





Facial Emotion Recognition with Varying Poses and/or Partial Occlusion Using Multi-stage Progressive Transfer Learning

Sherin F. Aly¹  and A. Lynn Abbott² 

¹ Information Technology Department, Institute of Graduate Studies and Research, Alexandria University, Alexandria, Egypt

igers.sherin@alexu.edu.eg

² Bradley Department of Electrical and Computer Engineering, Virginia Tech, Blacksburg, VA, USA

abbott@vt.edu

Abstract. This paper describes the use of multi-stage Progressive Transfer Learning (MSPTL) to improve the performance of automated Facial Emotion Recognition (FER). Our proposed FER solution is designed to work with 2D images, and is able to classify facial emotions with high accuracy in 6 basic categories (happiness, sadness, fear, anger, surprise, and disgust) for both frontal and (more challenging) non-frontal poses. We perform supervised fine-tuning on an AlexNet deep convolutional neural network in a three-stage process, using three FER datasets in succession. The first two training stages are based on FER datasets containing frontal images only. The final training stage uses a third FER dataset that includes non-frontal poses in images that are relatively low in resolution and/or with partial occlusion. Experimental results demonstrate that our proposed MSPTL approach outperforms typical TL and other PTL systems for FER in both frontal and non-frontal face poses. These results are demonstrated using two different testing datasets (VT-KFER and 300W), which corroborates the generality of the proposed solution and its robustness for handling a wide range of varying poses, occlusion, and expression intensities.

Keywords: Facial Emotion Recognition · Deep learning · Transfer learning · Progressive transfer learning

1 Introduction

The automatic detection of emotional cues from the face has many applications in psychology, human-computer interaction, games, and other areas. However, automated analysis of the face is still a challenging problem due to the wide variety of expressions the human face can make and the relatively small size of available datasets. Several researchers have addressed the problem of automated



Fig. 1. Examples of the six basic facial expressions in our testing datasets. The examples include various poses, expression intensities, and partial occlusion.

Facial Emotion Recognition (FER), typically with emphasis on six common emotions: happiness, sadness, fear, anger, disgust, and surprise. (See examples of these emotions in Fig. 1.) Even though strong progress has been made on FER, and most existing FER systems work well on frontal poses and/or with small variations in head pose [7, 8, 25, 27], increased accuracy and robustness against pose variations is still desired. The latter has proven to be more challenging [9, 18] with only a few researchers having explicitly tested their approach on non-frontal data [1, 6, 11, 18, 22].

In this paper we introduce a novel method that, by adding new levels to proven transfer learning techniques, successfully achieves higher accuracy in both frontal and non-frontal poses. The next section of this paper contains related work. Section 3 presents the proposed method. Section 4 presents the experimental setup. Section 5 presents the experimental results. Finally, Sect. 6 presents concluding remarks.

2 Related Work

Several researchers have employed deep convolutional neural networks (CNNs) for the task of FER [4, 5, 12, 14, 19–21, 30] to avoid the traditional feature extraction approaches such as HOG which are relatively complex and time-consuming [1, 25, 26]. However, training CNN from scratch requires a large amount of labeled training data. To address this problem, it is common to use the weights of a pre-trained CNN as the initial state for further training in what is called transfer learning (TL) [19]. TL has provided a reasonable compromise that enhances the accuracy of CNN-based FER systems without requiring very large datasets. However, the “forgetting effect” of TL and the model initialization process, given learned weights from a sequence of related tasks, limit the use of TL [23].

Therefore, a progressive transfer learning (PTL) approach was recently developed to alleviate these limitations [23]. Progressive networks are used to train sequences of tasks by freezing the previously trained tasks and using their intermediate representations as inputs into the new network. PTL therefore prevents the “forgetting effect” of TL by freezing and preserving the source task weights. Our approach extends the use of PTL for FER by adding additional stages that, in turn, increase the accuracy of the overall system.

