A Bayesian Filtering Framework for Continuous Affect Recognition from Facial Images

Ercheng Pei*, Yong Zhao*, Meshia Cédric Oveneke, Dongmei Jiang, Hichem Sahli

***Abstract*—Continuous affective state estimation from facial information is a task which requires the prediction of time series of emotional state outputs from a facial image sequence. Modeling the spatial-temporal evolution of facial information plays an important role in affective state estimation. One of the most widely used methods is Recurrent Neural Networks (RNN). RNNs provide an attractive framework for propagating information over a sequence using a continuous-valued hidden layer representation. In this work, we propose to instead learn rich affective state dynamics. We model human affect as a dynamical system and define the affective state in terms of valence, arousal and their higher-order derivatives. We then pose the affective state estimation problem as a jointly trained state estimator for high-dimensional input images, combining an RNN and a Bayesian Filter, i.e. Kalman filters (KF) and Extended Kalman filters (EKF), so that all weights in the resulting network can be trained using backpropagation. We use a recently proposed general framework for designing and learning discriminative state estimators framed as computational graphs. Such approach can handle high dimensional observations and efficiently optimize, in an end-to-end fashion, the state estimator. In addition, to deal with the asynchrony between emotion labels and input images, caused by the inherent reaction lag of the annotators, we introduce a convolutional layer that aligns features with emotion labels. Experimental results, on the RECOLA and SEMAINE datasets for continuous emotion prediction, illustrate the potential of the proposed framework compared to recent state-of-the-art models.**

***Index Terms*—recurrent neural network, Bayesian filter, adaptive alignment, continuous affect recognition**

This work is supported by the Chinese Scholarship Council (CSC) (grant 201706290115), the Shaanxi Provincial International Science and Technology Collaboration Project (grant 2017KW-ZD-14), the VUB Interdisciplinary Research Program through the EMO-App project, the Agency for Innovation by Science and Technology in Flanders (IWT) PhD grant nr. 131814, the Flemish Government (AI Research Program), and the Special Construction Fund for Key Disciplines of Shaanxi Provincial Higher Education.

Ercheng Pei is with School of Computer Science and Technology, Xi'an University of Posts and Telecommunications, Xi'an 710121, Shaanxi, China, with Shaanxi Key Laboratory of Network Data Analysis and Intelligent Processing, Xi'an 710121, Shaanxi, China, with Xi'an Key Laboratory of Big Data and Intelligent Computing, Xi'an 710121, Shaanxi, China and also with School of Computer Science, Northwestern Polytechnical University (NPU), Youyi Xilu 127, Xi'an 710072, China.
E-mail: peiercheng@mail.nwpu.edu.cn

Yong Zhao is with the NPU-VUB Joint AVSP Research Lab, Shaanxi Key Laboratory on Speech and Image Information Processing, National Engineering Laboratory for Integrated Aero-Space-Ground-Ocean Big Data Application Technology, School of Computer Science, Northwestern Polytechnical University (NPU), Youyi Xilu 127, Xi'an 710072, China.
E-mail: yong.zhao@mail.nwpu.edu.cn

Meshia Cédric Oveneke is with Department ETRO, Vrije Universiteit Brussel (VUB), Pleinlaan 2, 1050 Brussels, Belgium.
E-mail: mcovenek@etrovub.be

Dongmei Jiang is with the NPU-VUB Joint AVSP Research Lab, Shaanxi Key Laboratory on Speech and Image Information Processing, National Engineering Laboratory for Integrated Aero-Space-Ground-Ocean Big Data Application Technology, School of Computer Science, Northwestern Polytechnical University (NPU), Youyi Xilu 127, Xi'an 710072, China, and also with Peng Cheng Laboratory, Vanke Cloud City Phase I Building 8, Xili Street, Nanshan District, Shenzhen, Guangdong, 518055, China.
E-mail: jiangdm@nwpu.edu.cn

Hichem Sahli is with VUB-NPU joint AVSP Research Lab, Department of Electronics & Informatics (ETRO), Vrije Universiteit Brussel (VUB), Pleinlaan 2, 1050 Brussels, Belgium, and also with Interuniversity Microelectronics Centre (IMEC), Kapeldreef 75, 3001 Heverlee, Belgium.
E-mail: hsahli@etrovub.be

*These authors contributed equally to this work.

Manuscript received in YYY 2020; revised KKK 2021.(Corresponding author: Dongmei Jiang.)

I. Introduction

Automatic emotion recognition has a huge potential in human-computer interaction applications in fields such as human robot interaction and game entertainment interaction. In recent years, various automatic emotion recognition systems have been developed. They can be categorized as discrete emotion recognition and continuous affect recognition systems. Discrete emotion recognition assumes that the human emotional state can be described by several basic discrete emotions, such as the six basic emotions - happy, anger, sad, disgust, fear, and surprise [1]. However, the discrete categorical representation has an obvious drawback that it usually cannot cover the entire emotional space. To describe the subtle changes of emotions, continuous dimensional spaces have been proposed and a widely used dimensional model is the two dimensional affect model proposed by Russell in [2]. It consists of the arousal and valence dimensions. The arousal dimension denotes the level of excitement that the emotion depicts, and it ranges from sleepiness or boredom to wild excitement. The valence dimension defines the positivity or negativity of an emotion and it ranges from unpleasant feelings to pleasant feeling (sense of happiness).

Continuous affect recognition aims at estimating a series of emotional states (e.g. arousal and valence) of a person from an observed sequence of audiovisual data. Such sequence-to-sequence generation problem is very challenging mostly due to its complex temporal structure. Long short-term memory recurrent neural networks (LSTM) based models have been often used to model such complex temporal structure in continuous affective state recognition [3], [4], [5]. Often, the arousal and valence outputs (predictions) from LSTM are smoothed via a post-processing step involving median filtering, moving average or weighted moving average filtering [6], [7], [8], [9]. Other works posed the problem of affective state prediction as a time series filtering task, and proposed using Kalman filter [10], [11], [12], [13], switching Kalman

filter [14], and particle filter [15], [16] to continuously track each affective dimension. These approaches first adopt a static regression model, such as support vector regression or linear regression, on the input audio-video features to obtain the initial predictions of the affective states. Then the initial predictions are used, along with the ground-truth labels, as observations to learn the parameters of the considered Bayesian filters (BF), such as Kalman filter or particle filter, which models the dynamics of affective states as a state space model (SSM). When human perceive and label the real valued continuous affective dimensions, there exists uncertainty due to human expression and cognition of emotions. The post-processing methods based on moving average or weighted moving average try to reduce the uncertainty by smoothing the initial predictions. However, the process is equivalent to assuming an infinitely strong prior on the temporal structure of continuous affective states. On the other hand, Bayesian filter, based on state space model, models the uncertainty of continuous affective states more reasonably.

In our previous work [17], we addressed the problem of continuous affective state estimation from facial expressions by leveraging the Bayesian filtering paradigm, i.e. considering human affect as a latent dynamical system and recursively estimating its state using a sequence of visual observations. We modeled human affect as a dynamical system and defined the affective state in terms of valence, arousal and their higher-order derivatives. To address the non-linearity of real-world affect data, we proposed a deep extended Kalman filtering (DEKF) framework, in which stationary state transition and observation models of EKF are modeled using neural networks (NN). In addition, we introduced a sensor model implemented by a convolutional neural network (CNN) for extracting the visual features. The visual features from the sensor model are input to the DEKF for estimating the affective states. It needs to be noted that, the sensor, transition model and observation model are trained separately.

In the above Bayesian filter based affect recognition approaches, Bayesian filter was utilized to model the dynamics of the high-level affective states. However, the dynamics of the low-level visual data was ignored. In this work, we propose staking a Recurrent Neural Network (RNN) and a Bayesian filter to learn rich, dynamic representations of the visual data and affective states simultaneously. As a deep learning black box, RNN models the temporal structure of complex visual data, outputting the observations of affective states. Bayesian filter, as a "hand-designed" strong relation induction bias [18], models the spatio-temporal relationship between the observations and affective latent states. We propose learning the dynamics of both high-level affective states and low-level visual features by viewing the staked RNN-BF as a computational graph [19]. Specifically, the parameters of the BF module can be optimized based on a deterministic computational graph for training discriminative state estimators following the work of [20], therefore the RNN and BF can be jointly trained by the backpropagation through time (BPTT) algorithm [21]. The BF module, with a very sparse network structure designed with specialized domain knowledge, has very a small number of parameters. It is more suitable than LSTM to model the dynamics of the low-dimensional high-level affective states, more adaptable to scenarios with small amount of training data and low computing resources, and also benefits the generalization ability of the whole model. Compared to the affect recognition methods with LSTM and post-processing, our proposed RNN-BF framework can deal with the unreliability within the labels of the continuous affective dimensions. Compared to the methods with static regression model and Bayesian filter, as well as the DEKF model, our proposed RNN-BF framework can model the dynamics of the high-dimensional low-level visual data, and the dynamics of the low-dimensional high-level affective states, simultaneously.

When people annotate the affective dimensions, delays often exist between the labels and the input videos due to the reaction time. To deal with this challenge, in [22], [23], before training the regression models, grid search strategies were adopted to obtain the optimal delay by considering the linear correlation between the features and the affective labels, and the audiovisual signals were delayed to align with the labels. These methods actually performed a hard alignment to compensate for the delay. In fact, the mapping relationship between the features and affective labels is extremely complex, the hard alignment methods ignore the non-linear correlation between the features and affective labels. In this work, we propose to embed a Gaussian filter in the proposed RNN-BF model to achieve adaptive alignment (AA) on the features and labels, with the peak position of the Gaussian filter representing the delay time. The parameters of the Gaussian filter can be jointly optimized within the whole RNN-BF-AA framework. Overall, the proposed RNN-BF-AA framework models the dynamics of both low-level features and high-level affective states with uncertainty, and avoids the hard pre-/post-processing procedure such as hard smoothing and hard alignment.

The contributions of this work are:

(i) We model human affect as a dynamical system and define the affective state in terms of valence, arousal and their higher-order derivatives.
(ii) We frame the learning of the state estimators as a computational graph by stacking a Recurrent Neural Network (RNN) and a Bayesian filter to learn rich, dynamic representations of the visual data and the affective states simultaneously.
(iii) A highly synergistic RNN-BF framework is proposed for continuous affect recognition, where RNN models the dynamics of high-dimensional low-level visual data, while BF, with a very sparse network structure, models the dynamics of low-dimensional high-level affective states with uncertainty. The parameters of RNN and BF are jointly optimized as a computational graph for continuous affect recognition.
(iv) We further propose to embed a Gaussian filter in the RNN-BF framework for adaptively learning and aligning the input features with the labels of the affective dimensions. The resulting RNN-BF-AA framework is optimized jointly.

