Affective State Recognition from EEG with Deep Belief Networks

Kang Li*, Xiaoyi Li*, Yuan Zhang† and Aidong Zhang*
* State University of New York at Buffalo,
kli22,xiaoyili,azhang@buffalo.edu
† Beijing University of Technology
zhangyuan@emails.bjut.edu.cn

***Abstract*—With the ultimate intent of improving the quality of life, identification of human's affective states on the collected electroencephalogram (EEG) has attracted lots of attention recently. In this domain, the existing methods usually use only a few labeled samples to classify affective states consisting of over thousands of features. Therefore, important information may not be well utilized and performance is lowered due to the randomness caused by the small sample problem. However, this issue has rarely been discussed in the previous studies. Besides, many EEG channels are irrelevant to the specific learning tasks, which introduce lots of noise to the systems and further lower the performance in the recognition of affective states.**

To address these two challenges, in this paper, we propose a novel Deep Belief Networks (DBN) based model for affective state recognition from EEG signals. Specifically, signals from each EEG channel are firstly processed with a DBN for effectively extracting critical information from the over thousands of features. The extracted low dimensional characteristics are then utilized in the learning to avoid the small sample problem. For the noisy channel problem, a novel stimulus-response model is proposed. The optimal channel set is obtained according to the response rate of each channel. Finally, a supervised Restricted Boltzmann Machine (RBM) is applied on the combined low dimensional characteristics from the optimal EEG channels. To evaluate the performance of the proposed Supervised DBN based Affective State Recognition (SDA) model, we implement it on the *Deap Dataset* and compare it with five baselines. Extensive experimental results show that the proposed algorithm can successfully handle the aforementioned two challenges and significantly outperform the baselines by 11.5% to 24.4%, which validates the effectiveness of the proposed algorithm in the task of affective state recognition.

I. INTRODUCTION

Affective state recognition is the process of objectively identifying the subjective feelings through learning on the related biological signals. The recognized affective states can make external devices understand the emotions of the users and hence react in more appropriate ways to increase the quality of the service. For instances, in [1], researches found that empathic feedback could reduce user arousal while hearing interviewer questions; and in [2], researchers discussed the related emotion of each music which enables retrieving music according to users' affective states.

In the current studies of affective state recognition, EEG plays a major role in manipulating the emotion related biological signals. Specifically, multiple electrodes are spread over the scalp to obtain voltage fluctuations resulting from the neurons of the brain during various affective states. The simultaneously sampled features from the multiple electrodes then form the multi-channel EEG signals. In the task of affective state recognition, each segment of multi-channel EEG signals is divided into predefined affective state classes, such as happy and unhappy, and like and dislike.

In the existing methods of affective state recognition from EEG, two challenges are rarely discussed. First, to capture the details of brain activities w.r.t. different emotions, the segments of the multi-channel EEG signals usually have more than thousands of features in each channel. Due to the expense of labeling each emotion sample, only a few labeled emotion samples are available for the learning. Using the few labeled samples to guide the learning on the over thousands of feature may cause severe small sample problem. In these cases, unrelated features gain significance in the learning due to the randomness while important features may lose focus in the similar way. Second, most of the channels in the multi-channel EEG signals are irrelevant to the specific learning task. These irrelevant channels introduce lots of noise to the recognition of affective states and could significantly reduce the performance of the learning methods.

To tackle these two challenges, in this paper, we propose a novel supervised DBN based model. Specifically, a DBN is firstly utilized to extract the low dimensional characteristics of the data in each channel. The extracted low dimensional and deep characteristics can well reproduce the features of each channel, and are used in the learning to avoid the small sample problem. To filter out irrelevant channels in the multi-channel EEG signals, we measure the response rate of each channel according to a novel stimulus-response model, and select the channels that actively response to the specific emotions. Finally, the deep characteristics of the optimal channel set is combined and input into a supervised RBM for the purpose of discriminatively learning the affective states.

