Multi-modality Network with Visual and Geometrical Information for Micro Emotion Recognition

Jianzhu Guo$^{+1,2}$, Shuai Zhou^{+3}, Jinlin Wu^{1,2}, Jun Wan^{*1,3}, Xiangyu Zhu^{1}, Zhen Lei^{1},
Stan Z. Li^{1,3,4}
[1] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences
[2] University of Chinese Academy of Sciences
[3] Macau University of Science and Technology, Macau
[4] Authenmetric Inc.

Abstract— **Micro emotion recognition is a very challenging problem because of the subtle appearance variants among different facial expression classes. To deal with the mentioned problem, we proposed a multi-modality convolutional neural networks (CNNs) based on visual and geometrical information in this paper. The visual face image and structured geometry are embedded into a unified network and the recognition accuracy can be benefic from the fused information. The proposed network includes two branches. The first branch is used to extract visual feature from color face images, and another branch is used to extract the geometry feature from 68 facial landmarks. Then, both visual and geometry features are concatenated into a long vector. Finally, the concatenated vector is fed to the hinge loss layer. Compared with the CNN architecture only used face images, our method is more effective and has got better performance. In the final testing phase of Micro Emotion Challenge[1], our method has got the first place with the misclassification of 80.212137.**

I. INTRODUCTION

A facial expression is more motions or positions of the muscles beneath the face. These motions can express the emotional or psychological state of an individual. Humans are skilled at adopt various kinds of expressions voluntarily or involuntarily. The observers can also well judge one's emotion and mental state by watching his face, not listening to words. Researches about facial expression can trace back to nineteen century, Darwin claimed the universality of emotions [2]. In 1971, Paul Ekman present six basic expression: anger, disgust, fear, happiness, sadness, and surprise [3].

The field of facial emotion recognition has been researched for several years, and many novel methods were proposed. Among them the CNN based method has reach the state-of-art performance recently [6], [12], [16], [17]. But the current camera equipment has high resolution commonly, which gives a challenge that whether the proposed model can distinguish more details about emotion. In human-computer interaction, it will be a huge improvement if the computer is capable of recognizing more subtle expressions of the human interacting with it.

Therefore, Iiris et al. [11] has developed a new huge facial expression database called iCV Multi-Emotion Facial Expression Dataset (iCV-MEFED), designed for micro emotion recognition with 50 classes. In this paper, we also focus on micro emotion recognition and the experimental results are also provided on the iCV-MEFED dataset.

The rest of the paper is organized as follows. Related works are reviewed in Section II. The proposed method is presented in Section III. Then, experiments are provided in Section IV to evaluate our method. Section V gives some discussions about the proposed method. Finally, a conclusion is drawn in Section VI.

II. RELATED WORKS

Automatic facial expression recognition has been an active field in academic community for many years. In early years, the size of dataset was small and computation power was very low, which is a huge obstacle of the advancement of this field. For these limitations, the early works mainly focused on geometrical representations and hand-crafted features extracted from the face, which were fed to train classifiers to be capable of distinguish different facial expression.

In recent years, with the great computation ability, deep learning methods have been widely applied in computer vision tasks, natural language processing, and have made great success in many specific domain of these fields. In computer vision field, convolutional neural network architecture is the most common fundamental architecture. In 2012, Alex Krizhevsky et al. [8] won the ILSVRC (ImageNet Large-Scale Visual Recognition Challenge). In 2014, DeepID [14] convolutional neural network architecture were proposed by Tang et al. for face verification. In 2015, Kaiming He proposed deep residual networks [4] and won the 1st places in ILSVRC. For face detection, a convolutional neural network cascade [9] were proposed by Li et al.

Owing to deep learning's great performance in computer vision tasks, such deep architecture has been used to handle facial emotion recognition problem. Tang [15] reported a deep CNN jointly learned with a linear support vector machine (SVM) output. This method achieved the first place on both validation and testing subset on the FER 2013 Challenge [1]. Liu et al. [10] proposed a facial expression recognition framework with 3D CNN and deformable action parts constraints in order to jointly localizing facial action parts and learning part-based representations for expression

$^{+}$The authors are made an equal contribution; *The author is the corresponding author.

