Available online at www.sciencedirect.com

ScienceDirect

IFAC PapersOnLine 50-1 (2017) 6940–6945

Physiological Feature Based Emotion Recognition via an Ensemble Deep Autoencoder with Parsimonious Structure

Zhong Yin*, Yongxiong Wang*, Wei Zhang*, Li Liu*, Jianhua Zhang, Fei Han*, Wenjie Jin***

*Engineering Research Center of Optical Instrument and System, Ministry of Education, Shanghai Key Lab of Modern Optical System, University of Shanghai for Science and Technology, Shanghai, 200093, P. R. China (Tel: 86-21-55271064; e-mail: yinzhong@usst.edu.cn; wyxiong@usst.edu.cn).

** Department of Automation, East China University of Science and Technology, Shanghai, 200237 P. R. China (Tel: 86-21-64253808; e-mail: zhangjh@ecust.edu.cn)

Abstract: Since the deep learning classifier has the capability to hierarchically abstract the useful information from the physiological signals, it receives more attention in human emotion recognition in recent studies. Considering the structure of the deep network is required to be independently determined for multiple physiological modalities, we propose an ensemble deep learning framework by integrating multiple stacked autoencoder with parsimonious structure (M-SAE) to reduce the model complexity and improve the recognition accuracy. In M-SAE framework, the physiological feature abstractions from the deep hidden neurons of each signal modality are separately extracted via a group of member SAEs. The structural hyper-parameters are identified by minimizing the loss of the data distribution similarity across the original features and activation potentials in the hidden layer. The performance comparison on DEAP database validates the competence of the M-SAE against several classical emotion classifiers.

Keywords: Human-machine interaction, deep learning, electroencephalogram, affective computing, pattern classification, feature fusion.

1. INTRODUCTION

Emotions can be defined as a set of human affective states generated by the responses to the transferable information from the environment or other individuals. Emotions can be categorized into several representative states via an arousal-valence model, where each state is described by the coordinates on arousal and valence axes (Soleymani *et al.,* 2012). In various human-machine interaction (HMI) environments, human intensions and commands may carry different emotions (Fanelli *et al.,* 2010). To develop an intelligent HMI system in which the machine agents share a human-centered collaborative capability, the prerequisite is to achieve the machine adaptability regarding the comprehension of human affective reactions.

The machine emotional intelligence requires an recognizer to extract and fuse the hidden affective information from human and provide the corresponding emotion estimation. In recent field studies, multimodal physiological signals received increasingly attention as sensitive emotional cues because of the better robustness against the expressionless users and the capability of exploiting the latent cognitive information from central and peripheral nervous systems (Zeng *et al.,* 2009; Hanjalic and Xu, 2005). Typically, emotional change can be detected by using the power spectra on multiple frequency bands of electroencephalogram (EEG). Verma and Tiwary (2014) reported the EEG power in alpha (8-13 Hz) can differentiate arousal and valence degrees. Balconi and Lucchiari (2006) employed an asymmetry metric of the alpha power across left and right scalps to measure emotional variations. In addition, the effectiveness of the peripheral physiological signals, for instance, electrocardiogram (ECG) (Agrafioti *et al.*, 2012), ocular signals (EOG) (Chanel *et al.*, 2011), galvanic skin response (GSR), and electro-muscular signals (EMG) (Zhang *et al.*, 2015), have been also extensively investigated.

To establish a link between the multimodal physiological data and the discrete emotional states, the machine learning methodologies have been applied as effective data-driven modelling approaches. In particular, deep learning principles have been investigated in recent studies due to its capacity of eliciting high-level physiological abstractions from hidden neuron activations. Wang and Shang (2013) used one of the deep learning primitives, i.e., deep belief networks (DBN) to fuse raw physiological signals. Li *et al.* (2015) employed a improved DBN model to integrate EEG features. Two binary classifiers on arousal and valence scales have been built, respectively. In above reported works, the structure of the deep network is determined based on the prior knowledge from other domains. However, the size of the subject-specific physiological data is usually limited for training so that the emotion classifier with too much hidden neurons may lead to the overfitting of the input feature and cannot inherit the advantages of the network structure in another domain. Hence, a physiological-data-driven approach that identifies the optimal structural hyper-parameters of the deep emotion classifier is required.

