This ICCV Workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Dynamic Facial Models for Video-based Dimensional Affect Estimation

Siyang Song[1], Enrique Sánchez-Lozano[2], Mani Kumar Tellamekala [1], Linlin Shen[3], Alan Johnston[1] and Michel Valstar[1]
[1] University of Nottingham, UK
[2] Samsung AI Centre Cambridge, Cambridge, UK
[3] College of Computer Science and Software Engineering, Shenzhen University, China

Abstract

Dimensional affect estimation from a face video is a challenging task, mainly due to the large number of possible facial displays made up of a set of behaviour primitives including facial muscle actions. The displays vary not only in composition but also in temporal evolution, with each display composed of behaviour primitives with varying in their short and long-term characteristics. Most existing work models affect relies on complex hierarchical recurrent models unable to capture short-term dynamics well. In this paper, we propose to encode these short-term facial shape and appearance dynamics in an image, where only the semantic meaningful information is encoded into the dynamic face images. We also propose binary dynamic facial masks to remove 'stable pixels' from the dynamic images. This process allows filtering of non-dynamic information, i.e. only pixels that have changed in the sequence are retained. Then, the final proposed Dynamic Facial Model (DFM) encodes both filtered facial appearance and shape dynamics of a image sequence preceding to the given frame into a three-channel raster image. A CNN-RNN architecture is tasked with modelling primarily the long-term changes. Experiments show that our dynamic face images achieved superior performance over the standard RGB face images on dimensional affect prediction task.

1. Introduction

The face is an important asset for automatic human behaviour understanding, as it displays a wide range of cues about our cognitive state, including our affective state. Analysing human emotions by their face would find application in many cross-disciplinary fields, such as medicine [44], security, or entertainment [26]. Automatic emotion recognition by and large follows two main emotion theories: Ekman's six basic emotion model[6] or the dimensional affect model (a.k.a. Russel's Circumplex model [32]). The Circumplex model predicts values of emotional attributes such as arousal and valence on a continuous scale, where arousal is a physiological state of being alert, awake, attentive, and valence represents how negative or positive someone feels [8]. Although sometimes other dimensions are added, the combination of these two values represents a wide span of emotional states.

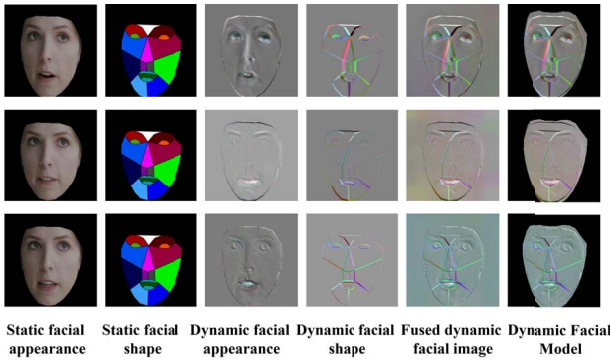


Figure 1: Illustrations of static facial appearance (SFA) and shape (SFS), dynamic facial appearance (DFA) and shape (DFS) and Dynamic Facial Model (DFM).

People express their affect through auditive, visual, and physiological signals, where the face is a highly valuable visual signal that can be sensed unobtrusively and that can also process many individuals at the same time in a shared scenario without the need for source-separation on a 1-dimensional signal. Motivated by this, existing affect analysis approaches build on analyzing the visual information provided by facial expressions. While some studies [17, 7, 33] analyze affect on a frame-by-frame basis, without exploiting the relationships between frames, the progression of affect has distinct temporal patterns that span multiple frames, and the values of arousal and valence are therefore highly correlated over time. Thus, temporal models should be used.

The standard approach for modelling dynamics is through sequential latent models, such as Recurrent Neural

Networks (RNN). These models exploit the temporal information by applying a set of latent variables that are supposed to model the intrinsic correlation that exists between the input and the output at a given frame, conditioned to the latent states at previous frames. However, they are generally used to learn dynamics from extracted features rather without considering the context of the face. Other works proposed to encode dynamics at the input level, by extracting features from a image sequence [25], constructing spectrum maps [40, 14] or an encoded image sequence [15]. However, these methods also have practical drawbacks, as the learning of later approach can become quite complex for long-term sequences. For instance, [15] proposed a temporal CNN approach that needs as many input channels as the number of frames being considered. This results in growing models and limited capacity: in [15] the number of input frames (and channels) is set to 5, which limits the temporal modelling of longer-term expressions at the input side.

To learn dynamics in the context of the face and avoid the limited capacity of encoding long sequence, this paper applies the dynamic image algorithm [1] to encode the short-term facial dynamics at the image level, which are further forwarded to a CNN-RNN-based model to re-encode both long-term and short-term variations at the feature level. Importantly, in doing so it keeps the framework simple. The dynamic image consists of a 3-channel raster image (similar to an RGB image) displaying a "summary" of an image sequence. This idea itself is very similar to temporal templates as introduced by Bobick & Davis in 2001 [2] but have proved its better ability in action recognition [1]. The use of a summarising image allows CNN-based architectures designed to take still images as input to process a video of variable length. While the dynamic image algorithm has been successfully applied for human action recognition, its extension to model the dynamics of facial actions is not straightforward. Bilen et al. made use of whole images to generate dynamic appearance [1], without segmentation of specific, semantically meaningful regions of objects (the human body, or the face), nor shape information, both of which are highly valuable for face analysis. In order to consider such information, this paper extends the dynamic image algorithm to account for shape domain by combining facial landmarks to produce a dynamic facial appearance (DFA) and shape image (DFS). After that, a Laplacian pyramid-based multi-scale transform is applied for the fusion of facial appearance and shape in order to retain maximum correlation between them.