The remainder of this paper is organized as follows. In Section II we give an overview of the related works. The proposed model is detailed in Section III. Experimental results on the RECOLA [24] and the SEMAINE datasets [25], illustrating the performance of the proposed RNN-BF-AA framework, compared to state-of-the art, are presented in Section IV. Finally, conclusions are drawn in Section V.

II. Related work

A. RNN and Bayesian Filtering

Modeling the temporal structure of sequential data is a fundamental challenge in a range of tasks dealing with sequence-to-sequence learning such as speech recognition and synthesis, tracking, and continuous emotion prediction. RNN and Bayesian filtering have been used to deal with such tasks.

RNNs, such as LSTMs [26], [27], gated recurrent units (GRUs) [28] capture the temporal structure of sequential data via recurrent weights and gating mechanism. Their flexible functional form can support large scale models. BF models the dynamics between the hidden state and the measurement as a stochastic Markov process [29], [30]. Given a sequence of noisy measurement, the goal of Bayesian filtering is to estimate the optimal states. Similarly, RNN learns the prediction from a sequential input to the sequential output via passing information overtime through the hidden state. As noted by Gu *et al.* [31], one obvious difference between RNN and BF models is that for BF, the system transition and observation matrices change over time, indicating that it is an adaptive estimator; while for RNN, after the training stage, the learned weight matrices are usually fixed.

Recently, there have been a few studies that discuss the relationship between a Kalman filter and an RNN. To alleviate RNN training issues, several authors introduced Bayesian filtering-based updates for learning the parameters of RNN architecture [32], [33], [34], [35]. Downey *et al.* [36] proposed the Predictive State Recurrent Neural Networks (PSRNN) for filtering and prediction in dynamical systems. PSRNN uses bilinear transfer functions to combine information from multiple sources. Such bilinear functions arise naturally from state updates in Bayesian filters. Other authors proposed integrating an RNN with a Kalman filter. Coskun *et al.* [37] proposed long short-term memory Kalman filter as recurrent neural estimators for pose regularization. Instead of using a Kalman filter as a motion model and measurement model of a pose, they proposed to learn rich, dynamic representations of the motion and noise models. In particular, they proposed learning these models from data using long short-term memory. Krishnan *et al.* [38] introduced the use of an RNN as a component in a Kalman filter, while Haarnoja *et al.* [20] proposed to train a Kalman filter as a type of RNN with backpropagation. [20] combines a convolutional neural network with a Kalman filter to model a generative dynamical system without having to explicitly create a generative model of images. Essentially, as we did in [17], one can train a CNN on images to predict some target information, e.g., the coordinates of a target, and then use the representation as an input to a Kalman filter. The approach in [20] is to instead treat this process as one single computational graph, and to backpropagate through the Kalman filter to tune everything together. The standard Kalman Filter equations have been expressed in a computational graph, thereby integrating probabilistic state estimation with a per-frame CNN.

In this work, we follow the same ideas and propose a novel model for continuous affect recognition by staking an RNN and a BF to learn rich, dynamic representations of the affective states with uncertainties. Different from the model proposed in [37], we consider the staked RNN-BF as a computational graph and use the approach of [20] to train the proposed architecture in an end-to-end fashion using backpropagation.

B. Continuous affect recognition

Affective analysis from facial video sequences, comprising facial expression recognition and continuous affect recognition, has attracted research interest within the affective computing community. For example, [39] proposed an incremental facial expression recognition network, which can learn a competitive multi-class classifier with a lower requirement of computing resources. For the more recent work about facial expression recognition, readers can refer to [40], which well summarized deep learning based methods. It's should be noted that, affective image content analysis (AICA) is a related but different branch, which focuses on understanding the semantics at a higher level – the affective level, i.e. understanding the emotions that can be induced by the images in viewers [41]. Our work focuses on continuous affect recognition from facial images. In the following we introduce recent related works.

Recently, a series of new deep neural network (DNN) architectures have been used for sequence-to-sequence modeling. The emotion recognition domain has highly benefited with the advent of such DNNs[42], [43], [44], [45], [46]. In order to capture the temporal and contextual structure of continuous affect recognition, GRU [47], LSTM [48], bidirectional LSTM (BLSTM) [49], and Echo State Networks [50], Bidirectional Convolutional Recurrent Sparse Network (BCRSN) [51], have been often used. For example, [47] proposed Dynamic Facial Models (DFM) and Bidirectional GRU (BGRU) based continuous affect recognition method. To capture more low-level dynamic information encoded in multiple adjacent frames, [52] proposed 3DCNN and Convolutional LSTM (CLSTM) based continuous affect recognition method.

Other works posed the problem of affect state prediction as a time series filtering task and proposed the Bayesian filtering as a post-processing step to smooth the (track) the estimated affective states. Kalman filter [10], [11], [12], [13], switching Kalman filter [14], and particle filter [15], [16] have been used. Such approaches first adopt a static regression model, such as support vector regression or linear regression, on the input audio-video features to obtain initial predictions of affective states. Then the initial predictions are used, along with the ground-truth labels, as observations to learn the parameters of the considered Bayesian filter.

In our previous work [17], estimating the human affective state given an incoming stream of video sequences has been posed as a Bayesian filtering problem, i.e., estimating the

latent state of a dynamical system based on a sequence of noisy measurements related to the state of the system. This way of posing the human affective state estimation problem opens new horizons in terms of computation techniques used for designing an automated system for continuous emotion recognition. Leveraging the Bayesian filtering paradigm requires at least the two following models: a *transition model* for describing how the latent state, $\mathbf{z}_i$ (in our case valence, arousal and higher-order derivatives) evolves in time and an *observation model* for describing how the noisy measurements, $\mathbf{o}_i$, are related to the latent state. In classical engineering problems, the noisy measurements are directly obtained from the sensor, while in our case we only have access to raw video sequences. We, therefore, added a third model, *sensor model*, which describes the relationship between the raw video sequences $\mathbf{x}_i$ and the noisy measurements $\mathbf{o}_i$ related to the latent state. We proposed using a CNN representation to extract $\mathbf{o}_i$, which are then used along with the ground-truth labels and a DEKF, modeled as neuronal networks, to estimate the transition and observation models [17]. The sensor, transition and observation models have been trained separately.

In this work, to consider both the temporal structure of low-dimensional high-level affective states and the temporal structure of the high-dimensional low-level input visual streams, we propose an RNN-BF framework by stacking an RNN model in a BF model, where the output of RNN are fed BF as its observations. The BF module is implemented as a deterministic computational graph for training discriminative state estimators following the work of [20]. Also, the BF module can be regarded as a special type of RNN. So, the parameters of the proposed RNN-BF framework can be jointly optimized by the BPTT algorithm.

Besides, to deal with the annotation reaction lags, we introduce a time delay estimation by a Gaussian filter which is embedded in our model architecture to align the visual features to the affective states. Our approach follows the recent work of Khorram *et al.* [53], who introduced a multi-delay sinc network that can simultaneously align and predict labels in an end-to-end manner. Their network is a stack of convolutional layers followed by an aligner network that aligns the speech signal and emotion labels.

III. RNN - Bayesian Filtering Framework

In recognizing the continuous affective state from face video, there exists two temporal dynamic processes: the dynamic process of the low-level facial expressions (visual features), and the dynamic process of the high-level affective states. When human perceive and label the real valued continuous affective state, there exists uncertainty due to human expression and cognition of emotions, moreover, there are delays between the labels and the input video due to the reaction time. In this paper, to model all the factors above in one nutshell for continuous affect recognition, we propose an RNN-BF-AA framework, in which RNN models the dynamics of high-dimensional low-level visual data, while BF, with a very sparse network structure, models the dynamics of low-dimensional high-level affective states with uncertainty, and a Gaussian filter is embedded to learn the delay and adaptively align the visual features to the affective states. The parameters of the proposed RNN-BF-AA framework are jointly optimized by viewing the framework as a computational graph. In the RNN-BF-AA framework, the BF module captures the sparse structure of real-world generation processes and supports effective learning and reasoning algorithms, enables the proposed framework to adapt to low data and computing resources scenarios, and improves the generalization performance of the framework.

In this section, we first present the proposed RNN-BF framework for latent affective state estimation. As in [17], we formally define the affective state as a continuous time-dependent state vector $\mathbf{z}(t)$, with an associated affective state-space $\mathcal{A} \subset \mathbb{R}^{(n+1)}$ [17]:

$$\mathbf{z}(t) \triangleq \begin{pmatrix} z & \frac{dz}{dt} & \frac{d^2z}{dt^2} \end{pmatrix}^T \tag{1}$$

consisting of a level of valence (or arousal) and its higher-order derivatives. In this work, we consider the highest order to be $n = 2$, yielding to a state vector composed of the valence (or arousal) and its first and second order derivatives. The dynamics of the affective state can be modeled by a state space model which can be described by the follwing dynamic system:

$$\begin{aligned} \mathbf{z}_t &= f(\mathbf{z}_{t-1}) + \mathbf{q}_t && \text{dynamics update,} \\ \mathbf{o}_t &= g(\mathbf{z}_t) + \mathbf{r}_t && \text{observation inferred from RNN,} \end{aligned} \tag{2}$$

where f and g is a function, $\mathbf{q}_t$ and $\mathbf{r}_t$ is random variables that have particular probability distributions. This state space model is a probabilistic graph based generative model as illustrated in Figure 1(a). Figure 1(a) represents the generating process of data, for example, $\mathbf{o}_t$ are generated based on $\mathbf{z}_t$. The affective state estimation problem can be posed as the estimation of the state $\mathbf{z}_t$ at each discrete time step t given noisy observations $\mathcal{O}_t \triangleq \{\mathbf{o}_1, \ldots, \mathbf{o}_t\}$.

We formulate the general affective state estimation problem as computing (at each time step t) a *belief* (posterior distribution) of the affective state $\mathbf{z}_t$ given all available observations up to and including time step t:

$$\begin{aligned} p(\mathbf{z}_t|\mathcal{O}_t) &= p(\mathbf{z}_t|\mathcal{O}_{t-1}, \mathbf{o}_t) \\ &\propto p(\mathbf{o}_t|\mathbf{z}_t, \mathcal{O}_{t-1}) p(\mathbf{z}_t|\mathcal{O}_{t-1}) \\ &= \underbrace{p(\mathbf{o}_t|\mathbf{z}_t)}_{\text{update}} \underbrace{p(\mathbf{z}_t|\mathcal{O}_{t-1})}_{\text{prediction}} \end{aligned}. \tag{3}$$

Computing such belief is also known as *Bayesian filtering* [29]. In this work, the observations $\mathcal{O}_t$ are the outputs of an RNN model which maps the high-dimensional input visual data $\mathbf{x}_t$ to a lower dimensional vector $\mathbf{o}_t$.

