End-to-end Facial and Physiological Model for Affective Computing and Applications

*Joaquim Comas, *Decky Aspandi and Xavier Binefa
Department of Information and Communication Technologies
Pompeu Fabra University
Barcelona, Spain
{*joaquim.comas, decky.aspandilatif, xavier.binefa*}*@upf.edu*

***Abstract*—In recent years, affective computing and its applications have become a fast-growing research topic. Furthermore, the rise of deep learning has introduced significant improvements in the emotion recognition system compared to classical methods. In this work, we propose a multi-modal emotion recognition model based on deep learning techniques using the combination of peripheral physiological signals and facial expressions. Moreover, we present an improvement to proposed models by introducing latent features extracted from our internal Bio Auto-Encoder (BAE). Both models are trained and evaluated on AMIGOS datasets reporting valence, arousal, and emotion state classification. Finally, to demonstrate a possible medical application in affective computing using deep learning techniques, we applied the proposed method to the assessment of anxiety therapy. To this purpose, a reduced multi-modal database has been collected by recording facial expressions and peripheral signals such as electrocardiogram (ECG) and galvanic skin response (GSR) of each patient. Valence and arousal estimates were extracted using our proposed model across the duration of the therapy, with successful evaluation to the different emotional changes in the temporal domain.**

基于使用外围生理信号与面部表情的深度学习技术的多模态情绪识别模型

使用BAE完成潜在特征提取

数据集：AMIGOS

用于焦虑治疗

外围生理信号：EGC心电图、GSR皮肤电反应

***Keywords*-Affective Computing; Multi-modal; Physiological signals; Auto-encoder; Deep Learning.**

I. Introduction

Emotions are essential factors in human behaviour which influence every social action [1–4]. The affective computing, which in its core originated from the efforts to understand these factors [5] has attracted considerable attentions lately due to its application on numerous venues, such as education [6], healthcare [7], etc.

In literature, a common approach to infer emotion states is by utilizing several modalities such as expressions, speech, body gestures, physiological signals, etc. Though the facial expressions are gaining more popularity due to their intuitive nature [8–10], the physiological signals offer unique advantages to other modalities. First is their growing availability supplemented by recent arise of wearable devices usages. Secondly, they are quite invariant against external visual noises, such as illumination therefore are quite robust and versatile. Third advantage includes their fidelity qualities, since it is very hard to replicate or to mask these signals to simulate specific emotions. Lastly, they have relatively low dimensional structure allowing more efficient processing.

Nowadays, the emergence of large dataset such as AMIGOS [11] has opened a new possibility to use powerful deep neural networks to this field of affective computing [10], [12]. However, compared to other computing fields, the adoption of deep learning techniques to process these physiological signals is still sparse [13], with the only the recent works of [14], [15] are being the exceptions. Most studies use the raw signal as their input to their models with the assumption that they able to learn relevant features for their estimations directly [13, 16]. However, we argue that this approach may lead to sub-optimal results since raw signals prone to contain a substantial amount of noises [17, 18]. This problem can be circumvented by the efficient use of auto-encoder [19], which is able to construct these relevant features in a compact way. Thus, we may have a less noisy features. This has been shown to improve the model estimates on other computing fields, such as facial recognition [20], object classification [21], music generation [22], data compression [23], dimensionality reduction [24] etc. Nonetheless, in the affective computing, it remains largely unexplored, with the only examples are the works of Tang et al. [16] and Yildirim et al. [25]. Unlike these approaches, however, our models fully differentiate the stream input from bio-signal and facial features, due to the large size of each modality which we argue will require specific processing. We realize this by adopting single modality auto-encoder to achieve compact representation of bio-signal [25] that simultaneously allows us to quantify its individual contribution on improving the quality of models estimates, which currently is still largely unexplored.

In this work, we propose end-to-end models for automatic affect estimations using multiple modalities. In conjunction, we also introduce new dataset collected from patients who have been exposed to anxiety treatments with aim to expand the application of affective computation. Different to previous approaches, we capitalize on the individual use of latent features extracted from our internal Bio Auto-Encoder alongside stream of spatial features to improve the accuracy of our models estimates. Using recently published

* Both authors contributed equally to this work

AMIGOS [11] dataset, we will perform an analysis to reveal the effectiveness of each modality. Then, the combination of extracted latent features will be used to improve our model estimates. Next, we will present a relative comparison of our results against related studies. Lastly, we will demonstrate a real-world application of our trained models by evaluating our model estimates using recorded patient data, collected before and after treatment. Specifically, the contributions of our work are as follow :

1) We propose improvements by incorporating of latent features extracted by means of our internal Bio Auto-Encoder to our bio multi-modal network.
2) We present a new dataset collection from a medical therapy to expand the application of affective computing model.
3) We perform a thorough analysis to confirm the benefit of the utilisation of multi-modal inputs, which will be complemented by the correspondent latent features.
4) We present our competitive results against other related studies on AMIGOS [11] dataset and its real-world capability of our models to estimate the patient emotion state during a therapy.

II. Related work

One of the earliest use of physiological data for automatic emotion estimation has been in the work of Fridlund et al. [26], where they applied linear discriminant analysis on the facial electromyographic (EMG) activity. They reported that there is a correlation between personal biological signal and their emotional activity. A multi-modal approach was introduced by Picard et al. [27], combining four separate modalities: heart rate, skin conductance, respiration and facial electromyogram. Using this more extensive data, they managed to achieve relatively better results.

