community because they are less stereotypical and not always full-fledged expressions [32]. In addition, the dynamic of these expressions is more complex and changes more slowly than acted expressions. The discrete emotional space has shown to be too limited to capture the complexity and variety of these expressions. The field is now moving towards a continuous space characterized by emotional dimensions. A continuous space not only allows for a more complete description of a complex emotional state [33] but also leads itself better to continuous tracking and classification of expressions and their temporal dynamics.

This has also led researchers to create datasets ([34], [35], [25], [36], [37], [38], [39]) that challenge the community to focus more closely on real-life situations and compare their results. Whilst these datasets raise the bar for the creation of an automatic emotion recognition system as they require to address the inherent complexity of an expression, they also provide the possibility to exploit this complexity to improve the classification process.

The work presented in this paper aims to contribute to this research area by proposing a novel framework for automatic emotional state prediction from facial expressions in a continuous space. In a previous work [40], we showed that, by taking into account temporal information on the decision level of a multi-stage system, the classification of a unit (e.g. a video-frame) of an emotion expression significantly improved. We extend this work in two ways. Firstly, the proposed framework reflects a more real-life situation where only past information is used for modeling the decision level. Secondly, the classification is not only continuous over time but also continuous over affective dimensions rather than binary. A Time-Delay Neural Network (TDNN) is used to capture the temporal relationship between predictions on continuous instances of a facial expression video recording. The system is fully evaluated on a purposely created dataset of facial expressions of people watching videos [41] as well as on the public AVEC2012 [35] and AVEC2013 [25] datasets. The results show that significant improvements are achieved in comparison with a one-stage method where temporal relationships are considered. In addition, the results show that by decoupling the modeling of the temporal dynamics of emotional states from high variability contained in the low-level features, we obtain an increase in performance and a decrease in computational cost. Whilst the system is tested on videos, it is modality independent.

The rest of the paper is organized as follows. Section 2 provides an overview of related research to highlight the motivation for suggesting the proposed approach. Detailed description of the TDNN and of the two-stage automatic emotional state prediction system is given in Sections 3 and 4 respectively. This is followed by the evaluation of the system on the three datasets. Finally, we conclude by discussing the lessons learnt and how the system could be further developed.

II. Related work

With the emergence of real-life datasets, the research field has moved from building systems that recognize preselected instances of expressions to continuously track and classify such expressions over time. Recently, the survey by Sariyanidi et al. [42] on registration, representation and recognition of facial expressions highlighted some open issues in this area and future directions for designing real-world affect recognition systems. In particular, they highlighted the need for work that makes a better use of temporal information. Indeed, initial approaches treated videos as sequences of independent facial expressions and focused on improving the classification performance of each independent expression. [43] proposed, for example, a system that detects, at run-time, video frames containing frontal faces and for each frontal face detects action units that relate to emotional expressions. It uses a combination of Support Vector Machine (SVM) and AdaBoost to increase both accuracy and speed. The work by [34] instead models each unit of expression (e.g. a video frame, a word) independently and makes it a standard classification problem at frame level. In [44], [45], the authors proposed to use the relationships between identity features and facial expression features to improve the detection of pain expressions over time. In [44], the algorithm modeled the quasi-orthogonality between these two types of features to optimize performance in identity recognition and facial expression recognition. In [45], a new multilinear multitask learning approach that facilitates transfer between tasks and limits negative transfer is used to classify frame by frame the activation of action units of facial expressions for continuous pain estimation. This traditional frame-by-frame approach is also used in building the baseline systems for the AVEC2012 [35] and AVEC2013 [25] challenges although different advanced features were extracted, feature selection process was added and optimized kernel-based SVMs for classification and regression were used. The results from these works were very positive, however, they missed the opportunity to exploit the temporal relations that exist between consecutive instances of an expression.

Other approaches have used spatio-temporal representations of an expression by computing features over a temporal window [46] rather than over a single frame (see Sariyanidi et al. [42] for a review). However, as this review discusses, most of these approaches make use of simple registration approaches and the variability of texture from frame to frame may be more apparent than the expression activity itself. Even if a few approaches have proposed more accurate registration techniques [47], these still lack in their capability to effectively perform a temporal registration of the frames. Sariyanidi et al. [42] suggest that there is a common unsaid assumption that within a window of expression there are no head pose variations but only facial activity changes. Sariyanidi et al. [48] tried to overcome these issues by proposing a probabilistic subpixel temporal registration method that measures registration errors and makes use of this information to improve its performance.

Other researchers explored other modeling techniques able to take advantage of this information. Modality-independent approaches attempted to use modeling techniques that have the inherent capability to model temporal information without exploiting modality-specific knowledge. A typical method used is the Hidden Markov Model (HMM). Already used for

this purpose in speech recognition [49] and body movement tracking and classification [50], it is now increasingly used in emotion recognition.

An extension of these works is in the multimodal recognition of emotional expressions and audiovisual affect recognition [51], [52], [53]. Other than considering hierarchical structures to facilitate a refinement of the initial predictions, another important point raised by these recent studies is that attention should be paid to the level of granularity of the modeling (unit of an expression). In the case of vocal modeling, [54] showed that different phonemes contribute differently to a vocal emotional expression. Their 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 sub-units that compose it as the expression develops and ends (phonemes in their case).

Most of these works are still using datasets that are not really continuous and where the expressions have already been pre-segmented or defined through controlled recording. However, the encouraging levels of performance reached by all these systems suggest that the temporal relationship may be even more informative in the case of non-acted expressions as the expressions do not always start from a predefined neutral state. As researchers tackle naturalistic expressions, they are starting to take advantage of the knowledge available about naturalistic expressions and results have shown that this information is indeed very beneficial [55], [40]. Modelling techniques used to make use of the knowledge and constraints of the muscular structure of the face to reduce the complexity of the modeled phenomenon include Dynamic Bayesian Networks [56], [57], [58], [59], restricted Boltzmann machines [60] and Latent-Dynamic Conditional Random Fields [61], [62]. In these approaches, the temporal relationship is represented by the transition probabilities between hidden states. The main shortcoming of this 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 (EM) algorithm, could be used, the estimation is not always accurate because the data might violate the assumptions.

Nicolaou et al. [63] exploits the temporal dependencies over a dimensional domain by extending the Relevance Vector Machine (RVM) regression framework to capture the output structure and the covariance within a predefined time window. Baltrusaitis et al. [64] proposed continuous Conditional Neural Fields for structured regression for dealing with all the affect dimensions together. The aim is to improve performance by using both temporal information and correlation between affective dimensions. Indeed the literature shows that arousal and valence are correlated as the physiological processes they relate to appear to be correlated (for a review see [32]). Long Short-Term Memory (LSTM) is one type of Recurrent Neural Network (RNN) that has been successfully used for modeling the relationship between observations [65], [66], [67], [68] by making use of past classifications. Wöllmer et al. [65] first proposed a method based on LSTM RNN 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. [66] 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. [67] also used LSTM networks to outperform SVR due to their ability to learn past and future contexts. Wöllmer et al. [68] used Bidirectional Long Short-Term Memory (BLSTM) networks to exploit long-range contextual information for modeling the evolution of emotions within a conversation.

Whilst the methods discussed above make use of modelling techniques that are able to capture temporal information, they are still very tied to the feature level. Unfortunately, as already suggested in Sariyanidi et al. [42], there is a significant gap between feature level and semantic information in the data. For example, face images can change fast and dramatically in naturalistic videos, even if the emotional state of the person will change at a slower speed [40]. Many of the expressions changes may be due to information that is not always relevant to the emotional expressions (e.g., head pose, illumination). Multi-stage approaches have been proposed to overcome this problem [69], [70], [40]. Nicolaou et al. [69] trained a multi-layer hybrid framework composed of a temporal regression layer for predicting emotion dimensions, a graphical model layer for modeling valence-arousal correlations, and a final classification and fusion layer exploiting informative statistics extracted from the lower layers. In [70] and [40], a multi-stage approach was proposed to separate the feature level and the decision level. In the feature level, traditional classification methods were used to predict the emotion labels. In the decision level, the transitions (over time) between consecutive affective dimension levels were modeled as a first-order Markov Model. The temporal sequences of affective dimension levels (i.e. binary labels) were defined as the hidden states sequences in the HMM framework. The probabilities of these hidden states and their state transitions were computed from the labels of the training set. The rationale behind this approach was to transform the continuous binary classification problem into a best path-finding optimization problem in the HMM framework. The approach won the AVEC2011 audio sub-challenge [34]. The results showed that a multi-stage approach decoupling the classification at feature level from the classification at semantic level could further improve the recognition performance for slow-changing emotional dimensional aspects of the expressions by exploiting the temporal relationships only at the decision level. The main limitation of the approach was that it could be used only for categorization and not for regression. The other limitation was that the categorization of each frame was based on the whole sequence of frames (i.e., video) rather than just on past information, making it not useful in real-time applications.

