2016 23rd International Conference on Pattern Recognition (ICPR)
Cancún Center, Cancún, México, December 4-8, 2016

Facial Expression Recognition by Re-ranking with Global and Local Generic Features

Duc Minh Vo
Faculty of Information Technology
University of Science, VNUHCM, Vietnam
Email: vmduc@fit.hcmus.edu.vn

Akihiro Sugimoto
National Institute of Informatics
Tokyo, Japan
Email: sugimoto@nii.ac.jp

Thai Hoang Le
Faculty of Information Technology
University of Science, VNUHCM, Vietnam
Email: lhthai@fit.hcmus.edu.vn

***Abstract*—Recognizing the facial expression plays an important role in human computer interaction. Following the recent success of the Convolutional Neural Network (CNN) in image classification and object recognition, this paper proposes a facial expression recognition method that makes full use of CNNs to detect face features globally and locally and that combines global and local generic features for improving accuracy in recognition. Our method uses global generic features with the Support Vector Machine (SVM) classifier to generate most plausible candidates in expression class while local generic features with the SVM classifier to look into the candidates to re-rank them for recognition. Experimental results using data-sets available in public support the effectiveness of our proposed method by demonstrating improved accuracy against the state-of-the-arts.**

I. Introduction

Facial expression is a way of non-verbal communication that has been analyzed for long time across various disciplines such as psychology, neuroscience, computer science, engineering or sociology. In computer vision, facial expression recognition has been an attractive topic in recent years due to many potential applications such as human computer interaction, education software, automobile safety, and telecommunications.

Various disciplines in studying facial expression recognition reveal that the human vision recognizes facial expressions via three types of facial perception: holistic, components-based and configure-based perception [1]. Holistic perception perceives faces as a single entity with discarding facial components. In contrast, components-based perception focuses on the role of facial components such as eyes, nose, and mouth. Configure-based perception, on the other hand, relies on spatial relations among facial components (e.g. left-eye and right-eye, mouth and nose).

Following the three types of facial perception, recognizing facial expressions has two main approaches [2]: appearance based approach and 3D shape model based approach. The former corresponds the holistic and components-based perceptions while the latter does to the configure-based perception. Because the 3D shape model based approach requires a variety of accurate 3D shape models in advance and constructing the 3D shape models of faces is tedious, it is less common than the appearance based approach [2]. We thus direct our attention to the appearance based approach in this work.

A Facial Expression Recognition (FER) method employing the appearance based approach has two main steps: feature extraction from face appearances and feature classification. A FER method can have a pre-processing to detect the whole face from an input image by removing non-facial parts in order to make the method robust against illumination variation, background variation, etc.

The first step detects a feature vector representing the face from a face image. Two types of features exist for FER: the global feature and the local feature. Global features are detected using the whole face image. In some methods, global features are computed directly from the coordinates of facial points [3] or the distance between facial points [4]. The high-level data-driven representations of global features are also proposed using the non-negative matrix factorization [5] or the sparse coding [6]. Local features are detected only from local regions of the face. Local features are computed as the histogram of low-level features such as Local Binary Patterns (LBP) [7], Histogram of Gradients (HoG) [8] or the Bag-of-Words (BoW) [9]. Local features are tolerant against illumination, rotation variation but they are sensitive to local region detection.

In image classification and object recognition, on the other hand, deep-learning (more specifically, the Convolutional Neural Network (CNN) [10]) based approaches are said to extract the generic feature that describes the semantic meaning of images [11], and the generic feature obtained via CNN together with the soft-max classifier or the Support Vector Machine (SVM) classifier has been recently reported to achieve remarkably high accuracy and to outperform existing methods [12], [13]. Since the generic feature are also effective for FER [14], [15], we employ the generic feature for our FER method.

