More Accurate Facial Emotion Recognition

Zhiyuan Wu
Department of Electronic Engineering
Tsinghua University
wuzy14@mails.tsinghua.edu.cn

Yuwei Qiu
Department of Electronic Engineering
Tsinghua University
qyw14@mails.tsinghua.edu.cn

Xuechao Wang
Department of Electronic Engineering
Tsinghua University
xuech14@mails.tsinghua.edu.cn

***Abstract*—Automatic analysis of human emotion from images is a challenging problem and has attracted interests of researchers for a long time. Many sophisticated models with impressively high accuracy have been reported, but most of which are not able to generalize well. In this paper, we describe a multi-model approach for image-based automatic facial emotion recognition. We combine various feature extractors including many pixel level descriptors and deep CNN features through multi-stage transfer learning. We also investigate different model fusion approaches. We employ a combinational training on many different datasets and get state-of-art accuracy as well as good generalization capability.**

I. Introduction

Automatic facial emotion recognition is a popular and challenging problem in the research fields of computer vision, human-computer interaction, pattern recognition and so on. Specifically, the goal of facial emotion recognition task is to classify a picture that contains a facial expression into one of eight emotion classes: Angry, Disgust, Fear, Sadness, Happy, Surprise, Contempt, and Neutral.

Facial emotion recognition task on image datasets collected in "Lab-controlled" environment has been well studied and many impressive methods and results have been reported. Compared to "In-wild" datasets that are mainly captured from film clips, these datasets have advantages like larger dataset size and higher signal-to-noise ratio, bringing the benefits that the cognitive mechanism is more straightforward and different methods can be implemented and compared easily.

This task has attracted many researchers, and numerous outstanding methods have been reported. These methods can be divided into several categories according to different approaches to extract discriminative features. Many researchers use different pixel level descriptors in order to catch the shape of facial parts like eyes or mouth, as well as the shape and direction of skin wrinkle, which intuitively contains rich emotion information. Beside widely used descriptors like HOG, SIFT and LBP etc., many hand crafted features like geometry feature are also explored [12], [13], [26]. Some researchers are interested in Facial Action Coding System(FACS), which was developed in Anatomy and try to explain the complex mechanism of human facial expressions and the connections between facial expressions and emotions. They analyze facial expressions and transfer it into a combination of several Action Units(AUs), and then translate it into certain emotion [1], [30]. By recent years, more and more researchers preferred deep learning approaches to extract deep level description of emotion. Deep models like Convolutional Neutral Networks(CNN), Deep Belief Networks(DBN), and their structural variation are widely used. These models outperform previous method by a large margin [7], [11], [29]. There also exist other methods like to rebuild 3-D face model.

Related works also include studies that focus on emotion recognition in videos. In the challenge like The EmotiW [6] researchers need consider not only the feature extraction and recognition of single frame images, but also the combination of spatial and temporal information, as well as the analysis of audio. The emotion analysis of single frame is the core of these works and method they used are very similar to methods mentioned above. And relevance analysis method like Three Orthogonal Planes (TOP) [12] or deep neural networks based method like RNN [10] or LSTM are used to add temporal information into consideration. These works are beset by the complex environment change and high noise level in data, thus it is convenient and important to focus on facial emotion recognition in static images.

However, these method suffer a common flaw. In order to get a high recognition accuracy, researchers have to be an expert at parameters tuning and use many sophisticated tricks. This often result in poor generalization capability of the model. People have to build different model for different datasets. In this paper, we try to solve this problem by combining various model and deploying a combinational training on various datasets, and achieve more accurate facial emotion recognition.

More precisely, our contributions are as follows:

- We explore a multi-model approach for facial emotion recognition task, combining many popular and powerful models all-in-one, including pixel level descriptors HOG, Dense-SIFT, LBP, LPQ, and a multi-stage fine-tuning CNN. Furthermore, we explore many different model fusion method.
- We deploy a combinational training on many different datasets, which enable our model to generalize well. We further test our model on a new dataset collected by our self.
- We achieve state-of-the-art accuracy on some publicly available datasets like CK+.