To train the full model, we follow the computational graph framework of Haarnoja *et al.* [20], the Bayesian filtering process is formalized into a BF network layer, whose input is the observation of the current affect state, the affect state of the previous moment and its covariance, namely its uncertainty measure. Its output is the affect state and the uncertainty of the affect state at the current moment. The parameters of the Bayesian filter are taken as part of the parameters of

the neural network. Based on the error back propagation and gradient descent method, the whole framework is optimized jointly for continuous affect recognition. In this way, the network framework of RNN-BF is obtained, as illustrated in Figure 1(b).

The input of our model is the image features sequence $\mathbf{x}_t$, and its outputs are the estimated latent states $\mathbf{z}_t$, corresponding to the valence or arousal affective dimension. First, to model the temporal dynamics of the image features sequence $\mathbf{x}_t$, we use an RNN model followed by a fully connected (FC) layer to get outputs $\mathbf{o}_t$:

$$\begin{aligned} h_t &= RNN(x_t, h_{t-1}), \\ \mathbf{o}_t &= FC(h_t), \end{aligned} \tag{4}$$

with h_0 set equal to $\mathbf{0}$. Then $\mathbf{o}_t$ is regarded as observations for the BF, providing the state space $\mathbf{z}_t$ and it covariance matrix $\mathbf{P}_t$, which represents a measure of uncertainty for affect state estimation:

$$(\mathbf{z}_t, \mathbf{P}_t) = BF(\mathbf{o}_t, \mathbf{z}_{t-1}, \mathbf{P}_{t-1}). \tag{5}$$

Actually, one realization of the framework consisting of a CNN and a Kalman filter has been evaluated on a visual tracking task [20]. The CNN is introduced to estimate the observable parts of the true state given the input images. The basic idea of the approach is that the state labels $\mathbf{y}_t$, representing noiseless observations, are not incorporated into the Bayesian Filter update, but instead enter the model at training through the cost function. This enables optimizing the CNN, as an observation function, and the resulting low-dimensional features, being the filtered state, directly on the input images using standard backpropagation.

In this work, we follow the same ideas, however, use an RNN model, instead of a CNN, to fully consider the temporal structure of the input visual stream. So, the whole framework allows jointly and efficiently training the proposed RNN-BF model using BPTT algorithm. As an instance of BF, we summarize hereafter the used Kalman filter module, more details can be found in [20].

A. KF Module

The Kalman filter represents the exact solution of the Bayesian filtering problem (Equation(3)) when the process and observation models are linear and Gaussian. In the case of linear dynamics, the dynamic system functions (Equation 2) reduce to $f(\mathbf{z}_t) = \mathbf{F}\mathbf{z}_t$ and $g(\mathbf{z}_t) = \mathbf{H}\mathbf{z}_t$. The dynamics noise, $\mathbf{q}_t$, and observation noise $\mathbf{r}_t$, are assumed to be IID zero mean Gaussian random variables with covariances $\mathbf{Q}$ and $\mathbf{R}$. Then, the computing process of the Kalman filtering module is given by

$$\begin{aligned} \widehat{\mathbf{z}}_t &= \mathbf{F}\mathbf{z}_{t-1} && \text{state dynamics update,} \\ \widehat{\mathbf{P}}_t &= \mathbf{F}\mathbf{P}_{t-1}\mathbf{F}^T + \mathbf{Q} && \text{covariance dynamics update,} \\ \mathbf{K}_t &= \widehat{\mathbf{P}}\mathbf{H}^T(\mathbf{R} + \mathbf{H}\widehat{\mathbf{P}}_t\mathbf{H}^T)^{-1} && \text{Kalman gain,} \\ \mathbf{z}_t &= \widehat{\mathbf{z}}_t + \mathbf{K}_t(\mathbf{o}_t - H\widehat{\mathbf{z}}_t) && \text{observation update,} \\ \mathbf{P}_t &= (\mathbf{I} - \mathbf{K}_t\mathbf{H})\widehat{\mathbf{P}}_t && \text{covariance observation update.} \end{aligned}$$

In the case of non-linear dynamics, Kalman filter can be extended to extended Kalman filter (EKF), whose computing process can is obtained by setting $\mathbf{F} = \frac{\partial f}{\partial z}|_{z=\mathbf{z}_t}$ and $\mathbf{H} = \frac{\partial g}{\partial z}|_{z=\mathbf{z}_t}$. Note that in [20], the authors only consider the case of linear dynamics (Kalman filtering), and include an intermediary ground truth observation, $\mathbf{y}_t = \mathbf{C}_t\mathbf{z}_t$, by adding an extra fully-connected layer to obtain the KF module output, which is not the case in our implementation.

In the prediction step, the time update is based on the latent state estimate of the previous moment. In the correction step, the measurement is updated based on the observations at the current moment. In this work, the EK/EKF process is implemented as a EK/EKF (BF) layer, whose computational graph is showed in the red dashed rectangle of Figure 1(b). And here, the latent state $\mathbf{z}_t$ is inferred based on the observations $\mathbf{o}_t$.

While the KF/EKF process and measurement noise are typically formulated with full covariance matrices, in this work we consider them as isotropic. So, $\mathbf{Q}$ and $\mathbf{R}$ are diagonal matrices. To make sure that they also are positive definite matrices, we parameterize them with vectors $\mathbf{q}$ and $\mathbf{r}$, with $\mathbf{Q} = Diag(Softplus(\mathbf{q}))$ and $\mathbf{R} = Diag(Softplus(\mathbf{r}))$, with $Diag()$ the diagonalization of a vector, and $Softplus()$ is an activation function.

B. RNN Module

In this work, we adopt the long short-term memory recurrent neural networks (LSTM) as RNN module, The input image features, $\mathbf{x}_t$, are first fed to a feedforward network of two hidden layers:

$$\begin{aligned} h_t^1 &= \sigma(W_{h^1,x}\mathbf{x}_t + b_{h^1}), \\ h_t^2 &= \sigma(W_{h^2,h^1}h_t^1 + b_{h^2}). \end{aligned} \tag{6}$$

Then, the output h_t^2 is fed to an $LSTM$ layer for modeling the temporal dynamics of the sequential data.

$$(h_t, c_t) = LSTM(h_t^2, h_{t-1}, c_{t-1}), \tag{7}$$

where c_t the *cell* activation vectors at time t. h_0 and c_0 are set equal to $\mathbf{0}$.

The outputs h_t are further input to a fully connected layer with activation function $\sigma = \tanh$ to obtain the observations $\mathbf{o}_t$ for the BF module:

$$\mathbf{o}_t = \sigma(W_z h_t + b_z), \tag{8}$$

with W_z and b_z the weights and biases, respectively.

C. Adaptive alignment

Continuous emotion labels are generally not synchronized with the input video stream due to delays caused by reaction-time, which is inherent in a human annotation. To deal with this challenge, most previous works adopted grid search strategies to align the feature sequence to the annotation sequence in a pre-processing step before the training phase [22], [23]. These methods actually used a linear correlation based hard alignment method to compensate the delay, which directly considers the linear correlation between the features and

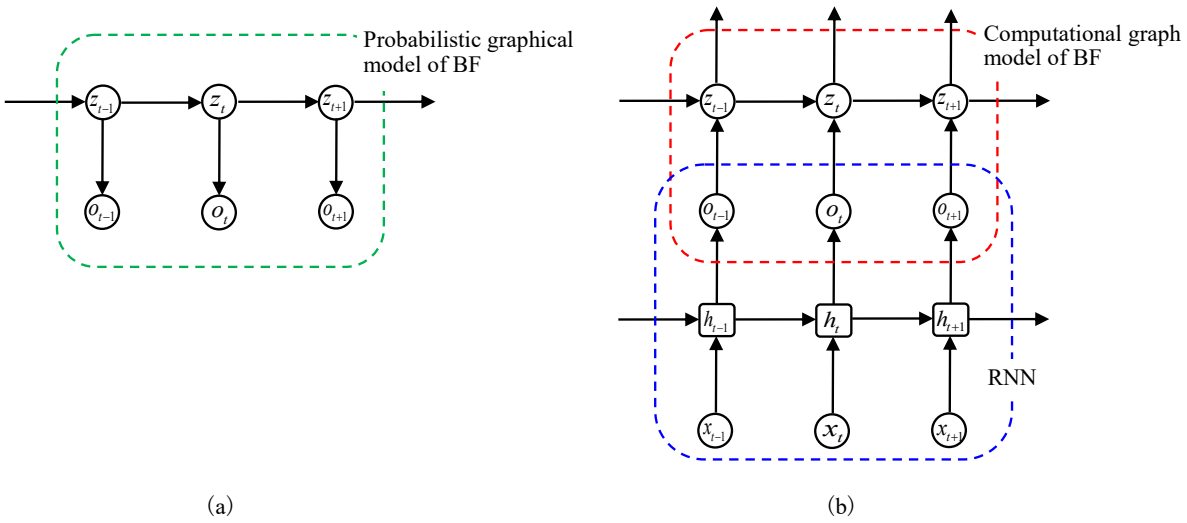

Fig. 1. (a): The probabilistic graphical model of the BF; (b): the proposed RNN-BF model which is formed by stacking a BF and an RNN for valance/arousal prediction. Green dashed rectangle: the probabilistic graphical model (generative model) of the BF, red dashed rectangle: the computational graph model (inference network) of the BF, blue dashed rectangle: RNN model.

affective labels to estimate the delay time. However, the mapping relationship between the features and affective labels is extremely complex, so these methods ignored the non-linear correlation between the features and affective labels. To solve this problem, this paper proposes to embed a Gaussian filter in the continuous affect recognition model to achieve automatic alignment of the affective labels and features, in which the peak position of the Gaussian filter indicates the delay time. The Gaussian filter is embedded in the proposed network model, which enables simultaneous alignment and prediction of affective states in an end-to-end manner.

