Patch-Gated CNN for Occlusion-aware Facial Expression Recognition

Yong Li[1,2], Jiabei Zeng[1], Shiguang Shan[1,2,3] and Xilin Chen[1,2]
[1]Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS),
Institute of Computing Technology, CAS, Beijing 100190, China
[2]University of Chinese Academy of Sciences, Beijing 100049, China
[3]CAS Center for Excellence in Brain Science and Intelligence Technology
{yong.li, jiabei.zeng}@vipl.ict.ac.cn, {sgshan, xlchen}@ict.ac.cn

***Abstract*—Facial expression recognition in the wild is challenging due to various un-constrained conditions. Although existing facial expression classifiers have been almost perfect on analyzing constrained frontal faces, they fail to perform well on partially occluded faces that are common in the wild. In this paper, we propose an end-to-end trainable Patch-Gated Convolution Neutral Network (PG-CNN) that can automatically percept the occluded region of the face and focus on the most discriminative un-occluded regions. To determine the possible regions of interest on the face, PG-CNN decomposes an intermediate feature map into several patches according to the positions of related facial landmarks. Then, via a proposed Patch-Gated Unit, PG-CNN reweighs each patch by the unobstructed-ness or importance that is computed from the patch itself. The proposed PG-CNN is evaluated on two largest in-the-wild facial expression datasets (RAF-DB and AffectNet) and their modifications with synthesized facial occlusions. Experimental results show that PG-CNN improves the recognition accuracy on both the original faces and faces with synthesized occlusions. Visualization results demonstrate that, compared with the CNN without Patch-Gated Unit, PG-CNN is capable of shifting the attention from the occluded patch to other related but unobstructed ones. Experiments also show that PG-CNN outperforms other state-of-the-art methods on several widely used in-the-lab facial expression datasets under the cross-dataset evaluation protocol.**

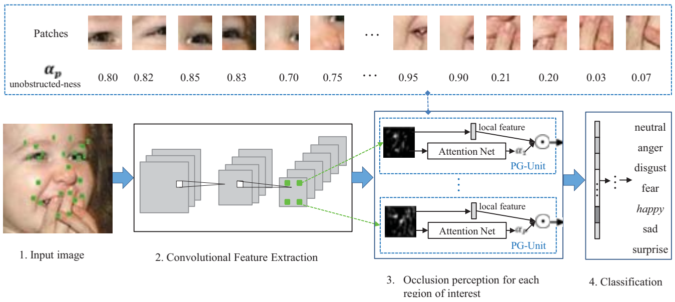


Fig. 1. Illustration of the proposed PG-CNN for occlusion-aware facial expression recognition. During Part 3, PG-CNN extracts 24 regions of interest from the intermediate feature maps. For each region, a specific Patch Gated Unit (PG-Unit) is learnt to reweigh the local representation according to the region's "*unobstructed-ness*" (to what extent the patch is occluded). Then, the weighted representations are concatenated and passed to the classification part.

I. Introduction

Facial expression recognition (FER) has received significant interest from computer scientists and psychologists over recent decades, as it holds promise to an abundance of applications, such as human computer interaction, affect analysis, and mental health assessment. Although many facial expression recognition systems have been proposed and implemented, majority of them are builded on images captured in controlled environment, such as on CK+ [1], MMI [2], Oulu-CASIA [3], and other lab-collected datasets. The controlled faces are frontal and without any occlusions. The FER systems that perform perfectly on the lab-collected datasets, is highly possible to perform poorly when recognizing human's expressions under natural and un-controlled conditions.

To fill the gap between the recognition accuracy on the controlled faces and un-controlled faces, researchers make efforts on collecting large-scale facial expression datasets in the wild ([4], [5]). Despite the usage of data from the wild, facial expression recognition is still challenging due to the existence of partially occluded faces. It it non-trivial to address the occlusion issue because occlusion varies in the occluders and their positions. The occlusion may caused by hair, glasses, scarf, breathing mask, hands, arms, food, and other objects that could be placed in front of the face in daily life. These objects may block the eyes, mouth, part of the cheek, and any other part of the face. The variability of occlusion cannot be fully covered by restricted amounts of data and will inevitably lead the recognition accuracy to decrease.