Other research attempt to concentrate on finding the correlation between emotion and physiological signal, such as the work of Nasoz et al. [28] and Feng et al [29]. Nasoz et al. proposed an experiment by eliciting certain types of emotions contained in the movie clip presented to the participant. They utilised k-nearest neighbour, discriminant function analysis and Marquardt back-propagation algorithm, using features from several modalities: galvanic skin response, temperature and heart rate. While Feng et al. adopted support vector machines (SVM) classifier to perform an automated method for emotion classification using EDA signals with wavelet-based features. In other hand, Soleymani et al. [30] introduced new modality of electroencephalography (EEG), pupillary response and gaze distance with decision level fusion (DLF) technique to produce more accurate results on each modality than previous methods.

Due to the lack of huge data to facilitate extensive comparisons, Koelstra et al. created DEAP dataset [31] which offers large number of response features such as electrocardiogram (ECG), galvanic skin response (GSR) or electrodermal activity (EDA), electrooculargram (EOG) and electromyogram (EMG) to expand the possibility to analyse such responses. Using this dataset, Tang et al. [16] further introduced bimodal deep denoising autoencoders to extract high level representations of both bio-signal and visual information with bimodal LSTM as bottleneck to analyze the impact of additonal temporal information. This possibility to construct the latent features from bio-signal has been studied further on the work of Yildirim et al. [25] which introduces efficient compression on similar ECG data. Albeit still to date, none of these approaches fully analyze and quantify the specific importance of the extracted bio-latent features on the final model estimations.

Currently, the biggest dataset so far that allows the analysis of the effect of physiological signals in emotion recognition is the AMIGOS dataset [11]. This dataset enables more extensive investigations of specific peripheral signals of ECG and GSR, and has been used on multiple literatures. First example is the work of Gjoreski et al. [32] who presented an inter-domain study for arousal recognition using RR-intervals and GSR. Other works tried to adopt variational auto-encoders network(VAE) to learn personality-invariant physiological signal representation by Yang et al. [33] with improved accuracy. Another study is the work of Santamaria et al. [34], where they introduced deep convolutional neural networks largely outperform previous results. Finally, Siddarth et al. [13] applied a novel deep learning method to different physiological and video data, and currently holdings state of the art accuracy on this dataset. Although these models have been shown to work well in their specific domains, there still has not been any investigation to asses their capability on real-life settings, such as in medical application.

III. Methodology

In this section, we will explain datasets used in our experiments, including our data pre-processing steps. Further, we describe our proposed multi-modal network, which operates by incorporating existing bio-physiological responses and facial features for affect estimations.

A. Datasets and pre-processing

1) AMIGOS Dataset: AMIGOS database [11], which stands for mood, personality and affect research on individuals and groups, was collected using two different experimental settings. The first setting involved 40 participants watching 16 short videos (trial length varying between 51 and 150 seconds). In the second part, some participants watched four long videos on different scenarios, i.e. individually and in groups. During these experiments, EEG, ECG and GSR signals were recorded using wearable sensors (all signals pre-processed at 128 Hz), while face and depth data were recorded by separate types of equipment.

In total, there are 640 instances (16 trials x 40 participants) along with their respective tag for valence, arousal, liking,

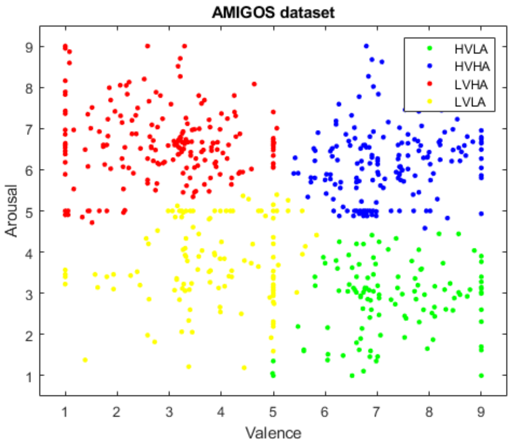

Figure 1: Distribution of valence and arousal label from self-assessment labels of AMIGOS dataset (scale from 1 to 9) classified in four quadrants (HVLA: high-valence / low-arousal, HVHA: high-valence / high-arousal, LVHA: low-valence / high-arousal), LVLA: low-valence / high-arousal.

familiarity and seven basic emotions (neutral, disgust, happiness, surprise, anger, fear, and sadness) available. The first four affect dimensions are measured in a continuous scale of 1 to 9 with basic emotions represented using one-hot label. Figure 1 visualises the distribution of the self-assessment label in terms of valence and arousal of the dataset. Notice that most of the emotion examples are located around the centre of each quadrant, which are relatively close to the neutral emotion state.

2) Medical Therapy Dataset: In this work, we introduce a new multi-modal data collection from several patients who have undergone an anxiety treatment. Using this data, we would like to apply and provide more objective analysis based on the patient bio-responses while being subjected to a complementary polarisation treatment [35]. In this case, our hypothesis is that there will be changes in terms of valence and arousal over the therapy.