Regarding the use of CNN for FER, most of the existing approaches have been tested on frontal poses only, non-frontal poses expressions with small datasets, non-varying expression intensity datasets, and/or using 1 or 2 levels of knowledge transfer [19, 23, 27, 29].

Mavani et al. [19] have fine-tuned the AlexNet CNN model [15] using two widely used facial expression datasets (CFEE and RaFD) of the 6 basic expressions plus neutral. Their model yielded test accuracies of 74.79% on CFEE and 95.71% on RaFD. In [23], Ng et al. proposed a PTL scheme for the facial expression recognition of the basic expressions presented by the EmotiW challenge. They also used AlexNet CNN to progressively transfer knowledge to the facial expression task. They first performed fine tuning of AlexNet using the FER28 dataset [10]. Then, a second fine-tuning step on the dataset of interest was performed. Their approach showed significant improvement in accuracy (up to 16%) over the baseline.

This paper proposes a multi-stage progressive transfer learning-based approach of 3 stages of fine-tuning for the recognition of the six basic expressions which goes beyond the TL and PTL approaches described above. We trained and tested our approach on four widely used FER datasets, JAFEE, CK+, VT-KFER and 300W. These datasets include facial expressions with several head poses, intensities and occlusion. We employ the AlexNet CNN, trained initially on the ImageNet database, as our base architecture. We selected AlexNet as the underlying Deep Neural Network (DNN) architecture due to its known reputation and quality for image classification. Moreover, this is an architecture that has been thoroughly studied and its performance is generally well understood in the community. The use of the proposed methodology is, however, not limited to AlexNet. Configuration and use of other DNN architectures is straightforward.

Contributions: There are three main contributions of this work: (1) A novel FER system that is more robust to pose variations; (2) in contrast to existing TL [19] or PTL approaches [23], our approach transfers knowledge progressively using 3 stages with varying content in gender and pose, and has been explicitly tested on non-frontal poses; and, relying only on 2D images, (3) our proposed system outperforms other 2D-based and 3D-based FER systems by more than 10% and some 2D+3D-based systems by more than 17%, when tested on VT-KFER.

3 Proposed Method

Our proposed approach is composed of three main steps: (1) preprocessing the input datasets to extract the face data and prepare the extracted faces for the deep learning step, (2) multi-stage progressive transfer learning and CNN fine tuning, and (3) classification. Each of these steps is described next.

3.1 Preprocessing

We employed 4 FER datasets in our experiments, two for training (JAFEE [17] and CK+ [13, 16]) and two for testing (VT-KFER [2] and 300W [24]). A hierarchical face detection (HFD) approach is applied on each image of the datasets. As part of this approach, a group sparse learning method [28] that automatically selects the most salient facial landmarks and thus extracts the face location is applied. If no face is detected, then an approach based on mixtures of trees with a shared pool of parts [31] is applied. This approach models every facial landmark as a part and uses global mixtures to capture topological changes due to viewpoint. Finally, if no face is detected in a particular image, we manually select the face region. (Note that manual processing, if any, is done on the training datasets to ensure that the largest dataset is used; manual processing is not done online.) The Kinect SDK is employed for automatic face detection on the VT-KFER testing dataset. Similarly, cropped faces are provided for the 300W testing dataset. All extracted faces are resized to $227 \times 227 \times 3$. Gray scale images are concatenated into 3 channels to be compatible with the input data size of AlexNet.

3.2 Multi-stage Progressive Transfer Learning (MSPTL)

We propose a multi-stage progressive fine-tuning transfer learning FER system of 3 stages based on AlexNet architecture. Our system not only uses the transferred knowledge from training over 1M images of the ImageNet dataset, but also fine tunes this knowledge, progressively, using widely used FER datasets. The progressive fine-tuning here means to adjust the weights of the pretrained network by continuing the backpropagation first on a simple FER dataset, and then repeating the same process on a larger dataset with more variations in gender, pose, and lightening conditions.

