CONVOLUTIONAL NEURAL NETWORKS FOR POSED AND SPONTANEOUS EXPRESSION RECOGNITION

Chang Xu[1]*, Tao Qin*[2]*, Yalong Bai*[3]*, Gang Wang*[1] *and Tie-Yan Liu*[2]

[1]College of Computer and Control Engineering, Nankai University, Tianjin, China
[2]Microsoft Research, Beijing, China, [3]Harbin Institute of Technology, Harbin, China
{changxu, wgzwp}@nbjl.nankai.edu.cn, {taoqin, tie-yan.liu}@microsoft.com, ylbai@mtlab.hit.edu.cn

ABSTRACT

Differentiating posed expressions from spontaneous ones is a more challenging task than conventional facial expression recognition. There are many methods proposed to differentiate posed and spontaneous expressions based on pixel level information. However, these methods still have some limitations : (1) Most of the studies use the difference between onset (the early stages of an expression) and apex (the most intense stages of an expression) pixel-level raw images as inputs, which may not only contain noisy information, but also lose some useful information. (2) A lot of previous work uses hand-crafted features designed by rules, which suffers from inadequate capability of abstraction and representation. Considering that the high-level image representations usually have less noisy information, we propose a special layer named "comparison layer" for convolutional neural network (CNN) to measure the difference between onset and apex images of high-level representations (instead of pixel-level difference). We add the comparison layer to a group of CNNs, and combine the learned representations from those CNNs to form inputs of a classifier for differentiating posed and spontaneous expressions. Experiments on USTC-NVIE database (so far the largest database for this task) show that our method significantly outperforms the state-of-the-art methods (91.73% to 97.98%).

1. INTRODUCTION

Perceiving each other's mood from their facial expressions benefits both sides of the communication. But, sometimes people disguise their true feelings by pretending a posed expression, which is hard to recognize by ordinary people without professional training. An automatic recognition system for posed versus spontaneous expression has many potential applications, such as human-computer interaction, polygraph test and medical diagnose. For instance, police can use such a system to detect deceptive facial expressions for analyzing testimony. Psychological consultants can offer help to the people under pressure or in sadness but pretending to be fine.

Researchers have proposed various effective statistical and machine learning methods based on analyzing data of physiological indices such as blood pressure, pulse, respiration and skin conductivity from sensing equipment. Furthermore, a series of studies in psychology provide possibilities of discriminating posed and spontaneous expressions using visual information. For example, posed smiles often involve movement of the mouth while spontaneous smiles also include movement of muscles surrounding the eyes [1]. The asymmetry of zygomatic majour actions occurs more frequently in posed smiles than in spontaneous smiles [2]. Recently, pioneering studies focusing on discriminating posed and spontaneous expressions by a pure machine vision system have been attracting more and more attentions. Compared with the physiological indices based methods, a simple vision system does not require various expensive external sensing equipments except a video camera, which is easy to use and embed to other systems.

Many vision based methods are proposed to distinguish posed and spontaneous expressions. Among them, most work uses hand-crafted features designed by rules [7, 8, 9, 10], such as displacement of facial key points, which may suffer from inadequate capability of abstraction and representations. Furthermore, compared with the conventional facial expression recognition task, differentiating posed expressions from spontaneous ones depends on subtler information, and therefore is more sensitive to noise. To relieve the impact of noise such as appearance, background, lighting conditions, etc., it is straightforward to make use of the difference between onset and apex images. A common practice is to take the pixel difference of the onset and apex raw images as input features. Doing so, however, a lot of useful information may get lost and new noise would be introduced because the difference images are directly computed based on low-level pixels without further abstraction and processing.