[1] http://icv.tuit.ut.ee/fc2017_facialexpr

recognition. Yu and Zhang [16] proposed a method containing a face detection module based on the ensemble of three state-of-the-art face detectors, followed by a classification module with the ensemble of multiple deep convolutional neural networks (CNN). They used two schemes to learn the ensemble weights of the network responses: by minimizing the log likelihood loss, and by optimizing the hinge loss.

Inspired by these works, we adapt CNN based methods trying to solve the micro emotion recognition problem. We also embed the geometrical information into the CNN, which have gained the best performance on the micro recognition recognition challenge.

III. The Proposed Method

For micro emotion recognition task, our method is based on visual and geometry information. In this section, we introduce how we extract the geometrical representation and visual feature.

A. Geometrical representation extraction

Many of early works on facial expression use the geometrical representation, including geometry distances and other geometry properties. Though they are not better than the CNN-based methods, geometrical representation still works and can be used in the CNNs. Therefore, we try to keep the geometrical information and visual face image in a unified CNNs. Face landmark is good geometrical representation, we first use only the face landmark represented as a vector, and feed it to a classifier like SVM (linear). The performance is not good as expected. For joint dominant and complementary emotion recognition, it just achieves about accuracy of 0.09 in iCV-MEFED dataset. While it shows the effectiveness of the landmark geometry information. The joint emotion recognition having 50 ground truth labels with not much data is really a hard task.

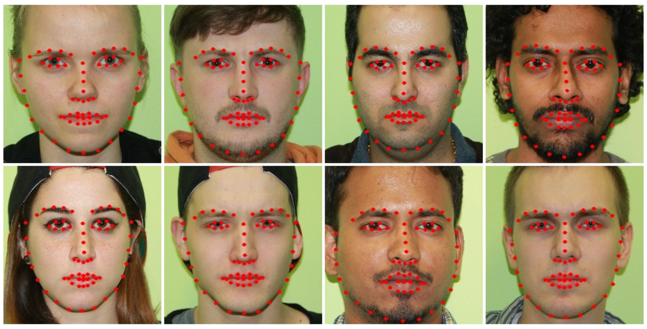

Fig. 1. Different faces (after cropped and aligned to 224 × 224 size) expressing angry emotion. The landmarks of the same emotion expression on different person ID's face are quite different, due to the difference of each person's face shape.

To reduce the landmarks' variance of the same emotion in different faces, we propose to subtract each face id's mean landmark of different emotion. In detail, we first calculate each face id's mean landmark by Eq. 1

$$lm^{(i)} = \frac{1}{N}\sum_{j=1}^{N} l_j^{(i)} \tag{1}$$

where lm means the average of the landmarks, i indicates face id number, N is each face id's number of samples, which is about 250 in iCV-MEFED dataset. l represents the flattened landmark, where (x, y) point is placed in one axis.

Then we extract the landmarks displacement by Eq. 2

$$lr^{(i)} = l^{(i)} - lm^{(i)} \tag{2}$$

where lr means landmark residual named landmark displacement.

We feed lr geometrical feature to SVM(linear) classifier, and it reaches about 0.15's accuracy on validation set, which is such a huge improvement compared to the landmark only. This indicates the geometrical information such as landmark is an valid representation for micro emotion recognition. So we try to keep it in our methods, which is one of the most important motivation of our final multi-modality model framework.

B. Visual feature extraction

The CNN based methods have made great results on vision tasks including emotion recognition. On many CNN-based method, CNN is treated as a feature extractor better than the hand-craft as well as an good classifier, which make it unified and end to end. In last subsection, we discuss about the geometrical representation of facial emotion, and realize landmark displacement is an valid geometrical feature for micro emotion recognition task. So we want to study the visual feature's effectiveness and performance.