To address the above issue, we propose an ensemble classifier of stacked autoencoder with parsimonious structure

10.1016/j.ifacol.2017.08.1220

(M-SAE) to reduce the model complexity and improve the accuracy for emotion recognition. The high-level representations of physiological features in each modality are separately abstracted via SAE based deep architectures (Bengio *et al.*, 2007). A similarity metric for data distribution is proposed and applied to identify the optimal model hyper-parameters which include the number of hidden layers and the number of the artificial neurons in each layer. Then, we also employ a fusion network to integrate the activation potential from the hidden neurons in each SAE. Emotional states can be thus discriminatively indicated. The final decision layer is constructed by using a Bayesian classifier. The performance of the M-SAE is validated by using the DEAP database and compared against several existing emotion classifiers.

The organization of the paper is as below. The methodology of the M-SAE for emotion recognition is described in Sect. 2. The DEAP database and the corresponding feature extraction schemes of physiological data are introduced in Sect. 3. In Sect. 4, the comparison and analysis of the classification results are shown. Finally, we conclude the contribution of the work in Sect. 5.

2. ENSEMBLE DEEP LEARNING MODEL WITH PARSIMONIOUS STRUCTURE

2.1 Autoencoder and Stacked Autoencoder

The classical SAE is generated by cascading several autoencoders (Bengio *et al.*, 2007). An autoencoder is constructed with a three-layer, feed-forward network in which the input and output neuron activations are the same. Let denote the input physiological feature vector and its activation potential in the hidden layer as $\mathbf{x} \in R^D$ and $\mathbf{h} \in R^d$, respectively, a mapping from the input layer to the hidden layer can be described by the following equation,

$$\mathbf{h} = f_s(\mathbf{W} \cdot \mathbf{x} + \mathbf{b}), \tag{1}$$

where $\mathbf{W}$, $\mathbf{b}$, and $f_s(\cdot)$ denote the weight matrix, bias vector, and the logistic sigmoid function, respectively. Then, the input feature of the autoencoder is rebuilt in the output layer as $\hat{\mathbf{x}}$ by the tied weight $\mathbf{M} = \mathbf{W}^T$,

$$\hat{\mathbf{x}} = f_s(\mathbf{M} \cdot \mathbf{h} + \mathbf{c}) = f_s[\mathbf{M} \cdot f_s(\mathbf{M} \cdot \mathbf{h} + \mathbf{c}) + \mathbf{c}]. \tag{2}$$

Based on a squared-error cost function J_s, the parameters of the autoencoder, $\{\mathbf{W}, \mathbf{M}, \mathbf{b}, \mathbf{c}\}$, can be pre-trained via the back-propagation algorithm as,

$$\{\mathbf{W}, \mathbf{M}, \mathbf{b}, \mathbf{c}\} = \arg\min J_s\{\mathbf{x}, f_s[\mathbf{M} \cdot f_s(\mathbf{W} \cdot \mathbf{x} + \mathbf{b}) + \mathbf{c}]\}. \tag{3}$$

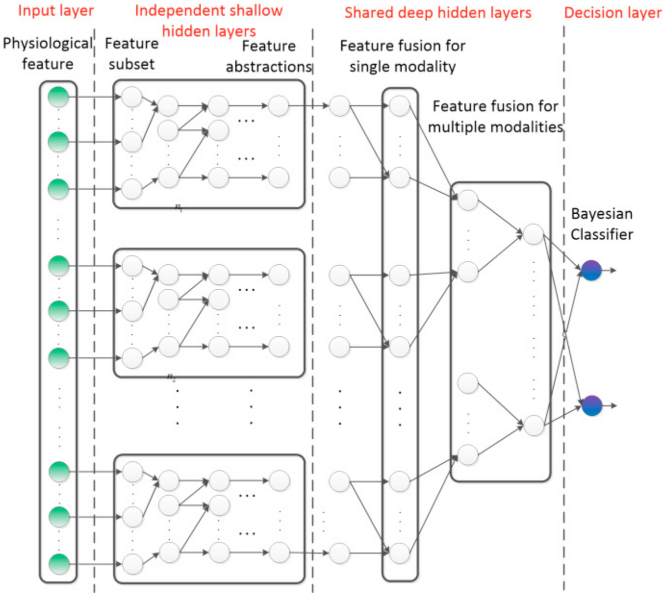
Let feed $\mathbf{h}$ to another autoencoder's input layer and repeat for *n* times. The high-level abstractions are elicited in the hidden layer of the new autoencoder. Thus, a SAE network is generated as below,

$$\begin{aligned}\mathbf{h}^{(n)} = f_s(\mathbf{W}^{(n)} \ldots f_s(\mathbf{W}^{(2)} f_s(\mathbf{W}^{(1)}\mathbf{x} + \\ \mathbf{b}^{(1)}) + \mathbf{b}^{(2)}) \ldots + \mathbf{b}^{(n)}) = \phi^{(n)}(\mathbf{x}).\end{aligned} \tag{4}$$

Finally, the output activations are fed to a Bayesian classifier to predict the affective classes. Note that the fining tuning can be applied on the whole network based on the pre-trained model parameters.

