

Region Based Robust Facial Expression Analysis

Zheng Lian

National Laboratory of Pattern Recognition,
Institute of Automation Chinese Academy
of Sciences, School of Artificial
Intelligence, University of Chinese
Academy of Sciences
Beijing, China
lianzheng2016@ia.ac.cn

Ya Li

National Laboratory of Pattern Recognition,
Institute of Automation Chinese Academy
of Sciences
Beijing, China
yli@nlpr.ia.ac.cn

Jianhua Tao

National Laboratory of Pattern Recognition,
CAS Center for Excellence in Brain Science
and Intelligence Technology, Institute of
Automation Chinese Academy of Sciences,
School of Artificial Intelligence, University
of Chinese Academy of Sciences
Beijing, China
jhtao@nlpr.ia.ac.cn

Jian Huang

National Laboratory of Pattern Recognition,
Institute of Automation Chinese Academy of
Sciences, School of Artificial Intelligence,
University of Chinese Academy of Sciences
Beijing, China
jian.huang@nlpr.ia.ac.cn

Mingyue Niu

National Laboratory of Pattern Recognition,
Institute of Automation Chinese Academy of
Sciences, School of Artificial Intelligence,
University of Chinese Academy of Sciences
Beijing, China
numiguye2017@ia.ac.cn

Abstract—Facial emotion recognition is an essential aspect in human-machine interaction. In the real-world conditions, it faces many challenges, i.e., illumination changes, large pose variations and partial or full occlusions, which cause different facial areas with different sharpness and completeness. Inspired by this fact, we focus on facial expression recognition based on partial faces in this paper. We compare contribution of seven facial areas of low-resolution images, including nose areas, mouth areas, eyes areas, nose to mouse areas, nose to eyes areas, mouth to eyes areas and the whole face areas. Through analysis on the confusion matrix and the class activation map, we find that mouth regions contain much emotional information compared with nose areas and eyes areas. In the meantime, considering larger facial areas is helpful to judge the expression more precisely. To sum up, contributions of this paper are two-fold: (1) We reveal concerned areas of human in emotion recognition. (2) We quantify the contribution of different facial parts.

Index Terms—facial emotion recognition, facial areas, class activation map, confusion matrix

I. INTRODUCTION

With the development of artificial intelligence, there is an explosion of interest in realizing more natural human-computer dialogue systems, which have wild applications such as the movie booking [1], chatbots [2] and smart homes. Inspired by psychological findings, [3] and [4] point out that addressing emotion information in the conversation agents or the dialogue systems can enhance satisfaction and cause fewer breakdowns in the dialogue. Therefore, emotion recognition, as an essential aspect in human-computer interaction, has received a large amount of attention recently [5,6,7].

Facial expression recognition is a hot research topic in the emotion recognition, which changes deeply under the influence

This work is supported by the National Natural Science Foundation of China (NSFC) (No.61425017, No. 61773379), the National Key Research & Development Plan of China (No. 2017YFB1002804) and the Major Program for the National Social Science Fund of China (13&ZD189).

of deep learning (DL). The previous method is a multi-step process, where handicraft features, i.e., Histogram of Oriented Gradient (HOG) [8], Local Binary Patterns from Three Orthogonal Planes (LBP-TOP) [9] and Scale Invariant Feature Transform (SIFT) [10], are extracted first, combined with various classifiers and fusion methods behind. However, the targets of multi-step are not consistent. Besides, there is no agreement on appropriate handicraft features for the emotion recognition. To solve these problems properly, the multi-step approach is replaced by the end-to-end method, which has gained state-of-the-art performance in the multiple tasks such as image classification [11], machine translation [12], scene classification [13], image caption generation [14] and speech synthesis [15].

Despite a large amount of efforts are made to improve the recognition performance of facial expression, many challenges still exist. In the real-world conditions, it's difficult to gather faces without the shade from other objects. In the meantime, faces are not always in the frontal pose and the proper light conditions. Inspired by the fact that facial areas have different sharpness and completeness in the wild, it is not a reasonable approach to treat every part equally.