In this paper, we propose to use a two-stage model with a TDNN model for continuous dimensional emotion prediction from facial expression image sequences. TDNN [71] is another neural network model with the capability of capturing the dynamic relationship between consecutive observations. In a TDNN, a current input signal is augmented with delayed copies of the previous input values and the neural network

is time-shift invariant since it has no internal state. In term of affect recognition this means that an instant of an emotional expression (e.g., a video frame) is classified by taking into account not only the input features describing that instant, but also the input features describing the previous instants, i.e., how the expression evolved over time to the current state. The delay, that is the number of past instants considered, is set as a parameter of the network. Its structure is much simpler than other RNN networks. For example, LSTM contains LSTM blocks instead of, or in addition to, regular network units. An LSTM block contains gates that determine when the input is significant enough to be remembered, when it should continue to be remembered or instead be forgotten, and when it should output the value. The simpler structure of the TDNN makes it less computationally expensive. Studies have in fact shown that TDNN are less computationally expensive than other RNNs [72]. Whilst the increased complexity of a RNN like the LSTM may be very beneficial at the first stage of emotion classification to deal with high dimensional and highly variable video features and their complex temporal relationship as in [73] [67], we propose to use the simpler TDNN structure when modeling the temporal relationship at the semantic level (second stage). In addition, it should be noted that to reach higher performance in modeling the temporal complexity presented by the low level features both [73] and [67] had to use both past and future information making the approach less useful in continuous real-life emotional state prediction. This is very important in the context of emotion recognition especially when dealing with video-based data.

The proposed two-stage TDNN-based method combines the benefits of the hierarchical approach proposed in [70] but overcomes its limitations. Firstly, it can deal with regression problems instead of categorizations due to its nature. Secondly, it only uses past knowledge gathered in real-time, rather than having to analyze the full sequence. Our method won the AVEC2013 affect sub-challenge. We also tested it on the AVEC2012 dataset and on our own recording dataset to verify its performance on different types of contexts.

III. Time-Delay Neural Network

TDNN is an artificial neural network model developed in the 1980s [71] in which all the neuron-like units (nodes) are fully connected by directed connections. Each unit has a time-varying real-valued activation and each connection has a modifiable real-valued weight. It has two special layers: hidden layer and output layer, in which the nodes are Time-Delay Neurons (TDNs) as shown in Figure 1 and described in the following.

A single TDN has M inputs $(I^1(t), I^2(t), \cdots, I^M(t))$ and one output $(O(t))$ where these inputs are time series with time step t. For each input $I^i(t)$ and $i = 1, 2, \cdots, M$, there is one bias value b_i, N delays (indicated as $D_1^i, , D_N^i$ in Figure 1) storing the previous inputs $I^i(t-d)$ with $d = 1, \cdots, N$, and the related N independent unknown weights $(w_{i1}, w_{i2}, \cdots, w_{iN})$. F is the transfer function $f(x)$ which is a non-linear sigmoid function here. A single TDN node can be represented using Equation 1:

$$O(t) = f(\sum_{i=1}^{M}[\sum_{d=0}^{N} I^i(t-d) * w_{id} + b_i]) \quad (1)$$

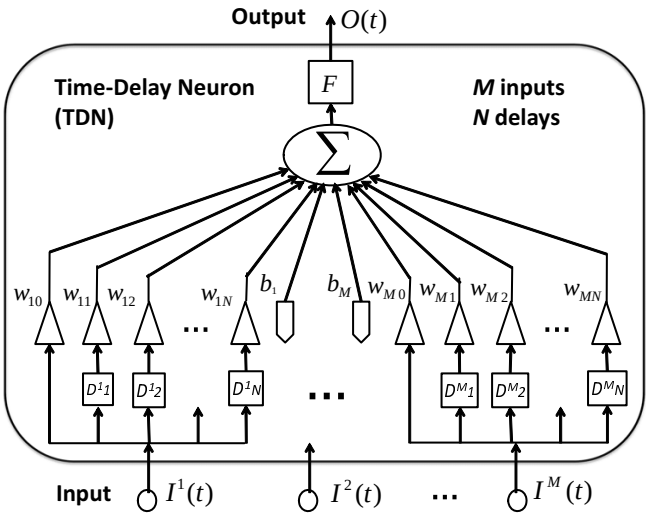

Fig. 1. Single TDN with M inputs and N delays for each input at time t. D_d^i are the registers that store the values of delayed input $I^i(t-d)$.

From Equation 1, it can be seen that both the inputs at current time step t and previous time steps $t-d$, with $d = 1, \cdots, N$ contribute to the overall outcome of the neuron. A single TDN can be used to model the dynamic non-linear behavior that characterizes series inputs.

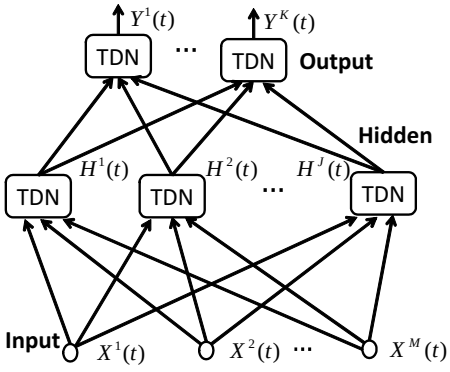

Fig. 2. The overall architecture of the TDNN neural network. It is a fully-connected two-layer neural network model with TDNs.

Figure 2 shows the overall architecture of the TDNN neural network, in which the hidden layer has J TDNs and the output layer has R TDNs with all the TDNs fully connected. The neural network model can be described using Equation 2 for the output layer and Equation 3 for the hidden layer:

$$O^r(t) = f(\sum_{j=1}^{J}[\sum_{d=0}^{N_1} H^j(t-d) * v_{jd}^r + c_j^r]), r = 1, 2, \cdots, R \quad (2)$$

$$H^j(t) = f(\sum_{i=1}^{M}[\sum_{d=0}^{N_2} X^i(t-d) * w_{id}^j + b_i^j]), j = 1, 2, \cdots, J \quad (3)$$

where, w_{id}^{j} and v_{jd}^{r} are respectively the weight of the hidden node H^j and of the output node O^r with b_i^j and c_i^r, the respective bias values. As seen from Equations 2 and 3, the TDNN is a fully-connected forward-feedback neural network model with delays in the nodes of the hidden and output layers. The number of delays for the nodes in the output layer is N_1 and that for the hidden layer is N_2. It is called Distributed Time-Delay Neural Network (DTDNN) if the delay parameter N varies between nodes.

For supervised learning in discrete time settings, the training set sequences of real-valued input vectors (e.g. representing sequences of video-frame features) are the sequences of activations of the input nodes, with one input vector at a time. At any given time step, each non-input unit computes its current activation as a non-linear function of the weighted sum of the activations of all units from which it receives connections. In supervised learning, the target labels at each time step are used to compute the error. For each sequence, its error is the sum of the deviations of the activations computed by the network at the output nodes from the corresponding target labels. For a training set, the total error is the sum of the errors computed for each individual input sequence. Training algorithms are designed to minimize this error.

A TDNN can be trained by using traditional methods for forward-feedback neural networks such as the Levenberg-Marquardt algorithm [74]. In the Levenberg-Marquardt algorithm, the training process optimises the weights W through iterations on the basis of the input time series $X(t)$ and the known labels $Y(t)$ for $t = 1, \cdots, T$ where T is the length of the sequence. During the testing process, the weights of the neural network are fixed and a predicted label is produced based on the input feature vectors only. Due to the delay property in the TDN nodes, the model can capture the dynamic behavior between consecutive elements of a sequence (e.g. video frames).

IV. Two-stage emotional dimension estimation system

Following the approach used in [40], we propose to integrate the TDNN into a two-stage architecture to predict the emotional state of a person along an affective dimension. We firstly describe the overall two-stage architecture and the rationale for it and then briefly present the algorithms used for the first-stage prediction. We then present the three datasets used to evaluate the architecture.

A. System overview

As discussed in the previous section, TDNN is a good candidate for real-time affective state prediction at unit level. It captures the dynamic relationship existing between consecutive units of expressions and utilizes it to improve the recognition performance. However, since facial expression features are generally very high dimensional, the TDNN model will have a large number of inputs and hence a large number of weights to be trained. This increases the model's complexity and the computational time. In addition, features between consecutive frames may show high variability due not only to change in emotional expressions but also to other factors such as head pose or illumination.

To overcome these problems, a two-stage system is proposed in this paper. In the first stage, a standard basic regression method is used to produce an initial prediction of the affective dimension level based on the highly variable and high-dimensional input features. Then, in the second stage, a TDNN is used to improve the accuracy of the prediction by taking into account past observations. This process mimics the method proposed in [40] with the difference that the new approach allows for real-time classification as it does not need to process the full sequence of observations (i.e. past, present and future units) to classify a unit of expression. It only requires a subset of the previous observations. This allows the recognition model to be used in real-life situations where prediction can be based only on already seen instances. In addition, by using a regression method at the first stage, the model is able to deal with continuous labels rather than just binary or discrete ones. This is an important requirement as real-life applications deal with complex emotional states that are better captured by continuous affective dimensional spaces. The problem to address is hence a regression problem rather than a classification one.