2.2 Ensemble SAEs with Parsimonious Structure

In the proposed M-SAE emotion classification framework, the single SAE model is used as the member feature encoder to separately process physiological features that share the same property. Then, the outputs of the member encoder are fused based on an additional deep model to improve the inter-class discrimination capacity for the high-level feature abstractions. The architecture of the M-SAE is presented in Fig. 1, where binary (low or high) arousal or valence levels can be estimated.



Input layer
Independent shallow hidden layers
Shared deep hidden layers
Decision layer
Physiological feature
Feature subset
Feature abstractions
Feature fusion for single modality
Feature fusion for multiple modalities
Bayesian Classifier

Fig. 1. Layout of the SAE ensemble for binary emotion recognition.

The independent shallow hidden layers can be initialized in the following way. Let denote the overall physiological feature set as U_0 with dimensionality of D_0, the non-overlapped feature subsets $\tilde{U}_i$ with D_i ($i > 0$) can be generated as,

$$\begin{aligned} & \tilde{U}_1^{D_1}, \tilde{U}_2^{D_2}, \cdots, \tilde{U}_m^{D_m} \\ s.t.\ & \tilde{U}_1^{D_1} \cup \tilde{U}_2^{D_2} \cup \cdots \cup \tilde{U}_q^{D_m} = U_0^{D_0}, \\ & \tilde{U}_i^{D_i} \cap \tilde{U}_j^{D_j} = \varnothing. \end{aligned} \tag{5}$$

Note that a feature subset contains a group of physiological features extracted by using a same approach. Then, the member SAE model with independent hidden layers can generate high-level feature representations in the following way.

$$\begin{aligned} \mathbf{h}_i &= \phi^{(n_i)}_{\tilde{\mathbf{W}}_i, \tilde{\mathbf{b}}_i}[\mathbf{x}(\tilde{U}_i^{D_i})] \\ &= f_s(\mathbf{W}_i^{(n_i)} \ldots f_s(\mathbf{W}_i^{(2)} f_s(\mathbf{W}_i^{(1)}\mathbf{x} + \\ &\quad \mathbf{b}_i^{(1)}) + \mathbf{b}_i^{(2)}) \ldots + \mathbf{b}_i^{(n_i)}), \\ & i \in \{1, 2, \cdots q\}. \end{aligned} \tag{6}$$

In Eqn. (6), $\phi^{(n_i)}_{\tilde{\mathbf{W}}_i,\tilde{\mathbf{b}}_i}(\cdot)$ denotes the member SAE with n_i hidden layers, $\tilde{\mathbf{W}}_i = \{\mathbf{W}_i^{(1)}, \mathbf{W}_i^{(2)}, ..., \mathbf{W}_i^{(n_i)}\}$ are the weight matrices between each two hidden layers, and the corresponding bias vectors are $\tilde{\mathbf{b}}_i = \{\mathbf{b}_i^{(1)}, \mathbf{b}_i^{(2)}, ..., \mathbf{b}_i^{(n_i)}\}$.

It is noted the structure of each member SAE must be identified when initializing the deep network parameters. To this end, we develop an objective function as a basis to optimize the deep model structure and then to find a parsimonious architecture. The structural loss function (SLF) for the deep emotion classifier is constructed as follows,

$$\tau = \lambda_1[1 - R_P^2(\mathbf{S}_\mathbf{X}, \mathbf{S}_\mathbf{H})] + (1-\lambda_1)[1 - R_S^2(\mathbf{S}_\mathbf{X}, \mathbf{S}_\mathbf{H})], \qquad (7)$$

In Eqn. (7), $\mathbf{S}_\mathbf{X}$ and $\mathbf{S}_\mathbf{H}$ are denoted as similarity matrices of SAE inputs and hidden neuron activations. Each entry of the matrix is the Euclidean distance between two vectors in the feature or abstraction space. The terms, $R_P^2(\mathbf{S}_\mathbf{X}, \mathbf{S}_\mathbf{H})$ and $R_S^2(\mathbf{S}_\mathbf{X}, \mathbf{S}_\mathbf{H})$, are the Pearson and Spearman correlation coefficients, respectively. The value of λ_1 is between the range of 0 to 1 in order to facilitate weighting the two terms in Eqn. (7). Since the linearity can lead to a simpler model structure while the nonlinearity mapping always exists between the activation potentials across two hidden layers, it is necessary to combine the Pearson and Spearman coefficients to build the loss function.