II. METHOD

A. Notation and Problem Definition

As listed in **Table I**, suppose there are n samples of EEG signals $\{G_1,\ G_2,\ G_3,\ ...,\ G_n\}$, and each sample contains the simultaneously sampled data in c channels as $G_i = \{X^j_{i,:} | j \in [1, c]\}$, where $X^j_{i,:}$ denotes the data in the channel j of the sample i. Specifically, in all the cases, $X^j_{i,:}$ has f features. Reshaping the data according the channels, we obtain another representation of the EEG signals as $\{X^j_{n\times f}\}^c_{j=1}$,

978-1-4799-1310-7/13/$31.00 ©2013 IEEE

TABLE I. NOTATION

n	number of samples
m	number of labeled samples
a	number of affective states
c	number of channels in the multi-channel EEG signals
f	number of features in each channel
G_i	the i-th EEG sample
$X^j_{i,:}$	the data in the j-th channel of the i-th EEG sample

in which $X^j_{n\times f}$ represents the data of different samples in channel j. We define the label matrices for the n samples as $\{Y^1_{1\times a},\ Y^2_{1\times a},\ ...,\ Y^n_{1\times a}\}$. In this notation, Y^i_k is the probability that the i-th sample is in the k-th affective state, and a is the number of affective states in consideration.

In the paper, we assume m of the n samples are already labeled in the a affective states, and denote the set of labeled EEG samples as L. The problem of affective state recognition can be expressed as: given the n EEG samples $\{G_i\}^n_{i=1}$ and the existing label matrix $\{Y^i\}_{i\in L}$, learn a mapping $\mathcal{F}: G \rightarrow Y$.

B. Deep Feature Extraction

In most cases, on the learning of affective states from EEG signals, only limited labeled samples are provided while each sample contains over thousands of features. Using the small number of training instances to supervise the learning of the large number of features cause severe small sample problems.

To handle the small sample problems, in the paper, we propose to extract high level and latent characteristics of each sample through the DBN model [3]. Specifically, the extracted deep characteristics could be used to well reconstruct the initial sample features , which favors minimizing the information loss in this process. By this step, high dimensional features in the samples are integrated into low dimensional latent characteristics. In the task of affective state recognition, learning on the low dimensional latent characteristics can avoid the small sample problem which caused by learning on the high dimensional features.

In general, DBN models are viewed as stacked RBM. For better clarity, we first give out the basic concepts of RBM, then introduce how to expand RBM to DBN.

RBM is a generative stochastic neural network that can learn the high level and nonlinear representations of its input variables. As demonstrated in Fig.1, the RBM model contains a set of visible units V representing the input initial variables, a set of hidden units H representing the learned latent layer, the symmetric weights W representing the connections between the visible and latent layers, and biases B and C to the visible and hidden layer, respectively. In the structure of an RBM, no connection exists among units in V or in H, and W connects the hidden and visible units to form a bipartite network.

The learning process of RBM seeks to obtain a distribution over its set of inputs. In details, an RBM has a joint probability distribution over the visible units V and hidden units H as:

$$P(V,H) = \frac{e^{-E(V,H)}}{z}. \tag{1}$$

In Eq.1, z is a normalizing factor. $E(V,H)$ is the energy function, which is usually defined as:

$$E(V,H) = -B^\top V - C^\top H - H^\top WV. \tag{2}$$

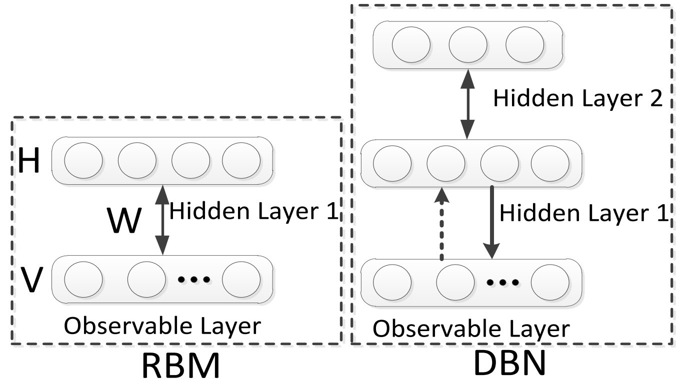

Fig. 1. **Models of RBM and DBN**. In RBM, weight matrix W connect the visible layer V and the hidden layer H to form a bipartite network. In DBN, multiple RBM models stack up to form a stochastic and deep graphical mode.

In the commonly studied cases, the above model is simplified by using binary input variables. As a result, a probabilistic version of activating the neurons in visible and hidden layers is formulated as:

$$\begin{aligned} P(H_i = 1|V) &= sigm\left(C_i + W_i V\right), \\ P(V_j = 1|H) &= sigm\left(B_j + W_j^\top H\right). \end{aligned} \tag{3}$$

In Eq.3, $sigm(x)$ is the logistic sigmoid function.