In the typical TL paradigm, the transferred knowledge is directly fine-tuned on the dataset of interest (either alone or combined with other datasets in training as is shown in Fig. 2). As opposed to the typical TL paradigm, in MSPTL, the pretrained CNN weights are adjusted using two FER related datasets, JAFEE and CK+, in two separate stages, before being fine-tuned to the dataset of interest in the final stage. With every fine tuning step, the network weights are adjusted with progressive knowledge from the corresponding dataset. In contrast to [23] (shown in Fig. 3), which employs PTL fine-tuning using one large dataset, we employ 3 stages of gradual fine-tuning starting on relatively simple dataset and ending by fine-tuning the model on a dataset with varying poses, varying intensities, and occluded facial expressions. Proposed MSPTL is shown in Fig. 4.

PTL is applied on three stages as follows. First, we transfer the knowledge from AlexNet to a new CNN and fine tune it on a small FER dataset of female-only actors (JAFEE). Then, in stage 2, the resulting CNN model is used to transfer the knowledge gained from AlexNet and JAFEE to a new CNN which is further trained on a larger FER dataset (CK+) that has a larger number of both male and female subjects. Then the resulting model is further fine-tuned and then tested on the datasets of interest.

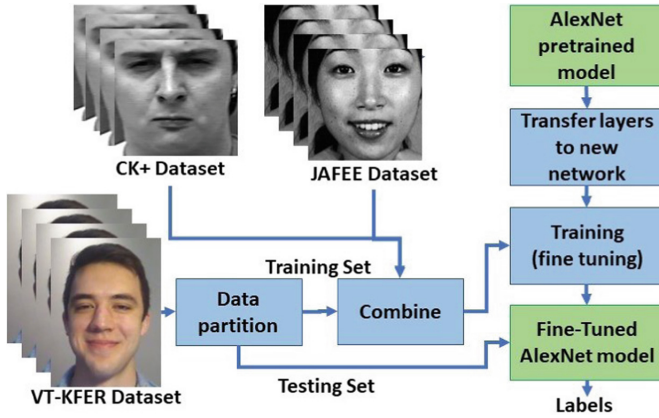


Fig. 2. Transfer learning (TL) approach as proposed in [19], where transferred knowledge is directly obtained from the dataset of interest.

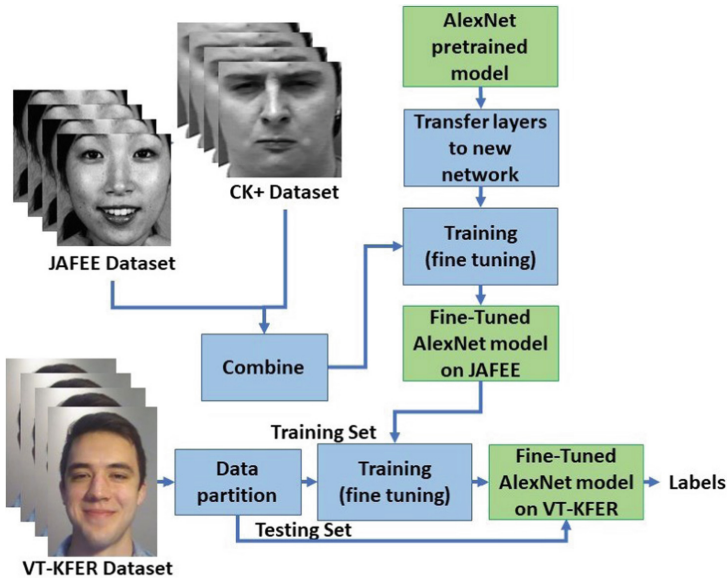


Fig. 3. Progressive transfer learning (PTL) approach as proposed in [23] where transferred knowledge is acquired indirectly using combined and related datasets before the one of interest.