Assume the output of RNN-BF is the predicted valence (arousal) $\mathbf{z}(t)$ without delay, if we know the time-delay, μ, the aligned predictions $\mathbf{y}(t)$ can be obtained by the following equation:

$$\mathbf{y}(t) = \mathbf{z}(t) * \delta(t-\mu), \tag{9}$$

with $*$ the convolution operator and δ the dirac-delta function. In this work, we approximate $\delta(t)$ with a Gaussian filter $f(t)$. The smaller the variance of the Gaussian filter is, the closer the approximation is to the $\delta(t)$ function:

$$\mathbf{y}(t) = \mathbf{z}(t) * f(t-\mu), \tag{10}$$

with

$$f(t) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{t^2}{\sigma^2}}. \tag{11}$$

In practice, we approximate the above equation by applying a rectangular window $h_w(t)$:

$$\mathbf{y}(t) = \mathbf{z}(t) * (f(t-\mu)h_w(t)), \tag{12}$$

where

$$h_w(t) = \begin{cases} 1, & t \in [-T, T] \\ 0, & otherwise \end{cases}, \tag{13}$$

and $[-T, T]$ is the range of the rectangular window $h_w(t)$. In our implementation, we discretize the above equation using the sampling frequency of $25Hz$ (the frame rate of labels in RECOLA), leading to the following discrete convolution.

$$\mathbf{y}_t = \mathbf{z}_t * (f_{t-\mu} h_{w,t}) = \sum_{k=-D}^{D} f_{k-\mu} \cdot \mathbf{z}_{t-k}, \tag{14}$$

where

$$f_{k-\mu} = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(k-\mu)^2}{\sigma^2}}, & k \in Z, -D \le k \le D \\ 0 & otherwise \end{cases}, \tag{15}$$

and $[-D, D]$ is the discrete time range. We further normalize the weights as

$$\mathbf{y}_t = \sum_{k=-D}^{D} \overline{f}_{k-\mu} \cdot \mathbf{z}_{t-k}, \tag{16}$$

with $\overline{f}_k = \frac{f_k}{\sum_{k=-D}^{D} f_k}$. As shown in Fig. 2, at time instance t, the peak of the Gaussian filter is at $t-\mu$, therefore the weighted sum $\mathbf{y}_t$ contains mainly the component of $\mathbf{z}_{t-\mu}$.

The parameters μ and σ are jointly learned with the other parameters of the model via gradient descent during the error backpropagation process. In the implementation, we re-parameter $\{\mu, \sigma\}$ as $\mu = 100 * tanh(\mu') \in (-100, 100)$ and $\sigma^2 = Softplus(\sigma'^2)$ where μ' and σ' are the learnable parameters of the Gaussian filter layer.

It should be noted that besides aligning the predicted valence/arousal to the annotations, as a side-effect, the Gaussian filter also introduces a further smoothing effect on the predicted values.

D. Model Learning

To optimize the proposed RNN-BF with adaptive alignment (RNN-BF-AA) we follow the above described approach of [20] [1]. All the parameter in the model are learned jointly. As

[1] code available at https://github.com/kyunghyuncho/backprop-kalman-filter

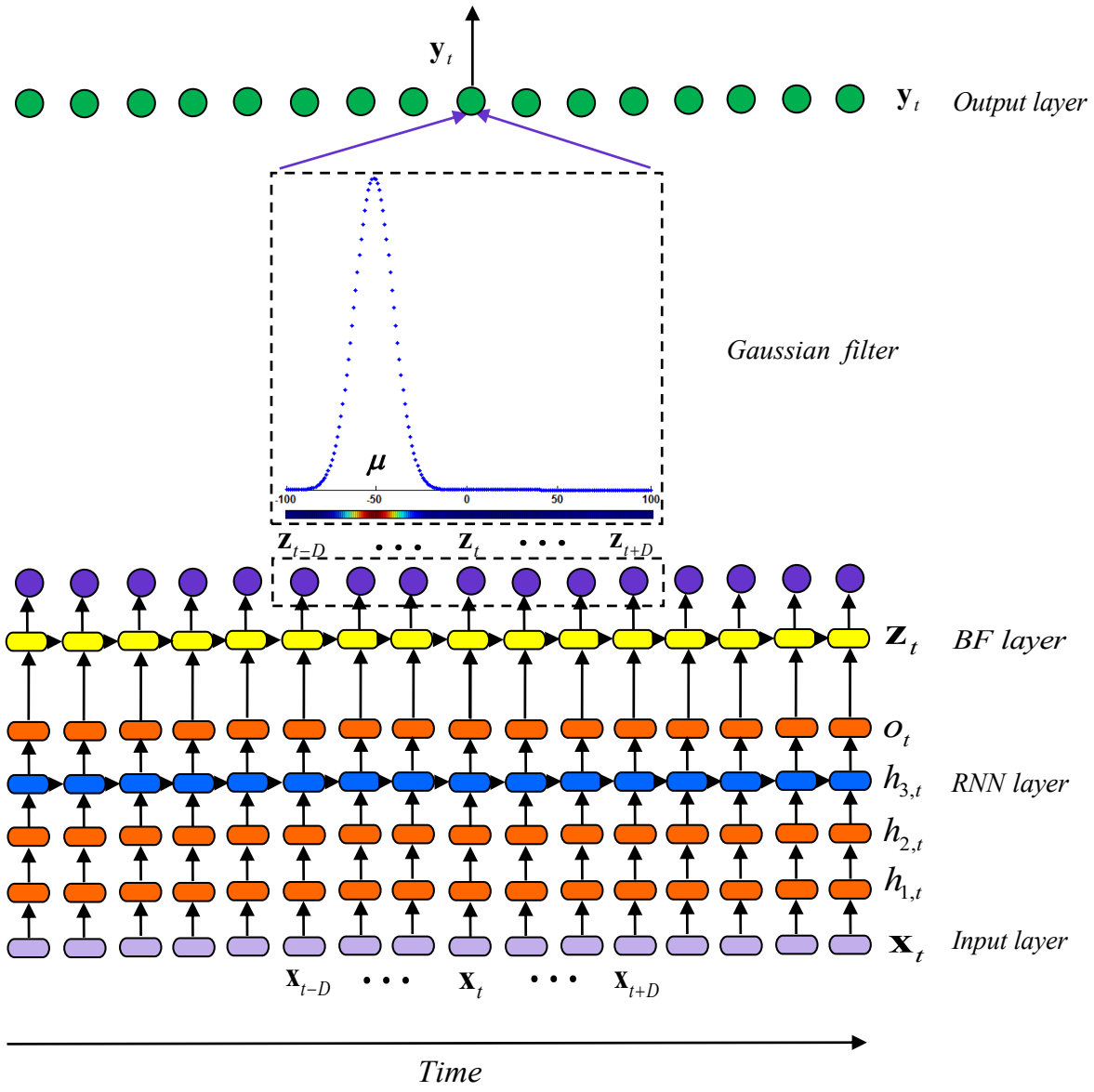

Fig. 2. The proposed computational graph framework for affective state estimation.

mentioned above, in our current model, the affective state vector is $\mathbf{z}_t = [z_t, \dot{z}_t, \ddot{z}_t]^T$, consisting of a level of valence/arousal and its first and second order derivatives, and the predicted delayed affective state vector is $\mathbf{y}_t = [y_t, \dot{y}_t, \ddot{y}_t]^T$. We train two models one for valence and one for arousal, with ground-truth $\widehat{\mathbf{y}}_t = [\widehat{y}_t, \widehat{\dot{y}}_t, \widehat{\ddot{y}}_t]^T$, using the following loss function:

$$\min_{\theta} O(\theta) = -CCC(\widehat{y}, y) + \lambda_1 RMSE(\widehat{\dot{y}}, \dot{y}) + \lambda_2 RMSE(\widehat{\ddot{y}}, \ddot{y}), \tag{17}$$

where θ includes all the parameters of the proposed model, and the parameters λ_1 and λ_2 are regularization parameters. In the above equation we omitted the alignment factor for simplicity.

The Concordance Correlation Coefficient (CCC) measure has been preferred to the Pearsons correlation coefficient (CC) because of its capability of better capturing the agreement between the estimated values and provided labels [54]:

$$CCC(\widehat{y}, y) = \frac{2\rho(\widehat{y}, y)\sigma_{\widehat{y}}\sigma_y}{\sigma_{\widehat{y}}^2 + \sigma_y^2 + (\mu_{\widehat{y}} - \mu_y)^2}, \tag{18}$$

where $\rho(\widehat{y}, y)$ is the Pearson's CC between ground truth and prediction, $\sigma_{\widehat{y}}$ and σ_y their respective standard deviations and $\mu_{\widehat{y}}$ and μ_y their respective means. The Root Mean Square Error (RMSE) is defined as:

$$RMSE(\widehat{\dot{y}}, \dot{y}) = \sqrt{E((\widehat{\dot{y}} - \dot{y})^2)}. \tag{19}$$

IV. Experimental results

To assess the effectiveness of our proposal, we conduct quantitative experiments on the RECOLA [24] dataset and SEMAINE [25] dataset. In this section we start by providing a brief description of the used datasets and features, then

we give the implementation details followed by a quantitative evaluation with comparison to state-of-the-art methods.

A. Datasets and Evaluation Metrics

RECOLA Dataset: To assess the performance of our proposed framework we conduct experiments on the AVEC2016 [23] dataset, being a subset of the RECOLA dataset [24]. RECOLA has been recorded to study socio-affective behaviours from multi-modal data in the context of remote collaborative work. Audio, video, electro-cardiogram (ECG) and electro-dermal activity (EDA) signals have been synchronously recorded from 27 French-speaking subjects. The dataset is equally divided in three partitions: training, development and testing, each having 9 recordings of 5 minutes. The dataset is labelled with arousal and valence at every $40ms$. Since the test set labels have not been released publicly, we use the provided validation set to compare our results to reported state-of-the-art results. The Concordance Correlation Coefficient (CCC) and the Root Mean Square Error (RMSE) are the commonly used metrics to evaluate affective state prediction performances. Higher CCC values indicate a better performance. Following the protocol of the AVEC2016 challenge, in our experiments, CCC and RMSE are calculated by averaging over all the sequences in the development set.

SEMAINE Dataset: We also conduct experiments on the AVEC2012 [55] dataset, being a subset of the SEMAINE dataset [25]. SEMAINE is a benchmark dataset of audio-video recordings of naturalistic human-agent interactions. Each video lasts about 3 to 5 minutes, recorded at 49.979 frames per second with 780×580 pixels as spatial resolution. The videos are provided with valence and arousal annotations. We follow the evaluation protocol of AVEC2012, in which the training set includes 31 videos, the development set consists of 32 videos, and 32 videos of test set. AVEC2012 used Pearsons correlation coefficient (CC) as evaluation metrics. Since the test set labels are not publicly released, we use the development set to compare our results to reported state-of-the-art results. In our experiments, the CC is computed by averaging over all the videos in the development set.