The second step classifies the features detected in the first step into expression classes. Used here are classification methods developed in machine learning such as SVM [16], Deep Brief Network (DBN) [17] or k-Nearest Neighbor (k-NN). In particular, the SVM classifier is known to work more effectively than the soft-max classifier [12], [13], [18], and is reported to be most promising for FER [2]. We remark that features used so far in the classification step are either from the whole face alone or from facial components alone. Combining features from the whole face and from facial components is not well exploited.

Current efforts devoted to FER are to seek as good features as possible and/or to improve the employed classifier as

978-1-5090-4847-2/16/$31.00 ©2016 IEEE

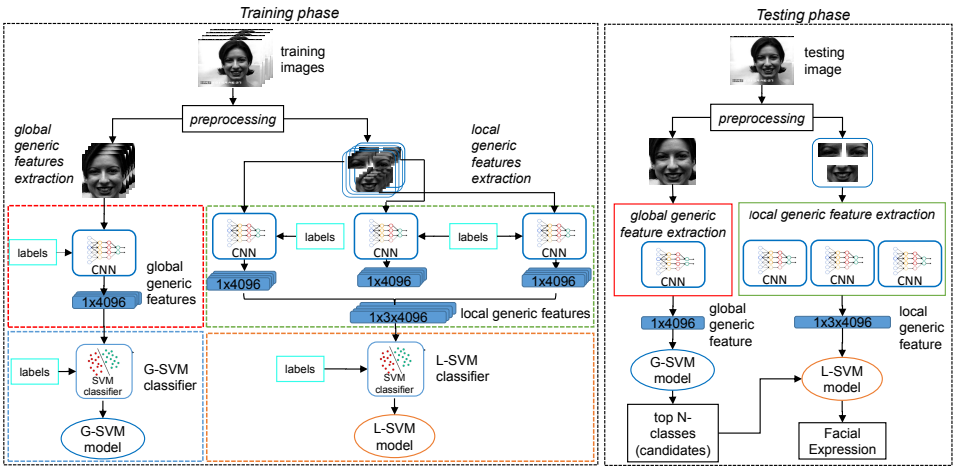

Fig. 1. Framework of our proposed method.

much as possible. However powerful the feature becomes, and however improved the ability of the classifier becomes, there always exists the limitation of accuracy in recognition. We should not rely too much on improved ability of the feature and/or the classifier. Rather than sticking to obtaining the correct expression at one try, we had better generate plausible candidates first and then look into the candidates to finally obtain the correct expression. With this approach, we can re-rank candidates via investigation from different aspects from the candidate generation for more accurate recognition. Namely, we can secure correct expression not taking the 1st rank in the candidates and, at the same time, drop incorrect expression that happens to take the 1st rank in the candidates.

Motivated by above, we propose an FER method where a global generic feature obtained by using a CNN and a local generic feature obtained by other CNNs are used to effectively classify face expressions. In our method, the global generic feature is used with a SVM classifier to generate most plausible candidates in expression classes while the local generic feature is used with another SVM classifier to look into the candidates and re-rank them to obtain the correct expression class. Our usage of global and local (generic) features is justified by the insight that facial expressions are recognized neither by global features coming from the whole face alone nor by local features coming from facial components, but by combing the both features [19]. (In our method, candidates are generated using global information while they are looked into using local information for re-ranking.)

II. Proposed Method

A. Overview of proposed method

Our proposed FER method includes two steps: (1) extract a global generic feature using a CNN (trained in advance) and generate most plausible candidates using the global SVM (G-SVM) classifier (trained using global generic features in advance), and (2) extract local generic features using CNNs (trained in advance) and look into the candidates with the local generic features using the local SVM (L-SVM) classifier (trained using local generic features in advance) to produce the expression class. As a pre-processing, we detect the whole face and facial components from an input image to reduce influence of background. Fig. 1 illustrates the framework of our proposed FER method.

B. Global and local generic features extraction

1) CNN used for generic feature extraction: A generic CNN representation extracted from the last fully-connected layer is explicitly trained to retain information relevant to semantic classes [11]. We thus use this representation as generic features.