Figure 3 shows the overview of the proposed dimensional affective state prediction system we proposed. In the first stage, basic regression methods can be used for a first prediction of affective dimension level for the unit of expression in input (e.g. a video frame of a facial expression). The second stage is performed by a TDNN, where the prediction is updated by taking into account the label assigned to the previous frames (units) by the first-stage prediction. For the basic regression step, any standard regression method can be used, such as k-Nearest Neighbor (k-NN) and SVR. During the training process, two models are produced. The first one is a direct output of the first-stage training process (basic regression) and can be directly used to make inferences. The output of this first-stage model (basic regression) is used to train the TDNN-based model (second stage). These two models are both built using the same training dataset. Once the two models have been trained, they can be used in tandem as a two-layered system where the predicted values are produced continuously as units of sequences of expressions are received.

It should be noted that in this paper, our focus is to investigate the contribution made by a two-stage approach embedded with the capability to exploit temporal information rather than optimize the feature extraction level. More complex features extraction methods may possibly lead to further improvement but this is outside the scope of this paper.

B. Data recording and labeling

The inputs to the system are continuous streams of data representing continuous levels of affective expressions (e.g. facial expressions, body or vocal expressions of a person in pain). Whilst the system is independent on the modality used to recognize the affective state of a person, we tested our system on three video datasets of facial expressions that were continuously labeled over time by multiple raters. The

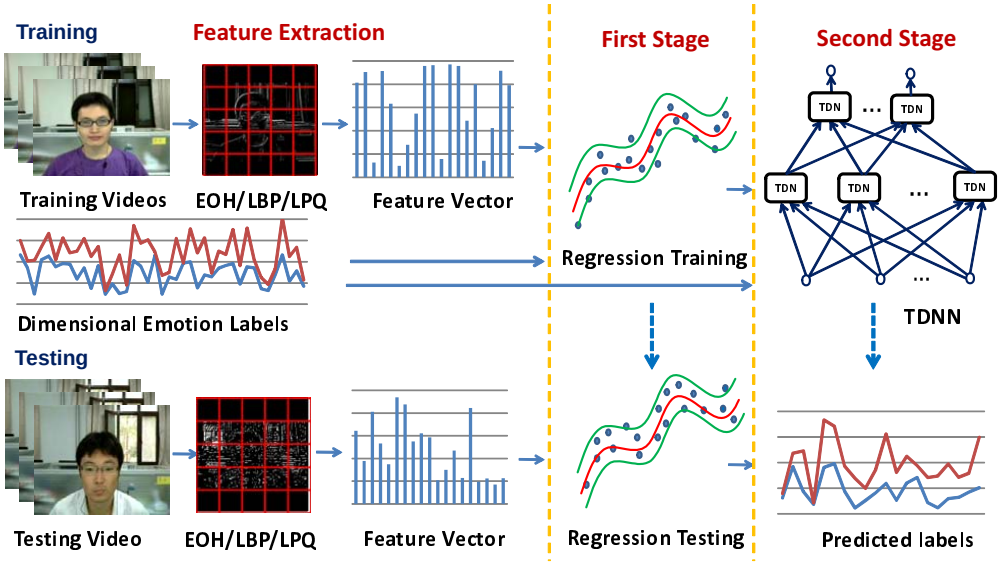

Fig. 3. A two-stage continuous prediction system: a basic regression model predicts the values of affective dimension conveyed by an individual video frame; the second level, based on TDNN, takes into account the relationship of predictions to finalize the labeling of frames.

labels used are continuous values over two or four of the following affective dimensions (according to the dataset): *Arousal*, *Expectation*, *Power*, *Valence*. These dimensions are well recognised in the psychological literature and account for most of the variability between everyday emotion categories [75]. *Arousal* is the individual's global feeling of dynamism or lethargy, including mental and physical activity. *Expectation* also subsumes various concepts such as expecting, anticipating, being taken unaware. The *Power* dimension combines two related concepts: power and control. It relates to the social experience of dominance and is also characterized by vocal and action tendency responses. The *Valence* dimension indicates the overall positive or negative feeling of an individual towards the object which is the focus of his/her affective state.

Each frame of a video is hence labeled with a vector of real values, one for each affective dimension. The ranges of values change in accordance with the affective dimension labeled and the protocol used in labeling the dataset.

C. Feature extraction

Whilst the framework is independent of the modality used and accepts any type of input features, for completeness, we briefly describe here some of the image features that were purposely developed for testing, or provided with the selected datasets.

1) LBP: The Local Binary Pattern (LBP) operator is defined as a gray-scale invariant texture measure derived from a general definition of texture in a local neighborhood. It was first described in [76]. It has since been found to be a powerful feature for texture classification [77] and has further been developed in different ways such as in [78] and in [79]. In this paper, only the basic LBP descriptor is used and the feature vector has a dimension of 256.

2) EOH: The second texture feature is computed by using the Edge Orientation Histogram (EOH) operator. The EOH is a simple, efficient and powerful operator that captures the texture information of an image. It has been widely used in a variety of vision applications such as hand gesture recognition [80] and object tracking [81]. It has also been used for facial expression analysis [82].

In the implementation of EOH, each image is scaled into a 40 x 40 size image and then divided into cells of 8 x 8 pixels. 2 x 2 cells form a block. The EOH feature is then computed for each cell and normalized within a block. This yields a 384-component vector [82].

3) LPQ: The Local Phase Quantization (LPQ) feature [83] is based on computing the Short-Term Fourier transform (STFT) on local image blocks. At each pixel the local Fourier coefficients are computed for four frequency points. Then the signs of the real and imaginary parts of each coefficient are quantized (binary scalar) to calculate phase information. The obtained eight bit binary coefficients are then represented as integers using binary coding like LBP. In this paper, this feature was provided by the AVEC2013 organizer, with face detection and normalization also used [25].

D. First-stage regression

The first stage of the architecture performs a typical regression process to provide a first classification of the unit of expressions (i.e. a video frame of a facial expression in our test case). In this paper, we explored both the k-NN regression and SVR methods. The first is very simple but very effective. The second is typically used for its generalization capabilities [84]. However, the architecture is general and any standard regression method could be used for this stage.

1) k-NN regression: k-NN is a lazy learning method for classifying objects based on the closest training examples in the feature space. Given a sample x, its predicted label $\hat{y}$ can

TABLE I
THE PEARSON CORRELATION COEFFICIENTS BETWEEN THE AFFECTIVE DIMENSIONS FOR THE THREE DATASETS USED FOR THE EVALUATION OF THE ARCHITECTURE. THE CORRELATION VALUES WERE COMPUTED ON THE TRAINING SUBSETS FOR EACH DATASET. ALL CORRELATION VALUES WERE STATISTICALLY SIGNIFICANT, I.E., P-VALUE < 0.0001

Dataset		Arousal	Valence	Expectation	Power
Video watching	Arousal	1	-0.4	-	-
	Valence	-0.4	1	-	-
AVEC2012	Arousal	1	0.4595	-0.0132	0.4288
	Valence	0.4595	1	0.0104	0.2731
	Expectation	-0.0132	0.104	1	-0.4428
	Power	0.4288	0.2731	-0.4428	1
AVEC2013	Arousal	1	0.2520	-	-
	Valence	0.2520	1	-	-

be computed as the average of the labels in its k neighbors $N(x) \subset (1, 2, \cdots, N)$ within the N training samples.

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 follows:

$$\hat{y} = \frac{1}{k} \sum_{l=1}^{k} y_l, l \in N(x) \tag{4}$$

For simplicity, k was set to 5 in all testing.

2) SVR: The SVR algorithm [84] can be considered the regression version of the SVM algorithm. The model produced by SVR depends only on a subset of the training data because the cost function for building the model ignores any training data close to the model prediction. In all our experiments, the linear kernel and default parameters were used for simplicity. In addition, no parameter optimization was carried out in order to provide a more fair comparison between the different architecture.

E. Second-stage prediction: TDNN modeling

The TDNN architecture was used in the second-stage prediction. The inputs are the predicted values from the first-stage regression method (i.e. the outputs of the first-stage regression). As the number of input and output to a TDNN can be any positive number, the architecture could be designed to model one affective dimension only or to model multiple affective dimensions at the same time. In the latter case, the TDNN will output a vector of values, one for each affective dimension modeled on the basis of the prediction of the units along the various dimensions. The latter approach could be useful when a certain relationship is known to exist between affective dimensions. The affective dimensions considered here are supposed to have minimum redundancies in modeling certain affective states [85]. In order to verify this assumption, we computed the Pearson correlation coefficients between the labels of the affective dimensions in the training set for each of the datasets presented below and these were overall quite low, as shown in Table I with only a few values reaching 0.4 (e.g., about 20% of variation in arousal is explained by either changes in valence or in power). We will discuss further this aspect when dealing with the specific datasets.