Unlike the task in face verification, the emotion recognition needs more texture details, so we choose AlexNet which is an simple CNN network architecture, as the prototype. AlexNet's input is 224×224, a trade-off between the texture details and parameter number. The original size of images in iCV-MEFED is 5184×3456, it will take up a lot of computation time during training or testing if we use them as training inputs without cropping.

We delete the last full connected layer, and decrease the dimension of full connected layer from 4096 to 256. The single branch network is in Fig. 2. We take it info account that the dataset of iCV-MEFED is not large and too many parameter will cause overfit early during training. To avoid overfit, we also add dropout parameter to the last full connected layer. The detail will be discussed in section IV.

It is important to note that we transform the original dominant and complementary label into a single label. The transform follows $n = 7(c - 1) + d$, where c indicates complementary emotion and d represents dominant. If original emotion is neutral, set $n = 0$. For example, the mirco emotion 3_6 meaning sad and disgust will be transformed to label $7(3 - 1) + 6 = 20$. So we map original emotion to integers in closed integer interval $[0, 49]$. This map is one to one, it is easy to do the inversion transformation.

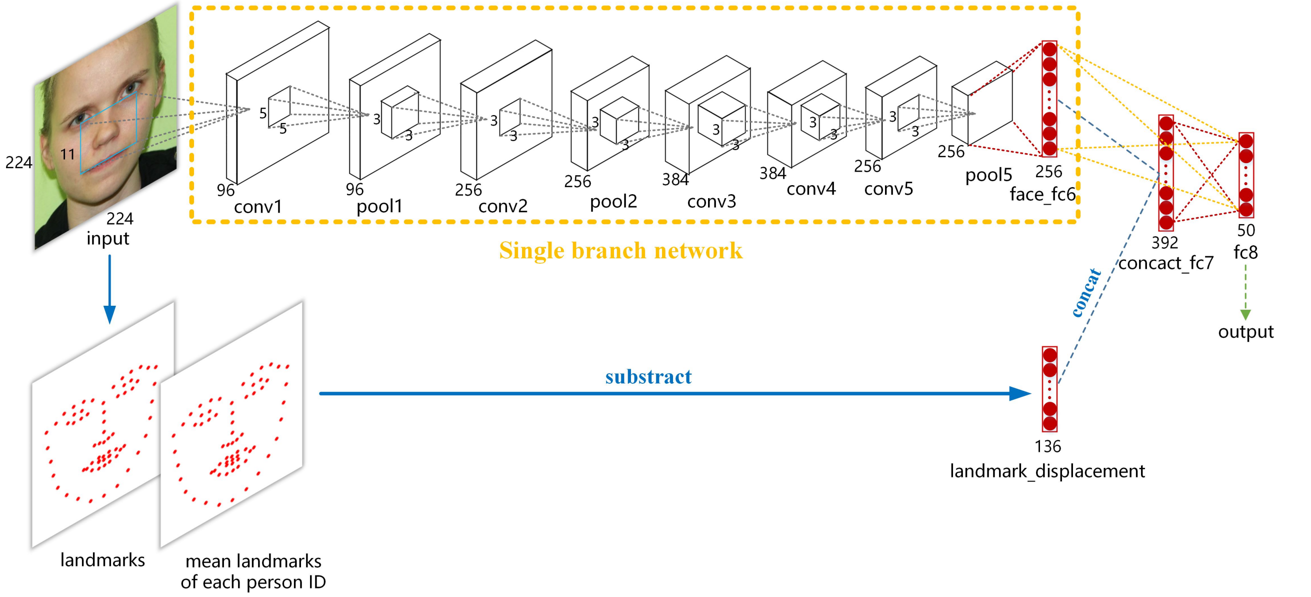

Fig. 2. The upper part is a single branch CNN network. The whole architecture constructs the multi modality network. It is similar to the single CNN, but one more input named landmark displacement is added into classifier in the last two layers of the network.