The Dynamic Facial Images (DFIs) (examples are shown in Fig. 1) are generated per video frame, summarising the content from a few frames prior to the current one and are computationally efficient (please see [1] for details on efficiency). Importantly, they are still images, and thus can be processed by standard CNN architectures whilst retaining the short-term temporal information. To learn long-term dynamics the CNN is followed by a RNN model, and RNN is trained individually, tasked with returning valence and arousal values at each time step. In summary, the main contributions of this paper are as follows:

1. We extend the dynamic image algorithm to the face domain, dismissing non-face related attributes and encoding face dynamics in the context of the face.

2. We propose a **Dynamic Facial Model (DFM)** encoding algorithms that allows to integrate the facial appearance and shape into a standard RGB image, summarising the variation of both along time.

3. We compared the dynamic face images to standard RGB face images on two datasets, where the proposed approach achieved superior results for both arousal and valence estimation tasks.

2. Related Works

Dimensional affect estimation is often regarded as the regression problem where both valence and arousal are continuous values lying in the range $[-1, 1]$. Its growing interest has been investigated by a series of AVEC Challenges [46, 45, 30, 28, 29], aiming to gather all efforts in a common benchmark of increasing difficulty. Like many other Computer Vision disciplines, existing approaches are generally divided into those that use hand-crafted features with general-purpose machine learning techniques, and those that built on the recent advances in Deep Learning.

As time-series data, temporal modeling is crucial for dimensional affects analysis. As shown above, traditional hand-crafted approaches have been [20, 16, 11, 22, 13, 23] frequently used kernel-based regression, such as SVR, which by nature cannot model contextual information. To overcome such limitations, some hand-crafted features have been extended to the temporal domain. In [25], global and local features are extended to the temporal domain through the magnitude of the Fourier transform of each of them. In order to capture both long and short-term dynamics, they applied the Fourier transform at different scales i.e. at sequences of one to four seconds long. Also, some features extended the spatial domain to the temporal dimension, referred as the Three Orthogonal Planes (TOP), were widely used by AVEC baselines. In [16] the LBP features are extended to the temporal domain as the LBP-TOP, and further combined with a novel sparse regression method, achieving excellent performance on the SEMAINE database [21]. In [17], histogram-based features, such as LPQ, LBP and LGBP, are extended to the temporal dimension, and were further combined with deep features.

However, the TOP extension of features grows drastically in complexity as the number of frames increases, and

thus learning temporal models is a better choice. While graphical models such as HMMs or CRFs are powerful temporal representations, they are prone to failure when modeling long-term dynamics. These drawbacks can be tackled with Recurrent Neural Networks, which are feed forward networks of latent states that can be learned through back-propagation. RNNs can be used with either hand-crafted features or in combination with CNNs. Some extensions handling back-propagation problems in RNNs have been proposed too. In [24] a Bidirectional Long Short Term Memory Network (BLSTM) is used with hand-crafted features, showing better results than SVR. Hasani et al. [12] extract features using Inception module [42]. Combining the it with an LSTM yields better results than using the Inception module only in a per-frame basis. A similar approach is adopted in [18], where a relatively shallow CNN is used in combination with a RNN. Kollias et al. [19] showed how pre-trained networks can be adapted to affect estimation tasks with great success, as training some networks end-to-end might not be affordable due to the lack of data or resources. In particular, When combined with RNN, the VGG-Face network, with only fully-connected layers fine-tuned, yielded the best results, showing the great potential of using existing CNNs to predict the intensity of continuous dimensional affects on data gathered "in-the-wild".

RNNs can also be combined with other non-temporal regression techniques. In [10], the output of a RNN is combined with an SVR, thus preventing the former to incur in overfitting, and the latter not to consider the temporal domain. The proposed approach, coined Strength Modeling algorithm, applies the two models in a hierarchical manner.

3. The proposed approach

The main novelty of our method resides in the encoding of the short-term facial shape and appearance dynamics of image sequences into a single raster image. Our work differs from that of Nicolle et al. [25] in that we do not rely on the frequency domain, as it contains nuisance factors that are hard to capture with a CNN, and in that we incorporate the temporal modelling of a RNN. Similarly, it differs from [15] in that the dynamics are encoded into single images, allowing the use of a flexible number of frames, rather than as a concatenation of frames, which in practice limits the time extent of the short-term encoding. Finally, our work differs from that of [39] in that we encode precise dynamic from a image sequence rather than estimate it from a single image. We further calculate the 'dynamic pixels' while removing 'stable pixels' of the encoded DFA and DFS, allowing both shape dynamics and appearance dynamics to be summarized in a single image without redundant information ('stable pixels') while the dynamic image of [1] only contains appearance dynamics without considering the effects of redundant information.

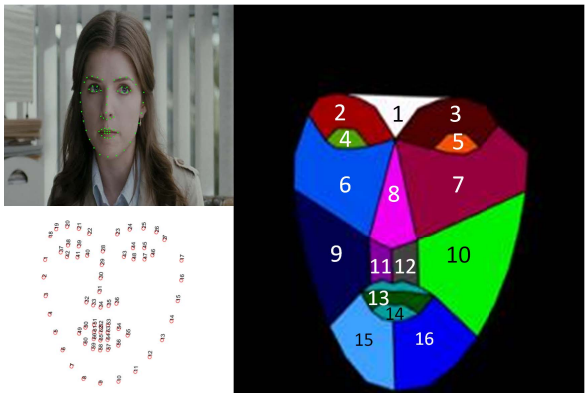


Figure 2: Generation of static facial shape (SFS) image

Our approach starts with detecting a set of 66 facial landmarks for each video frame. These landmarks, depicted in Fig. 2, are extracted using the publicly available code of iCCR [34]. These landmarks correspond to specific parts of the face, are then used to generate a static facial appearance (SFA) and static facial shape (SFS), per frame. Then, for each subsequence of T frames, the corresponding DFA and DFS are generated. These images are also generated per video frame, and subsequently fused into a sequence of DFM.