The remainder of this paper are as follows. In section II we introduce the overview of our automatic facial emotion recognition system, give a introduction to different feature extraction methods that we have used, and investigate different

model fusion method. In section III we put some experiment settings and main results. Section IV is a brief conclusion.

II. The Proposed Method

The proposed pipeline is a classical pattern recognition system. For an input image sampled from database or grabbed from video stream, after some basic image preprocessing, a face detector is used to find the face region, and different data augmentation methods are used to manually extend datasets. Then different feature extractors are used to compute discriminative features, and a set of classifiers compute corresponding posterior probability distributions based on those features in parallel. And at last a decision fusion layer give a emotion label to the input image. We will make a detailed discussion for each components in the following subsections.

A. Preprocessing and Face detection

Before performing face detection, we implement some basic image preprocessing procedure. We reduce the size of some high resolution images so that they can be processed faster, and we force the number of color channels to 3 by simply copying intensity value to each RGB channel, in order to fit the input requirement of deep network used. In training phase, many popular data augmentation method are used to extend datasets. We randomly flip the image horizontally and rotate the image by an angle between -15° and 15° in increasement of 15°, and random Gaussian noise is also added. These preprocessing method allow our model to avoid over fitting to some extent and increase the ability of generalization.

Face detection and registration is one of the most important steps in face image processing. The goal of this procedure is to propose a region that contain human face. There exist two strategies of face alignment that are mixture of parts(MoP) and deformable parts model(DPM) [12]. As we will show later, the DPM alignment is coarser and gives better results with CNN model, as in this case the images are more similar to the imaging conditions for pre-training and fine-tuning of the CNN model. But other visual descriptors work better with MoP.

Considering face detectors that is based on deep neural network architecture have reported better performance, we use work of [31] for actual face detection. In our test it outperform other detectors like [3], [24] or Haar Cascade Filter in both accuracy and compute speed.

B. Visual Descriptors

1) CNN: Because of the recent success of deep CNN approaches, we integrate pre-trained CNN models into our method. One of the important reason that CNN can be successfully used in various computer vision task like image classification is the support of huge-size datasets like ImageNet. But CNN suffer the over-fitting problem on relatively small datasets. One way of getting around this problem is using pre-trained CNN models for visual feature extraction and using transfer learning to adapt the models to the particular application. [22], [25] And many similar works on image emotion recognition like [14] confirm that this approach is indeed effective.

Inspired by [12], we propose a multi-stage fine-tuning strategy to get a CNN model having better performance and generalization capability. More specifically, we start from the VGG-Face model that is trained based on VGG-16 architecture for face recognition [23], and then apply the FER 2013 [8] emotion corpus for the fine tuning. Here we make a hypothesis that on this task VGG-Face will work better than models pre-trained on ImageNet, which is developed for general object classification. We then use Public Test set of FER2013 for 5 epochs stage-1 fine-tuning. After locking the parameters of layer before conv5, we use Private Test set of FER2013 for another 10 epochs stage-2 fine-tuning. Then we use a combination of some public datasets (See Experiment section for details) for a longer fine-tuning as last stage with all parameters be able to update. All procedure mentioned above are under the Dropout and weight decay regularization.

We use a Caffe [9] implementation of VGG-Face, where the main part of network has not been modified during whole fine-tune procedure except that the size of last two full connected layer is adjusted according to different class number on different datasets.

2) HOG: The histogram of oriented gradients (HOG) [4] feature describes the local shape and appearance of objects by capturing the distribution information of intensity gradients. The descriptor decomposes a local region into small squared cells, and computes the histogram of different bins of oriented gradients in each cell, and normalizes the result using block-wise style. Then features from different local regions are concatenated spatially.

HOG has been widely used in many computer vision task, especially in pedestrian detecting. It is a consensus that the shape of face components like eyes and mouth and the direction of skin wrinkles contains rich information of emotion. Because HOG can well capture the direction changes of edges so HOG is believed to be a good feature to represent emotion.