Transfer Layers to a New Network: The first step in MSPTL is to construct a new CNN of N layers where the first $N - i$ layers are transferred from the pretrained model and the last i layers are constructed according to the number of expressions that we want to recognize. In this work, we adopt the AlexNet network as our pretrained CNN. AlexNet comprises 25 layers (illustrated in Fig. 5). There are 8 layers with learnable weights: 5 convolutional layers, and 3 fully connected layers.

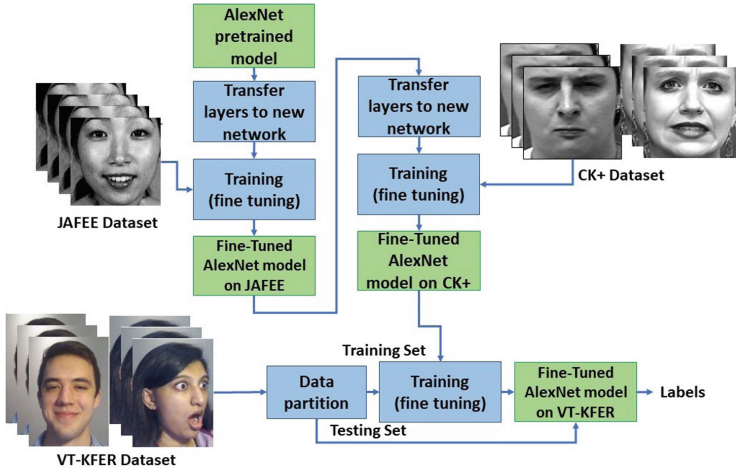


Fig. 4. Our proposed multi-stage progressive transfer learning (MSPTL) approach with separated tuning steps for each training dataset. Transferred knowledge is tuned progressively using small then bigger datasets. The progression is also with respect to dataset contents, where we first fine-tune the model using female-only frontal only FER data then using mix of gender and pose datasets.

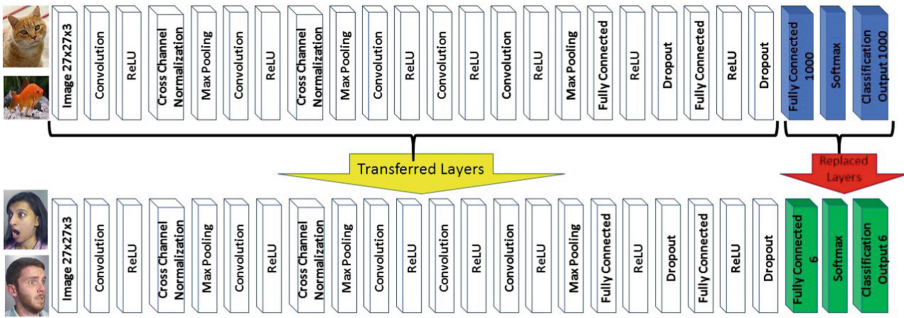


Fig. 5. Transferring layers to a new network. The original pretrained model architecture is shown on top, while the new architecture after transferring layers is shown at the bottom.

We transfer the first 23 layers from AlexNet and replace the last fully-connected layer with a new one that outputs the desired number of classes. Specifically, we replaced the last 3 layers of AlexNet that recognize 1000 classes with three layers for 6 expressions. The three new layers are (1) fully connected that classifies 6 classes, (2) softmax, and (3) classification to adjust the softmax output to a class labels format. The new network is illustrated in the lower part of Fig. 5. We apply this step at each stage in the PTL paradigm. At stage 1 we apply it on AlexNet pretrained network. In stage 2 we apply it on the pretrained network on JAFEE dataset. In stage 3 we apply it on the pretrained network on CK+ dataset.