3.1. Static Facial Image

Static Facial Shape Image: Based on the detected facial landmarks, the face is segmented into 15 semantic regions. These regions correspond to the left and right eyebrows, left and right eyes, nose, left and right cheekbones, left and right cheeks, mouth, lips, left and right philtrums, and left and right jaws. In static shape images, each region is represented by a unique colour. All pixels lying out of the convex hull of the face are set to 0 in each colour channel (black). An example of SFS is shown in Fig. 2.

Static Facial Appearance Image: Using the aforementioned landmarks, a binary mask is applied to the original face image, whereby only the pixels lying within the convex hull defined by the landmarks are set to one. This mask is applied to the input image to generate the static appearance image, which basically accounts for the facial appearance. This way, the background noise is removed before the feature extraction process.

3.2. Dynamic Facial Image

The dynamic image is a parameter matrix whose parameters are learnt to rank the position of the given frames from their features by implementing dot product between the per-frame features and the dynamic image. That is to say, it is an operator that contains the evolution information of frames and consequently can be treated as the representation of given frames. By extending this algorithm to

include shape and adapting it to the face domain by leveraging facial landmarks, we obtained two novel dynamic facial images (DFI): **dynamic facial appearance (DFA)** and **dynamic facial shape (DFS).**

Let $I_t \in \mathbb{R}^{m \times n}$ be the t-th image of a sequence composed of T consecutive face-aligned images, all of size $m \times n$, and let $V_t = \frac{1}{\tau}\sum_{\tau=1}^{t} I_\tau$ be the average value image up to frame t. V_t is defined as the average of a given feature mapping of the image, $\phi(I_\tau)$. The mapping chosen in this paper is the same as that which attained highest performance in the original paper by Bilen et al. [1], which defined ϕ to be the identity function. Let $\mathbf{d} \in \mathbb{R}^d$ be the raw DFI of the image sequence. The ranking score for frame t is defined as the dot product between $\mathbf{d}$ and V_t:

$$\begin{aligned} S(\mathbf{d}, V_t) &= \langle \mathbf{d}, V_t \rangle \\ &= \sum_{l=1}^{3}\sum_{i=1}^{N}\sum_{l=1}^{M} d_{lij} \times v_{lij}^t \\ &= d_{11} \times v_{t111} + \cdots + d_{LNM} \times v_{tLNM} \end{aligned} \tag{1}$$

where d_{lij} and v_{lij}^t are the values of pixel lij in the dynamic face image d and static face image v_t, respectively. Thus, the goal is to learn the DFI so that if $q > t$, then $S(\mathbf{d}, V_q) > S(\mathbf{d}, V_t)$ because those closer frames normally contribute more information to current face status. In other words, $\mathbf{d}$ is learned so that when projected into the aggregated kernel of the input image size, it returns a score that sorts frames by time. This kernel ranks the input SFIs, and hence contains temporal evolution of the face image sequence end at the last image, making it a good facial dynamic descriptor for the last image. In order to learn $\mathbf{d}$, we minimise the hinge loss between pairs of scores:

$$\mathbf{d}^* = \underset{\mathbf{d}}{\operatorname{argmin}}\, E(\mathbf{d}) \tag{2}$$

$$E(\mathbf{d}) = \frac{\lambda}{2}\|\mathbf{d}\|^2 + \gamma \sum_{q>t} \max\{0, 1 - S(\mathbf{d}, V_q) \ + S(\mathbf{d}, V_t)\} \tag{3}$$

where $\gamma = \frac{2}{T(T-1)}$, is the L2-norm regularised error. The second term in Eq. 3 defines the number of pairs on the subset that are incorrectly ranked by the score function. A pair $q > t$ is said to be correctly ranked if $S(\mathbf{d}, V_q) \geq S(\mathbf{d}, V_t) + 1$. The minimization of Eq. 3 is accomplished with RankSVM [38]. The parameters in the final learned kernel $\mathbf{d}$ are in the real space. It is worth highlighting that the RankSVM algorithm is also applied to learn the DFIs $\mathbf{d}$ at test time, i.e. it is learned on the go for each subsequence of images.

In order to generate a set of **DFA** and **DFS** for a video, we take a set of $T-1$ consecutive frames prior to each frame, for which we first obtained the SFA and SFS, respectively. Then, DFA and DFS for each frame are learned by a sliding window of T frames. Therefore, for a video of N frames we have $N - T + 1$ DFA and DFS images (From frame T to frame N).

3.3. Fusion of dynamic appearance and dynamic shape

Both DFA and DFS are separately generated for each frame. While this is a common approach, after which the two descriptors are combined before being input to a machine learning hypothesis (e.g. SVR, CNN), we propose to fuse them into a single dynamic image, unifying shape and appearance as a single input stream to the ML hypothesis, retaining the context of the face. To the best of our knowledge, despite that many reports of approaches combine facial appearance and shape information at the feature or decision level for affect analysis, no previous work has proposed to fuse DFA and DFS into a three channel image and then learn both features and their correlations at the input level, which is interesting to explore.

From Equation 1, we can see that variation in pixel values of a static face image results in differences in the final score across the image sequence, as the kernel matrix (dynamic image) is a constant matrix in each case. In particular, we found from the Equation 1 that pixels whose values remain fixed over the image sequence have no influence on the frame ranking, because the dot product between these pixels in each frame, and corresponding pixels in dynamic image, are the same. Thus, they are not discriminative. In this paper, we call these pixels as the "stable pixels", denoted as $(i_{\text{sta}}, j_{\text{sta}})$, while the reminder is called "dynamic pixels", defined as $(j_{\text{dy}}, j_{\text{dy}})$, as they can contribute different scores to different frames. In this sense, Equation 1 can be re-written as:

$$\begin{aligned} S(\mathbf{d}, V_t) &= \langle \mathbf{d}, V_t \rangle \\ &= \langle \mathbf{d}_{\text{dy}}, V_{\text{dy}}^t \rangle + \langle \mathbf{d}_{\text{sta}}, V_{\text{sta}}^t \rangle \\ &= \sum_{i \in \text{dy}} (d_i \times v_i) + \sum_{j \in \text{sta}} (d_j \times v_j) \end{aligned} \tag{4}$$