Observing the outstanding performance of deep convolutional neural networks in many vision tasks owing to their abilities of learning high-level image representations from raw pixels, we propose a deep convolutional neural network based method to discriminate posed and spontaneous expressions. By using our proposed comparison layer after abstracting

978-1-5090-6067-2/17/$31.00 ©2017 IEEE

middle or high level features from raw images, the spatial changes from onset to apex images are modeled and the noisy information is eliminated in different abstraction levels. Our recognition performance is further improved by combining all levels of abstraction feature , which represent the spatial changes from onset to apex images.

The main contributions of this paper resides in three folds: (1) We successfully use automatically learned features from CNNs rather than traditional hand-crafted features such as the displacement of facial key points. (2) We propose to model the differences of onset and apex images in different abstraction levels by applying comparison layers, rather than simply using pixel level difference. (3) A series of experiments validate the effectiveness of our proposed method, and the recognition results on USCT-NVIE database (so far the largest database for this task) and SPOS database significantly exceed those of the state-of-the-art.

2. RELATED WORK

A series of studies in nonverbal behavior show that it is possible to discriminate posed and spontaneous expressions using visual information [1, 2, 12, 13]. Inspired by that, some pioneering studies in computer vision have investigated this problem. These efforts mainly focus on smile [3, 4, 5, 14], eyebrow action[15], and pain [6].

In the work of Dibeklioğlu, Salah and Gevers [5], the authors proposed to track facial points to analyze the dynamics of eyelid, check and lip corner movement for differentiating between posed and spontaneous smile. Valstar *et al.* [15] distinguished between posed and spontaneous brow actions using velocity, duration and order of occurrence. Littlewort, Bartlett and Lee [6] investigated fake pain and real pain discrimination by detecting facial actions using Gabor features and employing a SVM classifier.

All of above researches only focus on one specific expression to recognize posed and spontaneous expressions. Recently, several works tried to solve this problem on multiple basic expressions (happiness, disgust, fear, surprise, sadness and anger). In the work of Zhang *et al.* [7], SIFT and FAP features are used to investigate the performance of a machine vision system for discrimination between posed and spontaneous expressions of six basic emotions. In the work of Pfister *et al.* [8] a spatiotemporal local texture descriptor (CLBP-TOP) was proposed to differentiate spontaneous expressions from posed expressions in both visible and infrared images. Wang *et al.* [9] proposed a method to differentiate posed and spontaneous expressions by modeling their spatial patterns such as facial shape and Action Unit variations. In their following work[10], the displacements of facial feature points between apex and onset images are extracted as features and two RBM models are trained for classification. In the work of Gan *et al.* [11], the authors proposed to use pixel-wise difference between onset and apex images as input features of a two-layer deep Boltzmann machine for differentiating posed and spontaneous expressions.

These pioneering works have explored this field in various aspects, such as feature selection, model design and learning. The performance of posed and spontaneous recognition has been improved a lot. By analyzing existing methods, we found that hand-crafted features designed by rules are used in most of the above studies [7, 8, 9, 10, 16] instead of automatically learned features, which may result in inadequate capability of abstraction and representations. Besides, these methods usually compute pixel level difference of onset and apex images as input features. In this way, however, a lot of useful information may get lost and new noise would be introduced.

Recently, deep convolutional neural networks have demonstrated outstanding performance in a variety of vision tasks such as face recognition [17, 18] and object classification [19, 20]. In this paper, we present a DCNN based framework to automatically extract features and differentiate between posed and spontaneous expressions. We propose a comparison layer for CNN to represent the spatial changes from onset to apex images in different levels

3. MODEL

In this section, we describe our CNN based framework for posed and spontaneous recognition. Since the posed and spontaneous expressions differ in a subtle way, the spatial facial change from the onset image to the apex image is important for this task. A naive approach is to directly train a CNN based on the difference image (see Figure 1 for an example) between the onset image and the apex image. As aforementioned, such a difference image is usually noisy and may miss useful information. Our proposal is to first abstract and process the onset and apex images respectively to get high-level feature representations and then compute the difference information based on the abstracted feature representations using a comparison layer.