According to this definition, by Bayesian theory, the target parameters $\{W, B, C\}$ can be obtained through stochastic gradient on the negative log-likelihood of the visible units V as: $-\frac{\partial \sum_H P(V,H)}{\partial \Theta}$, where Θ could be any target variable matrices in $\{W, B, C\}$. However, this naive solution is intractable since it involves the computation of the expectation $\mathbb{E}_{V,H}\left(\frac{\partial}{\partial\Theta}E(V,H)\right)$. For this problem, contrastive divergence gradient [4] technique is commonly used to approximate the expectation by a sample generated after a limited number of Gibbs sampling iterations. Current studies [5] show that even when using only one Gibbs sampling in the iteration, contrastive divergence can produce very reliable performance with significant speed-up in training time.

Through greedily stacking RBM models, DBN can be obtained as illustrated in Fig.1. By the hierarchical stacking the RBM models, deep and high level representation of the initial input variables can be extracted.

C. Critical Channel Selection

In brain computer interface, the signal of each channel in the multi-channel EEG is collected from an electrode attached to the scalp, which seeks to capture the activities in the attached area of the scalp. In biology, brain related activities, which include emotions, action and etc., are usually dominated by several specific areas of the brain. Therefore, the multi-channel EEG signals contains many channels that are irrelevant to the learning of affective states. To filter these irrelevant channels, in this section, we present a DBN based critical channel selection method.

Suppose there are c channels in the EEG signals. By applying the DBN introduced in Section II-B on the data of each channel, we can obtain c independent DBN models.

Obviously, data in irrelevant channels are irrelevant to the emotion activities, thus tend to be distributed randomly. In the contrast, data in critical channels are tightly associated with the specified affective states, thus tend to be distributed in certain patterns rather than random. Therefore, in learning the DBN on each channel, data in irrelevant channels randomly update the parameters in the DBN, and data in critical channels update the parameters in the DBN according to the related patterns. Therefore, each trained DBN encode the distribution pattern of the input channel.

Based on the above observation, we propose a novel approach that detects the critical channels from the c trained DBN models. Suppose the observable layer of each channel contains f features, we define **zero-stimulus** as $S_{1\times f} = \vec{0}$ which is a all zeros vector, and name the deepest feature vector of a DBN on the zero-stimulus as the **response** P.

The response can be calculated as:

$$\begin{aligned} &H^1 = sigm\left(C^1 + W^1 S\right), \\ &H^l = sigm\left(C^l + W^l H^{l-1}\right), \forall l \in [2, k], \\ &P = H^k. \end{aligned} \tag{4}$$

According to theory of DBN, when a channel is irrelevant to the learning task, the response of the zero-stimulus is close to a vector of 0.5, which indicating that each unit in the deepest layer H^k is randomly activated. In the contrast, for critical channels, the responses contain many features biased from 0.5. To measure the degree of the response of each DBN, we further define **response rate** as:

$$R = \sum_{i=1}^{d} \left(P_i - \frac{1}{2}\right)^2, \tag{5}$$

where R is the response rate and d is the dimension of the deepest/highest layer in the DBN.

Since R measures to how active the response is, the larger R is, the more critical the channel is. With a user-specified parameter u, the channels with the top u large response rates are selected as the critical channels. In the implementation, we fix $u = 5$ in all the cases.

D. Learning and Prediction

For the i-th channel, the deepest feature matrix of the DBN on the initial data is denoted as T^i. Let F denote the set of the selected critical channels, the deep features of the selected channels are $\{T^i\}_{i\in F}$. We combine these selected deep features as $T_{n\times(d\cdot u)} = \cup\{T^i\}_{i\in F}$, where T is the obtained matrix. By T, each of the n input samples is represented by the union of the corresponding deep features (length d) in the u selected channels.

By the problem definition of affective state recognition, among the n samples, labels of the m of them are provided for the training process. We denote $T = L \cup \tilde{L}$, in which L is the set of the m labels samples and $\tilde{L}$ is the set of the $n-m$ unlabeled samples. We further denote the labels for L and $\tilde{L}$ are Y and $\tilde{Y}$, respectively.