During the therapy session, we recorded several physiological signals using wearable sensors along with their face in a synchronous way. The recordings were performed with a sampling rate of 800 Hz for bio-signals collection, while 60 frames per second for the facial images. An example of data collection of a patient during the therapy session can be seen in Figure 2. We collected self-assessment questionnaires from each patient prior and after treatment that provide their subjective perceptions of therapy. Currently, we have successfully recorded a sample from five patients during a single therapy session, where we expect to gather more in our future works. This sample consists of three females and two males with the mean age of 42.8 years and 10.40 years of standard deviation.

To perform the treatment validation, we adopt the four-quadrant circumplex model [36] by combining the value of both valence and arousal of emotional states. In the ideal scenario, there should be a movement tendency of the patients emotional state across quadrants. Initially, the emotional state will be located around the second quadrant (low valence and positive arousal) indicating high-level stress of the patients. After the treatment, we expect its location to be around fourth quadrant, i.e. high valence and low arousal showing the patients tranquillity.

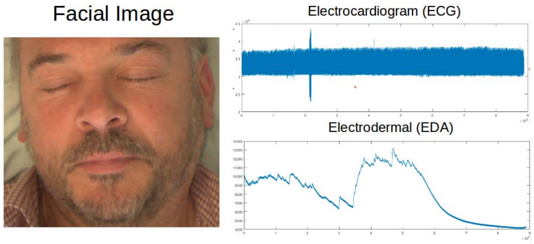

Figure 2: Data collection of a patient during therapy.

3) Data Preprocessing: To ease our model training and estimations, we performed multiple stages of data pre-processing to each modality used by our model : facial features, electrocardiogram (ECG) and electrodermal activities (EDA). The first stage was to locate the facial area to enable us to remove its respective backgrounds. Since they may contain other redundant parts of the scene which consequently slow down the training process. To do this, we used facial tracking model of [37] and cropped facial area given detected facial landmark.

The second stage involves scaling the physiological signals, which will be crucial to obtain a better reconstructed solution in the auto-encoder network. Considering that data is already down-sampled, pre-processed and segmented, no more signal pre-processing was needed. The final stage was to maintain the temporal relation of the physiological signals with facial images through data synchronisation. Specifically, for each frame, we segmented each bio-signals with a length of 1000 samples, that corresponds to eight heartbeats. We also performed padding to the initial and end of each bio-signal to accommodate this scheme.

B. Bio Multi-Modal with Internal Auto-Encoder Network

Our proposed model of Bio Multi-Modal Network (**BMMN**) operates by involving several modalities to estimate the people affect states while fully end-to-end trainable. BMMN model serves as the baseline of our results, which we will later improve by the incorporation of latent features captured by our Bio Auto-Encoder (**BAE**). Figure 3 depicts the whole structure of our proposed models.

1) Bio Multi-Modal Network: Our BMMN consists of three main parts: Bio Network, Spatial Network, and merging layer. Bio Network is dedicated to compute bio features given the raw physiological signals, which in this case are ECG and EDA. On the other side, the spatial network creates visual features given the cropped face. These features then merged on the Merge Layer and passed to the final bottleneck fully connected layer (FC). FC layer in the end produces the affect estimates: valence, arousal,

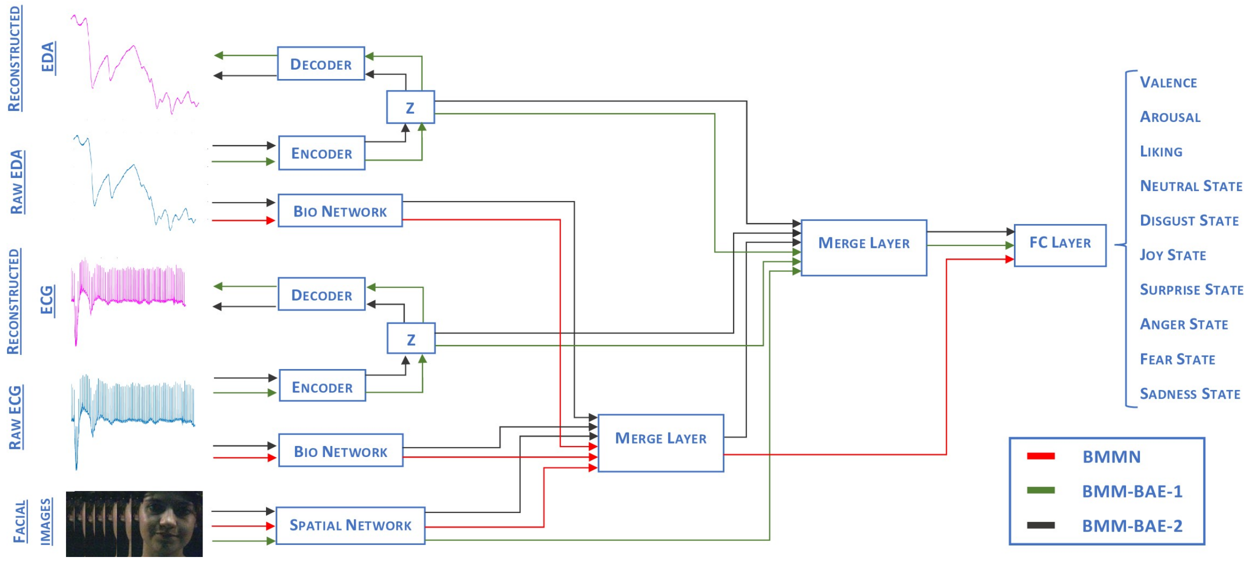

Figure 3: Overall structure of our proposed Bio Multi-Modal with Internal Auto-Encoder Network(BMMN). Red line shows the pipeline of our base network of BMMN, green line for BMMN-BAE-1 and black line for BMMN-BAE-2

liking and seven emotional states (neutral, disgust, joy, surprise, anger, fear, and sadness).