where the $\langle \mathbf{d}_{\text{sta}}, V_{\text{sta}}^t \rangle = \sum_{j \in \text{sta}} (d_j \times v_j)$ is the constant and thus only $\langle \mathbf{d}_{\text{dy}}, V_{\text{dy}}^t \rangle$ leads the difference of scores. Therefore, even we setting all 'stable pixels' as 0, making dynamic image as a sparse matrix, it can still rank frames correctly. Since the DFS mainly contains the edge dynamics of each semantic region while DFA contributes more details about the detailed facial texture dynamics in each region, the 'dynamic pixels' of them are expected to be largely independent in the space domain, allowing the fusion of them not to lose significant information or highly distort the dynamics of the original DFA and DFS. Motivated by this, assuming that the DFA and DFS are generated from sequence Seq of T face images, the framework applies the following steps to fuse DFA and DFS images. This process is also illustrated in Fig. 3.

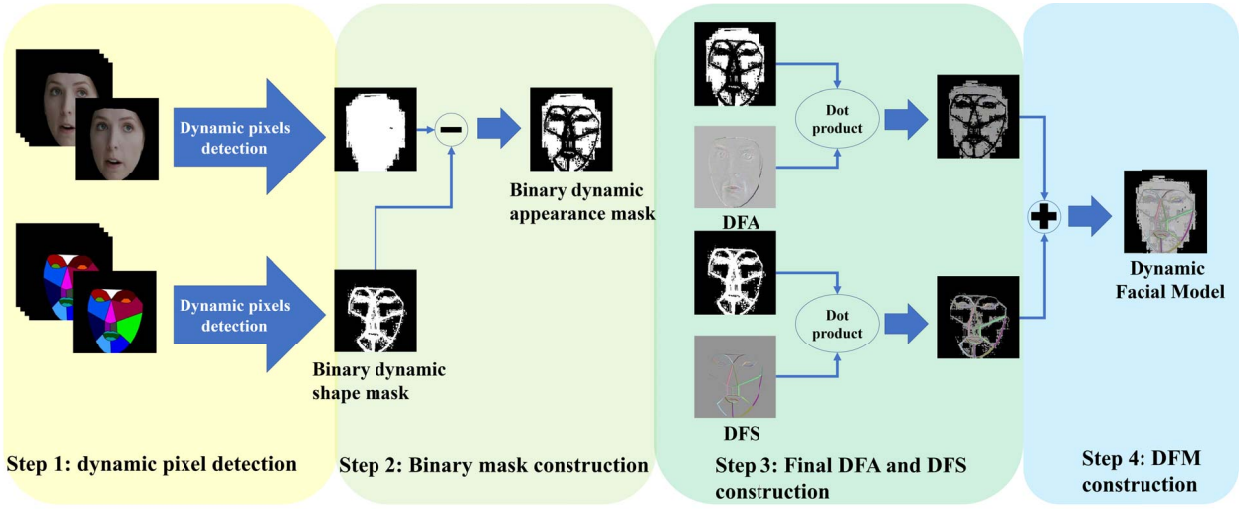


Figure 3: The process of DFM generation. All steps are corresponding to the process at the end of Sec. 3.3.

1. For T continuous SFS and SFA, we firstly find their 'stable pixels' whose R, G, B values keep stable over T frames. Specifically, for each of them, we calculate the absolute value of the difference between the given frame and other frames, respectively, resulting in $T-1$ maps. Then, a map representing the sum of these maps is obtained, of which the pixels (R, G, B) values equaling to 0 are defined as the 'stable pixels' while the reminder are denoted as 'dynamic pixels'.

2. Constructing binary dynamic shape mask and binary dynamic appearance mask, where the "dynamic pixels" are set as 1 and 'stable pixels' are set as 0. Since the location of some dynamic pixels in two masks may overlapped, we further set overlapped pixels in dynamic appearance mask to 0 to avoid distortion.

3. Generating a new DFS and a new DFA by conducting dot product between the binary dynamic shape mask/binary dynamic appearance mask and previously obtained DFS/DFA, respectively. Consequently, all redundant information and background noise would be removed from the generated new DFS and DFA as their pixels' value would equal to '0', while the new DFS containing all temporal shape information and new DFA containing most appearance dynamics.

4. Yielding the final fused dynamic facial image by simplely adding the new DFS to new DFA, which contains all temporal shape information and most appearance dynamics without any distortion. In this paper, we call this fused dynamic facial image as Dynamic Facial Model (DFM).

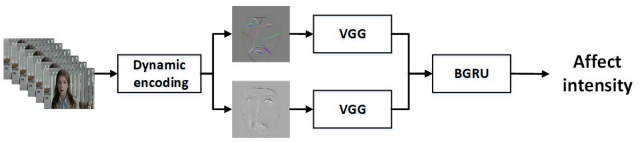
4. Deep Learning Dynamic Facial Features

As shown above, the DFM and DFIs are 3-channel raster images whose dimensions are same to the input SFIs. Therefore, it allows the information of a video to be learnt by existing CNN models for still images with fine-tuning. The features extracted from the CNN representation are subsequently forwarded to a Recurrent Neural Network (RNNs), which deals with dynamics at the feature level.

In this paper, we chosen VGG-16 network [37] pre-trained by VGG face datasets. We applied two simple structures to illustrate the benefit of each proposed DFI, which are shown in Fig. 4. The proposed approach, described above, is depicted in the top of the Figure. In particular, we investigate this approach against the use of two branches for the CNN-RNN structure, by which the shape/static face image and appearance/dynamic face image are not fused at the lowermost level, but are rather forwarded to two CNN networks, the output of which is fused by the RNN network.