Table 1. Pseudo codes for M-SAE model structural identification

```
Define the training set $T$ for the SAE model
 for $i = 1:m$
   $T(i) = \{[\mathbf{x}_1(\tilde{U}_i^{D_i}], y_1), (\mathbf{x}_2[\tilde{U}_i^{D_i}], y_2), ..., (\mathbf{x}_N[\tilde{U}_i^{D_i}], y_N)\}$
 End for
$m$ training sets are parepared
Set learning rate $L_{le}$ and batch size $B_{bs}$ of BP algorithm
 for $i = 1:m$
   for $j = 1:n_m$
     for $k = 1:d_j$
       Construct a initial SAE model $\phi_{(i,j,\mathbf{k}_{\min}(j)=k)}$
         with $j$ hidden layers, $\mathbf{k}_{\min}$ neurons
       Pre-train $\phi_{(i,j,\mathbf{k}_{\min}(j)=k)}$ via $T(i)$
       Add a decision layer
         and unfold $\phi$ to $\tilde{\phi}_{(i,j,\mathbf{k}_{\min}(j)=k)}$ as a feedforward network
       Finely tune $\tilde{\phi}_{(i,j,\mathbf{k}_{\min}(j)=k)}$
       Compute the SLF $\tau$ between $\mathbf{x}(\tilde{U}_i^{D_i})$ and $\phi^{(j)}_{\tilde{\mathbf{W}}_i,\tilde{\mathbf{b}}_i}[\mathbf{x}(\tilde{U}_i^{D_i})]$
     End for
     Find the minimum of $\sigma_R^2$ for $k$=1 to $d_j$ as $k_{\min}$
     Save $k_{\min}$ via $\mathbf{k}_{\min}(j) = k_{\min}$
   End for
   The $n_m$ entries in $\mathbf{k}_{\min}$ are the numbers of neurons
     for $n_m$ hidden layers in $\phi_{(i,j,\mathbf{k}_{\min}(j)=k)}$
 End for
```

Hence, the small value of SLF indicates that relative distances of the physiological features and their abstractions are more correlated. It also suggests they share the similar local geometrical structure. Here, λ_1 is set to 0.5 to balance linear and nonlinear correlation metrics. The yielded pseudo codes of identifying the SAE structure are listed in Table 1, where the learning rate and batch size of the back-propagation (BP) algorithm for pre-training and fining tuning are 1 and 10, respectively. The $\tilde{\phi}$ is denoted as the feed-forward network that is unfolded by a SAE.

The shallow hidden layer of M-SAE with identified structure can be trained and elicit high-level feature abstractions for each feature subset. Then, the abstraction fusion is performed based on an adjacent graph based mapping. More specifically, the graph G^{jk} is built at j^{th} shared hidden layer and k^{th} modality of physiological data with N training instances, where each two instances of the same class share an edge. Then, the corresponding weight matrix E^{jk} can be generated in which the value of the entry is set to 1 (or 0) with (without) having an edge between two instances.

Then, the feature fusion mapping in each shared hidden layers is performed in a hierarchical way. The weight matrix can be computed based on the generalized eigenvector problem as follows,

$$H^{(jk)}L^{(jk)}H^{(jk)T}\mathbf{c}^{(jk)} = \mathbf{e}^{(jk)}H^{(jk)}P^{(jk)}H^{(jk)T}\mathbf{z}^{(jk)}. \qquad (8)$$