We have tried to use our network to recognize dominant and complementary separately, then combine them to generate the prediction, but its performance is bad in practice. The multi-task architecture network which join dominant and complementary tasks is not good as well. So we choose to treat the dominant and complementary emotion as a single label. This explains why we do the label transformation.

It is known to all that one of shortcomings in CNN is weight initialization. We find it's hard to train the model, which means we have to choose the parameter such as learning rate carefully. After carefully adjusting parameters, we use the model to reach the accuracy about 0.14 on validation set, not as well as the landmark displacement representation. While it shows the effectiveness of CNN's extracted feature.

C. Multi modality of geometrical and visual information

In the previous two subsections, we introduce the geometrical and visual information based methods, they are both effective but performed not well. An general idea is to ensemble them, concatenating different feature from different domain is a very common way to handle such situation. The landmark displacement information belongs to geometry domain, while the features extracted from CNN come from visual domain. They are similar to multi-modality network, so we name our final proposed model as the title. In practice, it really has better performance, so we try do dig out more details and give more explanations.

The extracted feature of modified AlexNet is a vector $p_1 \in R^{256}$, the landmark displacement $p_2 \in R^{136}$, they are concatenated into $p \in R^{392}$. Then we feed p to a full connected layer as a classifier. We use hinge loss, to train the whole network. Full connected layer with hinge loss is equal to SVM to some degree, which is used to classify geometry representation before.

CNN's extracted feature p_1 spans a vector space $\mathcal{V}_1$, the decision surface can correctly divide some samples, but the decision ability may reaches the ceiling. Once the landmark displacement vector p_2 embedded into the lower vector space, $\mathcal{V}_1$ is mapped from a lower dimension into a higher dimension space $\mathcal{V}$. Because of the effectiveness of p_2, the $\mathcal{V}$ becomes more divisible. It is kind of similar to the kernel function in SVM, but not nonlinear.

IV. Experiment

A. Dataset

We evaluated the proposed method on iCV-MEFED dataset, which is available on the dominant and complementary multi-emotional facial expression recognition challenge [11]. Some samples from this dataset are shown in the fig below.

Fig. 3. Face emotion samples from iCV-MEFED dataset. Each figure is from different faces.

This dataset includes 31250 facial faces with different emotions of 125 subjects acts 50 different emotions and for each of these emotions 5 samples have been taken by Canon 60D camera under uniform lightening condition with relatively unchanged background and the original image resolution is 5184×3456. The images are taken and labelled under supervision of psychologists and the subjects are all trained to act these emotions, which makes the dataset convincing.

The emotion label in this dataset is represented as complementary_dominant, each has eight types: angry, contempt, disgust, fear, happy, sad, surprise, neutral. The ground truth emotion are labeled with number or character N. The number of total combinations is 50, and the label transformation rule during our training is covered in last section.

B. Data Preprocessing

In vision tasks about face attribute analysis, the face in the original image are usually cropped and aligned to eliminate the influence of pose.

We first use ensemble of regression trees method [7] to extract each face's landmarks, then use the two points of eyes center and upper lip to do similar transform to crop and align the face.

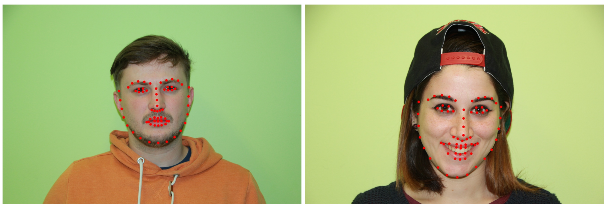

Fig. 4. Original images with extracted landmarks

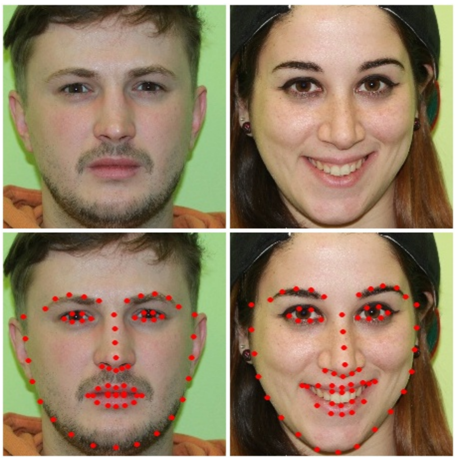

Fig. 5. The first row are cropped and aligned images with size 224×224, the second row are the faces with extracted landmarks.