Deep Learning (Fine Tuning) and Classification: The training is applied in three phases. First, we train the new CNN, composed of transferred layers from AlexNet, on the JAFEE dataset. This training step is for fine tuning the weights of AlexNet to the new classes in JAFEE. For transfer learning, we keep the features from the early layers of the pretrained network (the transferred layer weights) and initialize the weights of the replaced layers randomly. Then we conduct the training by setting a small global learning rate of 10^{-4} to slow down learning in the transferred layers based on the intuition that we may already be close to a good result. We also increased the learning rate for the fully connected layer to speed up learning in the new added layers. This combination of learning rate settings results in fast learning only in the new layers and slower learning in the other layers. We set the number of epochs (i.e., training cycles on the entire dataset) to 4. When performing transfer learning, we do not need to train for as many epochs. During training, the network is validated once per epoch, and automatically stops training if the validation loss stops improving. Stochastic Gradient Descent with Momentum (SGDM) is employed as the optimization algorithm. SGDM updates the weights and biases (parameters) by taking small steps in the direction of the negative gradient of the loss function to minimize the loss. It updates the parameters using a subset of the data every iteration.

At stages 1 and 2 the resulting fine-tuned networks are tested on images randomly selected from JAFEE and CK+, respectively, in order to select the best networks for stages 2 and 3. The final fine-tuned CNN in stage 3 is then used for classifying the six basic expressions in the testing dataset.

4 Experimental Setup

We employed the JAFEE [17] and Cohn-Kanade-plus (CK+) [13,16] datasets for training and VT-KFER [2] and 300W [24] datasets for testing.

JAFEE Dataset: It contains 213 posed images of 7 frontal facial expressions (6 basic facial expressions + neutral) by 10 Japanese female models. All images are in gray scale.

CK+ Dataset: It includes 4,001 posed images in 593 sequences from 123 subjects. Each sequence includes one of the 6 basic expressions in frontal pose only and starts with a neutral face. CK+ includes both gray-scale and RGB images.

VT-KFER Dataset: The dataset includes 11,619 posed images of the 6 basic expressions (plus neutral), in 3 different intensities, captured from 32 subjects by the Kinect 1.0. The data includes expressions performed in frontal, right, and left poses, with 4,732 frontal frames and 5,066 non-frontal. In the testing set, there are 1,005 frontal and 1,062 non-frontal images all in RGB format.

Faces-in-the-Wild (300W): The dataset consists of 300 indoor and 300 outdoor in-the-wild images. It covers a large variation of illumination conditions, poses, occlusion, and face size. For our experiments, we only selected the faces of the six basic expressions, happiness, sadness, surprise, disgust, fear, and anger. A total of 240 images were selected.

To train the first two stages, we used a holdout cross-validation strategy where 90% of the JAFEE and CK+ datasets were randomly selected for training and 10% for testing. To train the third stage on VT-KFER, we used “leave- p -sequence-out” cross-validation, where $100 - p\%$ ($p=20\%$) sequences of VT-KFER were randomly selected for training the fine-tuned model in stage 3 and the rest for testing. To train the third stage on 300W, we used hold-out cross validation where 80% of 300W were randomly selected for training and 20% for testing.

5 Results and Discussion

We compare our proposed approach to the typical TL paradigm [19] and to the PTL paradigm of [23]. For fairness of comparison, and since we used JAFEE and CK+ for training our model, we combined JAFEE and CK+ datasets with the dataset of interest to train the TL model of [19]. For PTL approach in [23], we used JAFEE and CK+ to fine-tune the model in the first stage while the dataset of interest is used to fine-tune the model in the second stage. To prove the generality of our approach, we tested our approach on two challenging datasets, namely, the VT-KFER and 300W. Experiments were conducted on frontal, non-frontal, and both frontal and non-frontal expressions in VT-KFER to compare the effect of the three tested approaches with respect to the pose.

Our experimental results, illustrated in Fig. 6, show that progressive MSPTL outperforms the typical TL paradigm [19] and the PTL paradigm of [23], especially for non-frontal poses. See Fig. 7 for example frontal and non-frontal expressions from 300W dataset where our MSPTL outperforms both TL and PTL. In the frontal pose, MSPTL approach achieved 84.2% vs. 81.2% for TL and 83% for PTL.