One of the solutions is facial expression analysis based on partial faces. The pioneer work by Paul Ekman [16] proposes Facial Action Coding System (FACS), which describes facial expression as combination of multiple action units. Followed with [16], [17] focuses on analyzing different facial parts, i.e., eyes, nose and mouth, and mapping them into FACS codes. However, those works are not suitable for the low-resolution images in real-world conditions as mapping facial parts into FACS is challenging. With the popularity of DL, [18] proposes to recognize facial expression based on regions of interest, which guides convolutional neural networks (CNNs) to focus on areas associated with the expression. However, it does not analyze the contribution of different facial areas for different emotions. Besides, it mainly focuses on emotion recognition in the lab-controlled environment.

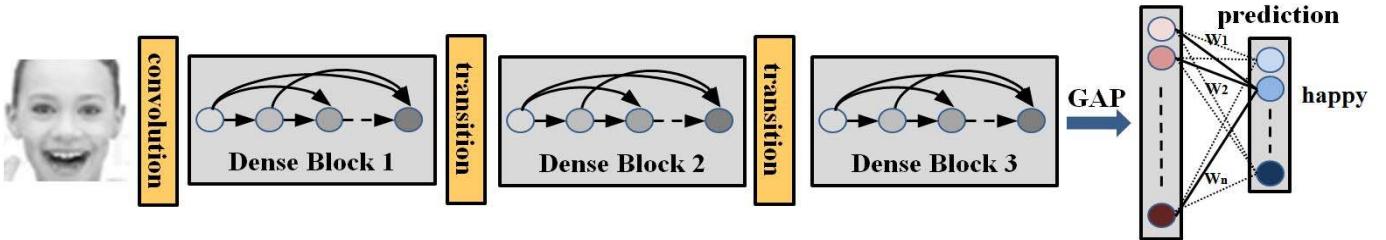


Fig. 1. Flowchart of the proposed system. Three dense blocks are followed behind the inputs, combining global average pooling (GAP) and a fully-connected layer in the end. The outputs of the system are normalized emotion probabilities.

Considering the limitation of previous works [16, 17, 18], in this paper, we analyze the impact of different facial areas for low-resolution images in the wild. Class Activation Mapping (CAM) [19] is adapted to visualize activation parts of different emotions. In the meantime, we analyze the contribution of seven facial areas for the robust emotion recognition, including nose areas, mouse areas, eyes areas, nose to mouse areas, nose to eyes areas, mouth to eyes areas and the whole face areas. To sum up, contributions of this paper are two-fold: (1) We qualitatively reveal concerned areas of human in emotion recognition. (2) We quantify the contribution of various facial parts further through the classification accuracy.

Our findings can be applied to establish systems for robust facial emotion recognition. While the whole face regions are difficult to extract, we analyze emotions through partial facial areas, treating the recognition performance of that areas as the confidence coefficient.

Paper is organized as follows. In Section II, we describe the proposed system in detail. Experiment setup and results are illustrated in Section III and Section IV, separately. Section V concludes the whole paper.

II. SYSTEM DISCRIPTION

In this section, system architectures and visualization methods are discussed in detail. We follow the DenseNet-BC in [20], which has already shown its performance in the Audio/Visual Emotion Challenges 2017 [21]. Through CAM, we visualize activation parts of different inputs. Two functions: expression classification and visualization, can be realized through one optimized system.

A. System Architecture

In our system, we follow the DenseNet-BC architecture, which has three dense blocks associated with the global average pooling (GAP) and the fully-connected layer (FC) behind. The inputs of the system are grey-scale images and the outputs are normalized emotion probabilities.

The system architecture is shown in Fig. 1. Before entering into the first dense block, the convolutional layer with 16 output channels is performed on the 64×64 grey-scale images. Three dense blocks are followed behind and each dense block has 16 layers. In each dense block, 3×3 convolutional filters are used combining zero-padding with one pixel to keep the feature-map size fixed. Batch Normalization is also added before convolutional layers to alleviate the gradient explosion problems. Between contiguous dense blocks, a transition block is applied to reduce the size and the channel of feature maps. The transition block is composed with a 1×1 convolutional

layer, followed with 2×2 average pooling behind. Finally, the GAP and FC are combined to generate emotion probabilities.