After cropped and aligned to size 224×224, we again extract the landmarks of each face and calculate each face's landmarks displacement relative to the its id's mean landmark. Then we use the images cropped and aligned and landmarks displacement as the inputs.

The alignment depends on only two points, so the faces are just weakly aligned. If aligned more accurately, the geometrical representation may perform better when fed to an classifier.

C. Model evaluation

In our experiments, the visual inputs are images resized into 224×224, and the geometrical input are landmarks displacement which can be represented as a vector $v \in R^{136}$. We use SGD optimization method and set mini-batch size of 32. The learning rate starts from 0.0005, and the model are trained for up to 100000 iterations. We choose a weight decay of 0.0005 and a momentum of 0.9. We use Caffe [5], a CNN framework to do our experiments.

In the final evaluation phase of this emotion challenge, we have 22969 images with ground truth from 92 persons, and the rest images without labels are the for testing. We split about ten percent of samples into validation set containing 2250 images from the last 9 persons, the left 20719 images are used in training phase. This guarantees that none of the persons will be trained and covered during training.

The remaining parts in this subsection are evaluations of our proposed model.

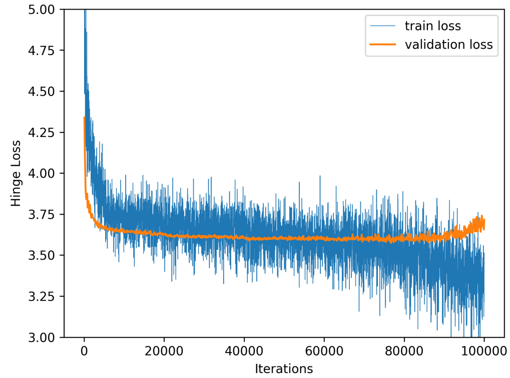

Fig. 6. The L_2 hinge loss during training and validation. From the loss curve, it begins to overfit at around 85000 iterations, simultaneously the accuracy curve does not increase. The lowest validation loss is 3.5736 at iteration 76700

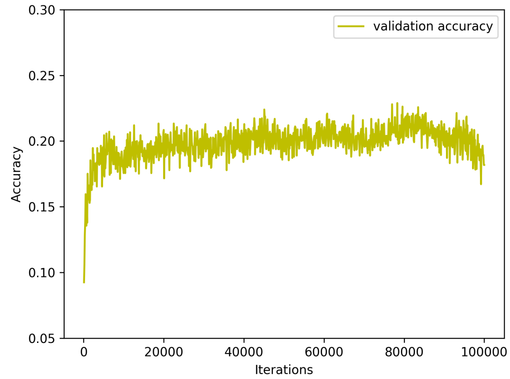

Fig. 7. Top-1 validation accuracy. The accuracy increases rapidly during the first several thousand iterations, but it becomes very slow afterwards. The highest accuracy is 0.2289 at iteration 76700.

During the above experiments, the dropout parameter is added into the last full connected layer, the value is 0.5 in

AlexNet, but we make it larger to 0.8 due to the the small size of this dataset. Dropout strategy [13] is effective on reducing overfit, while it may takes longer time to convergent to a sub-optimal place.

The above trained model is exactly the final submission on the final evaluation phase of micro emotion challenge. After that, we make some extra work discussed on section V.

We list the validation and final testing set results on table I. Our validation result is based only on geometrical information using SVM classifier, and it ranks No.2. While our final model gets the best results on both validation set and testing set.