In our model, we deploy HOG similarly to that described in [17], where we resize each image to 64×64 pixels, and divide them into $7 \times 7 = 49$ overlapping blocks with size of 16×16 pixels(i.e. the strides are 8 pixels in both horizontal and vertical directions). The descriptor is applied by computing histograms of oriented gradient on 2×2 cells in each block, and the orientations are quantized into 9 bins, which result in $2 \times 2 \times 9 \times 49 = 1764$ dimensions for the whole image.

3) Dense SIFT: The scale-invariant feature transform(SIFT) [18] combines a feature detector and a feature descriptor. The detector extracts a number of interested points from an image in a way that is consistent with some variations of the illumination or viewpoint. The descriptor associates to the region around each point a signature which identifies its appearance compactly and robustly. For dense SIFT, it is equivalent to performing SIFT descriptor directly on a dense grid of locations on a image at a fixed scale and orientation. SIFT is one of the most successful visual descriptors in various

computer vision task for decades, and many work in emotion recognition report its good performance.

In our model, again we divide each image into 49 overlapping regions as described in HOG. In each local block, we apply the SIFT descriptor to the center point, and finally get a $4 \times 4 \times 8 \times 49 = 6272$ dimensions feature for the whole image.

4) LBP: Local Binary Patterns (LBP) [2] computation amounts to finding the sign of difference with respect to a central pixel in a neighborhood, and transforms the binary pattern into an integer and finally converting the patterns in to a histogram. So LBP is believed to be able to capture the surface texture of object. A useful extension of original LBP is uniform LBP, which clusters 256 patterns into 59 bins, and takes into account occurrence statistics of common patterns.

In our model, LBP is also applied on 49 local regions grid, and finally get a $59 \times 49 = 2891$ dimensions feature for each images.

Note that there is a variant version of LBP method called Local Gabor Binary Patterns (LGBP), where the images are convolved with a set of 2D complex Gabor filters to obtain Gabor-pictures, then LBP is applied to each Gabor-picture and those features are concatenated as output. LGBP has been reported to have similar process style with human primary visual cortex [5] and it has slightly better performance in some task. We tried to add LGBP as another visual descriptor but it came to a problem that the parameters of Gabor filters is hard to be consolidated on different datasets.

5) LPQ: The Local Phase Quantization (LPQ) features are computed by taking 2-D Discrete Fourier Transform(DFT) of M-by-M neighborhoods of each pixel in the gray scale image. 2D-DFT is computed at four frequencies $\{[a,0]^T, [0,a]^T, [a,a]^T, [a,-a]^T\}$ with $a = 1/M$, which correspond to four of eight neighboring frequency bins centered at the pixel of interest. The real and imaginary parts of resulting four complex numbers are separately quantized using a threshold of zero, which gives an eight bit string. This string is then converted into an integer value in the range of $[0, 255]$. The pixel based values are Finally converted into a histogram of 256 bins. LPQ features have been successfully used in emotion recognition problems, and it serve as baseline in the AVEC 2013 Challenge. And therefore we believe it can be a powerful feature.

In our model, LPQ is also applied on 49 local regions grid(i.e. with $M = 16$), and finally get a $256 \times 49 = 12544$ dimensions feature for each images.

C. Classifiers

Once all visual features mentioned above are ready, numerous classifiers can be chosen to compute posterior probability distribution on different emotion class given those visual features, denoted by $p(c_i \mid \{x_j\})$, where c_i denote one of eight target emotions and $\{x_j\}$ denote the set of computed visual features from different descriptors.

In our model, we mainly use two popular classifiers. For pixel level descriptors HOG, Dense SIFT, LPB, and LPQ, Support Vector Machine (SVM) with radial basis function (RBF) kernel:

$$\mathcal{K}(x_i, x_j) = exp(-\gamma \|x_i - x_j\|^2) \tag{1}$$

are used to compute $p(c_i \mid x_j)$ correspondingly, and kernel parameters such as γ in Eq.(1) and $\mathcal{C}$ in loss function are greedily search on a grid of exponential powers of 2. Randomly pick a specific parameters setting, and move to a nearby point that have higher accuracy on validation set. Do this procedure for several times and record the parameters with highest accuracy on validation set.