For the purpose of learning the label $\tilde{Y}$ of $\tilde{L}$, we jointly train on T and Y in a generative supervised RBM [6]. In

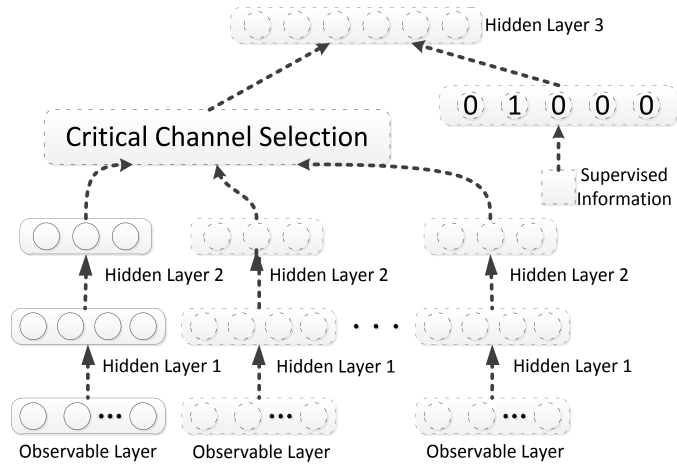

Fig. 2. **The Framework of the Proposed Model**. In the framework, the data of each channel are firstly processed by a DBN to extract highly level deep information. Through the channel selection model, the deep features of the selected channels are combined and feed into a supervised Restricted Boltzmann Machine to utilize the supervised information and make prediction on the unlabeled samples.

details, as demonstrated in Fig.2, the training features L and the existing labels Y are jointly mapped to the highest hidden layer of the whole model. Notice that here Y is preprocessed into sample-class matrix as $Y_{ij} = 1$ if sample i belongs to the affective state j, otherwise, $Y_{ij} = 0$.

The joint probability of the input L, label Y and the highest hidden layer J is:

$$P(L, Y, J) = \frac{e^{-E(L,Y,J)}}{z}. \tag{6}$$

In Eq.6, z is the normalizing factor. $E(L, Y, J)$ is the energy function defined as:

$$E(L, Y, J) = -B^\top L - D^\top Y - C^\top J - J^\top W_L L - J^\top W_Y Y. \tag{7}$$

In Eq.7, B, D and C are the bias matrices for L, Y and J, respectively. W_L is the weight matrix that connects input data L and hidden layer J into the bipartite network, and W_Y is the weight matrix that connects existing label Y and the hidden layer J. By the binary simplification, we have:

$$\begin{aligned} &P(L_i = 1|J) = sigm\left(B_i + \sum_j W_{L,ji} J_j\right), \\ &P(Y_y|J) = \frac{e^{D_y + \sum_j W_{Y,jy} J_j}}{\sum_{y^*} e^{D_y + \sum_j W_{Y,jy} J_j}}, \\ &P(J_j = 1|L, Y_y) = sigm\left(C_j + W_{Y,jy} + \sum_i W_{L,ji} L_i\right). \end{aligned} \tag{8}$$

According to the above equation, the model parameters $\{B, D, C, W_Y, W_L\}$ can be learned following the technique in [6]. The deep characteristics $\tilde{L}$ of the unlabeled data can then be put into the trained model to learn the target label matrix $\tilde{Y}$.

III. Experiments

A. Dataset and Evaluation Metric

The *DEAP* data set [2] is a database for emotion analysis using physiological signals. In the data set, the multi-channel EEG signals of 32 participants were recorded while each of them watched 40 one-minute long excerpts of music videos. According to the surveys, each music video was rated w.r.t. arousal, valence, like/dislike, dominance and familiarity of the participants. Specifically, the multi-channel EEG signals of each participant contains 40 channels, and the data of each channel have 8064 features. We process the data in the same way to [2], and focus on detecting whether each participant likes or dislikes the videos.

In signal detection theory, the receiver operating characteristic (ROC) curve plots the fraction of true positives out of positives vs. the fraction of false positives out of the negatives. In the experiments, we use area-under-the-curve (AUC), which measures the area under the ROC curve, to numerically evaluate the goodness of each result. Notice that AUC scores are in the range of $[0, 1]$. The higher AUC score a result achieves, the better the performance is.

B. Baselines

To evaluate the superiority of the proposed **S**upervised **D**BN based **A**ffective State Recognition (SDA) model, in the experiments, we compare it to five baselines.

Since the problem we study in this paper fall into the category of classification, we set support vector machine (SVM) [7] as the first baseline.