TABLE II
THE VALUES FOR THE TDNN PARAMETERS USED FOR THE VIDEO-WATCHING AND AVEC2013 DATASETS. M=NUMBER OF INPUT NODES, R=NUMBER OF OUTPUT NODES, J = NUMBER OF HIDDEN NODES, N_1 = NUMBER OF DELAYS PER INPUT NODE, N_2 = NUMBER OF DELAYS PER HIDDEN NODE

Dataset	Dimention	M	R	J	N_1	N_2	Iteration
Video watching	Arousal	1	1	10	2	2	20
	Valence	1	1	10	3	3	20
AVEC2013	Arousal	1	1	10	2	2	20
	Valence	1	1	10	2	2	20

TABLE III
THE VALUES FOR THE TDNN PARAMETERS USED FOR THE AVEC2012 DATASET. M=NUMBER OF INPUT NODES, R=NUMBER OF OUTPUT NODES, J = NUMBER OF HIDDEN NODES, N_1 = NUMBER OF DELAYS PER INPUT, N_2 = NUMBER OF DELAYS PER HIDDEN NODE

Feature	Dimention	M	R	J	N_1	N_2	Iteration
LBP	Arousal	1	1	4	3	2	20
	Expectation	1	1	4	3	3	20
	Power	1	1	2	1	3	20
	Valence	1	1	1	15	15	20
EOH	Arousal	1	1	10	2	2	20
	Expectation	1	1	1	15	15	20
	Power	1	1	2	15	1	20
	Valence	1	1	1	15	15	20
LBP+EOH	Arousal	1	1	1	30	30	20
	Expectation	1	1	10	15	15	20
	Power	1	1	2	3	3	20
	Valence	1	1	5	50	50	20

The implementation of the TDNN was simply achieved by using the TDNN function available from the MATLAB Neural Network toolbox and by experimentally setting its parameters. The parameters for the training of the TDNN were set as indicated in Table II for the Video watching and AVEC2013 datasets, and Table III for the AVEC2012 dataset.

V. EXPERIMENTAL EVALUATION

To test the performance of the systems, three datasets of videos of facial expressions were used. The first one was an in-house built dataset of people watching videos. The second and the third ones were respectively the AVEC2012 [35] and the AVEC2013 [25] audio-video datasets. Whilst the first dataset is part of our research, the second and third datasets allow us to compare our results with those of the research community.

A. Video-watching dataset

Fig. 4. On the left, the video the person is watching and on the right the person watching the video captured by a web camera.

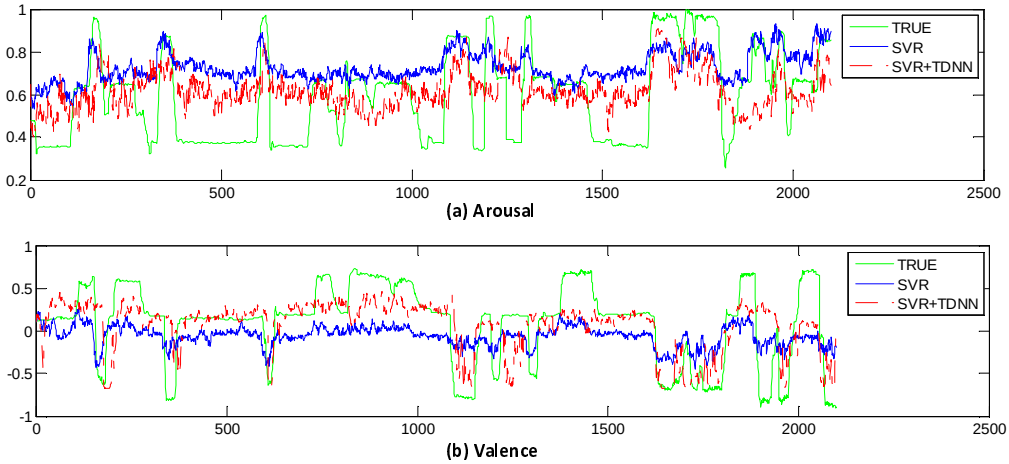

Fig. 6. Predicted dimensional affect labels for a video sample of the Video-watching dataset. (a) *Arousal* and (b) *Valence*. The ground truth label is shown by the green line, the blue line represents the first-stage prediction by SVR and the red line is the predicted labels for the second-stage prediction (SVR+TDNN).

Fig. 5. Some frame examples (facial expressions) from two participants of the video-watching dataset.

1) Data and labels: Facial expressions of people watching videos were continuously collected from a webcam [86], [87]. An example of the video watched and of the facial expressions gathered through the video camera are shown in Figure 4. A Logitech HD Webcam C270 was used for recording the facial expressions of the participants. Every video clip was recorded at a rate of 10 frames per second and a resolution of RGB24_160x120. The 'Motion JPEG AVI' was chosen as the compression format. A total of 2100 frames were recorded for each watching session.

The dataset consists of the recordings of 5 participants watching videos, with every participant recorded twice for a total of 21000 frames. The videos were selected to generate a variety of emotional responses such a disgust, fear, surprise, happiness and so on. Figure 5 shows examples of recorded facial expressions.

Only *Arousal* and *Valence* dimensions were used for the labeling. The Gtrace software [88] was used by two raters to annotate all the facial expressions reaching inter-rater Pearson's correlation values of 0.5538 for *Arousal* and 0.4814 for *Valence*. The ground truth was computed as the average of the raters' ratings. Detailed information of the dataset can be found in paper [41].

2) Features: Both basic LBP and EOH features were extracted for the testing experiments. For each frame, a basic LBP feature consisted of 256 values, whilst the EOH feature was formed of 384 values. Finally, LBP and EOH features were concatenated into a unique vector called LBP+EOH.

TABLE IV
VIDEO-WATCHING DATASET: PEARSON'S CORRELATION COEFFICIENTS (CORR) AND ROOT MEAN SQUARED ERROR (RMSE) AVERAGED OVER TWO-FOLD TESTING.

		KNN		KNN+TDNN	
Dimensions	Feature	CORR	RMSE	CORR	RMSE
Arousal	LBP	0.2317	0.8488	0.2658	0.1736
	EOH	0.2907	0.9247	**0.3017**	0.1917
	LBP+EOH	0.2905	0.9253	0.3	0.1915
Valence	LBP	0.1257	0.85	0.1497	0.3311
	EOH	0.3114	0.9323	0.3251	0.3002
	LBP+EOH	0.3119	0.9326	**0.3309**	0.2986
		SVR		SVR+TDNN	
Dimensions	Feature	CORR	RMSE	CORR	RMSE
Arousal	LBP	0.3205	0.3668	0.3646	0.1478
	EOH	0.4371	0.6356	**0.4720**	0.1504
	LBP+EOH	0.4262	0.6177	0.4699	0.1493
Valence	LBP	0.0398	0.1277	0.0415	0.2779
	EOH	0.3348	0.6406	**0.4037**	0.321
	LBP+EOH	0.3144	0.6176	0.3901	0.3361

3) Results: To test the proposed architecture, we compared the results of the one-stage regression system with either k-NN or SVR with the related two-stage architecture. Two-fold cross-validation method was used and the Pearson's correlation coefficients (CORR) and Root Mean Squared Error (RMSE) between the ground truth and the output of the four systems were computed.

The results are shown in Table IV. The table shows that for both *Arousal* and *Valence* dimensions, SVR performs better than k-NN. It also shows that the best results are obtained by using the two-stage architecture rather than just the single-stage regression approach for both SVR and k-NN versions of the system. Figure 6 provides an example of the predicted and ground truth values for *Arousal* and *Valence* dimensions for a sample video record for the SVR version of the architecture. The EOH feature seems to provide the best performance overall. This could be because EOH feature not only captures the texture information but also the spatial information.

We further evaluated the performance of a two stage-

TABLE V
VIDEO WATCHING DATASET: COMPARISON IN TERM OF PERFORMANCE (PEARSON'S CORRELATION COEFFICIENT - CORR) AS WELL AS COMPUTING SPEED TIME (SECOND) BETWEEN THE ONE-STAGE METHOD (EOH+TDNN) AND THE TWO-STAGE METHOD (EOH+SVR+TDNN).