For CNN model, we use a two layer full connected perception to convert deep CNN features of convolutional layers to posterior probability distribution. A combination of Dropout and weight decay is used to avoid overfitting.

Note that there exist some works where researchers use SVM or other classifiers to deal with CNN features and some of them report slightly better result. But in our test they are almost identical in performance, and for convenience, we directly use the full connect layer implementation of Caffe.

D. Model Fusion

[28] explores various approaches used in model fusion. Those method can be divided into three categories according to the fusion level. *Feature level fusion* concatenate the features derived from multi-source and apply an uniform classifier to get the result. Feature level fusion is able to exploit the most information of the original data, but it is beset by the problem that different kinds of features might be inconsistent and incompatible in dimensional variance, and the dimensionality might be too high to be classified efficiently. *Decision level fusion* is very easy to implement and different source classifiers vote for the final output. However, it does not allow the multi-source information to be fully exploited because the decision only contains a label number and too much detailed information are lost. *Score level fusion* is a compromise between two approaches above. It can be considered as the evaluation of $p(c_i \mid \{x_j\})$ according to $\{p(c_i \mid x_j)\}$ from multi-source classifiers.

As a result, fusion at the score level is a good way and is what we mainly focus on. More specifically, there are mainly two methods to apply score level fusion:

1) The Product Rule: If we make the assumption that the features form different source are conditionally independent given the ground truth label, i.e.

$$p(x_j \mid c) \perp p(x_i \mid c) \tag{2}$$

where $i \neq j$ and $\perp$ denote statistical independent. Then the relationship of label c and observed features set $\{x_i\}$ can be described by a tree structured Bayesian graphic model, where c is the root node and $\{x_i\}$ are leaf nodes. After converting it to a factor graph, the posterior probability distribution can be exactly inferred by the Product-Sum algorithm:

$$p(c_i \mid \{x_j\}) = \prod_j p(c_i \mid x_j) \tag{3}$$

2) The Sum Rule: Besides the same assumption described by Eq.(2), The Sum Rule also assumes that posterior distribution computed by the individual classifiers do not deviate much from the prior probabilities, then the posterior probability can be approximately inferred by:

$$p(c_i \mid \{x_j\}) = \frac{1}{C} \sum_j p(c_i \mid x_j) \tag{4}$$

There a variant of The Sum Rule in technical implementation which is popular and effective, called The Weighted Fusion approach. It add a weight term to Eq.(4) and can be written as:

$$p(c_i \mid \{x_j\}) = \frac{1}{C} \sum_j w_{ij} p(c_i \mid x_j) \tag{5}$$

with

$$\sum_j w_{ij} = 1$$

holds. A very popular setting in various pattern recognition tasks is $w_{ij} \equiv w_j$ which are manually selected as a set of hyperparameters. The Weighted Fusion approach can often make a big improvement on multi-model systems and many state-of-the-art work on facial emotion recognition rely on it.

However, w_{ij} is very hard to evaluate. Some researchers simply randomly generate and choose one [12], while some researchers optimize it on validation set or even train a more complex classifiers like SVM. Though this approach can significantly improve the performance like accuracy on certain dataset, it also brings the problem like over-fit to validation set and poorly generalize.

In our model, we tried all these fusion method and choose Weighted Fusion approach with $w_{ij} \equiv w_j$, for the consideration of tradeoff between accuracy and generalization capability.

Finally, our model choose a emotion label that can maximize the computed posterior probability distribution as the prediction:

$$\hat{c} = \arg\max_{c_i} p(c_i \mid \{x_j\}) \tag{6}$$

III. Experiment

We test our model on numerous popular facial emotion datasets, and compare our results with some state-of-art works. Furthermore, we introduce a new dataset provided by *Media and Cognition* course from Tsinghua University, China. We test the generalization capability of our method on this new dataset.