In the proposed SDA model, the idea of solving the small sample problem is to lower the dimension of the data in each channel by DBN. As one of the most classical methods in the area of dimension reduction, principle component analysis (PCA) [8] is combined with SVM as the second baseline. We denote this baseline as PSVM. In the implementation, data in different channels are independently processed by PCA, and then combined and fed into SVM for the learning task.

For the challenge of the noisy channel problem, in this paper, we proposed a novel stimulus-response based method on the trained DBN models to select the critical channels. In the current studies, Fisher Criterion [9] is widely used for this problem. Therefore, we set SVM + Fisher Criterion (FSVM) as the third baseline. In the implementation, the critical channels are selected by the Fisher Criterion, then data in the selected channels are combined and used in SVM for the learning task.

In the fourth baseline, both PCA and Fisher Criterion are combined with SVM for the recognition of affective states. We denote this baseline as PFSVM. In the implementation, data in each channel are firstly processed by PCA. We then apply Fisher Criterion on the processed data to select the u critical channels. The processed data in the u selected channels are combined and fed into SVM for the learning task.

Finally, we set DBN + Fisher + RBM (DFRBM) as the last baseline. After using DBN to lower the dimension of the data in each channel, we apply Fisher Criterion to select the top u critical channels. The deepest features of the data in the selected channels are then combined and fed into a supervised RBM for the task of affective state recognition.

C. Experiments and Discussions

In the experiments, in the data of each participant, we randomly pick 20 EEG segments as the training samples, and use the rest 20 EEG segments as the testing samples. For all the cases, we set the number of critical channels $u = 5$, and the feature number of each hidden layer to be 100.

The AUC scores of the experiment results are summarized in Table II. In the table, the results of the proposed method are listed under the "**SDA**" (**S**upervised **D**BN based **A**ffective State Recognition); the name of the data from each participant is listed under the "Data"; and the highest AUC score on each subset is marked in bold.

In the comparison between the proposed SDA model with SVM, the results of our method significantly outperform the results of SVM on the data of all of the 32 participants. The superior performance of our method over SVM verifies the benefits of handling the small sample problem and the noisy channel problem by the SDA model.

Among the five investigated baselines, the first thing we notice is that PSVM performs significantly worse than SVM in most of the cases. Please notice that the difference between PSVM and SVM is that PSVM processes the data set by PCA, which lowers the dimension of the data. Therefore, SVM has much severe small sample problem than PSVM on the recognizing of affective states. The reason behind the worse performance of PSVM is that, the extracted features by PCA can not well capture the characteristics of the information in the critical channels. According to the theory of the PCA technique, the features that dominate the data obtains higher percentage in the extracted features. For the task of affective state recognition, features of the most channels are usually randomly distributed and irrelevant to the learning task. By PCA, the randomly distributed irrelevant channels dominate the extracted features. As a result, PSVM performs badly in the experiments.

In the comparison between SVM and FSVM, we notice that FSVM achieves sightly better performance than SVM. By FSVM, the multi-channel EEG signals are processed with Fisher Criterion, which select critical channels according to their closeness to the existing labels. This sightly better performance of FSVM validates the effectiveness of critical channel selection in the recognition of affective states. Nevertheless, the AUC scores achieved by FSVM are still significantly lower than the results of the proposed SDA method for two reasons. On the one hand, since only 20 labels are available in the training process, the critical channels selected by Fisher Criterion are not reliable. On the other hand, after the channel selection process, there are still more than tens of thousands of features, which are way more than the number of labeled samples. The severe small sample problem limits the performance of FSVM.

PFSVM, which seeks to solve the small sample problem by PCA and the noisy channel problem by Fisher Criterion, performs the worse among all the investigated approaches. This bad performance is determined by the properties of the features in each channel. In the task of affective state recognition, the features in each channel are sampled data from electrodes over the time. Due to the difficulty in the segmentation of time series, even in the critical channels, many features may not be relevant to the affective states. As a result,