TABLE I

THE COMPETITION MISCLASSIFICATION RESULTS

User	Team Name	Validation Set	Testing Set
cleardusk	CBSR-CASIA	**79.328571** 84.957143[1]	**80.212137**
ntech	NTechLab	83.985714	-
Bekhouche	-	85.757143	-
szhou	-	86.014286	-
amitkumar05	-	86.228571	-
bknyaz	NTechLab	86.371429	-
charrin	-	87.414286	-
yuzhipeng	ZZXP	87.500000	-
csmath	DeepEmotion	87.571429	-
frkngrpnr	-	89.842857	-
icv	ICV Team	90.714286	-
pablovin	-	91.585714	-
iarganda	CVPD	91.800000	-
lld533	-	93.428571	-

[1] The result is obtained by landmark+SVM.

V. Discussion

A. Top-k accuracy

We are curious about why the accuracy is in such a low level, so we analysis the top-5 accuracy of this model. To accelerate the training procedure, we set dropout to 0.5.

Although the top-1 validation accuracy is not high, but the top-k's accuracy is convincing. It proves that our model have learned from the training samples, but it is a bottleneck on how to transform the top-k accuracy into top-1.

B. Hinge loss v.s. cross-entropy loss

We choose the full connected layer with L_2 hinge loss as an classifier before, for it is equal to SVM classifier. While during the classification tasks, the softmax classifier with cross-entropy loss is more common, and we do some experiments to evaluate its performance.

In conclusion, the cross-entropy loss has the same performance as hinge loss on validation dataset, but overfits faster than the latter one. Compared to cross-entropy loss, the hinge loss has better generalization ability.

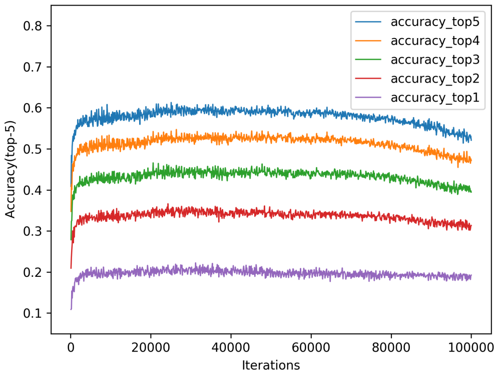

Fig. 8. Top-5 validation accuracy. When dropout is set to 0.5, the overfit comes earlier obviously. And the highest top-5 validation accuracy is 0.2222, 0.3662, 0.4658, 0.5471, 0.6129 in turn.

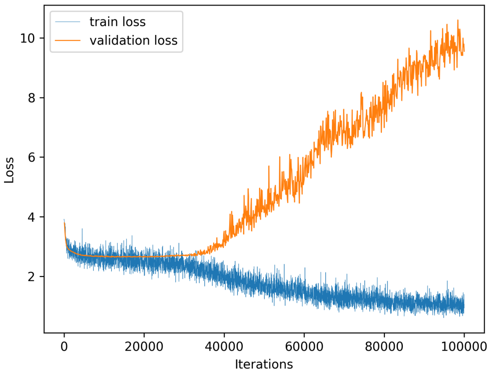

Fig. 9. The cross-entropy loss during training and validation. Although the performance almost the same as hinge loss, it is easily to overfit.

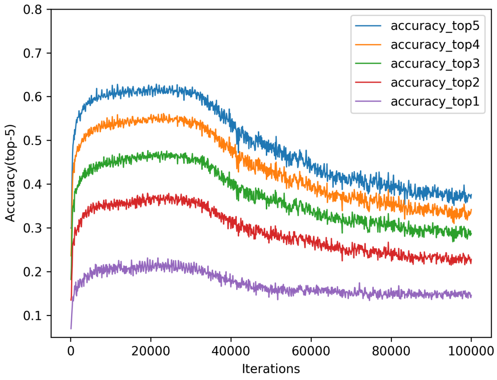

Fig. 10. Top-5 validation accuracy with cross-entropy loss.

C. Higher resolution

In our previous experiments, the landmarks displacement are extracted from images of size 224 × 224. We want to see if the landmarks extracted from higher resolution are more effective. So we use the same 224 × 224 size images but landmarks displacement extracted from 1200 × 1200 images cropped and aligned. And the performance is slightly better on our splitted validation dataset, the top-5 accuracy is listed on Table II. The table shows the higher resolution can provide more effective geometrical information.