A. Datasets

1) CK+: The Extended Cohn-Kanade AU-Coded Facial Expression Database, referred as CK+ [19], is a popular benchmark for research in automatic facial image analysis and synthesis and for perceptual studies. The dataset includes 486 sequences from 97 posers, and part of which are manually labeled into eight emotions without confusion. Finally 445 images at peak frames of labeled sequences are used in our experiment.

2) TFEID: The Taiwanese Facial Expression Image Database (TFEID) [15] is a large lab-collected facial emotion dataset, consisting of 7200 stimuli captured from 40 models (20 males), each with eight facial expressions. Models were asked to gaze at two different angles. Each expression includes two kinds of intensities (high and slight) and was captured by two CCD-cameras simultaneously with different viewing angles. In our experiment, only images captured at front view with high intensities are used.

3) JAFFE: The Japanese Female Facial Expression (JAFFE) Database [21] contains 213 images of 7 facial expressions (without *Contempt* Emotion in CK+ and TFEID) posed by 10 Japanese female models. In our experiment, it is extended to eight emotions in order to keep consistent with other datasets, by simply adding an empty label set.

4) KDEF: The Karolinska Directed Emotional Faces (KDEF) [20] is a set of totally 4900 pictures of human facial expressions of emotion, which was originally developed to be used for psychological and medical research purposes. The set contains 70 individuals, each displaying 7 different emotional expressions, each expression being photographed (twice) from 5 different angles. In our experiment, we only pick images captured at front view and also extend it to eight emotions by introducing an empty *Contempt* set.

5) Newly Collected Dataset: Many models introduced by previous works can not generalize well to other datasets or images that are "In-Wild" style. To test the generalization capability of our model, we use the dataset provided by *Media and Cognition* course from Tsinghua University, which contains images captured from students and images collected from Internet. Compared to datasets mentioned above, this newly collocated dataset is more authentic and living, rather than obvious spurious performance. So its a perfect dataset to test the generalization capability of facial emotion recognition systems. It consists approximately 1000 images and are manually labeled into same eight emotion classes.

B. Training

The amount of data provided to model training has been proved in many computer vision and pattern recognition tasks to be a significant factor that influences the performance of model. In consideration of that, we apply a combinational training procedure on our model. Instead of separately using each datasets, we combine all datasets mentioned above all-in-one to maximize the size of datasets used for training.

The number of images in five datasets adds up to approximately 2k, and 12k after augmentation. We use a random 5-folds cross validation approach for data segmentation. And all data are used for training with same hyperparameters in submission version of our systems. Compared to existing works where single dataset with size up to 1k is used for training, this combinational training approach can significantly improve the performance of our system, especially CNN part.

C. Result

1) Comparison of Different extractors: First we introduce some result obtained on a simplified version of our sys-

TABLE I
COMPARISON OF DIFFERENT EXTRACTORS

Method	Accuracy
HOG	87.64%
LPQ	82.02%
LBP	83.15%
Dense SIFT	89.89%
CNN	85.55%

TABLE II
COMPARISON OF DIFFERENT FUSION METHOD

Method	Accuracy
HOG	78.84%
LPQ	74.90%
LBP	77.59%
Dense SIFT	82.16%
Sum Rule	84.65%
Product Rule	84.44%
Weighted Fusion	87.23%

tem. More specifically, in order to show the performance improvement introduced by multi-model approach, we first test different feature extractors. We only used CK+ datasets and no augmentation is applied in this test.

Result is shown in TABLE I. It can be concluded that different pixel level feature extractors have similar performance on small dataset, and SIFT is slightly better than others. Deep neural networks is not good enough because the limit of dataset size (0.5k only).

2) Comparison of Different fusion method: In order to show the improvement obtained by different fusion method, we combine several pixel level by different fusion style. This time the model is trained on the combination of five datasets mentioned above, but still without any augmentation process.