Eqn. (8) is motivated based on the linear discriminating analysis that has the capability to improve the inter-class discrimination and reduce the intra-class diversity. The matrix of $H^{(jk)}$ denotes the feature abstractions of N training instances at j^{th} shared hidden layer and k^{th} modality. The diagonal matrix $P^{(jk)}$ with N entries is computed from E^{jk}, i.e.,

$$P_n^{(jk)} = \sum_m E_{m,n}^{(jk)}. \qquad (9)$$

The matrix $L^{(jk)}$ is defined as,

$$L^{(jk)} = P^{(jk)} - E^{(jk)}. \qquad (10)$$

The derived Eigen values can be sorted by,

$$e_0^{(jk)} < e_1^{(jk)} < \cdots < e_{d^{(jk)}-1}^{(jk)} \qquad (11)$$

with $d^{(jk)}$ denoting the abstraction dimensionality in each physiological modality. Given the solutions of Eqn. (8) by $e_0^{(jk)}, e_1^{(jk)}, \cdots, e_{d^{(jk)}-1}^{(jk)}$, the transformation matrix of the fusion mapping between j^{th} and $(j+1)^{th}$ shared hidden layers are generate by,

$$\begin{aligned} \mathbf{h}_i^{[(j+1),k]} &= Z^{(jk)T}\mathbf{H}_i^{(jk)}, \\ Z^{(jk)T} &= [\mathbf{z}_0^{(jk)}, \mathbf{z}_1^{(jk)}, \cdots, \mathbf{z}_{d^{(jk)}-1}^{(jk)}]. \end{aligned} \qquad (12)$$

Finally, the low and high arousal or valence emotional states $\tilde{y}$ can be estimated via a Bayesian decision layer, i.e.,

$$\tilde{y} = \arg\max_y P(\mathbf{H}_i^{(J)} \mid y). \qquad (13)$$

In Eqn. (13), $\mathbf{H}_i^{(J)}$ denotes the fused physiological feature abstractions in the last (J^{th}) shared hidden layer. To this end, the hidden activations of the SAEs have been properly fused.

3. PHYSIOLOGICAL DATA AND FEARURE EXTRACTION

3.1 DEAP Database

To investigate the performance of the M-SAE emotion classifier, we employ the Database for Emotion Analysis with Physiological signals (DEAP). The DEAP database is constructed by Koelstra *et al.* (2012) and is publicly available for exploring human affective states induced by musical videos. The physiological data of 32 healthy subjects (19-37 years, mean=26.9) were collected. Each subject performed 40 trials of the experiment. For each trial, the participant watched the selected musical video of 1 minute with the EEG (32 channels) and peripheral physiological signals (13 channels) of 512 Hz sampling rate. The participant subjectively rates the trial by the arousal, valence, liking and dominance scales from 1 to 9. In this study, the arousal and valence markers are discretized via the threshold of 5 into low vs. high levels as the target classes.

Table 2. Physiological features

Index	Notations
#1-160	EEG power: average power in theta, slow-alpha, alpha, beta, and gamma bands, (4-7 Hz, 8-10 Hz, 8-13 Hz, 14-25 Hz, 26-45 Hz)
#161-216	EEG power right-left scalp differences
#217-344	EEG time-domain features: mean, variance, zero-crossing rate, and entropy
#345-349	EOG/EMG power features
#350-365	EOG/EMG time-domain features: mean, variance, zero-crossing rate, and entropy
#366-371	Skin temperature features: power, mean, variance, entropy, and mean of derivative
#372-396	GSR features: mean, mean of derivative, mean of negative derivative values
#397-403	Blood pressure power features
#404-425	Respiration features: mean, mean of derivative, centroid of PSD, respiration rate

3.2 Physiological Feature Extraction

The one-minute physiological signals are first down-sampled to 128 Hz. Then, muscular noise in EEG has been removed based on the independent component analysis. The muscular artefact can be removed via visual inspection on the specific independent IC. Table 2 lists the adopted physiological features with the dimensionality of 425. The z-score standardization for the data of each participant was employed to eliminate feature scale differences.

4. RESULTS FOR MODEL IDENTIFICATION AND EMOTION CLASSIFICATION

4.1 Model Identification for M-SAE

The parsimonious structure of the deep network is determined according to the minimum SLF. According to feature extraction approaches, 11 feature subsets are employed. Model selection results of three member SAEs for left, central, and right scalp EEG power feature subsets are shown in Fig. 2. Taken Fig. 2(a) as an example, an input layer for 70 power features (i.e., 70 input neurons) of left scalp is first initialized with the SLF computed from 1 to 70 neurons in the 1st hidden layer. The line-circle plot indicates the SLF minimum is at 47 hidden neurons. For shallow hidden layer 2, 31 neurons could be determined in the same way with 47 candidate neurons.