We construct our Bio Network by using stacked of 1D convolutional layers which later joined step-wise through residual connections [38]. Specifically, we propose to utilise different kernel layers size, which has been shown to learn more efficient representations using their sparse filter relationship [39]. Furthermore, we also introduce a skip connection to previous outputs to ease the gradient while training [38]. The full network architectures of Bio Network is summarised in Table I. As for our Spatial Network, we rely on pre-trained ResNet-50 [1] which is available on the standard PyTorch-torchvision [40] libraries by performing transfer learning.

Layer	Kernel	Activation	Filters	Stride	Output
Input	-	-	-	-	1000 x 1
Conv1	200 x 1	ReLU	4	1	805 x 4
Maxpool1	2 x 1	-	4	2	402 x 4
Conv2	100 x 1	ReLU	2	1	307 x 2
Maxpool2	2 x 1	-	2	2	153 x 2
Conv3	50 x 1	ReLU	2	1	357 x 2
Maxpool3	2 x 1	-	2	2	178 x 2
Conv4	25 x 1	ReLU	2	1	382 x 2
Maxpool4	2 x 1	-	2	2	191 x 2
Merge	-	ReLU	-	-	2652 x 1
Output	-	-	-	-	2652 x 1

Table I: The architecture of proposed 1D-CNN for Bio Network.

2) Bio Auto-Encoder Network: Bio Auto-Encoder Network (BAE) is responsible to extract a fixed length of latent features that represent the overall bio-physio signal input it received. We hypothesise that utilizing this vector representation can ease the model learning, and thus improve global estimations. We use similar stack of 1D convolutions used on Bio Network layer but arranged to follow standard encoder and decoder structure. We use max-pooling layers on the encoder part to reduce the dimension of the input features. With an intermediate fully connected layer (FC) to obtain a fixed length of latent features of z. Subsequently by using the z features, we get the reconstructed signal back with series of un-pooling layers with ReLU activation. The detail of layers used in BAE network can be seen on Table II.

Layer	Kernel	Activation	Filters	Stride	Output
Input	-	-	-	-	1000 x 1
Conv1	200 x 1	ReLU	16	1	801 x 16
Maxpool1	2 x 1	-	16	1	402 x 16
Conv2	100 x 1	ReLU	8	1	301 x 8
Maxpool2	2 x 1	-	8	1	150 x 8
Conv3	50 x 1	ReLU	4	1	101 x 4
Maxpool3	2 x 1	-	4	1	50 x 4
FC	-	-	-	-	128 x 1
Unpool1	2 x 1	-	4	1	101 x 4
Conv4	50 x 1	ReLU	8	1	150 x 8
Unpool2	2 x 1	-	8	1	301 x 8
Conv5	100 x 1	ReLU	16	1	400 x 16
Unpool3	2 x 1	-	16	1	801 x 16
Conv6	200 x 1	ReLU	1	1	1000 x 1
Output	-	-	-	-	1000 x 1

Table II: The architecture of the Bio Auto Encoder Network.

BAE network ultimately will generate compact z features of 128 length vectors, in which we consider its representation quality based on the reconstructed output. The degree of their similarity indicates the robustness of extracted features z, i.e. more similar means more robust the extracted features. This is due to inherent structures of auto-encoder which

[1] http://pytorch.org/docs/stable/torchvision/models.html

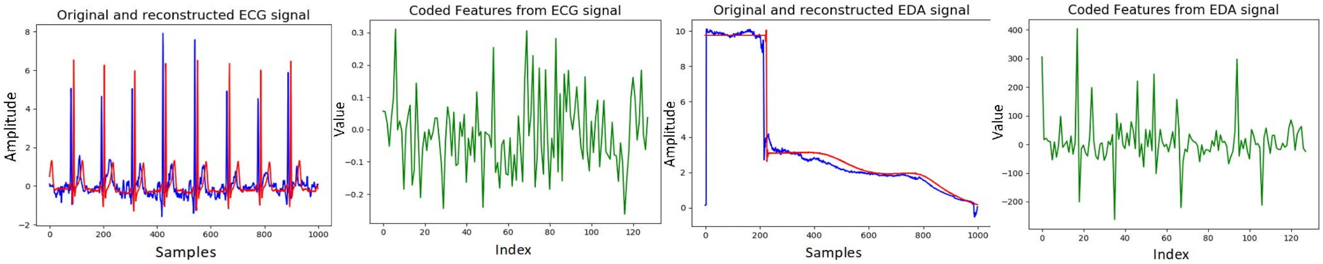

Figure 4: Reconstructed signals (red) overlaid against the original signals (blue), followed by respective latent features (green) for ECG and EDA sequentially.

forces the inner hidden layers to discover more relevant features to reconstruct original signal.