To avoid the over-fitting problem, we follow a recent training strategy [15], which shows the effectiveness of supervised fine-tuning with a small data-set on a CNN pre-trained with a large generic objects data-set. Among different CNN architectures, we choose as the pre-trained model, the popularly used modern CNN, i.e., *AlexNet* [20] (Fig. 2), because of two reasons: it has nice trade-off between speed and accuracy [21]; it is trained on the ImageNet data-set which has 15 million images belonging to 22000 categories.

Our used AlexNet comprises 5 convolution layers, 3 max-pooling layers, 2 fully-connected layers and 1 output layer. The input image size is $224 \times 224 \times 3$. The convolution layers contains 96 kernels of size $11 \times 11 \times 3$, 256 kernels of size $5 \times 5 \times 48$, 384 kernels of size $3 \times 3 \times 256$, 384 kernels of size $3 \times 3 \times 192$, and 256 kernels of size $3 \times 3 \times 192$ in this order. The max-pooling layer of size 5×5 and 3×3 follows the first, second, and fifth convolution layers, separately. Each fully-connected layer contains 4096 neurons. For capability with our method, we modify the number of neurons in the output layer from 1000 to the number of facial expression classes (7 for CK data-set and 20 for FaceWarehouse data-set as seen later). We note that our implementation is on Berkeley's Caffe tool (http://caffe.berkeleyvision.org/) using the stochastic gradient descent with momentum$= 0.9$, learning rate$= 0.001$, weight decay$= 0.0005$.

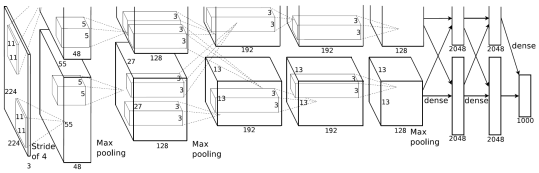

Fig. 2. AlexNet [20] used in our method (we modify only the number of neurons in the output layer).

2) Fine-tuning of AlexNet for generic feature extraction: We start from the pre-trained model of AlexNet to train our data-set. With training images together with the ground-truth in our data-set, AlexNet is finely tuned to our data-set. We can then extract the generic feature vector of 4096 dimensions from the last fully-connected layer of the finely tuned AlexNet.

When whole face images are used for fine tuning, we obtain the global generic feature vector of 4096 dimensions.

We use three facial components: left-eye, right-eye and mouth. To each component, we use one AlexNet for fine tuning; fine tuning is carried out independently of facial components. From one face image, we thus have three feature vectors, each of which is with 4096 dimensions and is extracted from the finely tuned AlexNet corresponding to left-eye, right-eye or mouth. We concatenate the three vectors to one local generic feature vector of 12288 $(= 3 \times 4096)$ dimensions. We note that the finely tuned AlexNets are used in the test phase to extract a global generic feature vector and a local generic feature vector from an input test image.

C. Training SVM classifier

The dimension of feature vectors is large (4096 for the global generic feature vector; 12288 for the local generic feature vector), and accordingly, we employ the linear kernel for SVM (we used LibLinear [22] in our implementation). Since our scenario is multi-class classification, we decompose our M-class problem into a series of two-class problems and employ the widely used one-against-all strategy to train the SVM classifier.

Using the ground-truth and global generic feature vectors extracted from training images in our data-set, we train one SVM classifier. We call the trained SVM classifier "G-SVM". Similarly, using the ground-truth and local generic feature vectors extracted from training images, we train one SVM classifier; we call "L-SVM". We note that the number and the order of classes in G-SVM and L-SVM are the same.

D. Recognizing expressions

We first use global features to generate candidates in expression classes and then use local features to investigate the candidates for re-ranking. This is because though the difference between facial expressions may be little [19], local features are expected to be more effective than global features for discriminating the difference once a small number of candidates are given.