Result is shown in TABLE II. We observe that these pixel level feature extraction methods have a obvious decrease in performance on the joint data set. This is mainly because these pixel level descriptors can not generalize well to newly collected dataset, so overall accuracy has decreased. Dense-SIFT is still slightly better than others. The Sum Fusion Rule has a very similar performance with the Product Fusion Rule, they improve the accuracy by 2.5% above Dense-SIFT. Weighted Fusion approach can further give another 3% improvement which indicates that the Weighted Fusion approach is an expert at accuracy improving.

3) Complete Model Test on CK+: Finally, we apply a complete version of our system described in Section II, i.e. we train our model on the joint dataset after fully augmentation. We observed a remarkable improvement of CNN approach under this setting ($87\% \rightarrow 92\%$), and finally Weighted Fusion is used for further boosting.

CK+ is one of the most popular dataset used as a benchmark for facial emotion recognition task. We report competitive performance on CK+ datasets. See TABLE III for detail.

TABLE III
OVERALL ACCURACY COMPARISON ON CK+

Method	Accuracy
[27]	88.80%
[16]	92.40%
Ours	**94.38%**

IV. CONCLUSION

Automatic analysis of human emotion from images is a challenging problem. In this paper, we emphasize the importance of deeper analysis of emotion recognition on lab-collected images rather than on other noisy way. We propose a multi-model approach for more accurate automatic facial emotion recognition on static images. In our system, we combine four different pixel level image descriptor: HOG, LBP, LPQ, and Dense-SIFT, and a multi-stage fine-tuned CNN. We apply a data augmentation and combinational training procedure in order to make fully use of large amount of data. We also investigate different model fusion methods. Our model is proved to effective and have a better generalization capability than previous work through many experiments. We report better generalization capability by testing our model on a newly collected dataset and competitive accuracy on CK+ to many other state-of-the-art works.

REFERENCES

[1] M. S Bartlett, Gwen Littlewort, M Frank, and C Lainscsek. Recognizing facial expression: machine learning and application to spontaneous behavior. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 568–573, 2005.

[2] C Frelicot C Silva, T Bouwmans. An extended center-symmetric local binary pattern for background modeling and subtraction in videos. *Visapp*, 2015.

[3] Dong Chen, Shaoqing Ren, Yichen Wei, Xudong Cao, and Jian Sun. Joint cascade face detection and alignment. In *European Conference on Computer Vision*, pages 109–122, 2014.

[4] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 886–893, 2005.

[5] JG Daugman. Two-dimensional spectral analysis of cortical receptive field profile. *Vision Research*, 1980.

[6] Abhinav Dhall, Roland Goecke, Jyoti Joshi, Jesse Hoey, and Tom Gedeon. Emotiw 2016: video and group-level emotion recognition challenges. In *ACM International Conference on Multimodal Interaction*, pages 427–432, 2016.

[7] Yin Fan, Xiangju Lu, Dian Li, and Yuanliu Liu. Video-based emotion recognition using cnn-rnn and c3d hybrid networks. In *ACM International Conference on Multimodal Interaction*, pages 445–450, 2016.

[8] Ian Goodfellow, Dumitru Erhan, Pierre-Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, Yingbo Zhou, Chetan Ramaiah, Fangxiang Feng, Ruifan Li, Xiaojie Wang, Dimitris Athanasakis, John Shawe-Taylor, Maxim Milakov, John Park, Radu Ionescu, Marius Popescu, Cristian Grozea, James Bergstra, Jingjing Xie, Lukasz Romaszko, Bing Xu, Zhang Chuang, and Yoshua Bengio. Challenges in representation learning: A report on three machine learning contests, 2013.

[9] Jia, Yangqing, Shelhamer, Evan, Donahue, Jeff, Karayev, Sergey, Long, and Jonathan. Caffe: Convolutional architecture for fast feature embedding. *Eprint Arxiv*, pages 675–678, 2014.

[10] Samira Ebrahimi Kahou, Vincent Michalski, Kishore Konda, Roland Memisevic, and Christopher Pal. Recurrent neural networks for emotion recognition in video. 13(5):467–474, 2015.