C. Feature Combinations

We propose two combination strategies to improve the estimation of our baseline models of BMMN by involving the latent features z generated by BAE:

1) The first strategy is to solely use z vector in place of features from Bio Network, and passing them to merge layer further for final estimation process. Since we assume original data could contain noisy or redundant information which can affect the emotion recognition process and a more compact feature of z alone is sufficient to provide relevant information. This combination produces our second network of **BMMN-BAE-1**.
2) The second approach involves combining latent z as auxiliary features along the essential features from the Bio Network, which later merged to follow further affect estimation pipelines. This is based on our assumption that latent features z can serve as complementary information, thus enriching the input features which may ease the model estimations as well as improving final accuracy. We name our third alternative network as **BMMN-BAE-2**.

Both of these network variants and their combination schemes can be seen on Figure 3.

D. Model Loss and Training Setup

We initially trained each network separately to ease the overall training processes, which we later perform joint training of all involving sub-networks to obtain our final estimations. We first trained BMMN for affect estimations to get our baseline results using standard ℓ^2 loss. In this phase, we also trained BMMNs using each modality separately, i.e. to use either facial feature or physiological signal only. This to provide analysis for each modality described in Ablation Analysis (Section IV-A).

Secondly, we pre-trained BAE to reconstruct the input signals to capture relevant features prior integration so that it has been conditioned enough to the characteristics of these signals. Figure 4 shows the reconstructed signals overlaid on top of the original bio-signal input, followed by its respective 128 coded latent features. As we can see, the reconstructed input signals are quite similar to original signals suggesting that our BAE able to construct corresponding latent features.

Thirdly, we combined the z features into BMMN network following the scheme explained on Section III-C and further jointly trained them. The total loss for both BMMN-BAE-1 and BMMN-BAE-2 is as follow:

$$L_{Total} = \lambda_{BMMN} L_{BMMN} + \lambda_{Bae} L_{Bae} \tag{1}$$

where L_{Bae} denotes the reconstruction loss between original and reconstructed signal and original signal, L_{BMMN} is the loss for emotion label estimatons, λ signifies the regularizer coefficients [41] and the L stands for standard ℓ^2 loss.

On all training instances, we used a uniform distribution for weights initialization and Adam optimizer with a learning rate of 0.0001 for optimization. Using single NVIDIA Titan X we were able to use a batch size of 115 and took approximately two days to train each model. All models were implemented using the Pytorch framework [40].

IV. Experiments

In this section, we will provide the results of our models using both pre-processed dataset of AMIGOS [11] and therapy dataset. With AMIGOS dataset, we first provide the analysis to see the impact of each modality used by our model to its final affect estimates, including the use of internally extracted latent features. Secondly, to see the quality of affect estimation from our proposed approach, we will compare our results with the reported results from related literature on this dataset. Then, we use our best performing model to validate the emotion state of each individual on our collected dataset. To evaluate the quality of affect estimates, we follow the original person-independence protocols of AMIGOS dataset. Finally, precision score (in per cent unit) is used to judge the quality of each estimated affect labels.

A. Ablation Studies

We provide two principal analysis in these ablation studies: modality and latent feature impact analysis. Modality analysis exemplifies the impact of individual modalities as input to our BMMN and to establish our baseline accuracy. While latent feature impact analysis substantiates the benefit of incorporating hidden features z for more accurate affect estimates of our joint models explained on section III-C.

1) Modality Analysis: To see individual contribution of each modality on the final estimates of our BMMN network, we trained BMMN for each modality separately by removing one modality to the other. This results in two other trained BMMN networks, with one, was trained only on physiological signals, and the other was using facial features only. We then compared their results against the normal BMMN, which received both bio-physio and facial features for a more complete analysis.

Label	Modality		
	Bio-signals	Faces	Multi-Modal
Valence	58,25 %	56,67 %	**67,91 %**
Arousal	55,65 %	78,28 %	**78,36 %**
Liking	69,01 %	75,16 %	**77,82 %**
Neutral	40,81 %	36,25 %	**48,91 %**
Disgust	55,08 %	26,01 %	**75,92 %**
Joy	45,88 %	40,48 %	**70,27 %**
Surprise	46,11 %	32,70 %	**77,27 %**
Anger	35,69 %	36,53 %	**51,18 %**
Fear	30,84 %	36,31 %	**62,09 %**
Sadness	29,88 %	26,24 %	**65,55 %**
Average	46,92 %	44,46 %	**67,53 %**

Table III: Results of BMMN utilizing bio-signals only, faces only, and both bio-signals and faces (multi-modal)

Table III provides results of trained models given specific modality of bio-signals only, faces only, and both (multi-modal). Based on these results, we can see that utilising only bio-signals and faces separately lead to comparable results. Mainly, bio-signals produced more stable accuracy across affects labels with 46,92 % total mean of accuracy. While using facial images results in similar, but unstable estimates with higher accuracy in some affect labels, such as arousal and liking while lower in other, with total accuracy of 44,46 %. This instability may the results of high parameter contained on our Spatial Networks. In general, these results are inferior when compared to our standard BMMN, i.e uses both modalities, with average accuracy of 67.53%. We can observe that in overall, it produces higher accuracy across all emotion with close to 10 % improvement on the valence estimates, indicating that bio-signals modality helps in classifying this particular affect dimension.