When an input test image is given, our FER method extracts its global generic feature vector using the finely tuned AlexNet (for global generic features) and feeds the vector to G-SVM to generate most plausible candidates in expression classes. We can take top N candidates depending on the scores produced by G-SVM.

Our FER method also extracts the local generic feature vector from the input (test) image using finely tuned AlexNet (for local generic features). Since we have already identified possible N classes in terms of candidates, we only have to investigate the N classes using the local generic feature vector with L-SVM. L-SVM gives us the class that our FER method recognizes. We note that $N = 2$ is sufficient because in our experiments, accuracy in recognition achieved by 2 top-ranked classes obtained via G-SVM was dominant compared with that by the other classes (i.e., degraded than the 2nd ranked class). We set $N = 2$ in our experiments, accordingly.

III. Experiments

A. Experiment setup

1) Data-sets: We evaluated our method using two data-sets available in public: Cohn-Kanade (CK) [23] and FaceWarehouse [24].

CK data-set includes 1917 sequences taken from 182 male and female subjects. The sequences have wide varieties in expression ranging from the neural to the maximum deformation. The expression consists of 8 classes: Anger, Disgust, Fear, Happiness, Sadness, Surprise, Contempt and Neural. In our experiments, for fair comparison with state-of-the-art methods, we eliminated the images classified into Contempt like other methods, and kept the remaining images, yielding 1276 images in total.

FaceWarehouse data-set collects images of 3D facial expressions for visual computing applications and includes 150 individuals aged 7–80 from various ethnic backgrounds. Each person provides 20 different facial expressions including neutral, mouth-opening, smile and kiss. The data-set also includes RGB-D images and 3D models; however, we used only RGB images. The total number of images we used is 3000.

2) Compared methods: We followed the k-fold method for evaluation (we set $k = 3$). Namely, we randomly divided each data-set into three groups ($1/3$ each) and then generated training ($2/3$) and testing ($1/3$) from the three groups iteratively, resulting in three cases of training and testing. We averaged three cases results for evaluation. Below are methods with which we compared our proposed method.

(a) "AlexNet": we input the whole face image to AlexNet [20] where we changed only the number of output neurons to that of facial expression classes. We trained the AlexNet using training images for fine-tuning and applied test images to the finely tuned AlexNet. Note that the soft-max classifier is employed in this case.

(b) "G-AlexSVM (Baseline)": We substitute the SVM classifier for the output layer of AlexNet before fine-tuning (the soft-max classifier is replaced by the SVM classifier). We then trained parameters of AlexNet and SVM together using whole face training images. We applied test images to this trained network. By changing the classifier to

SVM (known to be more effective), we can evaluate the limitation of using global generic features alone; in this sense, this method can be regarded as the baseine.

(c) "L-(AlexNet+SVM)": We prepare three AlexNets (one for one facial component) whose output neurons numbers are changed depending on expression classes. Like the baseline, we input facial component images to its corresponding AlexNet for fine-tuning. We then extract local generic feature vectors from the trained AlexNet. Concatenating local generic feature vectors extracted from training images, we trained the SVM classifier (this is equivalent with L-SVM). We applied test images to the combination of the fine-tuned AlexNets and L-SVM. This method can be used to evaluate the limitation of using local generic features alone and also our separate training for AlexNet and SVM.

(d) For the CK data-set, we compared our method with state-of-the-art methods: GPSNFM [5], LSH-CORF [3], DBN [17] and CSPL [25]. We note that because there are no methods yet that recognize 20 expressions on the FaceWarehouse data-set, we compared our method with only AlexNet, G-AlexSVM and L-(AlexNet+SVM) for the FaceWarehose data-set.

3) Pre-processing: To remove background and unnecessary parts of a face, we detect from an input image the whole face and facial components, separately.