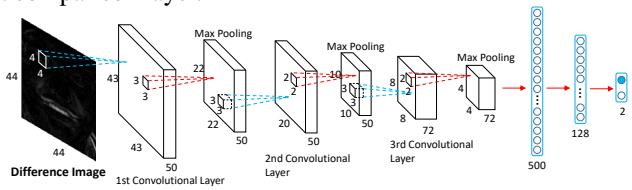

Fig. 1. The structure of DiffCNN, which directly uses difference images as input. The 3D convolution kernel sizes of the convolutional layers are shown as the small cuboids inside the feature maps. Neuron numbers of each fully connected layers are marked beside each layer.

3.1. CNN with Comparison Layer

Our key idea is to add a special layer named "comparison layer" to CNNs, which takes the abstracted feature representations from the onset and apex images as inputs and applies

a comparison operator to extract difference information between the two images. The comparison operator $f(\cdot)$ takes the two input feature maps X, Y from two images, and outputs feature map Z, where $X, Y, Z \in \mathbb{R}^{\mathrm{H}\times\mathrm{W}\times\mathrm{C}}$ with width W, height H and the number of channels C , satisfying $Z = f(X, Y)$. The comparison operator $f(\cdot)$ can take any form that can compute the differences between two inputs. In this paper, we explore two kinds of comparison operations:

- simple subtraction: $Z^r = X^r - Y^r$,
- linear combination: $Z^r = \sum_{k=1}^{C} \alpha_k^r \cdot X^k + \sum_{k=1}^{C} \beta_k^r \cdot Y^k$,

where r is used to index the channels of the feature map, α and $\beta \in \mathbb{R}^{\mathrm{C}\times\mathrm{C}}$ are parameters to be learnt. The simple subtraction can be considered as a special case of linear combination where $\alpha = I_C$ and $\beta = -I_C$. I_C is a $C \times C$ identity matrix, with ones on the main diagonal and zeros elsewhere. Simple subtraction explicitly computes the difference information in high-level feature maps, while linear combination can handle much more complicated relations beyond subtraction. After the comparison operation, a non-linear activation function, $max(0, \cdot)$ is applied element-wise on the output feature maps. Neurons with such nonlinearities are called rectified linear units [19].

Figure 2 shows the structure of one CNN with comparison layer used in this paper. It contains six layers with weights. The first three are convolutional layers and the remaining ones are fully connected layers. A pair of fixed-size 44 $\times$ 44 gray onset and apex images is taken as inputs of two-path convolutional layers for extracting mid-level or high-level features. A comparison layer lies after the first convolutional layer with max-pooling of the two paths. It is applied to computing the difference features followed by another convolutional layer for further abstraction. Finally, the features generated from convolutional layers pass two fully connected layers to the top layer neurons to predict the high-level concept of whether the expression is posed or spontaneous. Moreover, to prevent overfitting, we use dropout in all max-pooling layers and first two fully connected layers with rates of 0.05, 0.15, 0.25, 0.5, 0.25 respectively.

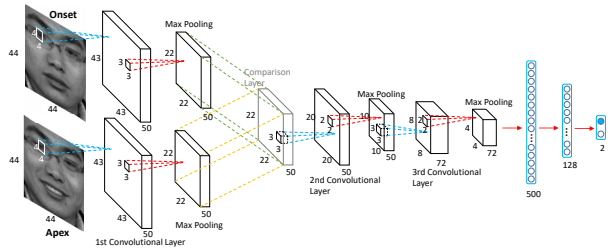

Fig. 2. The structure of one CNN which applies comparison layer at the output of first convolutional layer with max-pooling.

Moreover, weights of neurons (including convolution kernels and biases) in the same level of the two paths of convolutional layers are shared. Sharing the same set of weights between the convolutional layers of two paths allows the network to learn facial features from onset and apex using the same filters.