These findings suggest that both faces and bio-signals gives equal contribution to our BMMN models. However, we also note that our internal Spatial Network of BMMN has been externally pre-trained on other datasets, which shows the effectiveness of the structure of our Bio Network to be able to achieve comparable results. Furthermore, by utilising both modalities arranged in multi-task ways, we were able to improve overall affect estimates of our BMMN model outperforming their results when used separately.

2) Impact of Latent Feature Analysis: In this part, we present our baseline results from previous sections (BMMN) versus our other two model variants: BMMN-BAE-1 and BMMN-BAE-2. Table IV shows the overall comparisons of these models. Based on the total accuracy across affect labels, we can see that the introduction of z boosts the accuracy compared to the baseline. Notably when z is conflated together following the scheme of BMMN-BAE-2 models with quite a large margin of difference (71.57% vs 67.62%), while the observed improvement of BMMN-BAE-1 is negligible (less than 1%). In general, for the most considered important emotion dimensions of valence, arousal and liking [42], we found that the introduction of Z improves our models accuracy on both arousal and liking estimates, while we observe slight drop of accuracy for valence.

Label	Models		
	BMMN	BMMN-BAE-1	BMMN-BAE-2
Valence	**67,91 %**	66,19 %	65,05 %
Arousal	78,36 %	84,86 %	**87,53 %**
Liking	77,82 %	**78,76 %**	78,10 %
Neutral	48,91 %	44,29 %	**53,68 %**
Disgust	75,92 %	76,86 %	**92,62 %**
Joy	**70,27 %**	68,86 %	68,30 %
Surprise	**77,27 %**	73,90 %	72,94 %
Anger	51,18 %	55,62 %	**62,25 %**
Fear	62,09 %	**68,86 %**	65,15 %
Sadness	65,55 %	58,00 %	**70,10 %**
Average	67,53 %	67,62 %	**71,57 %**

Table IV: Results from variant of our models

Specifically, while following the scheme of BMMN-BAE-1 leads to the improved results on several affect estimates, such as liking and fear, however in overall, it produces comparable results. This may indicate that our baseline model is capable enough to indirectly captured its own internal features in their affect estimations, though not as compact as z used in BMMN-BAE-1 models. In other hand, we found noticeable improvements when z is used as complementary input along with raw signal as such BMMN-BAE-2 scheme, with minimal sacrifices in some affect estimates. Further comparison of the results of BMMN to the BMMN-BAE-2, we observe more than 10% accuracy improvement for anger and disgust emotions, and around 5% for sadness and neutral suggesting that the z may helps to estimates such emotions. Notice that these emotion states require a high level of arousal activation, which may explain close to 10% accuracy gain (87% from original 78.36%) in arousal estimation.

Based on these findings, we can conclude that the extracted latent features is beneficial on our model estimates, given that it is appropriately integrated. Which in this case is by using it as complementary information along with the raw signals for all modalities.

B. Comparison Against Other Studies

Table V presents the comparison of our best results from previous sections against other reported results on the

AMIGOS [11] dataset. We evaluate their results in terms of valence and arousal domain considering their extensive uses in consensus [32] and their availability on all compared studies. In addition, to investigate the importance of physiological signal for affect estimations, we also add state of the art, deep learning-based affect networks from Kollias et al. [10], which only uses facial features, i.e. without any bio-signal input in their emotion estimations pipeline.

Methods	Modality	Valence	Arousal
Kollias et al. [10]	Faces	48,76 %	60,35 %
Gjoreski et al. [32]	ECG, EDA	-	56,00 %
Yang and Lee [33]	EEG, ECG, EDA	68,80 %	67,00 %
Santamaria et al. [34]	ECG, EDA	76,00 %	75,00 %
Siddharth et al. [13]	Faces, EEG, EDA, ECG	**83,94 %**	82,76 %
BMMN-BAE-2 (ours)	Faces, ECG, EDA	65,05 %	**87,53 %**

Table V: Accuracy comparison with other related studies.

We can summarize several findings from these comparisons. Firstly, our model produces relatively comparable results against other approaches, with highest arousal estimation of 87.53% of accuracy. Even though it produces lower valence accuracy against the work of Gjoreski et al. [32] and Siddarth et al. [13], we need to note that their models require more elaborate EEG features, that demands more extensive instruments. Unlike others, however, our models only need to make efficient use of ECG and EDA signals alongside the face for inference, which is relatively easy to obtain.

Secondly, we notice that some approaches which do not involve joint modality of faces and bio-signals yield inferior results against other multi-modal approaches, including ours. The finding that complies by our conclusion in modality analysis suggesting the importance of multiple modalities input. Another important finding is relatively low results produced by Kollias et al. [10] that only uses facial features with the most moderate valence accuracy. This may be attributed to its lack of any bio-signal modality which may further help their estimations, which explains relatively higher results of other models that exploit them. Finally, we can see in overall a quite big margin of difference between the results of the handcrafted feature-based model, including Gjoreski et al. [32] to other compared models, which are deep-learning-based approach suggesting the superiority of the latter.

C. Assessment on Anxiety Therapy

We use our best performing model of BMMN-BAE-2 to evaluate the anxiety therapy using our collected dataset. Specifically, we run our model on the first 15 minutes subset of the data to obtain the mean emotions states of each patient prior the treatment. Then we collect another estimate for the last 15 minutes to represent the patients condition after the treatment. We accumulate these results for all patients and display them on each correspondent quadrant locations as shown on Figure 5 to see the existing changes.