TABLE II
HIGHEST TOP-k ACCURACY

Resolution	224 × 224	1200 × 1200
top-1	0.2222	**0.2227**
top-2	0.3662	**0.3800**
top-3	0.4658	**0.4764**
top-4	0.5471	**0.5613**
top-5	0.6129	**0.6293**

D. Future works

We believe that, higher resolution images should provide more geometrical and visual information. We have simply covered one branch at the landmarks. The geometrical information is effective, while its discriminability is limited. But the other one has no such restrictions, and is more attractive and challenging. So how to design a CNN network which can extract more discriminated and compact features from high resolution images should be the next step on mirco emotion recognition.

VI. CONCLUSIONS

In this paper, we proposed an effective multi-modality networks for micro emotion recognition. The proposed network contains two inputs: visual face images, and face geometry landmarks displacement. It fuses the visual and geometry features in a unified framework and can got very promising results. In the micro emotion challenge, we got the first place with the misclassification of 80.212137.

VII. ACKNOWLEDGMENTS

This work was supported by the National Key Research and Development Plan (Grant No.2016YFC0801002), the Chinese National Natural Science Foundation Projects ♯61502491, ♯61473291, ♯61572501, ♯61572536, Science and Technology Development Fund of Macau (No. 019/2014/A1, 112/2014/A3), NVIDIA GPU donation program and AuthenMetric *R&D* Funds.

REFERENCES

[1] Challenges in representation learning: Facial expression recognition challenge. https://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge. Accessed: 2016-06-30.

[2] Charles Darwin, Paul Ekman, and Phillip Prodger. *The expression of the emotions in man and animals*. Oxford University Press, USA, 1998.

[3] Paul Ekman and Wallace V Friesen. Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17(2):124, 1971.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *ArXiv e-prints*, 12 2015.

[5] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 675–678. ACM, 2014.

[6] Heechul Jung, Sihaeng Lee, Junho Yim, Sunjeong Park, and Junmo Kim. Joint fine-tuning in deep neural networks for facial expression recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2983–2991, 2015.

[7] Vahid Kazemi and Josephine Sullivan. One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1867–1874, 2014.

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

[9] Haoxiang Li, Zhe Lin, Xiaohui Shen, Jonathan Brandt, and Gang Hua. A convolutional neural network cascade for face detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5325–5334, 2015.

[10] Ping Liu, Shizhong Han, Zibo Meng, and Yan Tong. Facial expression recognition via a boosted deep belief network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1805–1812, 2014.

[11] Iiris Lüsi, Julio C S Jacques Junior, Jelena Gorbova, Xavier Baró, Sergio Escalera, Hasan Demirel, Juri Allik, Cagri Ozcinar, and Gholamreza Anbarjafari. Joint challenge on dominant and complementary emotion recognition using micro emotion features and head-pose estimation: Databases. In *Automatic Face and Gesture Recognition, 2017. Proceedings. 12th IEEE International Conference on*. IEEE, 2017.

[12] Masakazu Matsugu, Katsuhiko Mori, Yusuke Mitari, and Yuji Kaneda. Subject independent facial expression recognition with robust face detection using a convolutional neural network. *Neural Networks*, 16(5):555–559, 2003.

[13] Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[14] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation from predicting 10,000 classes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1891–1898, 2014.

[15] Yichuan Tang. Deep learning using linear support vector machines. *arXiv preprint arXiv:1306.0239*, 2013.

[16] Zhiding Yu and Cha Zhang. Image based static facial expression recognition with multiple deep network learning. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 435–442. ACM, 2015.

[17] Shuai Zhou, Yanyan Liang, Jun Wan, and Stan Z Li. Facial expression recognition based on multi-scale cnns. In *Chinese Conference on Biometric Recognition*, pages 503–510. Springer, 2016.