[11] Samira Ebrahimi Kahou, Christopher Pal, Xavier Bouthillier, Pierre Froumenty, ?aglar Gl?ehre, Roland Memisevic, Pascal Vincent, Aaron Courville, Yoshua Bengio, and Raul Chandias Ferrari. Combining modality specific deep neural networks for emotion recognition in video. In *ACM on International Conference on Multimodal Interaction*, pages 543–550, 2013.

[12] Heysem Kaya, Furkan Grp?nar, and Albert Ali Salah. Video-based emotion recognition in the wild using deep transfer learning and score fusion *. *Image and Vision Computing*, 2017.

[13] Heysem Kaya, Furkan Rpinar, Sadaf Afshar, and Albert Ali Salah. Contrasting and combining least squares based learners for emotion recognition in the wild. In *ACM on International Conference on Multimodal Interaction*, pages 459–466, 2015.

[14] Bo Kyeong Kim, Jihyeon Roh, Suh Yeon Dong, and Soo Young Lee. Hierarchical committee of deep convolutional neural networks for robust facial expression recognition. *Journal on Multimodal User Interfaces*, 10(2):173–189, 2016.

[15] YuShiuan Yen LiFen Chen. Taiwanese facial expression image database. 2007.

[16] Mengyi Liu, Shaoxin Li, Shiguang Shan, Ruiping Wang, and Xilin Chen. *Deeply Learning Deformable Facial Action Parts Model for Dynamic Expression Analysis*. Springer International Publishing, 2014.

[17] Mengyi Liu, Ruiping Wang, Shaoxin Li, Zhiwu Huang, Shiguang Shan, and Xilin Chen. Video modeling and learning on riemannian manifold for emotion recognition in the wild. *Journal on Multimodal User Interfaces*, 10(2):113–124, 2016.

[18] David G Lowe. *Distinctive Image Features from Scale-Invariant Keypoints*. Kluwer Academic Publishers, 2004.

[19] Patrick Lucey, Jeffrey F. Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *Computer Vision and Pattern Recognition Workshops*, pages 94–101, 2010.

[20] Daniel Lundqvist, Anders Flykt, and Arne Ohman. The karolinska directed emotional faces (kdef). 1998.

[21] M. J. Lyons, J. Budynek, and S. Akamatsu. Automatic classification of single facial images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(12):1357–1362, 2002.

[22] Hong Wei Ng, Viet Dung Nguyen, Vassilios Vonikakis, and Stefan Winkler. Deep learning for emotion recognition on small datasets using transfer learning. In *ACM on International Conference on Multimodal Interaction*, pages 443–449, 2015.

[23] A Zisserman OM Parkhi, A Vedaldi. Deep face recognition. *British Machine Vision Conference*, 2015.

[24] D. Ramanan and Xiangxin Zhu. Face detection, pose estimation, and landmark localization in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2879–2886, 2012.

[25] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. Cnn features off-the-shelf: An astounding baseline for recognition. In *Computer Vision and Pattern Recognition Workshops*, pages 512–519, 2014.

[26] A Saeed, A Al-Hamadi, R Niese, and M Elzobi. Effective geometric features for human emotion recognition. In *IEEE International Conference on Signal Processing*, pages 623–627, 2012.

[27] Ziheng Wang, Shangfei Wang, and Qiang Ji. Capturing complex spatio-temporal relations among facial muscles for facial expression recognition. 9(4):3422–3429, 2013.

[28] Yong Xu and Yuwu Lu. Adaptive weighted fusion: A novel fusion approach for image classification. *Neurocomputing*, 168:566–574, 2015.

[29] Anbang Yao, Dongqi Cai, Ping Hu, Shandong Wang, Liang Sha, and Yurong Chen. Holonet: towards robust emotion recognition in the wild. In *ACM International Conference on Multimodal Interaction*, pages 472–478, 2016.

[30] Anbang Yao, Junchao Shao, Ningning Ma, and Yurong Chen. Capturing au-aware facial features and their latent relations for emotion recognition in the wild. In *ACM on International Conference on Multimodal Interaction*, pages 451–458, 2015.

[31] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.