We employed the face detection [26] for the whole face detection. We used OpenCV face detector available in Matlab. We trained the detector using one-third of images in the data-set. When detecting facial components, we used the Face++ library [27] to have facial landmark points such as left_eye_bottom or left_eyebrow_left_conner. Using these landmark points, we detected left-eye, right-eye and mouth as our facial components in the rectangular form, separately. Finally, we re-sized the whole face image to $224 \times 224 \times 3$ and each facial component image to $244 \times 244 \times 3$ where we used the single image super-resolution method [28].

B. Results on CK data-set

Figure 3 illustrates overall accuracy for all the methods, and Fig. 4 shows accuracy of each expression class. Table I, on the other hand, shows the confusion matrix.

We observe that our method achieves higher accuracy than the other methods. For all the expressions, accuracy of our method is equal or higher than that of the others. From Table I, we see that there are some failures in recognition for some expressions, eg. neutral (incorrectly as surprise), fear (incorrectly as anger), and sadness (incorrectly as disgust). Fig. 5 gives some such examples, from which we see that they are sufficiently similar with each other and hard to discriminate.

Table II depicts analysis on which among 1st and 2nd ranked classes, was chosen by our re-ranking procedure. We observe from Table II that among 3.96% of incorrect results by G-AlexSVM (baseline), 1.8% come from images ranked as

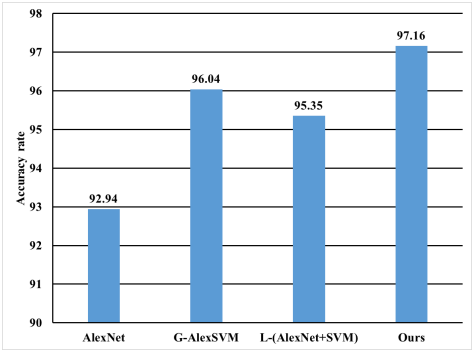

Fig. 3. Overall accuracy on CK data-set.

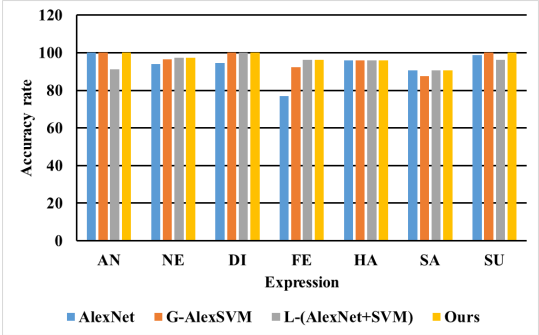

Fig. 4. Accuracy in each expression on CK data-set.

the 2nd class by G-AlexSVM. Among the 1.8%, 1.44% are secured by our re-ranking while we lose $0.32(= 96.04 - 95.72)\%$ from the correct G-AlexSVM results; the gain by our method is $1.12(= 1.44 - 0.32)\%$. As we see, (1) candidates generated by G-AlexSVM are indeed re-ranked using L-SVM, and (2) this re-ranking effectively works to improve the accuracy.

Table III shows accuracy averaged over 7 classes of the state-of-the-art methods. GPSNFM [5], LSH-CORF [3] and DBN [17] all use global features alone while CSPL [25] uses local features alone. Our method achieves 97.16% accuracy which is higher by from 4.16% to 10.36% than the other methods. This indicates that combining global and local features and re-ranking effectively work to outperform the other methods.

C. Results on FaceWarehouse data-set

We illustrate overall accuracy for all the methods, accuracy of each expression class, and the confusion matrix in Figs. 6, 7 and Table IV, respectively. Overall accuracy averaged over 20 classes of our method is 59.2% and it is higher by from 3.7% to 6.5% than the other methods.

When compared with AlexNet (best among the other methods), accuracy of our method is equal or higher for all expressions except for 9, 10, 11, 16 and 18, all of which are with small differences in expression between classes. Our method, on the hand, achieves equal or higher accuracy than G-AlexSVM (baseline) for 15 classes and than L-(AlexNet+SVM) (local features only) for 15 classes. In addi-

Fig. 5. Examples of similar expressions on CK data-set: Neutral and Surprise (top left), Fear and Anger (top right), Sadness and Disgust (bottom left), Happiness and Surprise (bottom right).

