Abstract—Current works on differentiating between posed and spontaneous facial expressions usually use features that are hand-crafted for expression category recognition. Till now, no features have been specifically designed for differentiating between posed and spontaneous facial expressions. Recently, deep learning models have been proven to be efficient for many challenging computer vision tasks, and therefore in this paper we propose using the deep Boltzmann machine to learn representations of facial images and to differentiate between posed and spontaneous facial expressions.

First, faces are located from images. Then, a two-layer deep Boltzmann machine is trained to distinguish posed and spontaneous expressions. Experimental results on two benchmark datasets, i.e. the SPOS and USTC-NVIE datasets, demonstrate that the deep Boltzmann machine performs well on posed and spontaneous expression differentiation tasks. Comparison results on both datasets show that our method has an advantage over the other methods.

Keywords—posed and spontaneous expressions; deep boltzmann machines; expression differentiation; feature learning

I. INTRODUCTION

Human-computer interaction, public security and many other areas will benefit from a better method that tells whether a facial expression is spontaneous or posed. For instance, a police can judge the suspect lies or not by recognizing his posed facial expressions. Robots can understand the users’ real emotions by differentiating between posed and spontaneous expressions. Hence, differentiating between posed and spontaneous facial expressions has attracted increasing attention. However, compared with a large amount of related work on expression category recognition, differentiating between posed and spontaneous facial expressions attracts less attention in spite of its wide application prospect.

Cohn and Schmidt [1] proposed using automatic feature tracking to measure the relation between amplitude and duration of smile onsets in spontaneous and posed smiles, and use a linear discriminant classifier for differentiation. Valstar [2] analyzed brow actions based on automatic detection of Action Units (AUs) to differentiate between posed and spontaneous facial behavior. Valstar [3] also differentiated between posed and spontaneous smiles by fusing head, face, and shoulder modalities, and show how to use these geometric features.

Dibeklioğlu et al. [4] proposed tracking 11 facial feature points to analyze the dynamics of eyelid, check and lip corner movement for differentiating between posed and spontaneous smile. Seckington [5] proposed using dynamic bayesian networks to model the temporal dynamics of expressions based on AUs and differentiate between posed and spontaneous smiles. Pfister et al. [6] proposed a new spatiotemporal local texture descriptor which helps in differentiating between posed and spontaneous expressions. Zhang et al. [7] differentiated between posed and spontaneous expressions using the SIFT appearance based features and FAP features. Wang et al. [8] proposed differentiating between posed and spontaneous expressions by modeling their spatial patterns. They also demonstrated that the the privileged information of gender and expression can help model the spatial patterns.

All the above researches use the hand-crafted features, such as AUs and CLBP-TOP, to differentiate between posed and spontaneous expressions. Most of the features used in the researches are designed for facial expression recognition. There are no hand-crafted features specifically designed for differentiating between posed and spontaneous facial expressions till now. Recently, the deep learning models, such as deep Boltzmann machines (DBMs) [9] and deep Belief networks (DBNs) [10], are proven to be efficient for some challenging computer vision tasks. Applications of deep learning in expression recognition has shown good results. He et al. [11] proposed using the DBM to recognize six categories of expressions from thermal infrared images. Susskind et al. [12] used deep belief network to generated expressions.

In this paper, we propose using a DBM to differentiate between posed and spontaneous expressions. First, we locate the face region from raw images and then normalize them. Next, a DBM which consists of two layers of restricted Boltzmann machines(RBMs), is proposed. The DBM is fine-tuned for posed and spontaneous expression differentiation after the pre-training phase. Experiments are conducted on the SPOS dataset and the NVIE dataset. Comparison experiments show the advantages of our proposed method.
II. THE PROPOSED METHOD

The framework of the proposed method is shown in Fig. 1, including extracting facial images and learning a multilayer generative model. The details are described as follows.