B. CAM Technique

To get prediction of the system, we utilize a weighted sum on the outputs of GAP, which are spatial average of the feature maps generated from the last dense block. In CAM technique, a similar idea is taken. We compute a weighted sum of the feature maps extracted from the output of the last dense block as CAM.

We formulize the process of GAP as:

$$F_k = \sum_{x,y} f_k(x,y) \quad (1)$$

where $f_k(x,y)$ represents value of (x,y) in the k^{th} feature map extracted from the last dense block. F_k represents the k^{th} output of GAP, which is spatial average of the k^{th} feature map.

To class c , the class score:

$$\begin{aligned} S_c &= \sum_k w_k^c F_k \\ &= \sum_k w_k^c \sum_{x,y} f_k(x,y) \\ &= \sum_{x,y} \left(\sum_k w_k^c f_k(x,y) \right) \end{aligned} \quad (2)$$

where w_k^c represents the value connected the k^{th} output of GAP to the class c .

Therefore, $\sum_k w_k^c f_k(x,y)$ is the CAM for inputs, which directly indicates importance of activation at spatial coordinate (x,y) related to class c . By upsampling the CAM to the size of inputs, we can identify image regions, which are the most relevant to the particular category.

III. EXPRIMENT SETUP

The system is tested on the FER+ database [22], which is an extension of the FER database [23]. They re-label each image in the FER database through ten crowd taggers to overcome the noise label issue.

The FER dataset is created to mimic real-world conditions through Google image search API. It consists of 35887 images: 28709 for the training, 3589 for the public testing and 3589 for the private testing. The dataset consists of 48×48 pixel grey-scale facial images. Each face is more or less centered and occupies about the same amount of space. The task is to categorize each face based on the emotion shown in the facial expression into one of seven categories, including neutral, happiness, surprise, sadness, anger, disgust and fear.

Compared with FER, FER+ has eight emotion categories adding contempt as well. We follow the same data selection method provided in [22]. If less than 50% of the votes are integrated, the sample will be removed. Then we combine the training data and the public testing set as the training set and evaluate the model performance on the private testing set. Data distribution of the training set and the testing set is shown in Table I.

TABLE I. CLASS CATEGORY DISTRIBUTION OF THE FER+ DATASET.

	<i>Train</i>	<i>Test</i>	<i>Total</i>
<i>Neutral</i>	11000	1219	12219
<i>Happiness</i>	8326	920	9246
<i>Surprise</i>	3807	429	4236
<i>Sadness</i>	3660	421	4081
<i>Anger</i>	2535	287	2822
<i>Disgust</i>	151	19	170
<i>Fear</i>	636	88	724
<i>Contempt</i>	153	21	174
<i>Sum</i>	30268	3404	33672

To divide facial images into different parts, facial landmark detection is essential. We utilize the open-source library, dlib library [24], to detect landmarks, which takes the now classic HOG feature set combined with a linear classifier, an image pyramid and the sliding window detection scheme [25]. After landmark extraction, we divide facial images into seven facial regions based on the position of corresponding landmarks, including nose areas, mouse areas, eyes areas, nose to mouse areas, nose to eyes areas, mouth to eyes areas and the whole face areas. An example of division results is shown in Fig. 2.



Fig. 2. Seven facial areas, including nose areas, mouse areas, eyes areas, nose to mouse areas, nose to eyes areas, mouth to eyes areas and the whole face areas.

IV. EVALUATION RESULTS

In this study, we analyze the impact of different facial areas in facial expression recognition through three experiments. First, we analyze emotion classification performance based on the whole faces, treating it as a comparison experiment. Then, we visualize activation parts of the inputs through CAM technique. Finally, we compare classification accuracy and the confusion matrix of seven facial areas.