The output of the CNN is taken from the first fully-connected layer of the corresponding VGG, which is a 4096-D vector. These features encode the short-term appearance and shape dynamics, constrained to the length of the time-window. In order to learn the long-term dynamics, we use a RNN on top of the CNN features. For this purpose, we adopt the Bidirectional Gated Recurrent Units (BGRU) [4] as our RNN model. BGRU is a simple version of Bidirectional Long-Short-Term-Memory networks (BLSTMs) due to its less complex structure. It has two multiplicative gates, i.e. reset gate and forget gate, to capture both long and short term dependencies in sequences, where the short-term dynamics will frequently have reset gates being active while the long-term dependencies will mostly update those forget gates. As a result, the use of BGRU allows

(a) The framework for inputing a single modality



(b) The framework for inputing two modalities

Figure 4: CNN-RNN framework: Top corresponds to the pipeline described throughout the paper, whereas bottom corresponds to the approach where shape/static and appearance/dynamic are fused after the CNN processing.

our framework to learn both long- and short-term temporal dynamics at the feature level. Thus, it compensates the drawback of DFIs that they only encode short-term dynamics in this paper.

5. Experiments

5.1. Database

To validate the proposed approach, we have carried out arousal and valence intensities estimation experiments on SEMAINE [21] and RECOLA [31] datasets. The SEMAINE dataset recorded uncontrolled facial expressions of participants who have a conversation with an operator, and it is annotated with valence and arousal dimensions in a continuous space within -1 and 1. In this paper, we have used the subset used in AVEC 2012[36], which contains 31 videos for training, 32 videos for development and 32 videos for test. The RECOLA dataset was recorded from 27 French-speaking participants to study socio-affective behaviours from video, audio, electro-cardiogram (ECG) and electro-dermal activity (EDA) in the context of computer supported collaborative work. Each video is around 300 seconds and labels are given with a rate of 25 Hz.

5.2. Evaluation measures

Three standard measures were used to assess the performance of the affect estimation; firstly the Mean Squared Error (MSE); secondly Pearson Correlation Coefficient (PCC); and thirdly the Concordance Correlation Coefficient (CCC, Eq. 5):

$$\rho_{ccc} = \frac{2\rho_{x,y}\sigma_x\sigma_x}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2}, \quad (5)$$

where $\rho_{x,y}$ is the PCC, μ_x and μ_y are mean values of predictions and labels while σ_x and σ_y are standard deviations.

5.3. Implementation details

DFIs generation: To generate frame-wise dynamic facial images for SEMAINE and RECOLA datasets, the lengths of time-windows are 20, 15 and 6, respectively, with the stride of 2. **Model training:** In this paper, VGG-16 networks pre-trained by VGG face database and BGRU with one hidden layer of 200 neurons were utilized. MSE was chosen as the loss function and standard SGD algorithm was applied as training method with learning rate of 5×10^{-3}, learning rate decay of 1×10^{-4}, and momentum of 0.85. For SEMAINE, the development partition was used to adjust model's hyper-parameters while test partition was used for reporting the final results. For RECOLA datasets, five-fold cross validation was conducted on training partition and reported results were yielded from the development partition.

5.4. Ablation studies

This section firstly conducts the ablation studies in terms of two experimental variables: 1. Temporal status of the input: static face images, (SFA, SFS, SFA+SFS) and dynamic face images (DFS, DFA, DFA+DFS); 2. Type of the input: appearance (SFA, DFA, SFA+DFA) and shape (SFS, DFS, SFS+DFS). All the experiments that have two inputs, e.g. SFA+SFS, DFA+SFA, DFA+DFS and SFS+DFS, were processed by the two branch architecture (Fig. 4(b)).

5.4.1 Facial appearance VS Facial shape

We firstly compared the average performance of facial appearance images to facial shape images in Fig. 5. For both dataset, the predictions yielded by shape inputs are more correlated to arousal and valence intensities labels than the appearance inputs, where the mean CCC values of shape inputs for arousal and valence are 0.354 and 0.304 for SEMAINE as well as 0.419 and 0.435 for RECOLA, which are outperformed the corresponding arousal and valence results obtained appearance inputs (0.302 and 0.283 for SEMAINE, 0.366 and 0.396 for RECOLA). Similarly, the predictions from facial shape features also achieved better MSE results than facial appearance features. When combining facial appearance and shape, it is obviously that each of them can benefit from the other, as the result achieved by 'Shape + Appearance' outperformed using shape or appearance independently for two tasks on both datasets.

5.4.2 Dynamic face VS Static face

As illustrated in Fig. 6, dynamic face images achieved higher average CCC and less average MSE results than static face images. In particular, the mean CCC value obtained by dynamic inputs on two datasets are 0.362 (arousal of SEMAINE), 0.302 (valence of SEMAINE) and 0.426 (arousal of RECOLA), 0.443 (valence of RECOLA),

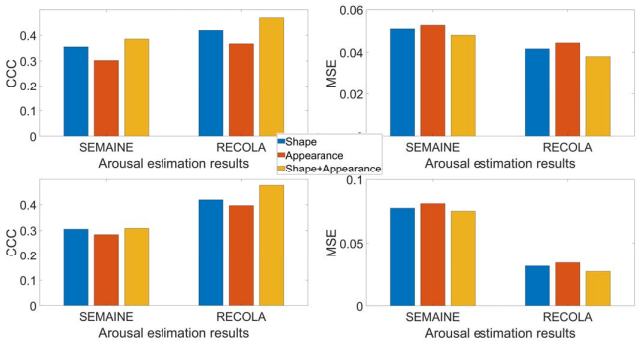


Figure 5: Comparison of the average results achieved by facial shape and facial appearance.