A. Extracting Facial Images

The facial expression datasets consist of many videos that record the changing process of expressions. Since the raw images are sampled from these videos, we can pick out onset frames and apex frames. The raw images contain much useless information except for the facial area, such as the background. We should extract the facial image from each raw image. A Harr-like features based cascade classifier proposed by Paul Viola and Micheal J. Jones[13] is used.

B. Learning a Multilayer Generative Model

We use a DBM that consists of two layers of RBMs in our experiment. The way to learn parameters of the DBM will be described in following sections.

1) Restricted Boltzmann machines: The basic components of a DBM model are RBMs. A RBM is an undirected graphic model, which consists of a set of hidden units $h$ and a set of visible units $v$, as shown in Fig. 2. Each of the visible units is independent of the other visible units. The hidden unit has the same property, which makes the posterior over hidden layer factorial. The visible layer corresponds to input data, and the hidden layer is seen as a stochastic, binary feature detector. These two layers connect to each other using symmetrically weighted connections. In this paper, we use the Gaussian-Bernoulli RBM (GRBM) [14] [15] for the real-valued input. Given the values of visible and hidden units, the energy function of the GRBM is given by:

$$E(v,h) = -\sum_i \sum_j \frac{w_{ij}v_i h_j}{N_v \sigma^2} - \sum_j b_j h_j + \sum_i \frac{(v_i-a_i)^2}{2\sigma^2/N_v}, \quad (1)$$

where $w, a$ and $b$ are parameters of the RBM model, and $\sigma$ is the standard deviation. $N_v$ is the times that we duplicate the visible layer. Given Eqn. 1, the distribution over the visible units $v$ can be calculated by:

$$p(v) = \frac{1}{Z} \sum_h e^{-E(v,h)}, \quad (2)$$

where $Z$ is the partition function given by:

$$Z = \sum_{h,v} e^{-E(v,h)}, \quad (3)$$

Calculating the partition function needs the sum over all the possible pairs of visible and hidden units, which is intractable.

A fast learning algorithm for a RBM was proposed by Hinton [10]. It is called contrastive divergence learning procedure. The algorithm starts with setting visible layer to an input vector. Then the states of hidden units can be calculated by:

$$p(h_j = 1 | v) = \sigma(b_j + \sum_i v_i w_{ij}) \quad (4)$$

The reconstructed visible units are sample from a Gaussian with unit variance and mean $a_i + \sum_j h_j w_{ij}$,

$$p(v_i = 1 | h) \sim N(a_i + \sum_j h_j w_{ij}, 1) \quad (5)$$

Finally, we can calculate the gradient in a weight by:

$$\Delta w_{ij} = \varepsilon(\langle v_i h_j \rangle_{data} - \langle v_i h_j \rangle_{recon}) \quad (6)$$

where $\varepsilon$ is a learning rate, and the angle brackets denote expectations under the distribution specified by the subscript. In contrastive divergence, $\langle v_i h_j \rangle_{data}$ is replaced with $\langle v_i h_j \rangle_{recon}$ because getting an unbiased sample of $\langle v_i h_j \rangle_{model}$ is intractable.

2) Deep Boltzmann machines:

a) Layer-wise pre-training: The DBM we used consists of a visible layer and two hidden layers. As shown in the lower part of Fig. 2, for the first layer, we double the input for the visible layer in order to eliminate the double-counting problem when top-down and bottom-up influences are combined. Thus, the first RBM is a GRBM model with the same two input. The training method of the GRBM is given in the previous section with setting $N_v = 2$.

The second layer is shown in the upper part of Fig. 2. The output of the pre-trained GRBM is considered as the input of the second layer. However, since the second layer is a binary RBM, we should preprocess the output of the first layer before using it.
The output of first layer is compared with a random-valued vector that is sampled from a Gaussian distribution, the function is defined as follows:

\[ h_i = \begin{cases} 1, & p(h_i = 1|v) > n_i \\ 0, & \text{otherwise} \end{cases} \]

\[ n_i \sim N(0, 1). \]  

In order to eliminate the double-counting problem, we double the hidden layer \( h^2 \) and label layer \( L \). Each of the two hidden layers \( h^2 \) is connected to a label layer so as to take posed and spontaneous labels into consideration.