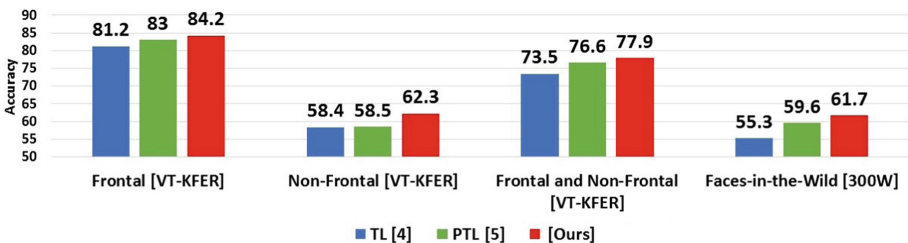


Fig. 6. Accuracy of TL [19] vs. PTL [23] vs. our proposed MSPTL approach tested on VT-KFER and 300W. Columns in groups 1 to 3 shows the testing results on VT-KFER when frontal, non-frontal, and both frontal and non-frontal poses were employed, respectively. The accuracy is based on training on JAFEE, CK+ and VT-KFER and tested using VT-KFER using leave- p -sequence-out cross validation. Columns in group 4 illustrates the testing results on the 300W dataset. The accuracy is based on training on JAFEE, CK+ and 300W and tested using 300W using holdout cross validation.



Fig. 7. Sample testing results of TL, PTL, and MSPTL [Ours] approaches on the 300W dataset. The text describes the predicted expression of each image below it with corresponding recognition probability.

MSPTL achieved a 62.3% accuracy on non-frontal poses while TL and PTL have 58.4% and 58.5%, respectively, with increase of around 4% compared to 1% in the frontal pose case. Overall, on VT-KFER, MSPTL showed better performance in all poses data with accuracy of 77.9% compared to 73.5% and 76.6% for TL and PTL, respectively. The results of experiments with 300W dataset show the superior performance of our MSPTL approach as well, with 61.7% accuracy vs. 55.3% obtained with TL, and 59.6% obtained with PTL approach in [23].

Quantitative comparison with the state-of-the-art FER systems tested on all poses data in VT-KFER is given in Table 1. Although MSPTL is based on 2D data only, it shows better performance than all other 2D-based and 3D-based FER systems. It also showed better performance than most of the 2D+3D-based systems with comparative results to the best 2D+3D-based system proposed in [1].

Table 1. Quantitative comparison with state-of-the-art FER systems tested on VT-KFER using leave- p -sequence-out cross validation, where $p=20\%$. Note that our MSPTL solution outperforms all 2D-only and 3D-only systems. Our approach also achieved competitive results to the state-of-the-art 2D+3D system [1] although it relied only on 2D images.

System	Modality	Leave- p -sequence-out
[3]	3D	49%
[25]	2D	59%
[2]	2D+3D	60%
[26]	2D	67%
[1]	2D+3D	80%
Proposed MSPTL	2D	78%

6 Conclusion

Although much progress has been made in recent years in the recognition of human facial expressions, most existing systems perform well only for frontal head poses. This paper has presented a new FER system that utilizes multi-stage progressive transfer learning (MSPTL) to classify the six basic expressions in varying poses, varying intensities, and with partial occlusion. Our MSPTL approach has led to a performance increase of about 1% and 4% for frontal and non-frontal head poses in VT-KFER, respectively, and 2% on 300W facial expressions dataset taken in the wild, as compared to previous transfer learning approaches. Overall, the system achieved an accuracy of 77.9% on all poses for the VT-KFER dataset, using leave- p -sequence-out cross validation and 62% on the 300W dataset using holdout cross validation. These results have surpassed exiting systems tested on VT-KFER dataset, utilizing 2D or 3D information only. In addition, the system is either better or almost as good as all published 2D+3D systems.