B. Features

For our experiments, we use the AVEC2016 baseline features (LGBP-TOP and Geometric) [23], and a recently proposed 3D scene flow features (3DSF) [56] on AVEC2016 and AVEC2012:

LGBP-TOP features. The LGBP-TOP features provided by the $AVEC2016$ challenge [23] are used for these experiments. First, the 50k LGBP-TOP features are reduced to 84 via principal components analysis (PCA) and keeping 99% of the total variance. Then, a window-based Avg & Std pooling is applied to the 84 dimensional features. The window size is set to 150 frames (6s) for arousal, and 100 frames (4s) for valence, resulting in a feature vector of dimension 168. The windows sizes are the ones that provided the best prediction accuracy on the development set.

Geometric features. We make use of the facial landmark provided by the $AVEC2016$ challenge [23]. Here also we apply window-based Avg & Std pooling. The window size is set to 100 frames (4s) for arousal, and 200 frames (8s) for valence, resulting in a feature vector of dimension 632.

3D scene flow features. First, we locate 2D facial landmarks [57], which are then used to fit a 3D morphable face model [58] to obtain for each frame a 3D point clouds. By calculating the displacements of the 3D point cloud for successive frames we obtain 3D scene flow features. Such features, representing facial muscle deformations, are robust to large facial pose, identity, illumination and background noise. Here also we apply window-based Avg & Std pooling. The window size is set to 7 frames, resulting in a 174 dimensional 3DSF feature vector.

C. Experimental setup

In our experiments we separately train two networks one for valence and one for arousal. Each network consists of (i) a feedforward network of two hidden layers, composed of a 29-unit fully-connected layer with $relu$ activation function, and a 16-unit fully-connected layer with $relu$ activation function, and (ii) a 16-unit LSTM layer followed by a 3-unit fully-connected layer with $tanh$ activation function; The output of the latter is regarded as observation and fed to (iii) the BF module. We evaluate both KF and EKF as BF modules. For the EKF, we adopt a linear mapping with $tanh$ activation function as observation model h and system model f. All the hyper-parameters of the model have been set empirically. The parameter of the adaptive alignment is set to $D = 100$, which is large enough to cover the annotation lag [7], [49]. Both λ_1 and λ_2 are set to 0.25. All the models are implemented based on PyTorch framework[59], and all the models are trained using Adam [60] based on a NVIDIA TITAN X GPU (12G), with learning-rate=0.001, β_1=0.9 and β_2=0.999 (coefficients used for computing running averages of gradient and its square).

D. Ablation study

To corroborate the effectiveness of the proposed RNN-BF-AA framework, we compare its performance with three models with different configurations: 1) a baseline model consisting of an LSTM followed by a Gaussian smoothing; 2) the RNN-BF frameworks without adaptive alignment which adopt KF and EKF as Bayesian Filters (BFs), respectively; 3) the RNN-BF frameworks with manual alignment(MA)[7], [49], using KF and EKF as BFs, respectively. We report the quantitative results of the ablation study for the different features in Table I, Table II and Table III. The reported mean and standard deviation are obtained by training the models 10 times with random initial network weights at each trial. For each trial the evaluation metrics, CCC and RMSE, are calculated by averaging over all the sequences in the development set.

From I and II, as can be observed, the proposed framework significantly improves the performance of continuous affect recognition compared to the baseline model consisting of an

TABLE I
ABLATION STUDY RESULTS ON THE DEVELOPMENT SET OF AVEC2016 USING LGBP-TOP FEATURES. THE SECOND AND THIRD COLUMN REPORTS THE MEAN AND STANDARD DEVIATION OF THE SCORES ON AROUSAL AND VALENCE, RESPECTIVELY. THE LAST COLUMN REPORTS THE AVERAGE CCC SCORE OVER-AROUSAL AND VALENCE.

Model	Arousal		Valence		Average	
	RMSE	CCC	RMSE	CCC	RMSE	CCC
Baseline	0.217±0.018	0.326±0.023	0.124±0.009	0.367±0.016	0.171±0.014	0.347±0.020
LSTM-KF	0.206±0.020	0.367±0.034	0.127±0.009	0.401±0.028	0.167±0.015	0.384±0.031
LSTM-EKF	0.209±0.024	0.367±0.017	0.127±0.005	0.420±0.020	0.168±0.015	0.394±0.019
LSTM-KF+MA	0.212±0.023	0.418±0.052	0.119±0.012	0.457±0.027	0.166±0.018	0.438±0.040
LSTM-EKF+MA	0.194±0.011	0.423±0.035	0.117±0.006	0.460±0.007	0.156±0.009	0.442±0.021
LSTM-KF-AA	0.198±0.005	0.424±0.026	0.118±0.003	0.474±0.014	0.158±0.004	0.449±0.020
LSTM-EKF-AA	**0.195**±0.012	**0.447**±0.029	**0.117**±0.004	**0.484**±0.014	**0.156**±0.008	**0.466**±0.022

TABLE II
ABLATION STUDY RESULTS ON THE DEVELOPMENT SET OF AVEC2016 USING GEOMETRIC FEATURES. THE SECOND AND THIRD COLUMN REPORTS THE MEAN AND STANDARD DEVIATION OF THE SCORES ON AROUSAL AND VALENCE, RESPECTIVELY. THE LAST COLUMN REPORTS THE AVERAGE CCC SCORE OVER-AROUSAL AND VALENCE.

Model	Arousal		Valence		Average	
	RMSE	CCC	RMSE	CCC	RMSE	CCC
Baseline	0.195±0.010	0.334±0.010	0.106±0.006	0.454±0.014	0.151±0.008	0.394±0.012
LSTM-KF	0.204±0.015	0.357±0.037	0.114±0.006	0.442±0.019	0.159±0.011	0.400±0.028
LSTM-EKF	0.198±0.011	0.359±0.027	0.116±0.004	0.441±0.011	0.157±0.008	0.400±0.019
LSTM-KF+MA	0.194±0.009	0.417±0.024	0.108±0.009	0.478±0.037	0.151±0.009	0.448±0.031
LSTM-EKF+MA	0.195±0.011	0.415±0.024	0.107±0.009	0.486±0.012	0.151±0,010	0.451±0.018
LSTM-KF-AA	0.184±0.007	0.417±0.022	0.111±0.004	0.480±0.022	0.148±0.006	0.449±0.022
LSTM-EKF-AA	**0.183**±0.003	**0.439**±0.026	**0.110**±0.005	**0.486**±0.020	**0.147**±0.004	**0.463**±0.023

TABLE III
ABLATION STUDY RESULTS ON THE DEVELOPMENT SET OF AVEC2016 USING 3DSF FEATURES. THE SECOND AND THIRD COLUMN REPORTS THE MEAN AND STANDARD DEVIATION OF THE SCORES ON AROUSAL AND VALENCE, RESPECTIVELY. THE LAST COLUMN REPORTS THE AVERAGE CCC SCORE OVER-AROUSAL AND VALENCE.

Model	Arousal		Valence		Average	
	RMSE	CCC	RMSE	CCC	RMSE	CCC
Baseline	0.172±0.007	0.426±0.071	0.116±0.008	0.432±0.024	0.144±0.008	0.429±0.048
LSTM-KF	0.163±0.012	0.563±0.057	0.120±0.007	0.450±0.028	0.142±0.010	0.507±0.043
LSTM-EKF	0.161±0.010	0.584±0.026	0.114±0.007	0.462±0.025	0.138±0.009	0.523±0.026
LSTM-KF-AA	0.167±0.009	0.574±0.027	**0.114**±0.007	**0.488**±0.014	0.141±0.008	0.531±0.021
LSTM-EKF-AA	**0.159**±0.009	**0.593**±0.030	0.115±0.008	0.480±0.024	**0.137**±0.009	**0.537**±0.027

LSTM with gaussian smoothing. Using LGBP-TOP features, LSTM-KF-AA improves the average CCC from 0.347 to 0.449, and LSTM-EKF-AA improves the average CCC from 0.347 to 0.466. Using Geometric features, LSTM-KF-AA improves the average CCC from 0.394 to 0.449, and LSTM-EKF-AA improves the average CCC from 0.394 to 0.463. In summary, the RNN-BF based methods consisting of LSTM and KF/EKF outperform the baseline method consisting of LSTM and Gaussian filter. This is because there exists dynamic uncertainly within the labels of the affective dimensions, and Bayesian filter can model such dynamic uncertainty more reasonably. After manual alignment, the performance is further improved. Compared to manual alignment, the proposed adaptive alignment obtains better results because it takes into account the non-linear relationship between the features and labels. All the results in Table I and Table II valid the effectiveness of our proposed method. It needs to be noted that the LSTM-EKF(-AA) model generally obtains better results than the LSTM-KF(-AA), meaning that there exists no-linear relationship in the dynamics of affective states. The same behaviour can be obtained when using the 3DSF features. From Table III, one can see that the 3DSF features provide the best performance in comparison to the other features. In fact, the 3DSF features can describe the natural facial muscle deformations, and are robust to self-occlusion and the variation of identity, face pose, and illumination, which bring the superiority for emotion recognition.

We also analyse the estimated annotation reaction lag (on AVEC2016) in terms of number of frames averaged on 10 trials, as listed in Table IV. The estimated delay μ is around $2s \sim 3s$ (50 frames ~ 75 frames), which corresponds to the delay time ($1s \sim 3s$) estimated in previous studies [7], [61]. In addition, the estimated Gaussian filter has a very sharp shape with a very small σ^2 approximating well the δ function.

The ablation study outcomes confirm that staking a Recurrent Neural Network (RNN) and a Bayesian Filter (BF) to learn rich, dynamic representations of the emotion dynamical model and noise models, along with the adaptive alignment of ground truth labels, perform well using different facial input feature sequences.

E. Comparison with the state-of-the-art methods

In the following we compare our outcomes with state-of-the-art methods using LGBP-TOP and Geometric features on

TABLE IV
ESTIMATED ANNOTATION REACTION LAG ON AVEC2016 (μ AND σ^2 IS CALCULATED IN TERMS OF FRAMES).

Model	Arousal				Valence			
	Geometric		LGBP-TOP		Geometric		LGBP-TOP	
	μ(seconds)	σ^2	μ(seconds)	σ^2	μ(seconds)	σ^2	μ(seconds)	σ^2
LSTM-KF-AA	54.7(2.2s)	1.03	53.3(2.1s)	0.84	59.2(2.4s)	0.89	51.8(2.1s)	0.90
LSTM-EKF-AA	60.8(2.4s)	1.00	70.6(2.8s)	0.82	53.2(2.1s)	0.81	59.5(2.4s)	0.78

AVEC2016. For these experiments we select the best models out of the 10 trials of the ablation study and used them to compare with the following state-of-the-art methods: using the same feature sets: SVR-KF approach of [11], the echo state networks approach (FESN) of [50], the DLSTM of [62], the bidirectional long short-term memory recurrent neural network (BLSTM) method of [49] and our previous work based on Deep Extended Kalman Filtering (DEKF) [17], using as inputs LGBP-TOP features. Table V and Table VI summarize the quantitative results of our proposed framework, along with the ones from state-of-the-art methods on AVEC2016. The scores of state-of-the-art approaches were directly taken from the original publications, except for the DEKF model [17], for which the scores were obtained using the LGBP-TOP features.