Given the structure of the second layer, the energy function is given by:

\[
E(h^1, h^2, L) = -\sum_i b_i h_i^1 - \sum_i c_i h_i^2 - \sum_i \sum_j h_i^1 h_j^2 w_{ij}^2 - \sum_i \sum_j h_i^2 L_j w_{ij}^{lab} - \sum_i d_i L_i
\]

where \( w^2 \) specifies the weights between the hidden layers \( h^1 \) and \( h^2 \). \( w^{lab} \) specifies the weights between the label layer \( L \) and the hidden layer \( h^2 \). \( b, c \) and \( d \) are the biases.

According to the energy function, we can get following equations:

\[
p(h_i^2 = 1|h^1, L) = \sigma(\sum_i w_{ij}^2 h_i^1 + \sum_k w_{ik}^{lab} L_k + c_j),
\]

\[
p(h_i^1 = 1|h^2) = \sigma(\sum_j 2w_{ij}^2 h_j^2 + b_i),
\]

\[
p(L_k = 1|h^2) = \frac{\sum_j w_{ik}^{lab} h_j^2 + d_k}{\sum_k \sum_j w_{ik}^{lab} h_j^2 + d_k},
\]

Equations (8), (9) and (10) will be used in the contrastive divergence to train the second RBM.

b) Fine-tuning for differentiation: After the pre-training phase, the DBM can be used to initialize a multilayer neural network which is shown in Fig. 4. \( q(h_2|v) \), an additional input for the deep neural network, can be obtained by using the mean-field inference. Then the standard backpropagation is used to fine-tune the deep neural network discriminatively for differentiating task [9].

III. EXPERIMENTS AND ANALYSES

A. Experimental Conditions

1) Datasets: Two benchmark datasets, i.e. the NVIE dataset [16] and the SPOS dataset [17] are adopted in this paper. The SPOS dataset consists of visible and near infrared expression datasets which are captured from 7 subjects (4 males and 3 females). The dataset contains onset frames and apex frames. Difference images are generated by taking apex frames minus corresponding onset frames, as shown in Fig. 5. Since difference images contain the information of the expression changing process to some extent, they are used as the input of the DBM. Finally 234 difference images are obtained from the SPOS dataset. The distribution of samples is shown in Table 1.

The USTC-NVIE dataset consists of natural visible and thermal infrared expression datasets, and both the datasets contain six expression categories captured from more than 100 subjects. The onset and apex frames are also provided. We also use difference images as the input data. We select 1028 difference images at last, which come from 80 subjects (55 males and 25 females). There are 514 spontaneous expression images and 514 posed expression images. The distribution of these samples is shown in Table 2.

From Fig. 5, we can find that the difference image contains the information about the difference between the apex frame and the onset frame, and the pixel value shows the magnitude of the variation. Another advantage of the difference image is that many factors which are against differentiation or have no effect on differentiation, such as hair, glasses, background and so on, can be eliminated through the subtraction.

\[
\text{TABLE I} \quad \text{DATA DISTRIBUTION ON SPOS DATASET}
\]

<table>
<thead>
<tr>
<th>Expression</th>
<th>Happiness</th>
<th>Disgust</th>
<th>Fear</th>
<th>Surprise</th>
<th>Anger</th>
<th>Sadness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Posed</td>
<td>14</td>
<td>14</td>
<td>14</td>
<td>14</td>
<td>14</td>
<td>14</td>
</tr>
<tr>
<td>Spontaneous</td>
<td>66</td>
<td>23</td>
<td>32</td>
<td>11</td>
<td>13</td>
<td>5</td>
</tr>
</tbody>
</table>
TABLE II
DATA DISTRIBUTION ON USTC-NVIE DATASET

<table>
<thead>
<tr>
<th>Expression</th>
<th>Happiness</th>
<th>Disgust</th>
<th>Fear</th>
<th>Surprise</th>
<th>Anger</th>
<th>Sadness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Posed</td>
<td>104</td>
<td>93</td>
<td>68</td>
<td>78</td>
<td>91</td>
<td>80</td>
</tr>
<tr>
<td>Spontaneous</td>
<td>104</td>
<td>93</td>
<td>68</td>
<td>78</td>
<td>91</td>
<td>80</td>
</tr>
</tbody>
</table>

Fig. 5. (a) the onset frame; (b) the apex frame; (c) the difference image between the apex frame and the onset frame.