References

1. Aly, S., Abbott, A.L., Torki, M.: A multi-modal feature fusion framework for Kinect-based facial expression recognition using dual kernel discriminant analysis (DKDA). In: IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 71–78 (2016)
2. Aly, S., Trubanova, A., Abbott, L., White, S., Youssef, A.: VT-KFER: a kinect-based RGBD+time dataset for spontaneous and non-spontaneous facial expression recognition. In: IEEE International Conference on Biometrics (ICB), pp. 90–97, May 2015. <https://doi.org/10.1109/ICB.2015.7139081>
3. Aly, S., Youssef, A., Abbott, L.: Adaptive feature selection and data pruning for 3D facial expression recognition using the Kinect. In: IEEE International Conference on Image Processing (ICIP), pp. 1361–1365 (2014)
4. Bazrafkan, S., Nedelcu, T., Filipczuk, P., Corcoran, P.: Deep learning for facial expression recognition: a step closer to a smartphone that knows your moods. In: International Conference on Consumer Electronics (ICCE), pp. 217–220, January 2017. <https://doi.org/10.1109/ICCE.2017.7889290>
5. Chen, J., Chen, Z., Chi, Z., Fu, H.: Emotion recognition in the wild with feature fusion and multiple kernel learning. In: International Conference on Multimodal Interaction, ICMI 2014, pp. 508–513. ACM, New York (2014)
6. Eleftheriadis, S., Rudovic, O., Pantic, M.: Discriminative shared gaussian processes for multiview and view-invariant facial expression recognition. IEEE Trans. Image Process. **24**(1), 189–204 (2015). <https://doi.org/10.1109/TIP.2014.2375634>
7. Fasel, B., Luetttin, J.: Automatic facial expression analysis: a survey. Pattern Recognit. **36**(1), 259–275 (2003)
8. Garg, A., Bajaj, R.: Facial expression recognition & classification using hybridization of ICA, GA, and neural network for human-computer interaction. J. Netw. Commun. Emerg. Technol. (JNCET) **2**(1) (2015)
9. Güney, F., Arar, N.M., Fischer, M., Ekenel, H.K.: Cross-pose facial expression recognition. In: IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FGR), pp. 1–6, April 2013. <https://doi.org/10.1109/FG.2013.6553814>