3.2. Combination of Multiple CNNs

As can be seen from Figure 2, the comparison layer transforms two set of input feature maps (from two images) to a set of output feature maps for further processing. A natural question is where to put the comparison layer. It is possible that putting it somewhere else could be better than after the first convolutional layer. Different placement of the comparison layer corresponds to different abstraction level of visual difference representations between onset and apex, which could be complementary to each other. So we try to leverage all of these different abstraction level of visual representations by combining the features extracted from all of the CNNs.

As shown in Figure 3, each CNN takes a pair of onset and apex images as input and places the comparison layer in different position. After training these CNNs, we extract features from them (e.g. the outputs of the last hidden layer) and train a final classifier through supervised learning.

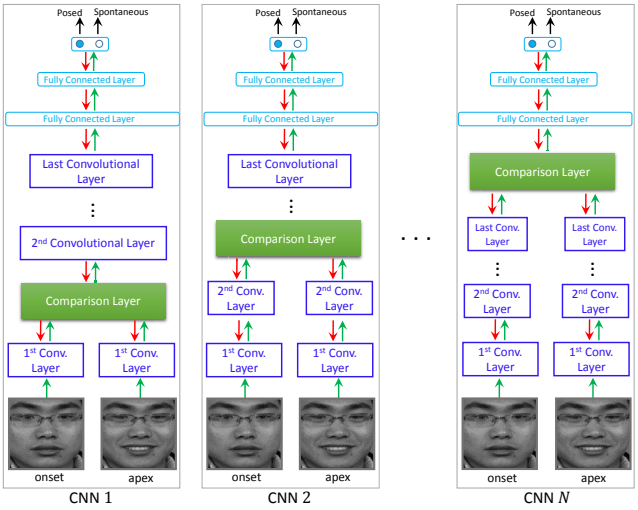

Fig. 3. Our framework for posed and spontaneous recognition contains a group of deep CNNs. A pair of onset and apex images is taken as input for each CNN. The second layer of each CNN is a two-way softmax to classify whether the input image pair is posed or spontaneous. Comparison layer is applied to the outputs of convolutional layers for different CNNs in the group. Finally, the outputs of the second fully connected layer of each CNN are extracted as features to be fed to a linear SVM for train a final classifier.

4. EXPERIMENTS

We conducted a set of experiments to test the effectiveness of our proposed framework, and compared it with several state-of-the-art methods.

4.1. Experimental Setup

We chose the USTC-NVIE database [21] and SPOS database [8] to evaluate our method. USTC-NVIE database is the largest database so far for this task and therefore is most representative and reliable to test machine learning algorithms. The USTC-NVIE database contains more than 100 subjects while SPOS dataset only contains 7 subjects. The onset and apex of an expression in USTC-NVIE database are manually labeled in the visible facial videos. More details about these two datasets can be found in appendix.

4.2. Experimental Results

We tested a set of CNN based methods.

(1) As a straightforward application of CNN for this task, one can feed the difference image between the onset and apex images to the CNN as input. We implemented a CNN with three convolutional layers and two fully connected layers, as shown in Figure 1. We call this baseline DiffCNN.

(2) For our proposed framework, we tested three ways of adding the comparison layer into the CNN, i.e., placing the comparison layer after the max-pooling of each of the three convolutional layers. The CNN with the comparison layer after the first convolutional layer is shown in Figure 2. Both simple subtraction and linear combination operators for the comparison layer are implemented. In total we got six modified CNNs, and we call them CompCNN-Sub@i and CompCNN-Comb@i for short, where $i = 1, 2, 3$.

(3) To make full use of the learned difference information from middle level to high level between onset and apex images, after training the three networks CompCNN-Sub@i, we extracted their outputs of the second fully connected layers and further trained a linear SVM based on those outputs for the task of posed and spontaneous expression recognition. We call this model CompCNN-Sub-SVM. Similarly, we got a model CompCNN-Comb-SVM using the three networks CompCNN-Comb@i.