From Table V, one can observe from the CCC error metric that our model obtains the best average performance compared to all other models when using the LGBP-TOP features. From Table VI, one can see that, with the Geometric features, our proposed models obtain better results than most of the state-of-the-art works and comparative results to the FESN method [50]. With the LFBP-TOP features, our method obtains much better performance than the FESN method. In terms of average performance on the LGBP-TOP features and Geometric features, our method outperforms the FESN method. It should be noted that [50] applied lots of post-processing procedures such as predictions bias and scaling estimation, while our method does not adopt any post-processing methods. We also note that the proposed RNN-BF framework outperforms our previous DEKF model [17] in which we decoupled the observation model and the latent state estimation, confirming that jointly training the RNN-BF model provides better estimation of the latent space model.

In Table VII, we use the 3DSF features and compare our proposed model to state-of-the-art end-to-end approaches [63], [62], [48], [64], [65], [66] on AVEC2016. The proposed 3DSF+LSTM-EKF-AA approach obtains the best results. In addition, our proposed model can be combined with a deep learning backbone which extracts the features, and also can be used in other kind of time sequence problems. Due to the hardware limitations, in some cases, this end-to-end training might be not feasible, for example, when the batch-size or the length of training sequence is too large. For a better comparison with the state-of-the-art results, we list the results of ResNet-50+LSTM-EKF-AA. First, we jointly fine-tune the ResNet-50+DLSTM model based on the pre-trained ResNet-50 model using RECOLA dataset following

TABLE V
QUANTITATIVE EVALUATION ON THE DEVELOPMENT SET OF AVEC2016 USING LGBP-TOP FEATURES. THE SECOND AND THIRD COLUMN REPORTS THE SCORES ON AROUSAL AND VALENCE, RESPECTIVELY. THE LAST COLUMN REPORTS THE AVERAGE CCC SCORE OVER-AROUSAL AND VALENCE. $*$ DENOTES A JOINT ESTIMATION OF VALENCE AND AROUSAL, I.E. STATE VECTOR COMPOSED BY BOTH VALENCE AND AROUSAL AND THEIR DERIVATIVES.

Model	Arousal		Valence		Average	
	RMSE	CCC	RMSE	CCC	RMSE	CCC
LGBP-TOP+SVR+KF[11]	—	0.481	—	0.474	—	0.478
LGBP-TOP+FESN[50]	—	0.503	—	0.357	—	0.430
LGBP-TOP+DLSTM[62]	0.188	**0.535**	0.121	0.463	0.155	0.499
LGBP-TOP+DNN-BLSTM[49]	**0.170**	0.367	**0.098**	0.485	**0.134**	0.426
LGBP-TOP+DEKF [17]	0.234	0.454	0.137	0.375	0.186	0.414
LGBP-TOP + LSTM-EKF-AA*	0.179	0.493	0.100	0.462	0.140	0.477
LGBP-TOP + LSTM-EKF-AA	0.178	0.500	0.109	**0.505**	0.144	**0.503**

TABLE VI
QUANTITATIVE EVALUATION ON THE DEVELOPMENT SET OF AVEC2016 USING GEOMETRIC FEATURES. THE SECOND AND THIRD COLUMN REPORTS THE SCORES ON AROUSAL AND VALENCE, RESPECTIVELY. THE LAST COLUMN REPORTS THE AVERAGE CCC SCORE OVER-AROUSAL AND VALENCE. $*$ DENOTES A JOINT ESTIMATION OF VALENCE AND AROUSAL, I.E. STATE VECTOR COMPOSED BY BOTH VALENCE AND AROUSAL AND THEIR DERIVATIVES.

Model	Arousal		Valence		Average	
	RMSE	CCC	RMSE	CCC	RMSE	CCC
Geometric+SVR+KF[11]	—	0.297	—	**0.612**	—	0.455
Geometric+FESN[50]	—	**0.516**	—	0.582	—	**0.549**
Geometric+DLSTM[62]	0.189	0.411	0.111	0.488	0.150	0.449
Geometric+DNN-BLSTM[49]	**0.162**	0.366	**0.095**	0.503	**0.129**	0.435
Geometric + LSTM-EKF-AA*	0.194	0.457	0.104	0.571	0.149	0.514
Geometric+ LSTM-EKF-AA	0.181	0.484	0.110	0.524	0.146	0.504

TABLE VII
COMPARATIVE EMOTION STATE PREDICTION RESULTS ON THE DEVELOPMENT SET OF AVEC2016. THE SECOND AND THIRD COLUMN REPORTS THE SCORES ON AROUSAL AND VALENCE, RESPECTIVELY. THE LAST COLUMN REPORTS THE AVERAGE CCC SCORE OVER-AROUSAL AND VALENCE. $^+$ DENOTES OUR REPRODUCED RESULTS OF RESNET-50+DBLSTM WITHOUT ANY POST-PROCESSING. $*$ DENOTES A JOINT ESTIMATION OF VALENCE AND AROUSAL, I.E. STATE VECTOR COMPOSED BY BOTH VALENCE AND AROUSAL AND THEIR DERIVATIVES.

Approach	Arousal		Valence		Average	
	RMSE	CCC	RMSE	CCC	RMSE	CCC
CNN_1+LSTM [63]	0.201	0.346	**0.107**	0.511	0.154	0.429
CNN_2+DLSTM [62]	0.203	0.336	0.116	0.538	0.160	0.437
ResNet-50+DLSTM [48]	—	0.371	—	**0.637**	—	0.504
ResNet-50+DLSTM$^+$	—	0.265	—	0.423	—	0.344
ResNet-50+ LSTM-EKF-AA	—	0.370	—	0.536	—	0.453
3DMM based Data Augmentation+ResNet-50+GRU[64]	—	0.312	—	0.554	—	0.433
Recurrent Attention Network[65]	—	—	—	—	**0.102**	0.546
FaceNet/DenseNet-201+DLSTM[66]	—	0.570	—	0.550	—	0.560
3DSF + LSTM-EKF-AA*	0.166	0.577	0.110	0.497	0.138	0.537
3DSF + LSTM-EKF-AA	**0.148**	**0.638**	0.113	0.520	0.131	**0.579**

the work in [48]. Then, the 640 dimensional feature vector is extracted based on the trained ResNet-50+DLSTM model. The extracted features are fed into LSTM-EKF-AA for continuous affect recognition. As can be seen from Table VII (ResNet-50+DLSTM$^+$ denotes our reproduction of [48] without any post-processing), our approaches obtain better results than CNN_1+LSTM [63] and CNN_2+DLSTM [62], while lower performance than the ResNet-50+DLSTM approach (Tzirakis et.al' work) [48]. These reasons are in two-folds: first, [48] applied a chain of post-processing to the predictions, such as centring (by computing the bias between gold-standard and prediction), scaling (using the ratio of standard-deviation of gold-standard and prediction as scaling factor); second, the fine-tuning of the ResNet-50 model is difficult. These outcomes validate our proposed RNN-BF model to learn rich, dynamic representations of the emotion dynamical model and noise models. We further compared our results to recent works. From Table VII we can see that, our proposed 3DSF+LSTM-EKF-AA approach obtains better results than the Recurrent Attention Network based approach [65] and the FaceNet/DenseNet-201+DLSTM based approach [66].

To better illustrate the performance of our proposed 3DSF+LSTM-EKF-AA approach, we show the estimated arousal curves on the AVEC2016 video 1 in Figure 3 and the estimated valence curves on the AVEC2016 video 9 in Figure 4. From Figure 3 and Figure 4 we can notice, when the subject's face is frontal or have a small pose, the peak or valley of the arousal/valence curves can be predicted very well. Meanwhile, we can see that the high estimation errors are usually within the video segments where the subject's face has a large pose with only little facial information can be utilized, which makes it challenging to predict the arousal/valence values.

To further assess the generalization of our proposed model, we also evaluate its performance on the AVEC2012 (part of the SEMAINE) dataset and compare its outcome to reported state-

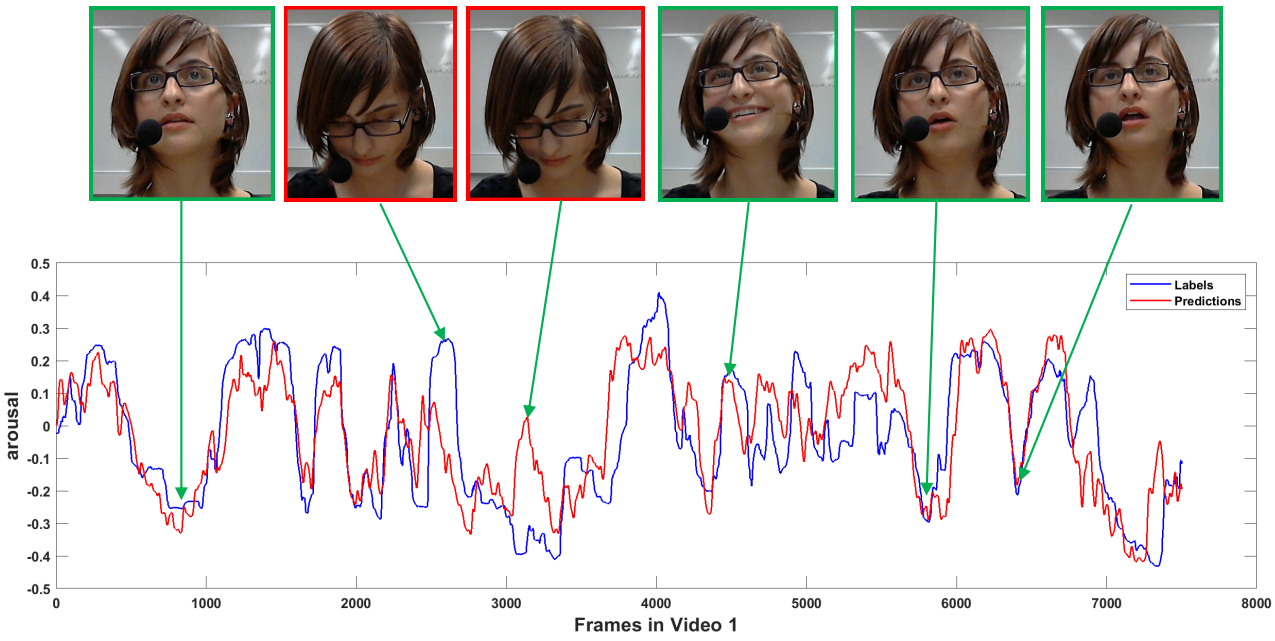

Fig. 3. Temporal curves of arousal on the AVEC2016 development set and some successful cases (see the green box) and failed cases (see the red box).