	One-stage TDNN		**Two-stage TDNN**			
Input Feature	EOH		EOH		EOH+SVR	
Prediction	TDNN		SVR		TDNN	
Measurement	CORR	TIME(s)	CORR	TIME(s)	CORR	TIME(s)
Arousal	0.2229	4833	0.4371	216	0.4720	4
Valence	0.1020	10901	0.3348	167	0.4037	4

architecture versus directly modelling the temporal relationship between the frame features, i.e., by applying the TDNN directly at the first level. This was tested only on the EOH features as these had shown better performance in Table IV. The results are reported in Table V. We can see from the first column that using a one-stage TDNN leads to very poor performance compared to a SVR approach that does not use temporal information and to the SVR+TDNN approach that exploits and decouples such information from the low level features. These results confirm that the temporal information is more effective when modelled at the semantic level rather than at the feature level. In addition, Table V shows a high decrease in computational cost when modelling the temporal relationship at semantic level rather than at feature level. For example, in the case of *Arousal*, the computing speed decreases from 4833 seconds to 216+4 seconds. The reduction in the case of *Valence* is even bigger due to the more complex structure of the TDNN used for it.

Finally, even if our results show a relatively small correlation between *Arousal* and *Valence* (Table I), for completeness we model these two affective dimensions together. The results show that modelling them together does not lead to further increase in performance but rather to a slight decrease: 0.4533 instead of 0.4720 for *Arousal* and 0.3574 instead of 0.4037 for *Valence*. The lack of increase in performance is possibly due to the fact that the correlation is not strong and may also vary between the subsets used in cross-validation process.

B. AVEC2012 dataset

1) Data and label: The AVEC2012 challenge [35] uses the SEMAINE corpus [89], which consists of a large number of emotional interactions between human participants and with Sensitive Artificial Listener (SAL) agents. This database is recorded to study natural social signals that occur during conversations in face-to-face interactions. In the data collection, participants were invited to engage in a conversation with other humans or with four emotionally stereotyped characters: Spike always angry, Poppy always happy, Obadiah gloomy and Prudence being the sensible one. The emotional traits of the characters aimed to induce emotional changes in the participants. The AVEC2012 challenge dataset consists of a subset of the SEMAINE dataset, with 95 video clips split into: 31 training sessions, 32 development sessions and 32 test sessions. The frame number of each session is different because of the variability of the conversations.

Each video is recorded at a frequency of 49.479 frames per second and has a resolution of 780 x 580 pixels and 8 bits per pixel. Whilst the dataset also contains the audio modality, only the visual modality was used to evaluate our architecture. The reason to focus on one modality only is that we can test the power of the TDNN-based architecture modeling independently of the power of data fusion techniques. The baseline result of the AVEC2012 dataset [35] for the video modality was used for comparison.

Labels for the four affective dimensions (*Arousal*, *Valence*, *Power* and *Expectation*) were provided with the AVEC2012 dataset. The labels for each dimension are real values at video frame level. More details about the dataset are provided in [35].

2) Features: In this experiment, EOH feature and uniform LBP feature were used. The Uniform LBP is an extension of the original LBP operator which reduces the length of the feature vector and implements a simple rotation-invariant descriptor. For each video frame, the EOH feature has a dimension of 384. To compute the uniform LBP, each frame was divided into 100 blocks producing an LBP vector of 5900 elements for each frame. In addition, LBP and EOH features were also concatenated as in the previous experiment.

3) Results: As with the previous dataset, the results of the four systems were compared. Only the SVR regression method was used for this dataset given its superior performance to k-NN.

The AVEC2012 training dataset was used for training the architecture and the AVEC2012 development subset and testing subset were used for testing. The Pearson correlation values (CORR) between the ground truth and the output of the systems were computed. The results are shown in Table VI in comparison with the AVEC2012 baseline and the winning method [90].

The results are similar to those obtained for the Video-watching dataset. As with the previous experiment, the two-stage architecture using TDNN made a significant improvement on the performance from basic regression in all cases. Again, the best performance was obtained with the EOH vector as input feature for all affective dimensions with the exception of *Power*. For *Power*, better results were obtained by using the combination of EOH and uniform LBP.

In comparison with the AVEC2012 baseline results on the video modality, it can be seen that our proposed approach obtained better results most of the time. In the development dataset, only *Valence* was just slightly lower than the AVEC2012 baseline, whilst, in the testing dataset, the results on *Expectation* were lower than the baseline results. The large discrepancy on *Expectation* may be due to the fact that temporal information may play a lower role than in the other dimensions as shown in [40]. This may be due to the higher entropy presented by the *Expectation* dimension.

The AVEC2012 winning system [90] produced better results for every affective dimensions. There are a couple of main reasons for this. First of all, [90] use optimized features with respect to those used in our work. Log-Magnitude Fourier Spectrum was used to modify the shape features, the global appearance feature and the local appearance features with

TABLE VI
AVEC2012 DATASET: COMPARISON BETWEEN RECOGNITION PERFORMANCE (PEARSON'S CORRELATION COEFFICIENT- CORR) OF THE PROPOSED SYSTEM FOR THE DEVELOPMENT SUBSET AND TESTING SUBSET. IT IS ALSO COMPARED WITH BASELINE [35] AND THE WINNING METHOD [90]. IT SHOULD BE NOTED THAT THE WINNING METHOD [90] USES AN OPTIMIZED SET OF FEATURES AND BOTH VIDEO AND AUDIO INFORMATION RATHER THAN JUST VIDEO.

Development subset		CORR				
Method	Feature	Arousal	Expectation	Power	Valence	Average
SVR	LBP	0.096	0.029	0.069	0.001	0.048
	EOH	0.160	0.169	0.047	0.178	0.138
	LBP+EOH	0.086	0.012	0.110	0.076	0.071
SVR+TDNN	LBP	0.108	0.029	0.079	0.011	0.057
	EOH	0.162	0.172	0.048	0.204	0.146
	LBP+EOH	0.113	0.018	0.131	0.090	0.088
Baseline [35]	LBP	0.151	0.122	0.031	0.207	0.128
Nicolle et al. [90]	Shape	0.538	0.365	0.429	0.319	0.413
	Global	0.498	0.347	0.431	0.281	0.389
	Local	0.470	0.323	0.432	0.354	0.395
Testing subset		CORR				
Method	Feature	Arousal	Expectation	Power	Valence	Average
SVR	LBP	0.152	0.004	0.058	0.070	0.071
	EOH	0.430	0.017	0.039	0.305	0.198
	LBP+EOH	0.105	0.005	0.058	0.029	0.049
SVR+TDNN	LBP	0.161	0.021	0.072	0.078	0.083
	EOH	0.444	0.025	0.040	0.308	0.204
	LBP+EOH	0.122	0.019	0.078	0.045	0.066
Baseline [35]	LBP	0.077	0.128	0.030	0.134	0.093
Nicolle et al. [90]	Video+Audio	0.612	0.314	0.556	0.341	0.456

dynamic information integrated. Second, a correlation-based measure was used for the feature selection process to increase the robustness of the labels. This boosted the performance further. In addition, in the testing set, [90] use fusion of video features and audio features to further boost the results. As indicated earlier in the paper, our focus has been on the modelling rather than the optimization of the features. In future work, it will be very interesting to integrate the new features and feature selection process of the winning system [90] in the first level of our system.

Finally, given that Table I highlighted a certain amount of correlation between all four affective dimensions of this dataset, we carried out further analysis to investigate if modeling the dimensions together could be of any interest. We first computed the Pearson correlation coefficients for each subset of the AVEC2012 dataset. Figure 7 shows that the correlation values between each pair of affective dimensions vary significantly between training, development and testing subsets. Only the correlation value between *Arousal* and *Valence* (AV) is still around 0.4 when considering all three subsets together. For completeness, we used our two-stage architecture with EOH feature to model these two affective dimensions together. As with the Video Watching dataset, the results show that modelling the two affective dimensions together does not lead to further increase in performance but rather to a slight decrease: 0.1318 instead of 0.162 for *Arousal* and 0.1942 instead of 0.204 for *Valence*. Again, the lack of increase in performance is possibly due to the fact that the correlation is not strong and mainly varies between subsets used for training and testing.

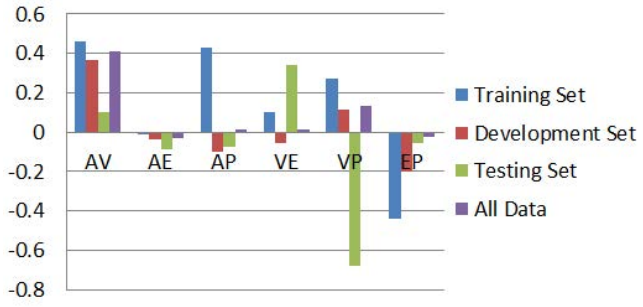

Fig. 7. The Pearson correlation coefficients between four dimensions for the AVEC2012 training, development and testing sets, p-values < 0.0001. A = Arousal, V = Valence, P = Power, E = Expectation. Each pair of capital letters indicates the two affective dimensions being correlated.

TABLE VII
AVEC2013 DATASET: PEARSON'S CORRELATION COEFFICIENTS (CORR) AVERAGED OVER ALL SEQUENCES, I.E. DEVELOPMENT DATASET AND TESTING DATASET.