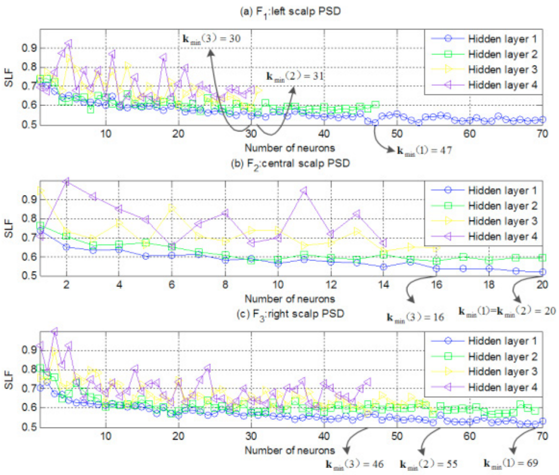


Fig. 2. Value of SLF vs. number of neurons in the first three member SAEs.

It is noted that the SLF variation shows unstable changes with four hidden layers employed. Such observation indicates stability of the training is reduced originated by too deep hidden layers. The sigmoid function is thus acceptable since the number of the hidden layers is relatively small due to the limited training instances of the physiological features. Table 3 lists the hidden neuron numbers of all 11 SAEs for each feature subset. It shows the number of physiological feature abstractions is reduced to 227 from 425 features.

Table 3. Identified SAE model structure

Hidden layer	1st	2nd	3rd
SAE1	47	31	30
SAE2	20	16	14
SAE3	69	55	46
SAE4	56	55	51
SAE5	54	22	10
SAE6	64	39	27
SAE7	20	20	19
SAE8	6	3	2
SAE9	23	15	15
SAE10	7	5	4
SAE11	14	14	9
Sum	380	275	227

After model identification of member SAEs, the integrated structure of the M-SAE is very complex. In order to validate

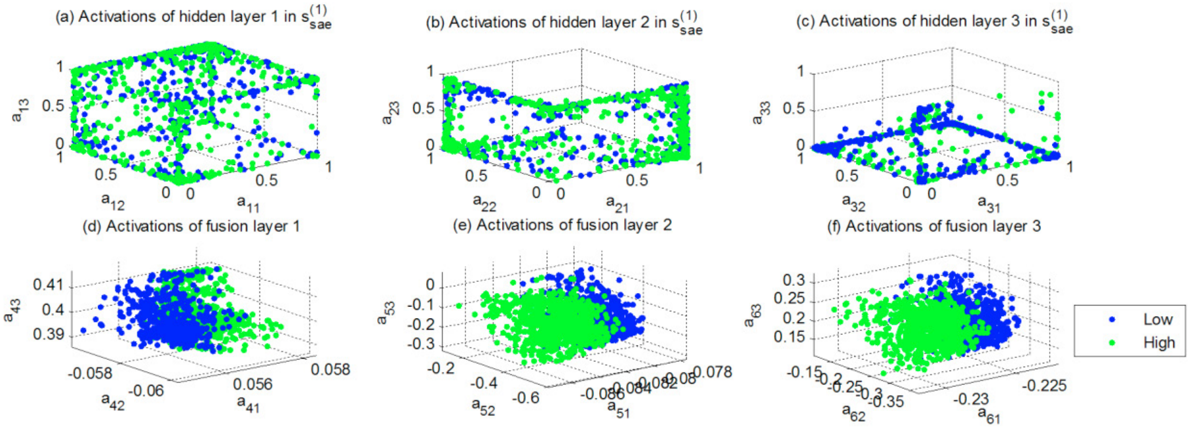


Fig. 3. Representative neuron activations in M-SAE for arousal recognition. a_{ij} is the j^{th} activation in the i^{th} hidden layer.

the effectiveness of the parsimonious deep network abstracting physiological features, the representative neuron activation potentials of each hidden layer are visualized by the 3-D scatter plots in Fig. 3 when the classifier is used to recognize binary arousal levels. In Fig. 3(a), the instances of feature abstractions for all participants of high arousal class are almost located on the edges. Those of low arousal state approximately lies in the middle. In Fig. 3(b) and (c), the abstractions become concentrated. In Fig. 3(d) to (f), an increased discrimination can be observed by two clear clusters when the shared fusion layer is applied.