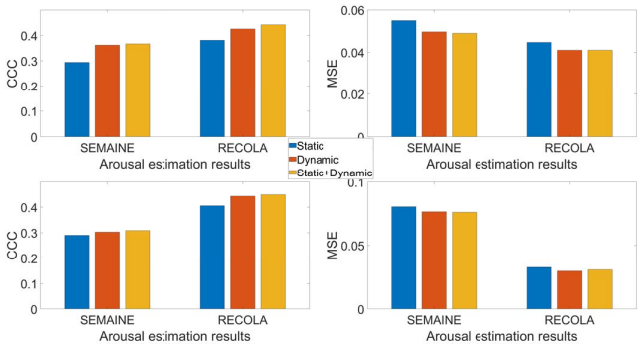


Figure 6: Comparison of the average results achieved by static face and dynamic face.

respectively, beating the corresponding results achieved by static face images, which are 0.294 (arousal of SEMAINE), 0.289 (valence of SEMAINE) and 0.369 (arousal of RECOLA), 0.405 (valence of RECOLA) respectively. These results indicate that the temporal dynamics encoded in the proposed DFIs can provide powerful clues for affect intensity estimation. We also reported the average results yielded by 'Static + Dynamic' which achieved similar result to dynamic face images, with slightly improvement.

To further investigate the property of DFIs, Fig. 7 compared some predictions of SFA, SFS, DFA and DFS on SEMAINE dataset. Obviously, dynamic predictions changed much heavier than static predictions as well as predictions from facial shape changed heavier than facial appearance because the difference between adjacent DFIs is much larger than SFIs. Another observation is that when groundtruth suddenly dropped or increased, e.g. 400th frame, 600th frame, namely high frequency dynamics, the amplitude of dropping or increasing of dynamic predictions were heavier than static predictions. This means that DFIs are more sensitive to affect changes.

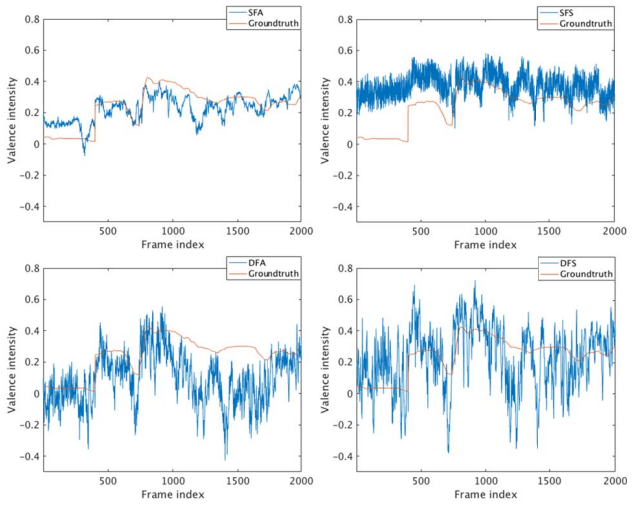


Figure 7: Predictions of SFIs and DFIs of the 6000th - 8000th frame of the 1st test video from SEMAINE.

5.4.3 Dynamic Facial Model VS Dynamic Facial Images

We also compared the proposed DFM with the best single input system, i.e. DFS, in Table. 1 and Table. 2. While DFS already yielded good performance, DFM achieved significant improvement on both datasets. As both system used the same structure, the only difference is that DFM combined shape and appearance dynamics while DFS only contains shape dynamics. Thus, it can be concluded that our fusion strategy can effectively encode facial shape dynamics and appearance dynamics for affect estimation. Another comparison is made between the result produced by DFA + DFS + BGRU and DFM + BGRU, as both of them input dynamic facial shape and appearance information for estimation. As reported, the results obtained by DFM + BGRU outperformed the results obtained by DFA + DFS + BGRU, except the arousal predictions in SEMAINE. Although DFA and DFS contain the original dynamic information rather than the reduced dynamic information in DFM, the trainable weights in CNN-RNN architecture used for DFA + DFS is at least as twice as it for DFM, resulting in higher computational cost. On the other hand, DFM has some advantages 1. it removes the redundant information ('stable pixels') from both DFS and DFA; 2. it fused shape and appearance information in the context of the face rather than at the feature level; 3. it can be learned by a simple network, where less weights need to be optimized.

5.5. Our methods VS state-of-the-art

This section compares our methods with state-of-the-art visual methods on SEMAINE and RECOLA datasets. As shown in Table. 1, our best system (DFM+BGRU) beats all state-of-the-arts for both arousal and valence estima-

tion tasks on SEMAINE dataset, especially for arousal estimation, which has 27.7% relative improvement compared to the second best system[16]. As shown in Table. 2, the baselines already generated very promising predictions in RECOLA dataset. However, features extracted from the proposed DFM still achieved excellent performance for both tasks. In terms of the seven recent works on RECOLA dataset that we have compared, our DFM+BGRU system yielded both better arousal and valence results than four of them. For the reminder, the DFM+BDFM+BGRU system either obtained better arousal predictions or valence predictions. In addition, the model of [27] were pre-trained by AFLW dataset, which may also an important factor for its excellent performance.

Table 1: State-of-the-art results on the SEMAINE dataset.

Method	Arousal		Valence	
	PCC	MSE	PCC	MSE
Baseline [36]	0.077	N.A.	0.134	N.A.
Kaltwang et al. [16]	0.310	0.042	0.310	**0.058**
Glodek et al. [9]	0.069	24.71	0.180	23.31
Savran et al. [35]	0.251	N.A.	0.210	N.A.
Cruz et al. [5]	0.227	N.A.	0.141	N.A.
Zhang et al. [47]	0.070	N.A.	0.241	N.A.
DFS+BGRU	0.376	0.049	0.310	0.075
DFA+DFS+BGRU	**0.385**	0.048	0.308	0.075
DFM+BGRU	0.381	**0.046**	**0.322**	0.073

Table 2: State-of-the-art results on the RECOLA dataset.