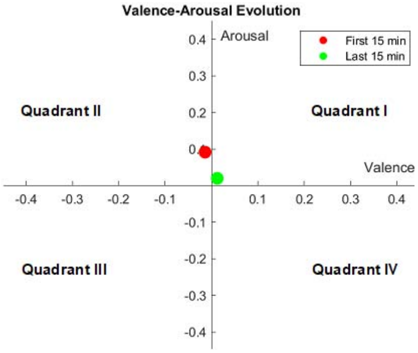

Figure 5: Changes on valence and arousal emotion of all patients, prior (red) and after the treatment (green). Scaled from -1 to 1.

We can see from the figure that our model able to produce quite sensible results, i.e. it shows the ideal tendency prior and post treatment on emotion changes (from second to four quadrant). We also notice that our models estimates are located close to the centre of each quadrant (neutral emotions). This may caused by our latent features, which are conditioned to specific characteristics of AMIGOS dataset it is trained upon, that we recall in the majority, all samples are located close to the neutral position (see Figure 1). Thus our training still lacks substantial amount of extreme examples. Nonetheless, our estimates still able to highlight the expected tendency of the treatment, though not too extreme, indicating its real-world capability.

V. Conclusion

In this paper, we present a novel multi-modal emotion recognition approach with Internal Auto-Encoder, which operates by extracting features efficiently from several modalities input. We propose a multi-stage pre-processing step to ease our model training on AMIGOS dataset, and we introduce a new therapy dataset to expand the real-world application of our models, and affective computing in general. We create our baseline models by the use of both physiological signals and facial features. Furthermore, we present an improvement by using bio latent features extracted using our internal bio auto-encoder.

The experiments using AMIGOS dataset provided us several findings: in the first place we observed the importance of multi-modality inputs to achieve higher accuracy compared to individual use of each modality independently. Secondly, we confirmed the benefit of using te latent features, notably by combining it with the original signal, which greatly improved our results from the baseline. Third, the comparison against other related studies in this dataset revealed our competitive results with higher arousal accuracy in overall. We also found that the current state of the art facial based on affect model, which ignores the importance of bio-signal modality, produced lower results than other multi-modal based approach. This finding further supported our conclusion regarding the benefit of multi-modalities input.

Finally, we applied our best performing models to our collected dataset to identify the emotion changes of the therapy to the patients. We later demonstrate that in fact, our model successfully shows the tendency of the ideal evolution of the patients emotion states, which reflects our models capability on the real-world application. In our future work, we seek to collect more examples to our dataset and adapt our models to its specific characteristics.

Acknowledgment

The authors would like to thank the research group of AMIGOS database used in this research and for granting us access to this dataset. We would also like to show our gratitude to Manel Ballester and Francesc Carreras from Hospital de la Santa Creu i Sant Pau (Barcelona) for their helps on data collections of medical therapy, and Oriol Pujol Martinez for the supports on developing the software for data acquisition. This work is partly supported by the Spanish Ministry of Economy and Competitiveness under project grant TIN2017-90124-P and the donation bahi2018-19 to the CMTech at the UPF.

References

[1] C. Darwin, *The expression of the emotions in man and animals*. New York ; D.Appleton and Co., 1916.

[2] W. James, "II. What is an emotion?" *Mind*, vol. os-IX, no. 34, pp. 188–205, 04 1884.

[3] W. B. Cannon, "The james-lange theory of emotions: A critical examination and an alternative theory," *The American Journal of Psychology*, vol. 39, no. 1/4, pp. 106–124, 1927. [Online]. Available: http://www.jstor.org/stable/1415404

[4] P. Ekman, R. W. Levenson, and W. V. Friesen, "Autonomic nervous system activity distinguishes among emotions," *Science*, vol. 221, no. 4616, pp. 1208–1210, 1983. [Online]. Available: https://science.sciencemag.org/content/221/4616/1208

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

[6] S. Duo and L. Song, "An e-learning system based on affective computing," *Physics Procedia*, vol. 24, 01 2010.

[7] C. Liu, K. Conn, N. Sarkar, and W. Stone, "Online affect detection and robot behavior adaptation for intervention of children with autism," *IEEE T Robot*, vol. 24, pp. 883 – 896, 09 2008.

[8] Y. Tian, T. Kanade, and J. F. Cohn, "Recognizing action units for facial expression analysis," *IEEE T Pattern Anal*, vol. 23, no. 2, pp. 97–115, Feb 2001.

[9] M. Yeasin, B. Bullot, and R. Sharma, "Recognition of facial expressions and measurement of levels of interest from video," *IEEE T Multimedia*, vol. 8, no. 3, pp. 500–508, June 2006.

[10] D. Kollias, P. Tzirakis, M. A. Nicolaou, A. Papaioannou, G. Zhao, B. Schuller, I. Kotsia, and S. Zafeiriou, "Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond," *IJCV*, pp. 1–23, 2019.

[11] J. Miranda, M. K. Abadi, N. Sebe, and I. Patras, "Amigos: A dataset for mood, personality and affect research on individuals and groups," *IEEE T Affect Comput*, vol. PP, 02 2017.

[12] Y. Kim, H. Lee, and E. M. Provost, "Deep learning for robust feature generation in audiovisual emotion recognition," in *ICASSP*, May 2013, pp. 3687–3691.