Table I. Confusion matrix obtained by our method on CK data-set (numbers indicate percentages) (AN: anger, NE: neural, DI: disgust, FE: fear, HA: happiness, SA: sadness, SU: surprise).

	AN	**NE**	**DI**	**FE**	**HA**	**SA**	**SU**
AN	100	0	0	0	0	0	0
NE	0	97.39	0	0	0	0	2.61
DI	0	0	100	0	0	0	0
FE	3.85	0	0	96.15	0	0	0
HA	0	1.35	0	0	95.95	0	2.7
SA	0	3.13	6.25	0	0	90.63	0
SU	0	0	0	0	0	0	100

tion, our method performs best among the all methods for 12 classes, and achieves higher than 90% accuracy for 4 classes. These indicate that global and local features combination and our re-ranking give better results than without doing so.

Table IV indicates that there are some classification failures, in particular, for expressions 2 and 11 (mutually), 6 and 9 (mutually), 10 and 7 (10 is incorrectly as 7). This is due to the fact that these expressions are sufficiently similar with each other and hard to discriminate (see Fig. 8).

Re-ranking generated candidates allows us to gain 4.6% accuracy where our method involves 10.1% among 15.8% that G-AlexSVM incorrectly ranked as the 2nd class while we lose $5.5(= 54.6 - 49.1)\%$ accuracy for the 1st ranked class by G-AlexSVM (baseline) (Table V).

IV. Conclusions

We proposed a CNN-based method for facial expression recognition. In developing our proposed method, we stand on the point that we do not stick to obtaining the correct expression at one try. Instead, we took the approach that we first generate most plausible candidates and then look into the candidates to finally obtain the expression. In our proposed method, the global generic feature is used with the SVM classifier to generate most plausible candidates in expression class while the local generic feature is used with the SVM classifier to re-rank the candidates for recognition. Our experiments using publicly available data-sets confirmed that our method outperforms in accuracy state-of-the-art methods.

In the future, we will investigate (1) the adaptation ability of our method to more realistic scenarios where face detection is not so reliable and (2) the extension of our method for dealing with videos to exploit temporal cues.

Table II. Contribution of 2 top-ranked classes in candidates to accuracy on CK data-set (numbers in the ground truth row mean ratios of candidates falling into classes).

	1st ranked	**2nd ranked**	**others**	**accuracy**
G-AlexSVM (baseline)	96.04%	–	–	96.04%
Our method	95.72%	1.44%	–	**97.16**%
ground truth	96.04%	1.8%	2.16%	100%

Table III. Accuracy with the state-of-the-arts on CK data-set.

	GPSNFM [5]	LSH-CORF [3]	DBN [17]	CSPL [25]	ours
accuracy	93%	86.8%	90.1%	89.89%	**97.16%**

Acknowledgements

This work was carried out under the NII International Internship Program. This work was in part supported by JST CREST in Japan and by the National Foundation for Science and Technology Development in Vietnam.

References

[1] M. J. A. Calder, G. Rhodes and J. Haxby, *Oxford Handbook of Face Perception*. Oxford, UK: Oxford Univ. Press, 2011.

[2] E. Sariyanidi, H. Gunes, and A. Cavallaro, "Automatic analysis of facial affect: A survey of registration, representation, and recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 6, pp. 1113–1133, 2015.

[3] O. Rudovic, V. Pavlovic, and M. Pantic, "Multi-output laplacian dynamic ordinal regression for facial expression recognition and intensity estimation," in *CVPR*. IEEE Computer Society, 2012, pp. 2634–2641.

[4] K.-C. Huang, S.-Y. Huang, and Y.-H. Kuo, "Emotion recognition based on a novel triangular facial feature extraction method," in *IJCNN*. IEEE, 2010, pp. 1–6.