A. Performance of the Whole Faces

To recognize the emotion from the whole faces, we train the system through the end-to-end method where the non-linear function is learned to map the whole faces into emotion states.

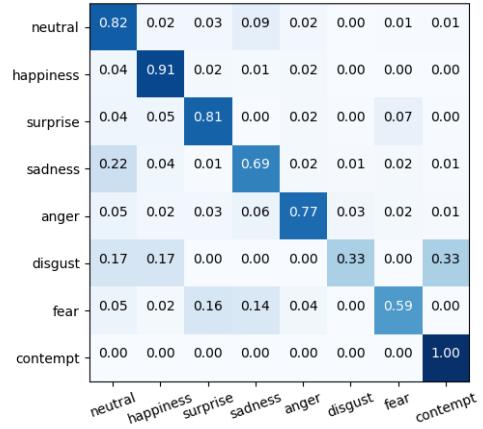


Fig. 3. The confusion matrix based on performance of the the whole face areas on the testing set.

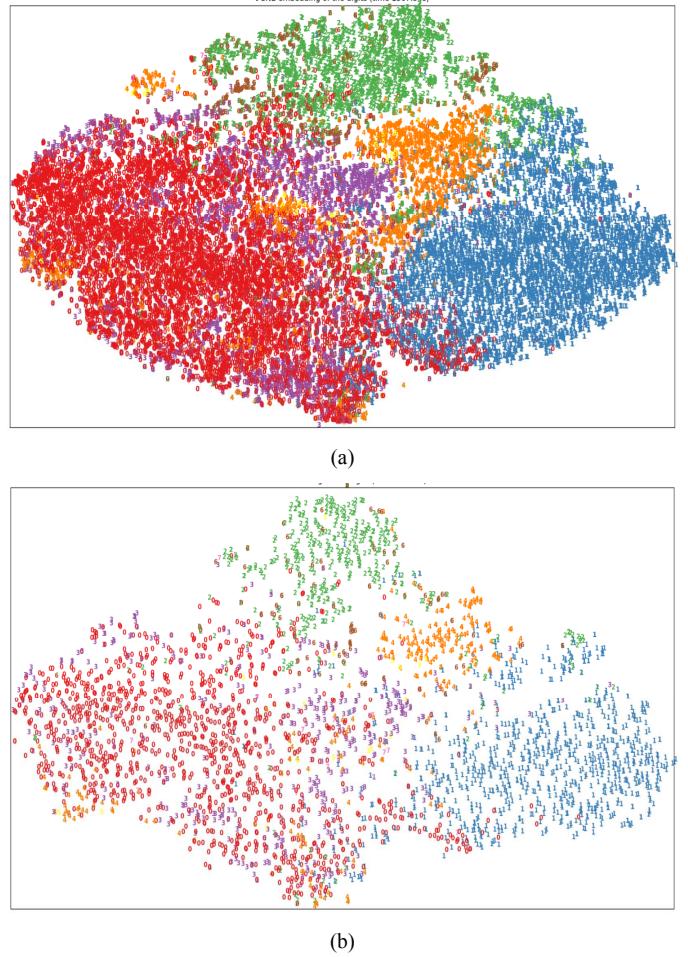


Fig. 4. Visualization bottleneck features through t-SNE. (a) is got from the training set, (b) is got from the testing set. [0(red): neutral, 1(blue): happiness, 2(green): surprise, 3(purple): sadness, 4(orange): anger, 5(brown): disgust, 6(brown): fear, 7(pink): contempt]

In the training process, Adam [26] optimizer is utilized to minimize the cross entropy loss, whose learning rate is chosen to be 10^{-2} . Four workers are opened at the same time to accelerate the input/output process. The data augmentation

methods, i.e., random crop and random horizontal flip, are considered to gain more robust emotion recognizers. The maximum training epoch is set to be 30 and early stopping is applied to alleviate overfitting problem. To alleviate the local minimum problem, we train the system five times and choose the best model according to the performance on the testing set. To analyze results, we visualize the confusion matrix on the testing set in Fig. 3. We also treat outputs of GAP as bottleneck features and visualize these features through t-SNE [27], which is realized under the open-source toolkit, scikit [28].