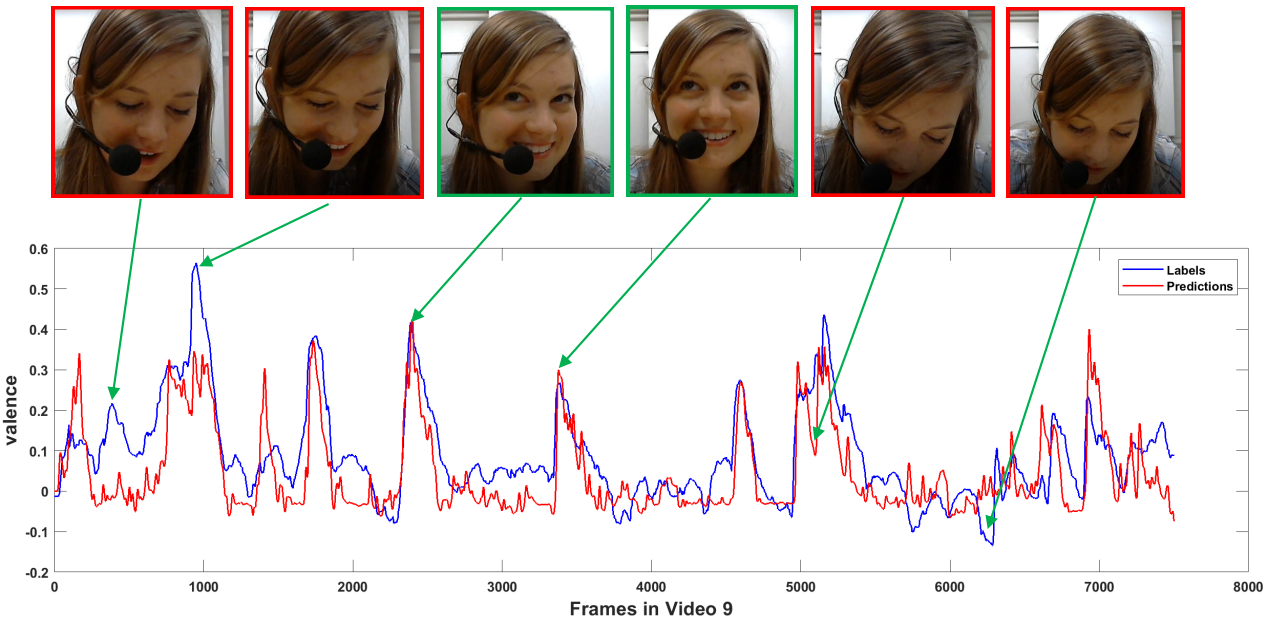

Fig. 4. Temporal curves of valence on the AVEC2016 development set and some successful cases (see the green box) and failed cases (see the red box).

of-the-art models for continuous affect recognition. Table VIII summarizes our outcome compared to the DFM+BGRU model of [47], the CNN+LSTM of [17], the convex unsupervised representation learning (CURL) and extended Kalman filtering (CURL-DEKF) of [17], and the 3D-CNN+LSTM of [52]. Here again, we obtain the best results for arousal prediction and valence prediction.

TABLE VIII
QUANTITATIVE EVALUATION, IN TERMS OF PEARSONS CORRELATION COEFFICIENT (CC), ON THE AVEC2012 DEVELOPMENT SET.

Method	Arousal	Valence	Average
DFM+BGRU [47]	0.38	0.32	0.35
CNN+LSTM [17]	0.36	0.39	0.38
3DCNN+CLSTM [52]	0.46	0.52	0.49
CURL-DEKF [17]	0.54	0.53	0.54
3DSF+LSTM-EKF-AA	**0.63**	**0.63**	**0.63**

V. CONCLUSIONS

In this work, to consider both the temporal structure of low-dimensional high-level affective states and the temporal structure of the high-dimensional low-level input visual streams, we propose an RNN-BF framework by stacking an RNN model in a BF model. The parameters of the proposed RNN-BF framework are jointly optimized as a computational graph for continuous affect recognition.

Besides, to deal with the annotation reaction lags, we introduce a time delay estimation by a Gaussian filter which is embedded in our proposed RNN-BF framework to adaptively align the input features with the labels of the affective states. The resulting RNN-BF-AA framework is optimized jointly.

The performance of the resulting RNN-BF-AA framework was empirically evaluated on publicly available benchmark for continuous affect recognition, for which we obtained state-of-the-art results. These facts manifest this work brought a potential direction and can inspire more powerful methods for continuous affect recognition. In the future work, we will implement more variations of KF, and do more experiments to further improve the performance of continuous affect recognition.

REFERENCES

[1] P. Ekman and W. V. Friesen, *Unmasking the face: A guide to recognizing emotions from facial clues*. Ishk, 2003.
[2] J. A. Russell, "A circumplex model of affect." *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.
[3] 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 *Proceedings of the Ninth Annual Conference of the International Speech Communication Association (INTERSPEECH 2008)*, 2008, pp. 597–600.
[4] 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, no. 2, pp. 92–105, 2011.
[5] M. Wöllmer, M. Kaiser, F. Eyben, B. Schuller, and G. Rigoll, "Lstm-modeling of continuous emotions in an audiovisual affect recognition framework," *Image and Vision Computing*, vol. 31, no. 2, pp. 153–163, 2013.
[6] E. Pei, L. Yang, D. Jiang, and H. Sahli, "Multimodal dimensional affect recognition using deep bidirectional long short-term memory recurrent neural networks," in *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2015, pp. 208–214.
[7] L. He, D. Jiang, L. Yang, E. Pei, P. Wu, and H. Sahli, "Multimodal affective dimension prediction using deep bidirectional long short-term memory recurrent neural networks," in *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, 2015, pp. 73–80.
[8] S. Chen, Q. Jin, J. Zhao, and S. Wang, "Multimodal multi-task learning for dimensional and continuous emotion recognition," in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*. ACM, 2017, pp. 19–26.
[9] J. Huang, Y. Li, J. Tao, Z. Lian, Z. Wen, M. Yang, and J. Yi, "Continuous multimodal emotion prediction based on long short term memory recurrent neural network," in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*. ACM, 2017, pp. 11–18.

[10] K. Markov, T. Matsui, F. Septier, and G. Peters, "Dynamic speech emotion recognition with state-space models," in *2015 23rd European Signal Processing Conference (EUSIPCO)*. IEEE, 2015, pp. 2077–2081.

[11] K. Somandepalli, R. Gupta, M. Nasir, B. M. Booth, S. Lee, and S. S. Narayanan, "Online affect tracking with multimodal kalman filters," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2016, pp. 59–66.

[12] Z. Huang and J. Epps, "An investigation of emotion dynamics and kalman filtering for speech-based emotion prediction." in *INTERSPEECH*, 2017, pp. 3301–3305.

[13] T. Dang, V. Sethu, and E. Ambikairajah, "Dynamic multi-rater gaussian mixture regression incorporating temporal dependencies of emotion uncertainty using kalman filters," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4929–4933.

[14] E. Pei, X. Xia, L. Yang, D. Jiang, and H. Sahli, "Deep neural network and switching kalman filter based continuous affect recognition," in *2016 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. IEEE, 2016, pp. 1–6.

[15] A. Savran, H. Cao, M. Shah, A. Nenkova, and R. Verma, "Combining video, audio and lexical indicators of affect in spontaneous conversation via particle filtering," in *Proceedings of the 14th ACM international conference on Multimodal interaction*. ACM, 2012, pp. 485–492.

[16] A. Savran, H. Cao, A. Nenkova, and R. Verma, "Temporal bayesian fusion for affect sensing: Combining video, audio, and lexical modalities," *IEEE transactions on cybernetics*, vol. 45, no. 9, pp. 1927–1941, 2014.

[17] M. C. Oveneke, Y. Zhao, E. Pei, A. D. Berenguer, D. Jiang, and H. Sahli, "Leveraging the deep learning paradigm for continuous affect estimation from facial expressions," *IEEE Transactions on Affective Computing*, 2019.

[18] P. W. Battaglia, J. B. Hamrick, V. Bapst, A. Sanchez-Gonzalez, V. Zambaldi, M. Malinowski, A. Tacchetti, D. Raposo, A. Santoro, R. Faulkner *et al.*, "Relational inductive biases, deep learning, and graph networks," *arXiv preprint arXiv:1806.01261*, 2018.

[19] J. B. Hamrick, K. R. Allen, V. Bapst, T. Zhu, K. R. McKee, J. B. Tenenbaum, and P. W. Battaglia, "Relational inductive bias for physical construction in humans and machines," *arXiv preprint arXiv:1806.01203*, 2018.

[20] T. Haarnoja, A. Ajay, S. Levine, and P. Abbeel, "Backprop kf: Learning discriminative deterministic state estimators," in *Advances in Neural Information Processing Systems*, 2016, pp. 4376–4384.

[21] R. Hecht-Nielsen, "Theory of the backpropagation neural network," in *Neural networks for perception*. Elsevier, 1992, pp. 65–93.

[22] F. Ringeval, F. Eyben, E. Kroupi, A. Yuce, J.-P. Thiran, T. Ebrahimi, D. Lalanne, and B. Schuller, "Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data," *Pattern Recognition Letters*, vol. 66, pp. 22–30, 2015.

[23] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic, "Avec 2016: Depression, mood, and emotion recognition workshop and challenge," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2016, pp. 3–10.

[24] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the recola multimodal corpus of remote collaborative and affective interactions," in *Proceedings of the 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG 2013)*. IEEE, 2013, pp. 1–8.

[25] G. McKeown, M. F. Valstar, R. Cowie, and M. Pantic, "The semaine corpus of emotionally coloured character interactions," in *2010 IEEE International Conference on Multimedia and Expo*. IEEE, 2010, pp. 1079–1084.

[26] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[27] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with lstm," in *Proceedings ICANN 1999, 9th International Conference on Artificial Neural Networks*. IET, 1999, pp. 850–855.

[28] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," *arXiv preprint arXiv:1409.1259*, 2014.

[29] Z. Chen *et al.*, "Bayesian filtering: From kalman filters to particle filters, and beyond," *Statistics*, vol. 182, no. 1, pp. 1–69, 2003.

[30] C. M. Bishop, *Pattern recognition and machine learning*. springer, 2006.