2) Data preprocessing and experimental setup: After obtaining samples from the datasets, we should preprocess these samples before using them. The samples are resized to $28 \times 28$ and normalized to zero mean and unit variance.

The architecture of the DBM is shown Fig. 4. The first layer is a GRBM which consists of $28 \times 28$ visible units and 500 hidden units. The output of the first layer is considered as the input of the second layer, and the number of the hidden units $h^2$ is set to 200. The number of the hidden units of layer $h^1$ and layer $h^2$ can make a good tradeoff between the training time and the classification capacity of the DBM in our experiment.

In order to conduct subject-independent experiments, for the SPOS dataset, the leave-one-subject-out cross validation is adopted. The subject number in the NVIE dataset is much larger than that in the SPOS dataset, therefore we divide subjects into several groups and each group contains 10 subjects, then we use leave-one-group-out cross validation on the NVIE database. Otherwise, the experimental time on the NVIE dataset will increase dramatically.

B. Experimental Results and Analyses

1) Analysis of the DBM model: In order to have a better understanding of what the algorithm do in training phases, we visualize the weights of the first RBM and analyze them. We sum up the weights between the visible unit $v_i$ and all the hidden units $h_j$’s for each visible unit. Then we can get a vector $V$ with size of $784 \times 1$ which is given by:

\[
V(i) = \sum_j w_{i,j},
\]

where $w_{i,j}$ is the weight between the visible unit $v_i$ and the hidden units $h_j$. After getting $V_i$ for all i’s, we reshape vector $V$ to $28 \times 28$. Fig. 6(a) shows the initial weights of the RBM. The weights are randomly generated from a standard Gaussian multiplied by $10^{-100}$. Fig. 6(b) shows the weights after pre-training. The pre-training phase gives the weights good enough starting points for later processes. From Fig. 6(b) we can find that the weight variation is roughly eudipleural if we consider the weights as a face image. Fig. 6(c) shows the fine-tuned weights on the basis of the pre-trained weights. Comparing Fig. 6(b) with Fig. 6(c), we can find that the weights of the two phases are roughly the same. It is because the fine-tuning phase just makes minor adjustments to the weights for differentiation.

In order to know the relationship between the weights and the facial area, we divide the face image into 15 subareas. The way of division is shown in Fig. 7. For each subarea we take the weights that are corresponding to the specified subarea minus weights of the former phase. Then these subareas are reshaped to be row vectors and shown in Fig. 8 and Fig. 9. From Fig. 8, we can see the variation between the pre-trained weights and the initial weights. The weights change a lot between these two phases. The weight variation corresponding to brows, eyes, mouth and chin are mostly positive, which means that the weights corresponding to these subareas are increasing. Since weights represent the importance of each
subarea to differentiation, we can infer that those subareas contain more information which is helpful to differentiate between posed and spontaneous facial expressions. Fig. 9 shows the weight variation between the fine-tuning phase and the pre-training phase. The variation is hundreds times smaller than that between the pre-trained weights and the initial weights, which will further confirm that the pre-training phase contributes more to the weight tuning and that the fine-tuning phase just makes minor adjustments to the weights for differentiation.