10. Goodfellow, I.J., et al.: Challenges in representation learning: a report on three machine learning contests. ArXiv e-prints, July 2013
11. Hu, Y., Zeng, Z., Yin, L., Wei, X., Tu, J., Huang, T.S.: A study of non-frontal-view facial expressions recognition. In: IEEE International Conference on Pattern Recognition (ICPR), pp. 1–4 (2008)
12. Kahou, S.E., et al.: Combining modality specific deep neural networks for emotion recognition in video. In: International Conference on Multimodal Interaction, ICMI 2013, pp. 543–550. ACM, New York (2013). <https://doi.org/10.1145/2522848.2531745>
13. Kanade, T., Cohn, J.F., Tian, Y.: Comprehensive database for facial expression analysis. In: IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580), pp. 46–53 (2000). <https://doi.org/10.1109/AFGR.2000.840611>
14. Karali, A., Bassiouny, A., El-Saban, M.: Facial expression recognition in the wild using rich deep features. In: International Conference on Image Processing (ICIP), pp. 3442–3446, September 2015. <https://doi.org/10.1109/ICIP.2015.7351443>
15. Krizhevsky, A., Sutskever, I., Hinton, G.: ImageNet classification with deep convolutional neural networks. In: Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q. (eds.) *Advances in Neural Information Processing Systems*, pp. 1097–1105. Curran Associates, Inc. (2012). <http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf>
16. Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I.: The extended Cohn-Kanade dataset (CK+): a complete dataset for action unit and emotion-specified expression. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, pp. 94–101, June 2010. <https://doi.org/10.1109/CVPRW.2010.5543262>
17. Lyons, M., Akamatsu, S., Kamachi, M., Gyoba, J.: Coding facial expressions with Gabor wavelets. In: IEEE Inter. Conference on Automatic Face and Gesture Recognition, pp. 200–205, April 1998. <https://doi.org/10.1109/AFGR.1998.670949>
18. Malawski, F., Kwolek, B., Sako, S.: Using kinect for facial expression recognition under varying poses and illumination. In: Ślęzak, D., Schaefer, G., Vuong, S.T., Kim, Y.-S. (eds.) *AMT 2014*. LNCS, vol. 8610, pp. 395–406. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-09912-5_33
19. Mavani, V., Raman, S., Miyapuram, K.: Facial expression recognition using visual saliency and deep learning. CoRR abs/1708.08016 (2017). <http://arxiv.org/abs/1708.08016>
20. Mayya, V., Pai, R., Pai, M.: Automatic facial expression recognition using DCNN. *Proc. Comput. Sci.* **93**(Suppl. C), 453–461 (2016). <https://doi.org/10.1016/j.procs.2016.07.233>, <http://www.sciencedirect.com/science/article/pii/S1877050916314752>. International Conference on Advances in Computing and Communications
21. Mollahosseini, A., Chan, D., Mahoor, M.H.: Going deeper in facial expression recognition using deep neural networks. In: IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1–10, March 2016. <https://doi.org/10.1109/WACV.2016.7477450>
22. Moore, S., Bowden, R.: Local binary patterns for multi-view facial expression recognition. *Comput. Vis. Image Underst.* **115**(4), 541–558 (2011)
23. Ng, H., Nguyen, V.D., Vonikakis, V., Winkler, S.: Deep learning for emotion recognition on small datasets using transfer learning. In: International Conference on Multimodal Interaction, ICMI 2015, pp. 443–449. ACM, New York (2015). <https://doi.org/10.1145/2818346.2830593>

24. Sagonas, C., Antonakos, E., Tzimiropoulos, G., Zafeiriou, S., Pantic, M.: 300 faces in-the-wild challenge: database and results. *Image Vis. Comput.* **47**, 3–18 (2016). <https://doi.org/10.1016/j.imavis.2016.01.002>
25. Shan, C., Gong, S., McOwan, P.: Facial expression recognition based on local binary patterns: a comprehensive study. *Image Vis. Comput.* **27**(6), 803–816 (2009)
26. Valstar, M., Jiang, B., Mehu, M., Pantic, M., Scherer, K.: The first facial expression recognition and analysis challenge. In: *International IEEE Conference on Automatic Face & Gesture Recognition & Workshops (FG)*, pp. 921–926 (2011)
27. Xu, M., Cheng, W., Zhao, Q., Ma, L., Xu, F.: Facial expression recognition based on transfer learning from deep convolutional networks. In: *International Conference on Natural Computation (ICNC)*, pp. 702–708, August 2015. <https://doi.org/10.1109/ICNC.2015.7378076>
28. Yu, X., Huang, J., Zhang, S., Yan, W., Metaxas, D.: Pose-free facial landmark fitting via optimized part mixtures and cascaded deformable shape model. In: *International Conference on Computer Vision (ICCV)*, December 2013
29. Zhang, F., Yu, Y., Mao, Q., Gou, J., Zhan, Y.: Pose-robust feature learning for facial expression recognition. *Front. Comput. Sci.* **10**(5), 832–844 (2016). <https://doi.org/10.1007/s11704-015-5323-3>
30. Zhang, T.: Facial expression recognition based on deep learning: a survey. In: Khafa, F., Patnaik, S., Zomaya, A.Y. (eds.) *IISA 2017. AISC*, vol. 686, pp. 345–352. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-69096-4_48
31. Zhu, X., Ramanan, D.: Face detection, pose estimation, and landmark localization in the wild. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2879–2886 (2012)