[13] S. Siddharth, T. Jung, and T. J. Sejnowski, "Utilizing deep learning towards multi-modal bio-sensing and vision-based affective computing," *IEEE T Affect Comput*, pp. 1–1, 2019.

[14] S. Tripathi, S. Acharya, R. D. Sharma, S. Mittal, and S. Bhattacharya, "Using deep and convolutional neural networks for accurate emotion classification on deap dataset," in *AAAI*, 2017.

[15] E. J. Choi and D. K. Kim, "Arousal and valence classification model based on long short-term memory and deap data for mental healthcare management," in *Healthcare informatics research*, 2018.

[16] H. Tang, W. Liu, W. L. Zheng, and B. L. Lu, "Multimodal emotion recognition using deep neural networks," in *Neural Information Processing*. Cham: Springer, 2017, pp. 811–819.

[17] A. F. Haag, S. Goronzy, P. Schaich, and J. Williams, "Emotion recognition using bio-sensors: First steps towards an automatic system," in *ADS*, 2004.

[18] S. Lahmiri and M. Boukadoum, "A weighted bio-signal denoising approach using empirical mode decomposition," *Biomedical Engineering Letters*, vol. 5, pp. 131–139, 2015.

[19] M. Kramer, "Nonlinear principal component analysis using autoassociative neural networks," *AIChE Journal*, vol. 37, pp. 233 – 243, 02 1991.

[20] S. Gao, Y. Zhang, K. Jia, J. Lu, and Y. Zhang, "Single sample face recognition via learning deep supervised auto-encoders," *IEEE T Inf Foren Sec*, vol. 10, pp. 1–1, 10 2015.

[21] B. Leng, S. Guo, X. Zhang, and Z. Xiong, "3d object retrieval with stacked local convolutional autoencoder," *Signal Processing*, vol. 112, pp. 119–128, 2015.

[22] A. M. Sarroff and M. A. Casey, "Musical audio synthesis using autoencoding neural nets," in *ICMC*, 2014.

[23] L. Theis, W. Shi, A. Cunningham, and F. Huszár, "Lossy image compression with compressive autoencoders," in *5th ICLR, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017.

[24] G. E. Hinton and R. Salakhutdinov, "Reducing the dimensionality of data with neural networks." *Science*, vol. 313 5786, pp. 504–7, 2006.

[25] Ö. Yildirim, R. S. Tan, and U. R. Acharya, "An efficient compression of ecg signals using deep convolutional autoencoders," *Cognitive Systems Research*, vol. 52, pp. 198–211, 2018.

[26] A. J. Fridlund and C. E. Izard, "Electromyographic studies of facial expressions of emotions and patterns of emotions," *Social psychophysiology: A sourcebook*, pp. 243–286, 1983.

[27] R. W. Picard, E. Vyzas, and J. Healey, "Toward machine emotional intelligence: analysis of affective physiological state," *IEEE T Pattern Anal*, vol. 23, no. 10, pp. 1175–1191, Oct 2001.

[28] C. L. Lisetti and F. Nasoz, "Using noninvasive wearable computers to recognize human emotions from physiological signals," *EURASIP Journal on Advances in Signal Processing*, vol. 2004, pp. 1–16, 2004.

[29] H. Feng, H. M. Golshan, and M. H. Mahoor, "A wavelet-based approach to emotion classification using eda signals," *Expert Syst Appl*, vol. 112, pp. 77–86, 2018.

[30] M. Soleymani, M. Pantic, and T. Pun, "Multimodal emotion recognition in response to videos," *IEEE T Affect Comput*, vol. 3, pp. 211–223, 04 2012.

[31] S. Koelstra, C. Mhl, M. Soleymani, J. S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "Deap: A database for emotion analysis using physiological signals," *IEEE T Affect Comput*, vol. 3, pp. 18–31, 12 2011.

[32] M. Gjoreski, M. Lustrek, M. Gams, and B. Mitrevski, "An inter-domain study for arousal recognition from physiological signals," *Informatica (Slovenia)*, vol. 42, pp. 61–68, 01 2018.

[33] H. C. Yang and C. C. Lee, "An attribute-invariant variational learning for emotion recognition using physiology," 2019, pp. 1184–1188.

[34] L. Santamaria-Granados, M. Organero, G. Ramirez-Gonzalez, E. Abdulhay, and A. N., "Using deep convolutional neural network for emotion detection on a physiological signals dataset (amigos)," *IEEE Access*, vol. PP, pp. 1–1, 11 2018.

[35] J. Oschman, *Energy Medicine: The Scientific Basis: Second Edition*, 01 2016.

[36] J. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, pp. 1161–1178, 12 1980.

[37] D. Aspandi, O. Martinez, F. Sukno, and X. Binefa, "Fully end-to-end composite recurrent convolution network for deformable facial tracking in the wild," in *14th IEEE FG*, May 2019, pp. 1–8.

[38] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CVPR*, pp. 770–778, 2015.

[39] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *CVPR*, June 2016.

[40] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.

[41] Y. L. Zhang and Q. Yang, "A survey on multi-task learning," *ArXiv*, vol. abs/1707.08114, 2017.

[42] P. J. Lang, "The emotion probe: studies of motivation and attention." *American psychologist*, vol. 50, no. 5, p. 372, 1995.