Development subset		CORR		
Method	Feature	Arousal	Valence	Average
SVR	LPQ	0.123	0.125	0.124
SVR	EOH	0.156	0.142	0.149
SVR	LPQ+EOH	0.179	0.132	0.156
SVR+TDNN	LPQ	0.124	0.130	0.124
SVR+TDNN	EOH	0.161	0.143	0.149
SVR+TDNN	LPQ+EOH	**0.184**	0.136	0.156
Baseline [25]	LPQ	0.157	**0.337**	**0.247**
Lozano et al. [91]	LBP+GABOR	0.119	0.154	0.137
Testing subset		CORR		
Method	Feature	Arousal	Valence	Average
SVR+TDNN	LPQ+EOH	**0.1548**	0.1269	**0.1409**
Baseline [25]	LPQ	0.134	0.076	0.1050
Lozano et al. [91]	Video+Audio	0.1318	**0.1352**	0.1335

C. AVEC2013 dataset

1) Data and labels: The third dataset that was used for testing is the AVEC2013 challenge dataset [25]. This is a subset of the audio-visual depressive language corpus (AViD-Corpus) [25]. The dataset is composed of 340 video recordings of people performing a Human-Computer Interaction task while being recorded by a webcam and a microphone. There is only one person per clip and the total number of subjects is 292, i.e. some subjects feature in more than one clip. Each person was recorded between one and four times, with a period of two weeks between the measurements. Five subjects appear in 4 recordings, 93 in 3, 66 in 2, and 128 in only 1 session. The length of the clips varies between 20 minutes and 50 minutes (mean = 25 minutes) and the frame rate is 30. Examples of the captured frames are shown in Figure 8.

Fig. 8. The video recording setting for the AVEC2013 dataset.

In this paper, we focus on the AVEC2013 affect sub-challenge (ASC). The sub-challenge required the prediction at frame level of the value of the affective dimensions (*Arousal* and *Valence*). The AVEC2013 dataset provides both audio and video modalities, but only the video modality was used here for the reason stated above. There are 50 videos for training, 50 videos for development and 50 videos for testing.

2) Features: For each frame, the texture features were extracted. In the AVEC2013 dataset, LPQ features were provided by the challenge organizer. In addition, the EOH feature was also computed as it provided the best results in the previous experiments.

3) Results: As with the previous experiments, we compared the results for the one-stage regression system with the two-stage prediction system. Only the SVR regression method was used for this dataset given its superior performance over k-NN. The AVEC2103 training dataset was used for training the architecture and the AVEC2013 development dataset was used for testing. The Pearson's correlation coefficients (CORR) between the ground truth and the output of the systems were computed. Table VII shows the results for the AVEC2013 development dataset. The results show that the combination of SVR+TDNN outperforms SVR alone. For the *Arousal* dimension, the combination of LPQ and EOH achieved the best result, while the EOH feature alone achieved best performance for the *Valence* dimension.

The results were also compared with the ones from the baseline [25] and the runner up [91] at AVEC2013. The TDNN based two-stage architecture obtained better results for *Arousal* but performed worse than the AVEC2013 baseline for *Valence* for the development dataset. However, when we compare the results of the two-stage architecture with the AVEC2013 baseline results for the testing dataset, the two-stage architecture reaches higher performance for both dimensions with a clear improvement for *Valence* (see Table VII). This may suggest that the system was better able to generalize to new datasets by using the temporal information. In the runner-up system [91], a three-stage system was proposed to perform multiple fusions but temporal information was not considered in the modeling process.

In addition, our results outperform those of the other participants in the AVEC2013 ASC sub-challenge, as can be seen from Figure 9[2]. Our team (Brunel-Beihang team) produced the lowest RMSE among all participants with 0.1829 and higher correlation value with 0.1409 when compared with the baseline and [91]. Whilst we took part and won the ASC challenge, the material produced here was not submitted for publication[3].

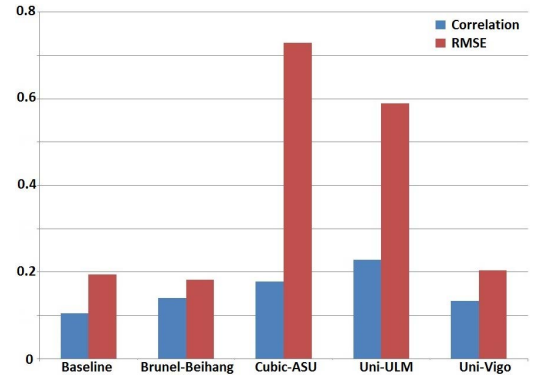

Fig. 9. Performance comparison on the Affect Recognition sub-challenge of AVEC2013. Correlation and RMSE values are used as measures of performance. Our approach is represented by the Brunel-Beihang Team with relative high CORR and lowest RMSE[2].

VI. Conclusion and discussion

In this paper, a two-stage architecture that combines a simple regression algorithm and a TDNN was proposed for automatic continuous affective state prediction from facial expressions in naturalistic contexts. In the second stage, the dynamic temporal relationship on the decision level was modeled by a TDNN model and significant improvement in performance was achieved. The TDNN receives input from a regression stage rather than the large and highly variable input features describing the sequence of expressive units. This reduces the computational complexity and facilitates training and generalization capabilities. The length of history to be taken into account was decided experimentally. In comparison with the HMM-based method [40], the proposed TDNN-based method can deal with regression problems instead of categorization problems at the level of unit of expression. It also allows for a continuous assessment over time without having to assess all the sequences at once.

The two-stage continuous affective state prediction system was tested on three different datasets of naturalistic facial expression videos. The three datasets varied depending on the type of tasks that the person recorded was engaged in. Across all datasets, the two-stage architecture performed better than the single-unit assessment approach. The results also outperformed the baseline set for the AVEC2013 challenge and the performance of other teams that participated in the challenge. The result of the *Valence* dimension on the development set was lower than that of baseline. However, the baseline method might have overfitted because the baseline *Valence* result on the AVEC2013 testing set was very low. Instead, our approach reached interesting results on the testing dataset showing possibly greater generalization capabilities through decoupling the modeling of the features from the modeling of the temporal relationship characterising the affective expression.

The results for the AVEC2012 dataset were also good with only the results on the *Expectation* dimension being worse than baseline results on the testing set. This could be due to the fact that for *Expectation* faster changes in expressions may lead to a lower contribution of temporal information during the modeling process, as shown in [40]. The results also showed that modeling the most correlated affective dimensions together did not lead to better results. This is possibly due to the fact that the correlation was not very high and that this may

[2]Taken from http://sspnet.eu/avec2013/

[3]Our paper [82] appearing in AVEC2013 is for the DSC sub-challenge only. Although we submitted testing results for both ASC and DSC sub-challenges, due to time limitation, only the paper on DSC was submitted for publication in the challenge proceedings.

even strongly decrease according to the dataset at hand. It is possible that to exploit their weak relationship, more complex approaches are needed when fusing them together as shown in other works (e.g. [69], [61], [62]).

It was also found that, overall, the use of the EOH feature only yielded better performance than the other image features in most cases and for both one-stage and two-stage types of approach. This is interesting as it reduces modeling complexity. A possible reason for this is that the EOH feature captures not only the edge information of the image but also its spatial information. However, the results were worse than those of the AVEC2012 winning system [90]. The main reason for this was probably the optimization of the features used in [90] and computed using advanced methods. These results together suggest that a combination of a two-stage approach proposed here and optimized features may lead to further improvements in the recognition rates.

In this paper, the proposed method was only tested on the facial expression image sequences. However, the modeling and affective dimension prediction method is independent of the affective dimension or affective modality used. However, as it was discussed above, it is possible that different delay parameters may be needed as different modalities or different affective dimensions may present different temporal dynamics and temporal dependencies.

In conclusion, the method proposed appears to be a good candidate for building automatic real-time affective state prediction systems thanks to its lower computational complexity during training and the fact that the predicted values depend only on past information. It is ideal for real-world applications where the signals are inputted in streams and continuous affective state levels are expected to be predicted in streams. TDNN can be regarded as a simple model of deep networks, other models (e.g., [92]) will be studied for facial expression analysis in future work.

REFERENCES

[1] D. R. Carney, A. J. Cuddy, and A. J. Yap, "Power posing: Brief nonverbal displays affect neuroendocrine levels and risk tolerance," *Psychological Science*, vol. 21, pp. 1363–1368, 2010.

[2] N. Bianchi-Berthouze, "Understanding the role of body movement in player engagement," *Human-Computer Interaction*, vol. 28, no. 1, pp. 40–75, 2013.

[3] A. Laszlo and K. Castro, "Technology and values: Interactive learning environments for future generations." *Educational Technology*, vol. 35, no. 2, pp. 7–13, 1995.