Since a lot of previous work [9, 10, 11] conducted experiments on a subset of the USTC-NVIE database, including 1028 expression samples from 80 subjects, for a fair comparison, we first evaluate our methods on this subset. Following the common practice [7, 8, 9, 10, 11], we conducted subject-independent experiments on the subset of USTC-NVIE with 1028 samples. Subjects were divided into 10 groups and our results were obtained through 10-fold cross validation on the subjects. The experimental results are shown in Table 1 .

We can find that all of the models with the comparison layer outperform the model based on difference images (DiffCNN) except CompCNN-Sub@3, which indicates that conducting comparison operations on middle or high level features is better than on low level features (i.e., the raw images). The comparison layer in our models measures the difference between onset and apex after processing and abstracting the pixel-level inputs, which can remove the noise and retain the key information. Since the comparison layer can be applied after the convolutional layer at different depths in the network, the spatial changes from onset to apex images can be modeled and the noisy information can be eliminated in different semantic levels.

Table 1. Error rate of different methods on the USTC-NVIE database and Full USTC-NVIE dataset.

Method	Subset	Full set
Zhang *et al.* 2011 [7]	20.57	-
Wu and Wang 2016 [10]	18.77	-
Wang, Wu and Ji 2016 [16]	10.99	-
Wang *et al.* 2015 [9]	8.37	-
Gan *et al.* 2015 [11]	8.27	20.30
DiffCNN	3.70	8.47
CompCNN-Sub@1	2.82	7.47
CompCNN-Sub@2	3.11	7.61
CompCNN-Sub@3	3.79	7.35
CompCNN-Sub-SVM	2.43	6.52
CompCNN-Comb@1	2.33	6.40
CompCNN-Comb@2	2.43	7.21
CompCNN-Comb@3	2.33	6.83
CompCNN-Comb-SVM	**2.04**	**5.95**

From the table, we also observe that using linear combination in the comparison layer can achieve better performance than using simple subtraction. Simple subtraction is a special case of linear combination. Linear combination is more expressive and can learn much more complicate relations from data than simple subtraction. Furthermore, we see that CompCNN-Sub-SVM and CompCNN-Comb-SVM achieve lower error rate than corresponding single CNN models. That is, the combination of difference information between onset and apex in different abstraction levels can achieve better performance than abstraction in a single level.

We then compared our proposed models with several state-of-the-art methods. In the work of Zhang *et al.* [7], they selected 3572 posed and 1472 spontaneous apex images and then used SIFT appearance based features and FAP features to train their model. Since we do not know which images they selected, we simply cite the numbers reported in [7] as a reference. The other four works use the same data and settings as us. As can be seen from the table, our model improves recognition accuracy of the state-of-the-art method by 6.23%, reaching an accuracy of nearly 98%. We conducted a further study to better understand our methods by visualizing the feature maps to check the internal states of CNNs with comparison layers in Appendix B.

Since the manually selected subset of the USTC-NVIE database may be biased, we collected all the onset and apex pairs of posed and spontaneous expressions from all subjects and discarded the sample if an onset or apex image of a pair is missing. Finally we got 4203 samples from 148 subjects in total. We evaluated our method on this full database. As no

previous work reports their performance on the full USTC-NVIE database, we tested the performance of the state-of-the-art method[11] on this database using source code provided by the author. From Table 1 we can see that our models achieve the best performance again.

We then conducted experiments on SPOS database, which is relatively smaller than USTC-NVIE datbase. SPOS database includes 84 posed expression samples and 150 spontaneous expression samples from only 7 subjects. In order to compare with other related works, leave-one-subject-out cross validation is used.