Method	Arousal		Valence	
	CCC	MSE	CCC	MSE
Baseline appearance [45]	0.483	N.A.	0.474	N.A.
Baseline shape [45]	0.379	N.A.	**0.612**	N.A.
Brady et al.[3]	0.346	0.040	0.511	**0.010**
Povolny et al.[27]	**0.617**	N.A	0.467	N.A
Tzirakis et al.[43]	0.363	N.A.	0.488	N.A.
Han et al.[10]	0.292	N.A.	0.592	N.A.
Sun et al.[41]	0.215	N.A.	0.366	N.A.
DFS+BGRU	0.432	0.040	0.441	0.030
DFA+DFS+BGRU	0.468	0.038	0.476	0.027
DFM + BGRU	0.498	**0.036**	**0.506**	**0.022**

5.6. Cross dataset evaluation of the DFA parameters

To assess the generalizability of the DFA parameters tuned for SEMAINE and RECOLA datasets, we trained dimensional affect recognition models on Aff-wild dataset [19]. Here our aim is to demonstrate that the parameters learned for generating DFAs of SEMAINE and RECOLA datasets could extract meaningful representations from a new dataset. For this reason, we do not include other state-of-the-art methods on the Aff-wild dataset and the models that were trained using either DFA or DFS alone. Firstly, the DFAs of the Aff-wild data were generated using the parameters that were tuned for SEMAINE and RECOLA datasets. Then we trained two different models with randomly initialized weights, one with SFAs as inputs and the other with stacked SFAs and DFAs as inputs. As shown in Table 3, on the Aff-wild test set, the CNN model trained with the stacked SFA plus DFA inputs outperformed the model trained with only the SFA. This performance improvement clearly demonstrates that the DFA parameters can generalize well and extract meaningful face representations that complement the static face inputs.

Table 3: Comparison of the CNN models trained on Aff-wild dataset using static face inputs and combined static and dynamic face appearance inputs **(with the dynamic appearance parameters tuned for SEMAINE and RECOLA datasets)**

Inputs	Arousal		Valence	
	CCC	MSE	CCC	MSE
SFA	0.153	0.097	0.283	0.167
SFA+DFA	**0.203**	0.090	**0.392**	0.134

6. Conclusion

This paper proposed a dynamic facial encoding method that allows a single raster image to encode facial appearance and shape dynamics of an image sequence. The features learned from these dynamic inputs using CNN-RNN models outperformed the static inputs on dimensional affect estimation task. The experimental results suggest the following conclusions: 1. facial shape features generate better affect predictions than facial appearance features; 2. combined static and dynamic face inputs perform better than the static face inputs alone on dimensional affect estimation; 3. the proposed DFM can effectively encode facial shape and appearance dynamics, as it achieved better results than either using a DFI or SFI as the input, or jointly using DFI and SFI as the input in most cases. Meanwhile, we believe DFM still haven't fully shown its ability on RECOLA database as the initial weights of VGG face network was pre-trained by RGB images rather than dynamic images, and the VGG structure is not specifically designed for dynamic face images, and thus the trained models may lack of ability to capture encoded dynamic information.

References

[1] Hakan Bilen, Basura Fernando, Efstratios Gavves, and Andrea Vedaldi. Action recognition with dynamic image networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[2] Aaron F. Bobick and James W. Davis. The recognition of human movement using temporal templates. *IEEE Transactions on pattern analysis and machine intelligence*, 23(3):257–267, 2001.

[3] Kevin Brady, Youngjune Gwon, Pooya Khorrami, Elizabeth Godoy, William Campbell, Charlie Dagli, and Thomas 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*, pages 97–104. ACM, 2016.

[4] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[5] Albert C Cruz, Bir Bhanu, and Ninad Thakoor. Facial emotion recognition with expression energy. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, pages 457–464, 2012.

[6] Paul Ekman. Emotional and conversational nonverbal signals. In *Language, knowledge, and representation*, pages 39–50. Springer, 2004.

[7] Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, et al. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE Transactions on Affective Computing*, 7(2):190–202, 2016.

[8] Nico H Frijda. The emotions: Studies in emotion and social interaction. *Paris: Maison de Sciences de l'Homme*, 1986.

[9] Michael Glodek, Martin Schels, Günther Palm, and Friedhelm Schwenker. Multiple classifier combination using reject options and markov fusion networks. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, pages 465–472, 2012.

[10] Jing Han, Zixing Zhang, Nicholas Cummins, Fabien Ringeval, and Björn Schuller. Strength modelling for real-worldautomatic continuous affect recognition from audiovisual signals. *Image and Vision Computing*, 2016.

[11] Jing Han, Zixing Zhang, Fabien Ringeval, and Björn Schuller. Prediction-based learning for continuous emotion recognition in speech. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5005–5009. IEEE, 2017.

[12] Behzad Hasani and Mohammad H Mahoor. Facial affect estimation in the wild using deep residual and convolutional networks. *arXiv preprint arXiv:1705.07884*, 2017.

[13] Zhaocheng Huang, Brian Stasak, Ting Dang, Kalani Wataraka Gamage, Phu Le, Vidhyasaharan Sethu, and Julien Epps. Staircase regression in oa rvm, data selection and gender dependency in avec 2016. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, pages 19–26. ACM, 2016.

[14] Shashank Jaiswal, Siyang Song, and Michel Valstar. Automatic prediction of depression, anxiety and personality traits from facial behaviour attributes. In *2019 eighth International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2019.

[15] Shashank Jaiswal and Michel Valstar. Deep learning the dynamic appearance and shape of facial action units. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*, pages 1–8, 2016.

[16] Sebastian Kaltwang, Sinisa Todorovic, and Maja Pantic. Doubly sparse relevance vector machine for continuous facial behavior estimation. *IEEE transactions on pattern analysis and machine intelligence*, 38(9):1748–1761, 2016.

[17] Heysem Kaya, Furkan Gürpınar, and Albert Ali Salah. Video-based emotion recognition in the wild using deep transfer learning and score fusion. *Image and Vision Computing*, 65:66–75, 2017.

[18] Pooya Khorrami, Tom Le Paine, Kevin Brady, Charlie Dagli, and Thomas S Huang. How deep neural networks can improve emotion recognition on video data. In *Image Processing (ICIP), 2016 IEEE International Conference on*, pages 619–623, 2016.