2) Differentiation results: After training the DBM, we can then evaluate the performance on the SPOS dataset and the USTC-NVIE dataset. The differentiation results on the USTC-NVIE and SPOS datasets are shown in Table 3 and Table 4. Comparing the differentiation results on the USTC-NVIE dataset with those on SPOS dataset, we can find that the accuracy and F1-score are higher on the USTC-NVIE dataset. It further confirms that the DBM model can perform better on larger data. From the two tables, we will find that the results are biased toward the spontaneous expressions.

3) Comparison with related works: Three related works, i.e. Zhang et al.’s [18], Pfister et al.’s [17] and Wang et al.’s [8] are compared with our proposed method to demonstrate the effectiveness of the features that are learned by the DBM. The three works conducted experiments on one or both of the SPOS and USTC-NVIE datasets and take 6 basic expression categories into consideration. Zhang et al. differentiated between posed and spontaneous expressions using the SIFT appearance based features and FAP features. They conducted experiments on the USTC-NVIE dataset, and 3572 posed images and 1472 spontaneous images were selected as the samples. Since Zhang’s paper hasn’t stated the information about the images they removed, we don’t have the same samples as theirs. We just compare the experimental results with theirs. Pfister et al. proposed the spatiotemporal local texture descriptor and use it for posed and spontaneous expression differentiation, they conducted experiments on the SPOS dataset. Wang et al. [8] proposed differentiating between posed and spontaneous expressions by modeling their spatial patterns.

The comparison results are shown in Table 5 and Table 6. Concerning the SPOS and USTC-NVIE dataset, the accuracy of the proposed method is higher than that of all related works. It demonstrates that our proposed method is more stable than Wang et al.’s when conducting experiments on small datasets. Both Zhang et al.’s and Pfister et al.’s papers propose some hand-crafted features which can be used to differentiate between posed and spontaneous expressions. From their experimental results, we can find that their proposed features are very useful for the differentiation, while these features may not contain all of the information that can help differentiate between posed and spontaneous expressions. Our method propose using a deep learning model, since the model can learn features automatically. From the improvement of the differentiation results, we can infer that the learned features contain more information that helps differentiate between posed and spontaneous expressions than the hand-
crafted features proposed by Pfister et al. and Zhang et al.

IV. Conclusion

In this paper, we propose using a DBM for posed and spontaneous expression differentiation. First, face images are located from datasets and difference images are formed by taking apex frames minus corresponding onset frames. Second, a two-layer DBM model is proposed to learn features and differentiate between posed and spontaneous facial expressions.

To prove the effectiveness of the proposed method, experiments are conducted on the USTC-NVIE and SPOS datasets. The experimental results show the advantage of the proposed method over other methods. It indicates that the DBM model is suitable for posed and spontaneous expressions differentiation and has a strong feature learning capacity.

In the future, we shall further validate our proposed posed and spontaneous expression differentiation method on more datasets. In addition, we will investigate why the DBM model outperforms other methods for posed and spontaneous expression recognition.

ACKNOWLEDGMENT

This work has been supported by the National Natural Science Foundation of China(61175037, 61228304, 61473270), and Project from Anhui Science and Technology Agency(1508085SMF223).

REFERENCES


TABLE V

Comparison with related work on the NVIE dataset

<table>
<thead>
<tr>
<th></th>
<th>Accuracy</th>
<th>F1-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>L. Zhang et al [18]</td>
<td>79.43</td>
<td>0.7</td>
</tr>
<tr>
<td>Wang et al. [8]</td>
<td>91.63</td>
<td>0.91</td>
</tr>
<tr>
<td>Proposed</td>
<td>91.73</td>
<td>0.91</td>
</tr>
</tbody>
</table>

TABLE VI

Comparison with related work on the SPOS dataset

<table>
<thead>
<tr>
<th></th>
<th>Accuracy</th>
<th>F1-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>T. Pfister et al [17]</td>
<td>72.0</td>
<td>0.7</td>
</tr>
<tr>
<td>Wang et al. [8]</td>
<td>74.79</td>
<td>0.6740</td>
</tr>
<tr>
<td>Proposed</td>
<td>84.62</td>
<td>0.78</td>
</tr>
</tbody>
</table>