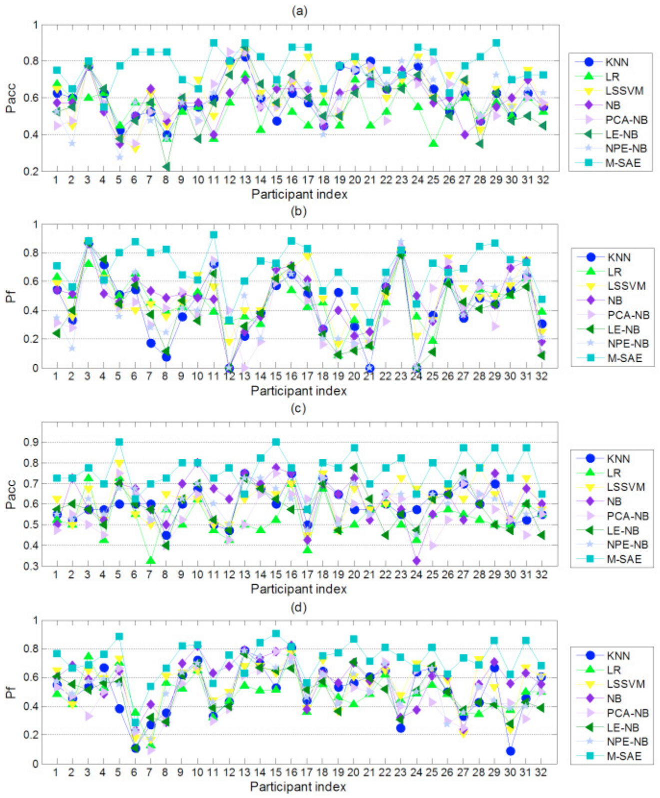


Fig. 4. Emotion classification accuracy and F1-score for arousal ((a) and (b)) and valence dimension ((c) and (d)). The subject-specific results are shown.

4.2 Classification Results Comparison

In this section, M-SAE's performance is compared with several state-of-the-art emotion classification model. Let denote the low and high arousal or valence levels as positive and negative classes, respectively. The classification performance can be described by the correct classification rate P_{acc} and the F1-score of low emotional level P_f .

The classification performance metrics of eight classifiers, i.e., KNN, LR, LSSVM, NB, PCA-NB, LE-NB, NPE-NB, and M-SAE are computed. KNN indicates the K-nearest neighbor classifier with the selected parameter of K=4, LR LSSVM, and NB are the logistic regression, linear least square support vector machine. and naive Bayesian classifiers, respectively. PCA-NB, LE-NB, and NPE-NB denote hybrid Bayesian classifiers with the physiological features mapped by principal component analysis, Laplacian eigenmaps, and neighbor preserving embedding, respectively. Considering the characteristic of individual difference that commonly exists in physiological features, all classifiers are evaluated via ten-fold cross validation method in a subject-specific manner.

The classification results of arousal dimension for each participant are illustrated in Fig. 4(a)-(b). It is shown M-SAE achieves the highest average classification rate (P_{acc}). In particular, the values of P_{acc} for participant 11, 13, 16, 17, 24, and 29 are approximately 90%. It is noted for specific participant, i.e., 5, 8, and 28, M-SAE has the capability to substantially improve the P_f values. In Fig. 4(b), M-SAE achieves the best P_f while PCA-NB achieves the lowest P_f for most of the participants. The observation suggests the unsupervised dimensionality reduction via a shallow linear mapping is not sufficient for generating stable physiological abstractions. Other classifiers show quite similar performance particularly for participant 3, 5, 15, and 23. For the data of specific participants, the high skewness of instances between low and high arousal classes and the classifier has taken all testing data as a single class. On the other hand, the classification results of valence dimension are shown in Fig. 4(c) and (d). The M-SAE also achieves the performance for most participants. It is found that recognizing low and high valence class is more difficult than that of arousal dimension for participant 6 and 23. On the contrary, the arousal classification performance for participant 12, 21, and 24 is much better. That is, the performance of M-SAE is highly dependent with the individual. The performance of the proposed M-SAE is compared against other classifiers by using the *t*-test. The *t*-test comparison is repeatedly carried out and the significant difference of the distributions are found for all cases with $p<0.001$. The summations of the confusion matrices from all 32 participants for each classifier and emotion dimension are listed in Table 4, it is shown the number of the corrected predicted instances for both of low and high emotional classes are higher than other classifiers.

Table 4. Total classification confusion matrices (summation over all participants).