Table 2. Results on SPOS database. [1]

Method	Error Rate (%)
Wu and Wang 2016 [10]	25.64
Wang *et al.* 2015 [9]	25.21
Wang, Wu and Ji 2016 [16]	23.93
Pfister *et al.* 2011 [8]	21.80
Gan *et al.* 2015 [11]	18.38
DiffCNN	21.37
CompCNN-Sub@1	18.38
CompCNN-Sub@2	17.95
CompCNN-Sub@3	20.51
CompCNN-Sub-SVM	**16.66**
CompCNN-Comb@1	19.49
CompCNN-Comb@2	19.91
CompCNN-Comb@3	19.91
CompCNN-Comb-SVM	19.06

Experimental results are shown in Table 2. CompCNN-Sub-SVM achieves the best performance. The performance of CompCNN-Comb-SVM is not as good as CompCNN-Sub-SVM on SPOS dataset. Our explaination is that it is more likely to overfit on this small dataset because of its larger model size. Impact of the size of training data is studied in next section.

4.3. Impact of the Size of Training Data

In this section, we analyze how the size of training data influence the performance of our models. The recognition accuracy of our methods with respect to different number of training subjects on the full USTC-NVIE dataset is shown in Figure 4. Note that in this study, we fixed the test set and only changed the size of training data. The performances of DiffCNN, CompCNN-Sub-SVM and CompCNN-Comb-SVM are compared.

From the figure we have several observations. (1) When the number of training subjects is small, e.g., less than 40,

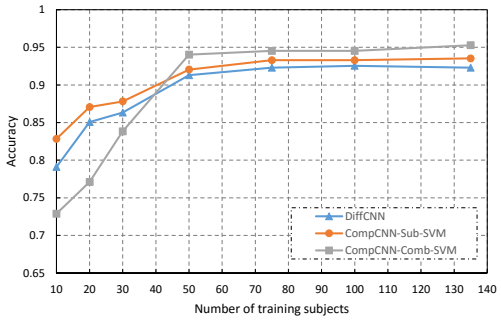

Fig. 4. Accuracy w.r.t sizes of training data.

the performances are heavily affected by model complexity, e.g., number of parameters of a model. CompCNN-Sub-SVM achieves the best performance for small training set because of its smallest model size which is not likely to be overfitted compared with CompCNN-Comb-SVM which has more parameters.(2) All the three models achieve steady performance when the number of training subjects is more than 50. (3) When the number of training subjects is large enough, models using comparison layers outperform the model using difference images since comparison layers can distinguish useful information from noisy information through comparison in abstraction levels. Linear combination based model performs better than simple subtraction based model because it is more expressive (due to more parameters) and can handle more complicate relations than subtraction.

5. CONCLUSIONS AND FUTURE WORK

In this work, we have applied deep CNNs for posed and spontaneous expression recognition, and proposed to add a new layer, the comparison layer, to CNNs, which can effectively extract difference information between onset and apex images by comparing their abstracted feature representations in middle and high levels.

There are several directions to explore in the future. First, we have invested simple subtraction and linear combination for the comparison operators. Nonlinear operators may further improve the recognition accuracy. Second, in this work, we have focused on the approach of using two images (onset and apex) for the task. While this is a popular approach, an alternative is to use the whole image sequence. We will study how to apply the comparison idea to image sequences for the task.

6. AKNOWLEDGEMENT

This work is partially supported by NSF of China (grant numbers: 61373018, 61602266 11550110491). Gang Wang is the corresponding author of this paper.

[1] Error rate reported in [11] was 15.38%, which was got by training several models using different random seeds and selecting the best model on test set. Our model can got 13.25% in this way. However, we think doing so cannot represent the real test accuracy. Instead, we report averaged performance over those models with different random seeds.

7. REFERENCES

[1] Paul Ekman and Maureen O'Sullivan, "Who can catch a liar?," *American psychologist*, vol. 46, no. 9, pp. 913, 1991.

[2] Paul Ekman, Joseph C Hager, and Wallace V Friesen, "The symmetry of emotional and deliberate facial actions," *Psychophysiology*, vol. 18, no. 2, pp. 101–106, 1981.