[5] R. Zhi, M. Flierl, Q. Ruan, and W. B. Kleijn, "Graph-preserving sparse nonnegative matrix factorization with application to facial expression recognition," *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, vol. 41, no. 1, pp. 38–52, 2011.

[6] M. H. Mahoor, M. Zhou, K. L. Veon, S. M. Mavadati, and J. F. Cohn, "Facial action unit recognition with sparse representation," in *FG*. IEEE, 2011, pp. 336–342.

[7] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 12, pp. 2037–2041, Dec. 2006.

[8] C. Orrite, A. Gan, and G. Rogez, "Hog-based decision tree for facial expression classification," in *IbPRIA*, ser. LNCS, H. Arajo, A. M. Mendona, A. J. Pinho, and M. I. Torres, Eds., vol. 5524. Springer, 2009, pp. 176–183.

[9] K. Sikka, T. Wu, J. Susskind, and M. S. Bartlett, "Exploring bag of words architectures in the facial expression domain," in *ECCV Workshops (2)*, ser. LNCS, vol. 7584. Springer, 2012, pp. 250–259.

[10] Y. Lecun and Y. Bengio, *Convolutional Networks for Images, Speech and Time Series*. The MIT Press, 1995, pp. 255–258.

[11] D. K. Nithin and P. B. Sivakumar, "Generic feature learning in computer vision," *Procedia Computer Science*, vol. 58, pp. 202–209, 2015.

[12] Y. Tang, "Deep learning using support vector machines," *CoRR*, vol. abs/1306.0239, 2013.

[13] J. Nagi, G. A. D. Caro, A. Giusti, F. Nagi, and L. M. Gambardella, "Convolutional neural support vector machines: Hybrid visual pattern classifiers for multi-robot systems," in *ICMLA (1)*. IEEE, 2012, pp. 27–32.

[14] B. Fasel, "Head-pose invariant facial expression recognition using convolutional neural networks," in *ICMI*. IEEE Computer Society, 2002, pp. 529–534.

[15] H. Ng, V. D. Nguyen, V. Vonikakis, and S. Winkler, "Deep learning for emotion recognition on small datasets using transfer learning," in *ICMI*. ACM, 2015, pp. 443–449.

[16] W. Chu, F. D. la Torre, and J. F. Cohn, "Selective transfer machine for personalized facial action unit detection," in *CVPR*. IEEE, 2013, pp. 3515–3522.

Table IV. Confusion matrix obtained by our method on FaceWarehouse data-set (numbers indicate percentages; large classification errors are highlighted with colors except for yellow; the same color means mutual classification failures).

class ID	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
0	38	0	2	2	16	2	6	0	8	2	2	2	2	0	0	10	2	0	0	6
1	0	98	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0
2	4	0	48	0	2	2	0	2	0	4	2	28	0	0	0	0	0	2	0	6
3	8	0	10	38	4	12	2	2	6	0	2	8	0	2	0	2	0	0	0	4
4	4	0	4	0	78	0	2	0	2	0	0	2	0	0	0	4	2	2	0	0
5	0	0	2	12	4	44	0	0	10	0	2	0	8	2	0	4	2	0	0	10
6	8	0	0	0	6	2	50	0	2	20	0	6	2	0	0	0	4	0	0	0
7	6	0	2	0	6	2	0	46	4	0	6	8	4	0	0	8	4	0	2	2
8	4	0	6	0	6	2	2	2	34	2	2	2	10	6	0	6	8	4	4	0
9	0	0	6	0	4	2	18	0	0	62	0	6	0	0	0	0	0	0	0	2
10	0	0	0	0	0	0	0	20	8	0	54	10	0	2	0	0	4	0	0	2
11	4	0	26	0	4	0	0	0	6	6	2	38	0	0	0	4	6	2	0	2
12	2	0	0	0	0	4	0	4	2	0	2	0	50	14	0	8	0	0	12	2
13	0	0	0	0	0	0	0	2	0	0	0	0	8	78	6	0	0	0	6	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0	6	92	0	0	0	2	0
15	24	0	6	2	0	2	4	4	6	2	2	10	14	0	0	18	0	2	2	2
16	0	0	4	0	2	2	2	4	4	0	2	10	0	0	0	2	68	0	0	0
17	0	0	0	0	2	0	0	0	0	2	0	0	0	0	0	0	2	94	0	0
18	2	0	0	0	2	4	0	2	2	0	0	0	14	16	0	0	0	0	56	2
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100