Although the classifier has 100% recognition accuracy on the contempt through Fig. 3, we cannot affirm that it has good performance on the contempt due to the lack of testing samples. Through Table I, we figure out the disgust and the fear are lack of training samples, causing the classifier has high possibilities to make incorrect prediction of them. Therefore, we only analyze anger, neutral, surprise, happiness and sadness in the experiment. Through Fig. 3 and Fig. 4, the same phenomenon can be found through the CAM technique and the confusion matrix. Happiness has the highest classification accuracy, 91%, among five emotions. However, sadness has the worst performance as it is confused with the neutral.

B. CAM Visualization

We visualize facial activation areas through CAM for CNNs with GAP. Heatmap is visualized through COLORMAP_JET color mapping realized under opencv [29], which varies from blue (low range) to green (mid range) to red (upper range). To show heatmaps on original images, we combine them together through weighted coefficient in [19]:

$$result = heatmap \times 0.4 + image \times 0.5 \quad (3)$$

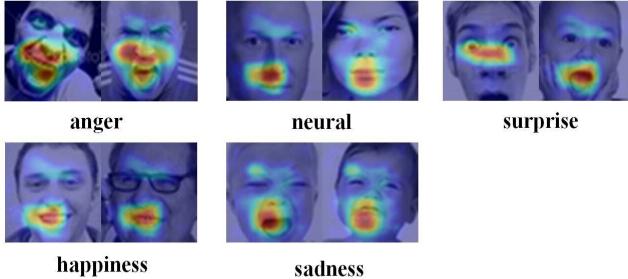


Fig. 5. Combination heatmaps and original images together through the weighted coefficient.

We only visualize and analyze five main emotions. As the heatmaps for mouth regions and lower part of nose areas are close to red, we infer that those areas convey more emotional information than other parts. Besides, as areas around eyes and eyebrows are colored, those areas also count.

C. Contribution of Different Facial Areas

To compare the impact of different facial areas, we train seven systems based on seven facial regions, including nose areas, mouse areas, eyes areas, nose to mouse areas, nose to eyes areas, mouth to eyes areas and the whole face areas. The training process and the testing process are similar with Sec. III except the inputs are different.

TABLE II. EMOTION CLASSIFICATION ACCURACY OF DIFFERENT FACIAL AREAS IN THE TESTING SET.

Facial areas	Accuracy (in %)
Mouth	73.17
Nose	68.81
Eyes	55.21
Nose and mouth	77.01
Nose and eyes	77.29
Mouth and eyes	81.57
The whole faces	81.93

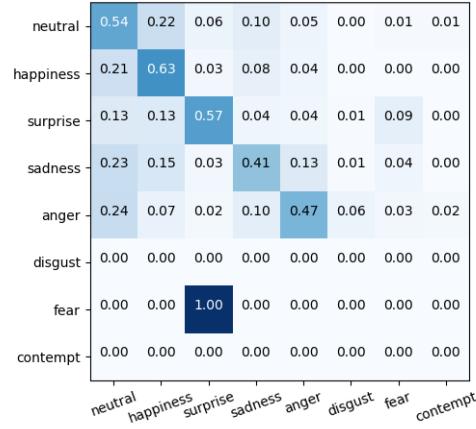
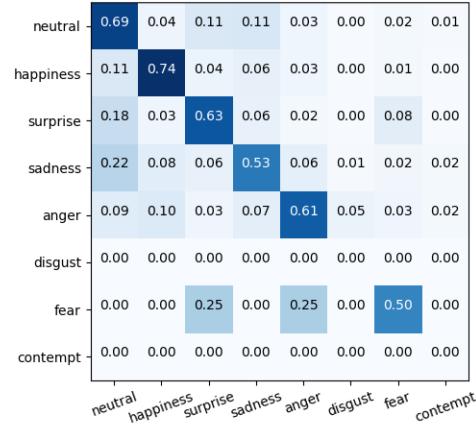
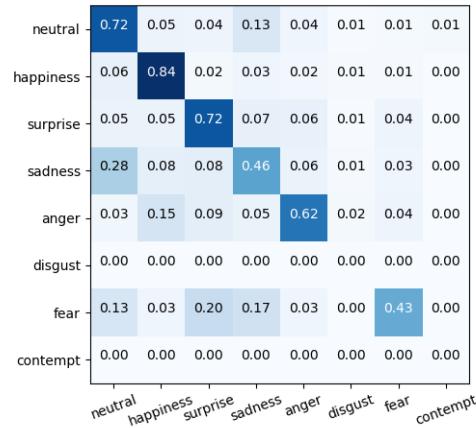