Data	SVM	PSVM	FSVM	PFSVM	DFRBM	SDA
$S01$	0.6768	0.6703	0.6264	0.6310	0.7198	**0.8438**
$S02$	0.6923	0.7381	0.7253	0.6374	0.7135	**0.7800**
$S03$	0.6800	0.5165	0.6465	0.6154	0.7240	**0.7500**
$S04$	0.6364	0.5354	0.7473	0.6900	0.6667	**0.7604**
$S05$	0.6042	0.6310	0.6000	0.5100	**0.7917**	0.7552
$S06$	0.7292	0.5938	0.6667	0.5152	0.6875	**0.7600**
$S07$	0.6566	0.6563	0.8056	0.5156	0.6373	**0.8611**
$S08$	0.5469	0.5521	0.6154	0.7024	0.6267	**0.7333**
$S09$	0.6164	0.5300	0.7500	0.5152	0.7083	**0.8700**
$S10$	0.6700	0.5253	0.6500	0.7473	0.5989	**0.7500**
$S11$	0.7083	0.6061	0.7262	0.6146	0.6264	**0.7374**
$S12$	0.5960	0.5657	0.6800	0.5600	0.7198	**0.7424**
$S13$	0.6429	0.6250	0.5952	0.5357	0.6406	**0.7292**
$S14$	0.6566	0.5313	0.6484	0.5714	0.5900	**0.7396**
$S15$	0.6374	0.7475	0.7143	0.5604	0.6813	**0.7828**
$S16$	0.6667	0.5556	0.6905	0.6000	0.7121	**0.7308**

Data	SVM	PSVM	FSVM	PFSVM	DFRBM	SDA
$S17$	0.7153	0.6250	0.6310	0.6300	0.5960	**0.7440**
$S18$	0.5833	0.5313	0.6800	0.5657	0.7292	**0.7448**
$S19$	0.5455	0.6566	0.6786	0.6042	0.6758	**0.8077**
$S20$	0.6190	0.5625	0.6263	0.6354	0.7253	**0.7400**
$S21$	0.6264	0.6250	0.6164	0.5938	0.6771	**0.6813**
$S22$	0.6566	0.6923	0.7273	0.6374	0.6300	**0.7323**
$S23$	0.6042	0.6593	0.6875	0.7071	0.6869	**0.7240**
$S24$	0.6000	0.6400	0.6484	0.5960	0.6350	**0.6850**
$S25$	0.6771	0.5313	0.6044	0.697	**0.7100**	0.6771
$S26$	0.6768	0.5521	0.6429	0.6354	0.6727	**0.7626**
$S27$	0.6566	0.6800	0.6465	0.5455	0.6254	**0.7828**
$S28$	0.6667	0.6061	0.6465	**0.7708**	0.6823	0.6970
$S29$	0.5700	0.6364	0.6566	0.5313	0.7679	**0.7912**
$S30$	0.7071	0.6263	0.5800	0.5253	0.7363	**0.7552**
$S31$	0.6813	**0.7969**	0.7738	0.6800	0.6616	0.7500
$S32$	0.6869	0.6667	0.7300	0.5758	0.6354	**0.8021**

TABLE II. **EXPERIMENT RESULTS ON THE *DEAP* DATASET**. IN THE TABLE, THE PARTICIPANTS ARE DENOTED AS $S01$ TO $S32$. THE RESULTS ARE PRESENTED IN AUC SCORES. ON THE AVERAGE, THE PROPOSED SDA MODEL OUTPERFORMS SVM, PSVM, FSVM, PFSVM AND DFRBM BY 16.9%, 23.1%, 12.8%, 24.4% AND 11.5%, RESPECTIVELY.

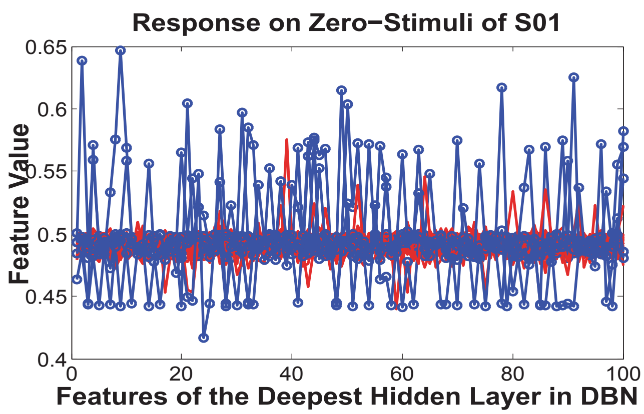

Fig. 3. **Response of the Zero-Stimulus on S01**. In the plot, blue and circled lines are the responses of the selected critical channels. Red lines are the responses from the other channels.