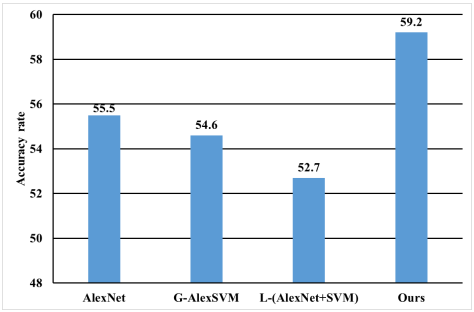

Fig. 6. Overall accuracy on FaceWarehouse data-set.

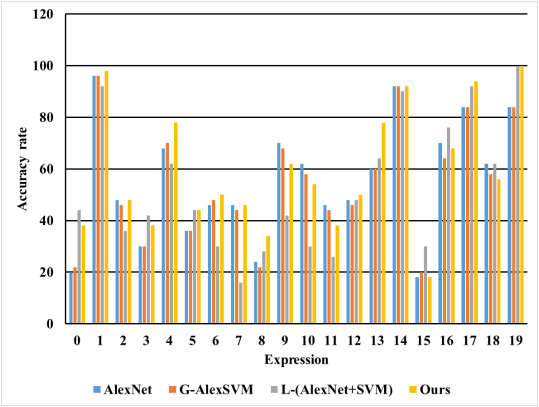

Fig. 7. Accuracy in each expression on FaceWarehouse data-set.

Fig. 8. Examples of similar expressions on FaceWarehouse data-set: 2 and 11 (left), 6 and 9 (right).

Table V. Contribution of 2 top-ranked classes in candidates to accuracy on FaceWarehouse data-set (numbers in the ground truth row mean ratios of candidates falling into classes).

	1st ranked	2nd ranked	others	accuracy
G-AlexSVM (baseline)	54.6%	–	–	54.6%
Our method	49.1%	10.1%	–	**59.2**%
ground truth	54.6%	15.8%	29.6%	100%

[17] M. Ranzato, J. M. Susskind, V. Mnih, and G. E. Hinton, "On deep generative models with applications to recognition," in *CVPR*. IEEE Computer Society, 2011, pp. 2857–2864.

[18] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," *CoRR*, vol. abs/1311.2524, 2013.

[19] M.White, "Parts and wholes in expression recognition," *COGNITION AND EMOTION*, vol. 14, no. 1, pp. 39–60, 2000.

[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[21] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," *CoRR*, vol. abs/1405.3531, 2014.

[22] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "Liblinear: A library for large linear classification," *Journal of Machine Learning Research*, vol. 9, 2008.

[23] T. Kanade, Y. li Tian, and J. F. Cohn, "Comprehensive database for facial expression analysis," in *FG*. IEEE Computer Society, 2000, pp. 46–53.

[24] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou, "Facewarehouse: A 3d facial expression database for visual computing," *IEEE Trans. Vis. Comput. Graph.*, vol. 20, no. 3, pp. 413–425, 2014.

[25] L. Zhong, Q. Liu, P. Yang, B. Liu, J. Huang, and D. N. Metaxas, "Learning active facial patches for expression analysis," in *CVPR*. IEEE Computer Society, 2012, pp. 2562–2569.

[26] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vision*, vol. 57, no. 2, pp. 137–154, May 2004.

[27] M.Inc, http://www.faceplusplus.com/, accessed 12, 2015.

[28] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *CoRR*, vol. abs/1501.00092, 2015.