Fig. 6. The confusion matrix on the testing set based on performance of mouth areas (the first figure), nose areas (the second figure) and eyes areas (the third figure).

Classification performance of seven facial areas in the testing set is shown in Table II. We also visualize the confusion matrix on the testing set for mouth, nose and eyes areas in Fig. 6.

Due to the lack of training samples, we ignore disgust, fear and contempt in the analysis process. Through Table II and Fig. 6, we figure out the mouth is the most important area among the nose and eyes area, which conveys much information for the facial expression. The same result can also be found in Fig. 5. Mouth areas lead to less confusion with the neutral, causing nice performance according to accuracy as the neutral contains 35.81% samples of the testing set. Through eyes areas, we have higher possibilities to distinct anger from happiness and surprise than other parts.

What's more, we find that considering larger facial areas is helpful to judge expressions more precisely. Through Table II, we find combination of two areas is better than single areas. Besides, consideration of the whole face areas has the highest classification accuracy.

V. CONCLUSION

Facial emotion recognition is an essential aspect in human-machine interaction. Despite great efforts have been made to improve the performance, many challenges still exist, i.e., illumination changes, large pose variations and partial or full occlusions, which cause different facial areas with different sharpness and completeness. Inspired by this idea, we focus on region based facial expression analysis in this paper. The FER+ dataset is chosen to conduct three experiments. In experiments, we analyze the impact of seven facial areas, including nose areas, mouse areas, eyes areas, nose to mouse areas, nose to eyes areas, mouth to eyes areas and the whole faces through the confusion matrix and the CAM. Through experimental results, we find mouth areas contain much robust emotional information compared with nose areas and eyes areas. And considering larger facial areas can judge expressions more precisely. We can gain the emotion classification accuracy close to the whole faces only through eyes to mouth areas. Our work can promote the understanding of emotion expression.