Classifier	Predicted	Target arousal		Target valence	
		low	high	low	high
KNN	low	276	230	316	259
	high	267	507	256	449
LR	low	290	330	298	324
	high	253	407	274	384
LSSVM	low	295	232	335	250
	high	248	505	237	458
NB	low	321	307	368	276
	high	222	430	204	432
PCA-NB	low	270	236	287	263
	high	273	501	285	445
LE-NB	low	261	260	292	251
	high	282	477	280	457
NPE-NB	low	278	231	307	233
	high	265	506	265	475
M-SAE	low	**402**	151	**433**	166
	high	141	**586**	139	**542**

6. CONCLUSIONS

This paper proposed a multimodal physiological signal fusion approach based on multi-fusion-layer based ensemble SAE network for emotion recognition. Optimal network layout of the independent SAE has been systematically identified via a structural loss function. An adjacent-graph based fusion network that consists of three fusion layers for merging feature abstraction was newly introduced to achieve final predictions emotional states. The DEAP multimodal database of EEG and peripheral signals with 32 participants were employed to validate the effectiveness of the proposed method. Analysis of neuron activations on different hidden layers demonstrated that the inter-class discrimination between binary emotions was hierarchically improved with the parsimonious model structure. In future work, the comparison between the proposed M-SAE classifier and the comparative deep learning methods, i.e., shallow auto-encoders and deep belief networks, will be investigated in detail.

ACKNOWLEDGEMENTS

This work is sponsored by the Shanghai Sailing Program (17YF1427000), the National Natural Science Foundation of China under Grant No. 61673276, No. 11502145, No. 61603256. We thank all reviewers for their helpful comments and constructive suggestions on an earlier draft of this paper.

REFERENCES

Agrafioti, F., Hatzinakos, D., Anderson, A.K. (2012). ECG pattern analysis for emotion detection, *IEEE Trans. Affect. Comput.*, 3, 102-115.

Balconi, M., Lucchiari, C. (2006). EEG correlates (event-related desynchronization) of emotional face elaboration: a temporal analysis. *Neurosci. Lett.*, 392, 118-123.

Bengio, Y., Lamblin, P, Popovici, D. (2007). Larochelle, H., Greedy layer-wise training of deep networks. *Adv. Neural. Inf. Process. Syst.* 19, 153-160.

Chanel, G., Rebetez, C., Bétrancourt, M., Pun, T. (2011). Emotion assessment from physiological signals for adaptation of game difficulty, *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, 41, 1052-1063.

Fanelli, G., Gall, J., Romsdorfer, H., Weise, T., Van Gool, L. (2010). A 3-D audio-visual corpus of affective communication, *IEEE Trans. on Multimedia*, 12, 591-598.

Hanjalic, A., Xu, L.-Q. (2005). Affective video content representation and modeling, *IEEE Trans. on Multimedia*, 7, 143-154.

Koelstra, S., Muehl, C., Soleymani, M., Lee, J.-S., Yazdani, A., Ebrahimi, T.E., Pun, T., Nijholt, A., Patras, I.Y. (2012). DEAP: a database for emotion analysis using physiological signals, *IEEE Trans. Affect. Comput.*, 3, 18-31.

Li, X., Zhang, P., Song, D., Yu, G., Hou, Y., Hu, B. (2015). EEG based emotion identification using unsupervised deep feature learning, *SIGIR2015 Workshop on Neuro-Physiological Methods in IR Research*, Santiago, Chile.

Soleymani, M., Lichtenauer, J., Pun, T., and Pantic M. (2012). A multi-modal database for affect recognition and implicit tagging, *IEEE Trans. Affect. Comput.,* 3, 42-55.

Verma, G.K., Tiwary, U.S. (2014). Multimodal fusion framework: a multiresolution approach for emotion classification and recognition from physiological signal, *NeuroImage*, 102, 162-172.

Wang, D., Shang, Y. (2013). Modeling Physiological Data with Deep Belief Networks, *Int. J. Inf. Educ. Technol.*, 3, 505-511.

Zhang, J., Yin, Z., Wang, R. (2015) Recognition of mental workload levels under complex human-machine collaboration by using physiological features and adaptive support vector machines, *IEEE Trans. Hum. Mach. Syst.*, 45, 200-214.

Zeng, Z., Pantic, M., Roisman, G.I., Huang, T.S. (2009). A survey of affect recognition methods: audio, visual, and spontaneous expressions, *IEEE Trans. Pattern Anal. Mach. Intell.,* 31, 39-58.