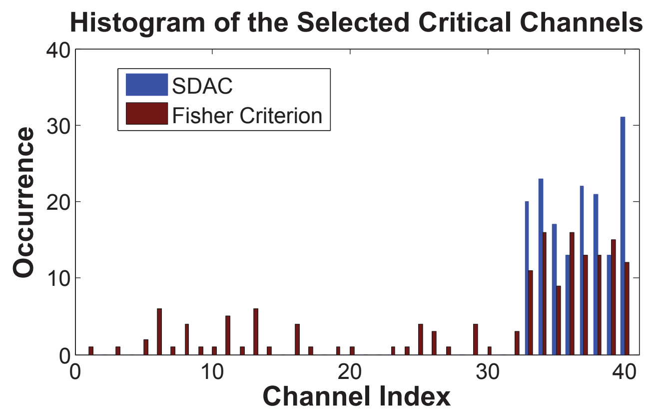

Fig. 4. **Histograms on the Selected Channels**. In the plot, blue bars are for the channels selected by the proposed channel selection method in SDA. Red bars are for the channels selected by Fisher Criterion.

the extracted features by PCA is significantly affected by these noise features in some cases. The critical channels selected by applying Fisher Criterion on these extracted features are thus not precise, which further lower the performance. Compared to PFSVM, the proposed SDA model performs significantly better. This fact validates the superiority of DBN in extract high level features from noise data.

Among the five baselines, DFRBM achieves the highest performance. There are two differences between DFRBM, which performs the best in the baselines, and PFSVM, which performs the worst in all the investigated methods. First, DBN is utilized to extracted high level features instead of PCA in PFSVM; and second, supervised RBM is implemented in DFRBM while PFSVM uses SVM. This comparison validates the effectiveness of RBM and DBN in the learning of affective states.

In the comparisons between the proposed method with the other 5 investigated approaches, our method performs significantly the best in 28 out of the 32 cases. In the other four cases, the AUC scores of the proposed SDA model are quite close to the best performance achieved by the baselines. Overall, on average AUC scores, the proposed SDA model significantly outperforms the investigated SVM, PSVM FSVM, PFSVM and DFRBM by 16.9%, 23.1%, 12.8%, 24.4% and 11.5%, respectively.

D. Analysis on Critical Channels

In this section, we discuss the effectiveness of the proposed critical channel selection method, and compare the method with the Fisher Criterion model.

In the proposed channel selection approach, on each channel, the response of the zero-stimulus on the trained DBN is calculated as demonstrated in Fig.3. In the plot, the responses of all the channels in $S01$ are included. Obviously, the blue and circled lines, which stand for the responses of the five selected critical channels, are significantly biased from 0.5 in many features. In the contrast, the red lines, which are the responses of the rest channels, are very close to 0.5 over all the features. This plot well fits the fact that in multi-channnel EEG signals, most channels are irrelevant to the affective states, and data in them are randomly distributed. Besides, as shown in Table II the good performance of these selected critical channels supports the effectiveness of the proposed channel selection approach.

To evaluate the stability of the proposed channel selection approach, we calculate the occurrences of the channels selected by the proposed method as well as by Fisher Criterion over $S01$ to $S32$. The results are shown in Fig.4. Obviously, the results (blue bars) of our method (denoted as SDAC) concentrate on the 33rd to the 40th channels. In the contrast, although majority of the channels selected by Fisher Criterion are also in the same range, there are also many other channels identified as critical channels by Fisher Criterion. Due to the fact that all of the data are for the same affective state recognition task, which is distinguishing whether the emotions are ”like” to ”dislike”, the critical channels should be the same in all the cases. Therefore, the performance of the proposed channel selection method is much more stable than the performance of Fisher Criterion.

To sum up, the proposed method can select meaningful critical channels for the task of affective state recognition, and achieve very stable performance across the data from different participants.

IV. Related Work

Although there are several existing methods on learning affective states from multi-channel EEG signals, the proposed method significantly differs from them in both the model and the focus. Here we summarize the difference between the existing methods and the proposed approach as follows.