[4] A. Singh, A. Klapper, J. Jia, A. Fidalgo, A. Jimenez, N. Kanakam, N. Bianchi-Berthouze, and A. Williams, "Motivating people with chronic pain to do physical activity: Opportunities for technology design," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2014, pp. 2803–2812.

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

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

[7] F. Schwenker, S. Scherer, Y. Magdi, and G. Palm, "The GMM-SVM supervector approach for the recognition of the emotional status from speech," in *ICANN*, ser. LNCS, 2009, vol. 5768, pp. 894–903.

[8] M. Pantic and L. J. Rothkrantz, "Automatic analysis of facial expressions: The state of the art," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, pp. 1424–1445, 2000.

[9] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39–58, 2009.

[10] H. Meng, B. Romera-Paredes, and N. Bianchi-Berthouze, "Emotion recognition by two view SVM_2K classifier on dynamic facial expression features," in *FG*, 2011, pp. 854–859.

[11] A. Kleinsmith and N. Bianchi-Berthouze, "Affective body expression perception and recognition: A survey," *IEEE Transactions on Affective Computing*, vol. 4, no. 1, pp. 15–33, 2013.

[12] Y. Gao, N. Bianchi-Berthouze, and H. Meng, "What does touch tell us about emotions in touchscreen-based gameplay?" *ACM Transactions on Computer-Human Interaction*, vol. 19, no. 4, 2012.

[13] 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.

[14] H. Martinez, Y. Bengio, and G. Yannakakis, "Learning deep physiological models of affect," *Computational Intelligence Magazine, IEEE*, vol. 8, no. 2, pp. 20–33, 2013.

[15] G. Molina, T. Tsoneva, and A. Nijholt, "Emotional brain-computer interfaces," in *3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, ACII 2009*, Sept 2009, pp. 1–9.

[16] S. Koelstra and I. Patras, "Fusion of facial expressions and eeg for implicit affective tagging," *Image Vision Comput.*, vol. 31, no. 2, pp. 164–174, Feb. 2013.

[17] H. Singh, M. Bauer, W. Chowanski, Y. Sui, D. Atkinson, S. Baurley, M. Fry, J. Evans, and N. Bianchi-Berthouze, "The brains response to pleasant touch: an eeg investigation of tactile caressing," *Frontiers in Human Neuroscience*, vol. 8, no. 893, 2014.

[18] N. Bianchi-Berthouze, L. Berthouze, and T. Kato, "Understanding subjectivity: An interactionist view," in *UM99 User Modeling*, ser. CISM International Centre for Mechanical Sciences, J. Kay, Ed. Springer Vienna, 1999, vol. 407, pp. 3–12.

[19] T. Kanade, J. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis," in *Proceedings of Fourth IEEE International Conference on Automatic Face and Gesture Recognition*, 2000, pp. 46–53.

[20] M. Pantic, M. F. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," in *Proceedings of IEEE Int'l Conf. Multimedia and Expo (ICME'05)*, Amsterdam, The Netherlands, July 2005, pp. 317–321.

[21] A. Kleinsmith, P. R. D. Silva, and N. Bianchi-Berthouze, "Cross-cultural differences in recognizing affect from body posture." *Interacting with Computers*, vol. 18, no. 6, pp. 1371–1389, 2006.

[22] A. Kleinsmith, P. Silva, and N. Bianchi-Berthouze, "Grounding affective dimensions into posture features," in *Affective Computing and Intelligent Interaction*, ser. Lecture Notes in Computer Science, J. Tao, T. Tan, and R. Picard, Eds., vol. 3784. Springer Berlin Heidelberg, 2005, pp. 263–270.

[23] A. Dhall, R. Goecke, J. Joshi, M. Wagner, and T. Gedeon, "Emotion recognition in the wild challenge 2013," in *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*, ser. ICMI '13. New York, NY, USA: ACM, 2013, pp. 509–516.

[24] 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 *FG*, 2011, pp. 57–64.

[25] M. F. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, and M. Pantic, "AVEC 2013 - the continuous audio/visual emotion and depression recognition challenge," in *International Conference on ACM Multimedia - Audio/Visual Emotion Challenge and Workshop*, 2013.

[26] T. A. Olugbade, M. H. Aung, N. Bianchi-Berthouze, N. Marquardt, and A. C. Williams, "Bi-modal detection of painful reaching for chronic pain rehabilitation systems," in *Proceedings of the 16th International Conference on Multimodal Interaction*, ser. ICMI '14. New York, NY, USA: ACM, 2014, pp. 455–458. [Online]. Available: http://doi.acm.org/10.1145/2663204.2663261

[27] 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 Transactions on Systems, Man, and Cybernetics, Part B*, vol. 41, no. 3, pp. 664–674, 2011.

[28] N. Savva, A. Scarinzi, and N. Bianchi-Berthouze, "Continuous recognition of player's affective body expression as dynamic quality of aesthetic experience," *IEEE Transactions on Computational Intelligence and AI in Games*, vol. 4, no. 3, pp. 199–212, 2012.

[29] H. Gurkok and A. Nijholt, "Affective brain-computer interfaces for arts," in *Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII)*, 2013, pp. 827–831.

[30] J. Nijhar, N. Bianchi-Berthouze, and G. Boguslawski, "Does movement recognition precision affect the player experience in exertion games?" in *Intelligent Technologies for Interactive Entertainment*, ser. Lecture Notes

of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, A. Camurri and C. Costa, Eds. Springer Berlin Heidelberg, 2012, vol. 78, pp. 73–82.

[31] P. Ekman and W. V. Friesen, *Facial action coding system*. Consulting Psychologists Press, 1978.

[32] H. Gunes and B. Schuller, "Categorical and dimensional affect analysis in continuous input: Current trends and future directions," *Image Vision Comput.*, vol. 31, no. 2, pp. 120–136, Feb. 2013.

[33] A. Metallinou and S. Narayanan, "Annotation and processing of continuous emotional attributes: Challenges and opportunities," in *10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, April 2013, pp. 1–8.

[34] B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie, and M. Pantic, "AVEC 2011 the first international audio/visual emotion challenge," in *International Conference on Affective Computing and Intelligent Interaction (ACII 2011)*, Oct. 2011.

[35] B. Schuller, M. F. Valstar, R. Cowie, and M. Pantic, "AVEC 2012: the continuous audio/visual emotion challenge - an introduction," in *ICMI*, 2012, pp. 361–362.

[36] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, and P. Liu, "A high-resolution spontaneous 3d dynamic facial expression database," *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, vol. 0, pp. 1–6, 2013.

[37] X. Li, T. Pfister, X. Huang, G. Zhao, and M. Pietikainen, "A spontaneous micro-expression database: Inducement, collection and baseline," in *10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, 2013, pp. 1–6.

[38] H. J. Griffin, M. S. Aung, B. Romera-Paredes, C. McLoughlin, G. McKeown, W. Curran, and N. Bianchi-Berthouze, "Laughter type recognition from whole body motion," in *Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII)*, 2013, pp. 349–355.

[39] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Collecting large, richly annotated facial-expression databases from movies," *IEEE MultiMedia*, vol. 19, no. 3, pp. 34–41, 2012.

[40] H. Meng and N. Bianchi-Berthouze, "Affective state level recognition in naturalistic facial and vocal expressions," *IEEE Transactions on Cybernetics*, vol. 44, no. 3, pp. 315–328, 2014.

[41] J. Cheng, Y. Deng, H. Meng, and Z. Wang, "A facial expression based continuous emotional state monitoring system with gpu acceleration," in *10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, April 2013, pp. 1–6.

[42] E. Sariyanidi, H. Gunes, and A. Cavallaro, "Automatic analysis of facial affect: A survey of registration, representation and recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PP, no. 99, pp. 1–1, 2014.

[43] 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," *Journal of Multimedia*, vol. 1, no. 6, pp. 22–35, 2006.

[44] B. Romera-Paredes, A. Argyriou, N. Berthouze, and M. Pontil, "Exploiting unrelated tasks in multi-task learning." in *AISTATS*, ser. JMLR Proceedings, N. D. Lawrence and M. Girolami, Eds., vol. 22. JMLR.org, 2012, pp. 951–959.

[45] B. Romera-paredes, H. Aung, N. Bianchi-Berthouze, and M. Pontil, "Multilinear multitask learning," in *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, S. Dasgupta and D. Mcallester, Eds., vol. 28, no. 3. JMLR Workshop and Conference Proceedings, May 2013, pp. 1444–1452.

[46] L. Shao, X. Zhen, D. Tao, and X. Li, "Spatio-temporal laplacian pyramid coding for action recognition," *Cybernetics, IEEE Transactions on*, vol. 44, no. 6, pp. 817–827, June 2014.

[47] B. Jiang, M. Valstar, and M. Pantic, "Action unit detection using sparse appearance descriptors in space-time video volumes," in *IEEE International Conference on Automatic Face Gesture Recognition and Workshops (FG 2011)*, March 2011, pp. 314–321.