[31] J. Gu, X. Yang, S. De Mello, and J. Kautz, "Dynamic facial analysis: From bayesian filtering to recurrent neural network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1548–1557.

[32] F. A. Gers, J. A. Pérez-Ortiz, D. Eck, and J. Schmidhuber, "Dekf-lstm." in *ESANN*, 2002, pp. 369–376.

[33] J. A. Pérez-Ortiz, F. A. Gers, D. Eck, and J. Schmidhuber, "Kalman filters improve lstm network performance in problems unsolvable by traditional recurrent nets," *Neural Networks*, vol. 16, no. 2, pp. 241–250, 2003.

[34] B. Todorović, C. Moraga, and M. Stanković, "Sequential bayesian estimation of recurrent neural networks," in *Claudio Moraga: A Passion for Multi-Valued Logic and Soft Computing*. Springer, 2017, pp. 173–199.

[35] T. Ergen and S. S. Kozat, "Efficient online learning algorithms based on lstm neural networks," *IEEE transactions on neural networks and learning systems*, vol. 29, no. 8, pp. 3772–3783, 2017.

[36] C. Downey, A. Hefny, B. Boots, G. J. Gordon, and B. Li, "Predictive state recurrent neural networks," in *Advances in Neural Information Processing Systems*, 2017, pp. 6053–6064.

[37] H. Coskun, F. Achilles, R. DiPietro, N. Navab, and F. Tombari, "Long short-term memory kalman filters: Recurrent neural estimators for pose regularization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 5524–5532.

[38] R. G. Krishnan, U. Shalit, and D. Sontag, "Deep kalman filters," *arXiv: Machine Learning*, 2015.

[39] J. Zhu, B. Luo, S. Zhao, S. Ying, X. Zhao, and Y. Gao, "Iexpressnet: Facial expression recognition with incremental classes," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 2899–2908.

[40] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE transactions on affective computing*, 2020.

[41] S. Zhao, X. Yao, J. Yang, G. Jia, G. Ding, T.-S. Chua, B. W. Schuller, and K. Keutzer, "Affective image content analysis: Two decades review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[42] T. Zhang, W. Zheng, Z. Cui, Y. Zong, J. Yan, and K. Yan, "A deep neural network-driven feature learning method for multi-view facial expression recognition," *IEEE Transactions on Multimedia*, vol. 18, no. 12, pp. 2528–2536, 2016.

[43] H. Li, J. Sun, Z. Xu, and L. Chen, "Multimodal 2d+ 3d facial expression recognition with deep fusion convolutional neural network," *IEEE Transactions on Multimedia*, vol. 19, no. 12, pp. 2816–2831, 2017.

[44] S. Xie and H. Hu, "Facial expression recognition using hierarchical features with deep comprehensive multipatches aggregation convolutional neural networks," *IEEE Transactions on Multimedia*, vol. 21, no. 1, pp. 211–220, 2018.

[45] Z. Xia, X. Hong, X. Gao, X. Feng, and G. Zhao, "Spatiotemporal recurrent convolutional networks for recognizing spontaneous micro-expressions," *IEEE Transactions on Multimedia*, vol. 22, no. 3, pp. 626–640, 2019.

[46] S. Lin, M. Bai, F. Liu, L. Shen, and Y. Zhou, "Orthogonalization-guided feature fusion network for multimodal 2d+ 3d facial expression recognition," *IEEE Transactions on Multimedia*, 2020.

[47] S. Song, E. Sánchez-Lozano, M. Kumar Tellamekala, L. Shen, A. Johnston, and M. Valstar, "Dynamic facial models for video-based dimensional affect estimation," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2019, pp. 0–0.

[48] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, "End-to-end multimodal emotion recognition using deep neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1301–1309, 2017.

[49] E. Pei, D. Jiang, M. Alioscha-Perez, and H. Sahli, "Continuous affect recognition with weakly supervised learning," *Multimedia Tools and Applications*, pp. 1–26, 2019.

[50] M. Amirian, M. Kächele, P. Thiam, V. Kessler, and F. Schwenker, "Continuous multimodal human affect estimation using echo state networks," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2016, pp. 67–74.

[51] Y. Dong, X. Yang, X. Zhao, and J. Li, "Bidirectional convolutional recurrent sparse network (bcrsn): An efficient model for music emotion recognition," *IEEE Transactions on Multimedia*, vol. 21, no. 12, pp. 3150–3163, 2019.

[52] J. Huang, Y. Li, J. Tao, Z. Lian, and J. Yi, "End-to-end continuous emotion recognition from video using 3d convlstm networks," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 6837–6841.

[53] S. Khorram, M. G. Mcinnis, and E. M. Provost, "Jointly aligning and predicting continuous emotion annotations," *IEEE Transactions on Affective Computing*, pp. 1–1, 2019.
[54] F. Weninger, F. Ringeval, E. Marchi, and B. Schuller, "Discriminatively trained recurrent neural networks for continuous dimensional emotion recognition from audio," in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*. AAAI Press, 2016, pp. 2196–2202.
[55] B. Schuller, M. Valster, F. Eyben, R. Cowie, and M. Pantic, "Avec 2012: the continuous audio/visual emotion challenge," in *Proceedings of the 14th ACM international conference on Multimodal interaction*, 2012, pp. 449–456.
[56] E. Pei, M. C. Oveneke, Y. Zhao, D. Jiang, and H. Sahli, "Monocular 3d facial expression features for continuous affect recognition," *IEEE Transactions on Multimedia*, 2020.
[57] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "Openface 2.0: Facial behavior analysis toolkit," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 2018, pp. 59–66.
[58] X. Zhu, Z. Lei, J. Yan, D. Yi, and S. Z. Li, "High-fidelity pose and expression normalization for face recognition in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 787–796.
[59] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *arXiv preprint arXiv:1912.01703*, 2019.
[60] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
[61] Z. Huang, T. Dang, N. Cummins, B. Stasak, P. Le, V. Sethu, and J. Epps, "An investigation of annotation delay compensation and output-associative fusion for multimodal continuous emotion prediction," in *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, 2015, pp. 41–48.
[62] L. Chao, J. Tao, M. Yang, Y. Li, and Z. Wen, "Long short term memory recurrent neural network based multimodal dimensional emotion recognition," in *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2015, pp. 65–72.
[63] K. Brady, Y. Gwon, P. Khorrami, E. Godoy, W. Campbell, C. Dagli, and T. S. Huang, "Multi-modal audio, video and physiological sensor learning for continuous emotion prediction," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2016, pp. 97–104.
[64] D. Kollias, S. Cheng, E. Ververas, I. Kotsia, and S. Zafeiriou, "Deep neural network augmentation: Generating faces for affect analysis," *International Journal of Computer Vision*, vol. 128, no. 5, pp. 1455–1484, 2020.
[65] J. Lee, S. Kim, S. Kim, and K. Sohn, "Multi-modal recurrent attention networks for facial expression recognition," *IEEE Transactions on Image Processing*, vol. 29, pp. 6977–6991, 2020.
[66] L. Schoneveld, A. Othmani, and H. Abdelkawy, "Leveraging recent advances in deep learning for audio-visual emotion recognition," *Pattern Recognition Letters*, 2021.

Ercheng Pei received his Master degree in Computer Application Technology and Ph.D degree in Computer Science and Technology from the Northwestern Polytechnical University (NPU), Xi'an China in 2015 and 2020, respectively. Since 2021 he has been an associate professor at the School of Computer Science & Technology, Xi'an University of Posts & Telecommunications. His current research interest focuses on facial affective analysis, affective computing, computer vision and machine learning.

Yong Zhao received his Bachelor and Master degree in Computer Science and Technology and Computer Application Technologies from Northwestern Polytechnical University (NPU), Xian China in 2010 and 2013 respectively. He is currently working toward a PhD degree in Engineering Sciences at the Joint VUB-NPU Audio-Visual Signal Processing (AVSP) Laboratory of the VUB Electronics and Informatics (ETRO) department, under the supervision of Professor Hichem Sahli and Dongmei Jiang. His main research interest includes facial expression synthesis, facial animation and facial expression analysis. He is also interested in image processing, machine learning and human computer interaction.

Meshia Cédric Oveneke received his BScEng degree in Electronics & Information Technology and MScEng degree in Computer Science (Artificial Intelligence) with great distinction from the Vrije Universiteit Brussel (VUB), Belgium in 2011 and 2013, respectively. In 2018, he obtained his PhD degree in Engineering Sciences at the Joint VUB-NPU Audio-Visual Signal Processing (AVSP) Laboratory of the VUB Electronics and Informatics (ETRO) department. He is a teaching assistant in computer vision and his current research focuses on developing machine learning based approaches to audio-visual signal processing, applied to affective computing and behavior analysis. His main research interest includes image, video and speech processing, large-scale optimization, machine learning and Bayesian inference.

Dongmei Jiang received her Bachelor and Master degree in Automatic Control, and Ph.D degree in Computer Science and Technology from the Northwestern Polytechnical University (NPU), Xi'an China in 1994, 1997 and 2000, respectively. Since 2003 she has been an associate professor at the School of Computer Science, NPU. She became a professor in 2010, and currently she is the Head of the Computer Information Engineering Dept. of the School of Computer Science, NPU. She was a visiting scholar at the ETRO Dept., Vrije Universiteit Brussel (VUB), from November 2001 to June 2002, and from June 2006 to October 2007, respectively. Since 2005, she is the NPU's team coordinator of the Joint NPU-VUB Audio Visual Signal Processing (AVSP) Lab. Her research interest focuses mainly on multi-modal affective computing from speech, face video and text, 2D or 3D facial expression synthesis, and speech recognition. She is the regular reviewer of international journals such as Multimedia Tools and Applications.

Hichem Sahli holds a degree in Mathematics and Computer Science, DEA in computer vision, a PhD degree in computer sciences from the Ecole Nationale Sup. De Physique Strasbourg - France. He was affiliated as a 'Charger de Recherche' at the CAD and Robotics Dept. of the Ecole des Mines de Paris. Since 1999 he is professor at the Dept. of Electronics and Informatics (ETRO) and team-coordinator at the Interuniversitair Micro-Elektronica Centrum vzw (IMEC). Within ETRO-IRIS, he coordinates the research team in computer vision. The research team is dealing with variational and partial differential equations and stochastic models in computer vision, image and motion analysis; pattern recognition and machine learning for object detection recognition and tracking. Since 2005 he is the coordinator of the AVSP research cluster within ETRO, as well as the Joint NPU-VUB AVSP Lab. AVSP main area of research are mathematical models and tools for the application of novel machine learning techniques for multi-modal content analysis and automatic detection of affect in audio-visual speech and body gestures. He is a regular reviewer for most of the major vision and multimedia conferences and journals.