REFERENCES

- [1] Y. Chen, J. Gao, X. Li, and L. Li, “End-to-End Task-Completion Neural Dialogue Systems,” CoRR, 2017.
- [2] A.D. Brébisson, Y. Bengio, A.P. Chandar, M. Germain, T. Kim, N.R. Ke, Z. Lin, S. Mudumba, V. Michalski, A. Nguyen, M. Pieper, J. Pineau, I. Serban, C. Sankar , S. Subramanian, J. Sotelo, D. Suhubdy, and S. Zhang, “A Deep Reinforcement Learning Chatbot,” CoRR, 2017.
- [3] M. Ishizuka, J. Mori, and H. Prendinger, “Using human physiology to evaluate subtle expressivity of a virtual quizmaster in a mathematical game,” Int. J. Hum.-Comput. Stud., vol. 62, pp. 231-245, 2005.
- [4] B. Martinovsky, and D.R. Traum, “Breakdown In Human-Machine Interaction,” 2003.
- [5] N. Asghar, J. Hoey, X. Jiang, L. Mou, and P. Poupart, “Affective Neural Response Generation,” CoRR, 2017.
- [6] H. Zhou, M. Huang, T. Zhang, X. Zhu, and B. Liu, “Emotional Chatting Machine: Emotional Conversation Generation with Internal and External Memory,” CoRR, 2017.
- [7] S. Ghosh, M. Chollet, E. Laksana, L. P. Morency, and S. Scherer, “Affect-LM: A Neural Language Model for Customizable Affective Text Generation,” CoRR, 2017.
- [8] N. Dalal, and B. Triggs, “Histograms of oriented gradients for human detection,” IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 886-893, 2005.
- [9] G. Zhao, and M. Pietikainen, “Dynamic texture recognition using local binary patterns with an application to facial expressions,” IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 915-928, 2007.
- [10] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” International Journal of Computer Vision, vol. 60, pp. 91-110, 2004.
- [11] Y. Chen, J. Feng, X. Jin, J. Li, H. Xiao, and S. Yan, “Dual Path Networks,” NIPS, 2017.
- [12] A.N. Gomez, L. Jones, L. Kaiser, N. Parmar, I. Polosukhin, N. Shazeer, J. Uszkoreit, and A. Vaswani, “Attention is All you Need,” NIPS, 2017.
- [13] Q. Huang, Z. Lin, and L. Shen, “Relay Backpropagation for Effective Learning of Deep Convolutional Neural Networks,” ECCV, 2016.
- [14] L. Chen, T. Chua, W. Liu, L. Nie, J. Shao, J. Xiao, and H. Zhang, “SCA-CNN: Spatial and Channel-Wise Attention in Convolutional Networks for Image Captioning,” IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6298-6306, 2017.
- [15] S. Dieleman, A. Graves, N. Kalchbrenner, K. Kavukcuoglu, A.V. Oord, K. Simonyan, A.W. Senior, O. Vinyals, and H. Zen, “WaveNet: A Generative Model for Raw Audio,” CoRR, 2016.
- [16] P. Ekman, and W. Friesen, “Facial action coding system: a technique for the measurement of facial movement,” Palo Alto: Consulting Psychologists, 1978.
- [17] J.F. Cohn, T. Kanade, and Y. Tian, “Recognizing Action Units for Facial Expression Analysis,” IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 97-115, 2001.
- [18] X. Sun, M. Lv, C. Quan, and F. Ren, “Improved Facial Expression Recognition Method Dased on ROI Deep Convolutional Neural Network,” International Conference on Affewctive Computing and Intelligent Interaction (ACII), pp. 256-261, 2017.
- [19] A. Khosla, Á. Lapedriza, A. Oliva, A. Torralba, and B. Zhou, “Learning Deep Features for Discriminative Localization,” IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2921-2929, 2016.
- [20] G. Huang, Z. Liu, and K.Q. Weinberger, “Densely Connected Convolutional Networks,” IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2261-2269, 2017.
- [21] S. Chen, Q. Jin, S. Wang, and J. Zhao, “Multimodal Multi-task Learning for Dimensional and Continuous Emotion Recognition,” AVEC, 2017.
- [22] E. Barsoum, C. Canton-Ferrer, C. Zhang, and Z. Zhang, “Training deep networks for facial expression recognition with crowd-sourced label distribution,” ICMI, 2016.
- [23] L.J. Goodfellow, ... and Y. Bengio, “Challenges in Representation Learning: A Report on Three Machine Learning Contests,” Neural networks : the official journal of the International Neural Network Society, vol. 64, pp. 59-63, 2013.
- [24] GitHub, <https://github.com/davisking/dlib>. Accessed December 29, 2017.
- [25] V. Kazemi, and J. Sullivan, “One millisecond face alignment with an ensemble of regression trees,” IEEE Conference on Computer Vision and Pattern Recognition, pp. 1867-1874, 2014.
- [26] J. Ba, and D.P. Kingma, “Adam: A Method for Stochastic Optimization,” CoRR, 2014.
- [27] Y. Bengio, G.E. Hinton, and L.V. Maaten, “Visualizing Data using t-SNE,” Journal of Machine Learning Research, pp. 2579-2605, 2008.
- [28] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, ... and J. Vanderplas, “Scikit-learn: Machine learning in Python,” Journal of Machine Learning Research, vol. 12, pp. 2825-2830, October 2011.
- [29] GitHub, <https://github.com/opencv/opencv>. Accessed December 1, 2017.