[48] E. Sariyanidi, H. Gunes, and A. Cavallaro, "Probabilistic subpixel temporal registration for facial expression analysis," in *Proceedings of the Asian Conference on Computer Vision*, 2014.

[49] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1993.

[50] C. Bregler, "Learning and recognizing human dynamics in video sequences," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 1997, pp. 568–574.

[51] L. De Silva and P. C. Ng, "Bimodal emotion recognition," in *FG*, 2000, pp. 332 –335.

[52] 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 *CVPR*, vol. 2, 2005, pp. 967–972.

[53] T. Kitazoe, S.-I. Kim, Y. Yoshitomi, and T. Ikeda, "Recognition of emotional states using voice, face image and thermal image of face," in *INTERSPEECH*, 2000, pp. 653–656.

[54] 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. ICSLP04*, 2004, pp. 889–892.

[55] 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 *FG*, USA, 2011.

[56] Y. Tong, J. Chen, and Q. Ji, "A unified probabilistic framework for spontaneous facial action modeling and understanding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 2, pp. 258–273, Feb 2010.

[57] Y. Tong, W. Liao, and Q. Ji, "Facial action unit recognition by exploiting their dynamic and semantic relationships," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 10, pp. 1683 –1699, oct. 2007.

[58] A. M. Rahman, M. I. Tanveer, and M. Yeasin, "A spatio-temporal probabilistic framework for dividing and predicting facial action units," in *International Conference on Affective Computing and Intelligent Interaction (ACII 2011)*, ser. LNCS, vol. 6975, 2011, pp. 598–607.

[59] 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 *International Conference on Affective Computing and Intelligent Interaction (ACII 2011)*, ser. LNCS, vol. 6974, 2011, pp. 609–618.

[60] Z. Wang, S. Wang, and Q. Ji, "Capturing complex spatio-temporal relations among facial muscles for facial expression recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013, pp. 3422–3429.

[61] G. A. Ramirez, T. Baltrušaitis, and L.-P. Morency, "Modeling latent discriminative dynamic of multi-dimensional affective signals," in *Proceedings of International Conference on Affective Computing and Intelligent Interaction (ACII 2011)*, ser. LNCS, vol. 6975. Springer, Oct. 2011, pp. 396–406.

[62] T. Baltrušaitis, N. Banda, and P. Robinson, "Dimensional affect recognition using continuous conditional random fields," in *IEEE Conference on Automatic Face and Gesture Recognition*, 2013, pp. 1–8.

[63] M. A. Nicolaou, H. Gunes, and M. Pantic, "Output-associative rvm regression for dimensional and continuous emotion prediction," *Image and Vision Computing, Special Issue on The Best of Automatic Face and Gesture Recognition 2011*, vol. 30, pp. 186–196, 2012, issue 3.

[64] T. Baltrušaitis, L.-P. Morency, and P. Robinson, "Continuous conditional neural fields for structured regression," in *European Conference on Computer Vision*, ser. Lecture Notes in Computer Science, vol. 8692, 2014, pp. 593–608.

[65] 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 *INTERSPEECH*, 2008, pp. 597–600.

[66] 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 *ICASSP*, May 2011.

[67] M. A. Nicolaou, H. Gunes, and M. Pantic, "Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space," *IEEE Transactions on Affective Computing*, vol. 2, pp. 92–105, 2011.

[68] 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 *INTERSPEECH*, 2010, pp. 2362–2365.

[69] M. A. Nicolaou, H. Gunes, and M. Pantic, "A multi-layer hybrid framework for dimensional emotion classification," in *Proceedings of the 19th ACM International Conference on Multimedia*, ser. MM '11. New York, NY, USA: ACM, 2011, pp. 933–936.

[70] H. Meng and N. Bianchi-Berthouze, "Naturalistic affective expression classification by a multi-stage approach based on hidden markov models," in *Proceedings of International Conference on Affective Computing and Intelligent Interaction (ACII 2011)*, ser. LNCS, vol. 6975, 2011, pp. 378–387.

[71] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. Lang, "Phoneme recognition using time-delay neural networks," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 37, no. 3, pp. 328–339, 1989.

[72] D. Molina, J. Liang, R. Harley, and G. Venayagamoorthy, "Comparison of TDNN and RNN performances for neuro-identification on small to medium-sized power systems," in *2011 IEEE Symposium on Computational Intelligence Applications In Smart Grid (CIASG)*, 2011, pp. 1–8.

[73] F. Eyben, M. Wöllmer, A. Graves, B. Schuller, E. Douglas-Cowie, and R. Cowie, "On-line emotion recognition in a 3-d activation-valence-time continuum using acoustic and linguistic cues," *Journal on Multimodal User Interfaces*, vol. 3, no. 1-2, pp. 7–19, 2010.

[74] K. Levenberg, "A method for the solution of certain non-linear problems in least squares," *Quart. J. Appl. Maths.*, vol. II, no. 2, pp. 164–168, 1944.

[75] J. R. J. Fontaine, K. R. Scherer, E. B. Roesch, and P. C. Ellsworth, "The World of Emotions is not Two-Dimensional," *Psychological Science*, vol. 18, no. 12, pp. 1050–1057, 2007.

[76] T. Ojala, M. Pietikainen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern Recognition*, vol. 29, no. 1, pp. 51 – 59, 1996.

[77] D. Huang, C. Shan, M. Ardabilian, Y. Wang, and L. Chen, "Local binary patterns and its application to facial image analysis: A survey," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 41, no. 6, pp. 765–781, Nov 2011.

[78] T. Ojala, M. Matti Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, pp. 971–987, 2002.

[79] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, pp. 915–928, June 2007.

[80] W. T. Freeman, W. T. Freeman, M. Roth, and M. Roth, "Orientation histograms for hand gesture recognition," in *International Workshop on Automatic Face and Gesture Recognition*, 1994, pp. 296–301.

[81] C. Yang, R. Duraiswami, and L. Davis, "Fast multiple object tracking via a hierarchical particle filter," in *Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*. Washington, DC, USA: IEEE Computer Society, 2005, pp. 212–219.

[82] H. Meng, D. Huang, H. Wang, H. Yang, M. AI-Shuraifi, and Y. Wang, "Depression recognition based on dynamic facial and vocal expression features using partial least square regression," in *Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge*, ser. AVEC '13. New York, NY, USA: ACM, 2013, pp. 21–30.

[83] V. Ojansivu and J. Heikkila, "Blur insensitive texture classification using local phase quantization," in *Image and Signal Processing*, ser. Lecture Notes in Computer Science, A. Elmoataz, O. Lezoray, F. Nouboud, and D. Mammass, Eds. Springer Berlin Heidelberg, 2008, vol. 5099, pp. 236–243.

[84] H. Drucker, Chris, B. L. Kaufman, A. Smola, and V. Vapnik, "Support vector regression machines," in *Advances in Neural Information Processing Systems 9*, vol. 9, 1997, pp. 155–161.

[85] J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, pp. 1161–1178, 1980.

[86] J. Han, X. Ji, X. Hu, D. Zhu, K. Li, X. Jiang, G. Cui, L. Guo, and T. Liu, "Representing and retrieving video shots in human-centric brain imaging space," *IEEE Transactions on Image Processing*, vol. 22, no. 7, pp. 2723–2736, July 2013.

[87] J. Han, K. Li, L. Shao, X. Hu, S. He, L. Guo, J. Han, and T. Liu, "Video abstraction based on fmri-driven visual attention model," *Inf. Sci.*, vol. 281, pp. 781–796, Oct. 2014.

[88] R. Cowie, E. Douglas-Cowie, S. Savvidou*, E. McMahon, M. Sawey, and M. Schröder, "'feeltrace': An instrument for recording perceived emotion in real time," in *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, 2000.

[89] G. Mckeown, M. Valstar, R. Cowie, and M. Pantic, "The semaine corpus of emotionally coloured character interactions," in *IEEE Int'l Conf. Multimedia, Expo (ICME'10)*, 2010, pp. 1079–1084.

[90] J. Nicolle, V. Rapp, K. Bailly, L. Prevost, and M. Chetouani, "Robust continuous prediction of human emotions using multiscale dynamic cues," in *Proceedings of the 14th ACM International Conference on Multimodal Interaction*, ser. ICMI '12. New York, NY, USA: ACM, 2012, pp. 501–508.

[91] E. Sánchez-Lozano, P. Lopez-Otero, L. Docio-Fernandez, E. Argones-Rúa, and J. L. Alba-Castro, "Audiovisual three-level fusion for continuous estimation of russell's emotion circumplex," in *Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge*, ser. AVEC '13. New York, NY, USA: ACM, 2013, pp. 31–40.

[92] L. Shao, D. Wu, and X. Li, "Learning deep and wide: A spectral method for learning deep networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 12, pp. 2303–2308, Dec 2014.