[19] Dimitrios Kollias, Panagiotis Tzirakis, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, Björn Schuller, Irene Kotsia, and Stefanos Zafeiriou. Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond. *International Journal of Computer Vision*, 127(6-7):907–929, 2019.

[20] Jean Kossaifi, Georgios Tzimiropoulos, Sinisa Todorovic, and Maja Pantic. Afew-va database for valence and arousal estimation in-the-wild. *Image and Vision Computing*, 2017.

[21] Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schroder. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing*, 3(1):5–17, 2012.

[22] Arianna Mencattini, Eugenio Martinelli, Fabien Ringeval, Bjorn Schuller, and Corrado Di Natlae. Continuous estimation of emotions in speech by dynamic cooperative speaker models. *IEEE Transactions on Affective Computing*, 2017.

[23] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *arXiv preprint arXiv:1708.03985*, 2017.

[24] Mihalis A Nicolaou, Hatice Gunes, and Maja Pantic. Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *IEEE Transactions on Affective Computing*, 2(2):92–105, 2011.

[25] Jérémie Nicolle, Vincent Rapp, Kévin Bailly, Lionel Prevost, and Mohamed Chetouani. Robust continuous prediction of human emotions using multiscale dynamic cues. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, pages 501–508, 2012.

[26] Alexandru Popescu, Joost Broekens, and Maarten Van Someren. Gamygdala: An emotion engine for games.

IEEE Transactions on Affective Computing, 5(1):32–44, 2014.

[27] Filip Povolny, Pavel Matejka, Michal Hradis, Anna Popková, Lubomír Otrusina, Pavel Smrz, Ian Wood, Cecile Robin, and Lori Lamel. Multimodal emotion recognition for avec 2016 challenge. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, pages 75–82. ACM, 2016.

[28] Fabien Ringeval, Björn Schuller, Michel Valstar, Roddy Cowie, Heysem Kaya, Maximilian Schmitt, Shahin Amiriparian, Nicholas Cummins, Denis Lalanne, Adrien Michaud, et al. Avec 2018 workshop and challenge: Bipolar disorder and cross-cultural affect recognition. In *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*, pages 3–13. ACM, 2018.

[29] Fabien Ringeval, Björn Schuller, Michel Valstar, NIcholas Cummins, Roddy Cowie, Leili Tavabi, Maximilian Schmitt, Sina Alisamir, Shahin Amiriparian, Eva-Maria Messner, et al. Avec 2019 workshop and challenge: State-of-mind, detecting depression with ai, and cross-cultural affect recognition. *arXiv preprint arXiv:1907.11510*, 2019.

[30] Fabien Ringeval, Björn Schuller, Michel Valstar, Jonathan Gratch, Roddy Cowie, Stefan Scherer, Sharon Mozgai, Nicholas Cummins, Maximilian Schmitt, and Maja Pantic. Avec 2017: Real-life depression, and affect recognition workshop and challenge. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, pages 3–9, 2017.

[31] Fabien Ringeval, Andreas Sonderegger, Juergen Sauer, and Denis Lalanne. Introducing the recola multimodal corpus of remote collaborative and affective interactions. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pages 1–8, 2013.

[32] James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.

[33] Enrique Sánchez-Lozano, Paula Lopez-Otero, Laura Docio-Fernandez, Enrique Argones-Rúa, and José Luis 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*, AVEC '13, pages 31–40, 2013.

[34] Enrique Sánchez-Lozano, Georgios Tzimiropoulos, Brais Martinez, Fernando De la Torre, and Michel Valstar. A functional regression approach to facial landmark tracking. *IEEE transactions on pattern analysis and machine intelligence*, 2017.

[35] Arman Savran, Houwei Cao, Miraj Shah, Ani Nenkova, and Ragini 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*, pages 485–492, 2012.

[36] Björn Schuller, Michel Valster, Florian Eyben, Roddy Cowie, and Maja Pantic. Avec 2012: the continuous audio/visual emotion challenge. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, pages 449–456, 2012.

[37] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[38] Alex J Smola and Bernhard Schölkopf. A tutorial on support vector regression. *Statistics and computing*, 14(3):199–222, 2004.

[39] Siyang Song, Enrique Sánchez-Lozano, Linlin Shen, Alan Johnston, and Michel Valstar. Inferring dynamic representations of facial actions from a still image. *arXiv preprint arXiv:1904.02382*, 2019.

[40] Siyang Song, Linlin Shen, and Michel Valstar. Human behaviour-based automatic depression analysis using handcrafted statistics and deep learned spectral features. In *Automatic Face & Gesture Recognition (FG 2018), 2018 13th IEEE International Conference on*, pages 158–165. IEEE, 2018.

[41] Bo Sun, Siming Cao, Liandong Li, Jun He, and Lejun Yu. Exploring multimodal visual features for continuous affect recognition. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, pages 83–88. ACM, 2016.

[42] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[43] Panagiotis Tzirakis, George Trigeorgis, Mihalis A Nicolaou, Björn W Schuller, and Stefanos Zafeiriou. End-to-end multimodal emotion recognition using deep neural networks. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1301–1309, 2017.

[44] Michel Valstar. Automatic behaviour understanding in medicine. In *ICMI Workshops*, 2014.

[45] Michel Valstar, Jonathan Gratch, Björn Schuller, Fabien Ringeval, Dennis Lalanne, Mercedes Torres Torres, Stefan Scherer, Giota Stratou, Roddy Cowie, and Maja Pantic. Avec 2016: Depression, mood, and emotion recognition workshop and challenge. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, pages 3–10, 2016.

[46] Michel Valstar, Björn Schuller, Kirsty Smith, Timur Almaev, Florian Eyben, Jarek Krajewski, Roddy Cowie, and Maja Pantic. Avec 2014: 3d dimensional affect and depression recognition challenge. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, pages 3–10, 2014.

[47] Ligang Zhang, Dian Tjondronegoro, and Vinod Chandran. Representation of facial expression categories in continuous arousal–valence space: feature and correlation. *Image and Vision Computing*, 32(12):1067–1079, 2014.