[3] Jeffrey F Cohn and Karen L Schmidt, "The timing of facial motion in posed and spontaneous smiles," *International Journal of Wavelets, Multiresolution and Information Processing*, vol. 2, no. 02, pp. 121–132, 2004.

[4] Michel F Valstar, Hatice Gunes, and Maja Pantic, "How to distinguish posed from spontaneous smiles using geometric features," in *Proceedings of the 9th international conference on Multimodal interfaces*. ACM, 2007, pp. 38–45.

[5] Hamdi Dibeklioğlu, Albert Ali Salah, and Theo Gevers, "Are you really smiling at me? spontaneous versus posed enjoyment smiles," in *Computer Vision–ECCV 2012*, pp. 525–538. Springer, 2012.

[6] Gwen C Littlewort, Marian Stewart Bartlett, and Kang Lee, "Automatic coding of facial expressions displayed during posed and genuine pain," *Image and Vision Computing*, vol. 27, no. 12, pp. 1797–1803, 2009.

[7] Ligang Zhang, Dian Tjondronegoro, and Vinod Chandran, "Geometry vs. appearance for discriminating between posed and spontaneous emotions," in *Neural Information Processing*. Springer, 2011, pp. 431–440.

[8] Tomas Pfister, Xiaobai Li, Guoying Zhao, and Matti Pietikäinen, "Differentiating spontaneous from posed facial expressions within a generic facial expression recognition framework," in *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*. IEEE, 2011, pp. 868–875.

[9] Shangfei Wang, Chongliang Wu, Menghua He, Jun Wang, and Qiang Ji, "Posed and spontaneous expression recognition through modeling their spatial patterns," *Machine Vision and Applications*, vol. 26, no. 2-3, pp. 219–231, 2015.

[10] Chongliang Wu and Shangfei Wang, "Posed and spontaneous expression recognition through restricted boltzmann machine," in *MultiMedia Modeling*. Springer, 2016, pp. 127–137.

[11] Quan Gan, Chongliang Wu, Shangfei Wang, and Qiang Ji, "Posed and spontaneous facial expression differentiation using deep boltzmann machines," in *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on*. IEEE, 2015, pp. 643–648.

[12] Paul Ekman and Erika L Rosenberg, *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*, Oxford University Press, 1997.

[13] Paul Ekman, "Darwin, deception, and facial expression," *Annals of the New York Academy of Sciences*, vol. 1000, no. 1, pp. 205–221, 2003.

[14] Hamdi Dibeklioğlu, Albert Ali Salah, and Theo Gevers, "Recognition of genuine smiles," *IEEE Transactions on Multimedia*, vol. 17, no. 3, pp. 279–294, 2015.

[15] Michel F Valstar, Maja Pantic, Zara Ambadar, and Jeffrey F Cohn, "Spontaneous vs. posed facial behavior: automatic analysis of brow actions," in *Proceedings of the 8th international conference on Multimodal interfaces*. ACM, 2006, pp. 162–170.

[16] Shangfei Wang, Chongliang Wu, and Qiang Ji, "Capturing global spatial patterns for distinguishing posed and spontaneous expressions," *Computer Vision and Image Understanding*, vol. 147, pp. 69–76, 2016.

[17] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lars Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. IEEE, 2014, pp. 1701–1708.

[18] Yi Sun, Yuheng Chen, Xiaogang Wang, and Xiaoou Tang, "Deep learning face representation by joint identification-verification," in *Advances in Neural Information Processing Systems*, 2014, pp. 1988–1996.

[19] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," in *Computer Vision–ECCV 2014*, pp. 346–361. Springer, 2014.

[21] Shangfei Wang, Zhilei Liu, Siliang Lv, Yanpeng Lv, Guobing Wu, Peng Peng, Fei Chen, and Xufa Wang, "A natural visible and infrared facial expression database for expression recognition and emotion inference," *Multimedia, IEEE Transactions on*, vol. 12, no. 7, pp. 682–691, 2010.