Most of the existing models on affective state recognition from EEG signals are not designated for handling the small sample problem and the noisy channel problem. For instances, in [10], the authors present the application of fractal dimensions on the task of emotion classification; similarly, in [11], self organized map (SOP) is utilized in the same task. Both of these above methods ignore the impact of the limited training samples in the learning of affective states. Besides, in [10], there is no discussion on how to select the optimal channel set; and in [9], channels are selected in favor of maximizing the Fisher Criterion between the labeled samples and the optimal channel set. The major drawback of these methods is that, without successfully handling the small sample problem, the limited labeled samples make the channel selection criterion unreliable. As a result, the selected optimal channel set may include many noise channels and miss important ones. Different from these methods, the proposed approach doesn’t rely on the labeled instances in selecting the optimal channel set. In the proposed stimulus-response model, the critical channels in each affective state recognition task are selected according to the response rates in the DBN. Moreover, to handle the small sample problem, the proposed method utilizes the DBN to reduce the dimensionality of the data in each channel while preserving their characteristics.

We also notice that there are several existing papers that apply DBN on the learning of EEG signals. For instances, semi-supervised DBN is applied in [12], [13] for the task of anomaly detection. Specifically, in these papers, DBN is utilized as a reconstructor, and samples with high reconstruction errors are classified to be anomalous. Compared to them, the proposed model in this paper focuses on affective state recognition instead of anomaly detection. Besides, we use DBN in this paper for the purpose of reducing dimensionality w.r.t. the small sample problem and of selecting critical channels w.r.t. the noisy channel problem in the multi-channel EEG signals.

V. Conclusions

In this paper, we proposed a Deep Belief Network based model for affective state recognition from multi-channel EEG signals. To solve the small sample problem, we proposed to use Deep Belief Networks to extract deep and low dimensional features from the data of each channel while preserving the characteristics of the channels. To avoid the noise caused by the irrelevant channels, the critical channels are selected according to their response rates to the input data in a novel stimulus-response model. Moreover, to utilize the existing supervised information, the extracted deep features of the critical EEG channels are combined into the training of a supervised Restricted Boltzmann Machine. Experiments on a real world data set validated that the proposed method significantly outperforms five baselines by 11.5% to 24.4%.

VI. ACKNOWLEDGMENT

The materials published in this paper are partially supported by the National Science Foundation under Grants No. 1218393, No. 1016929, and No. 0101244.

References

[1] H. Prendinger, J. Mori, and M. Ishizuka, “Recognizing, modeling, and responding to users' affective states,” *Proceedings of the 10th international conference on User Modeling*, 2005.

[2] S. Koelstra, C. Muehl, M. Soleymani, J. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, “Deap: A database for emotion analysis using physiological signals,” *IEEE Transaction on Affective Computing*, 2012.

[3] R. the Dimensionallity of Data with Neural Networks, “Hinton, g.e. and salakhutdinov, r.r.” *Science*, 2006.

[4] G. Hinton, “Training products of experts by minimizing contrastive divergence,” *Neural Computation*, 2002.

[5] M. Carreira-Perpinan and G. Hinton, “On contrastive divergence learning,” *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, 2005.

[6] H. Larochelle and Y. Bengio, “Classification using discriminative restricted boltzmann machines,” *Proceedings of the 25th International Conference on Machine Learning (ICML)*, 2008.

[7] C. J. Burges, “A tutorial on support vector machines for pattern recognition,” *Data Mining and Knowledge Discovery*, 1998.

[8] K. Pearson, “On lines and planes of closest fit to systems of points in space,” *Philosophical Magazine*, 1901.

[9] T. N. Lal, M. Schroder, T. Hinterberger, J. Weston, and M. Bogdan, “Support vector channel selection in bci,” *IEEE Transactions on Biomedical Engineering*, 2004.

[10] Y. Liu, O. Sourina, and M. K. Nguyen, “Real-time eeg-based emotion recognition and its applications,” *Transactions on computational science XII*, 2011.

[11] R. Khosrowabadi, H. Quek, A. Wahab, and K. Ang, “Eeg-based emotion recognition using self-organizing map for boundary detection,” *Proceedings of the 20th International Conference on Pattern Recognition*, 2010.

[12] D. Wulsin, J. Blanco, R. Mani, and B. Litt, “Semi-supervised anomaly detection for eeg waveforms using deep belief nets,” *Proceeding of the Ninth International Conference on Machine Learning and Applications*, 2010.

[13] D. Wulsin, J. Gupta, R. Mani, J. Blanco, and B. Litt, “Modeling electroencephalography waveforms with semi-supervised deep belief nets: fast classification and anomaly measurement,” *Journal of Neural Engineering*, 2011.