To address the issue of occlusion, we propose a Patch-Gated Convolution Neutral Network (PG-CNN), mimicing the way that human recognize the facial expression. Intuitively, human recognize the facial expression based on certain patches of the face. When some regions of the face are blocked (e.g., the lower left cheek), human may judge the expression according to the symmetric part of face (e.g., the lower right cheek), or other highly related facial region (e.g., region around the eyes or mouth). Inspired by the intuition, PG-CNN automatically percepts the blocked facial patch and pays attentions mainly to the unblocked and informative patches. Fig. 1 illustrates the main idea of PG-CNN. The patches of interest are cropped from the last convolution feature maps according to the positions of the related facial landmarks. For each patch, a Patch-Gated Unit (PG-Unit) is learned to reweigh the patch's local representation by its unobstructed-ness that is computed from the patch itself. As can be seen in Fig. 1, the last four visualized patches are blocked and thus they have low unobstructed-ness (α_p). Then, the weighted representations are concatenated and used in the classification part.

The contributions of this work are summarized as follows:

1) We propose a PG-CNN to recognize facial expressions with partially occluded faces. PG-CNN can automatically percept the occluded region of the face and focus on the most informative and un-blocked regions. To the best of our knowledge, PG-CNN is the first end-to-end trainable framework that addresses occlusions in facial expression recognition.
2) Visualized results show that PG-Unit (the crucial part of PG-CNN) is effective in perceiving the occluded facial patch. PG-Unit is capable to learn a low weight for the blocked patch and a high weight for an unblocked and informative one.
3) Experimental results demonstrate the advantages of the proposed PG-CNN over other state-of-the-art methods on two large in-the-wild facial expression datasets and several popular in-the-lab datasets, under settings with either partially occluded or non-occluded faces.

II. Related Work

We review the previous work considering two aspects that are related to ours, i.e., the similar tasks (facial analysis with occluded faces) and related techniques (attention mechanism).

A. Methods towards facial occlusions

For facial analysis tasks, occlusion is one of the inherent challenges in the real world FER and other facial analysis tasks, e.g., facial recognition, age estimate, gender classification, etc. Previous approaches that address facial occlusions can be classified into two categories: holistic-based or part-based methods.

Holistic-based approaches treat the face as a whole and do not explicitly divide the face into sub-regions. They usually improve the robustness of the features through designated regularization, e.g., L_1-norm [6]. This idea is also suitable for non-facial occlusions, for example, Elad et al. [7] proposed to mutually re-weight L_1 regularization in an end-to-end framework to deal with arbitrary occlusion in object recognition. Another holistic way is to reconstruct a complete face from the occluded one([8], [9]). These reconstruction based methods rely on the training data with varied occlusion conditions. Specially, Irene et al. [10] analysed how partial occlusion affects FER performance in detail.

Part-based methods explicitly divide the face into several overlapped or non-overlapped segmentations. To determine the patches on the face, existing works either divide the facial image into several uniform parts([11], [12]), or get the patches around the facial landmarks([13], [14]), or get the patches by a sampling strategy [15], or explicitly detect the occluders([16], [17]). Then, the part-based methods detect and compensate the missing part ([18], [19]), or re-weight the occluded and non-occluded patches differently([13], [17]), or ignore the occluded part([15], [16]). We adopt the way of the part-based methods because they successfully incorporate the priors information of the structure of human faces and have a better interpretation. The proposed PG-CNN is end-to-end trainable. It learns occlusion patterns from data and encodes them with model weights. Therefore, it is preferable to handle arbitrary kind of occluder at any position in front of the face.

B. CNN with attention

Recently, attention models have been successfully applied in many computer vision tasks, including fine-grained image recognition [20], image caption [21], visual question answering [22], person re-identification [23], etc. Usually attention can be modeled as a region sequence in an image. An RNN/LSTM model is adopted to predict the next attention region based on current attention region's location with visual features.

Moreover, zheng et al. [20] adopted channel grouping subnetwork to cluster convolutional feature maps into groups according to peak responses of maps, which do not need part annotations but is not suitable for FER under occlusion. For false responses caused by occluders will inevitably disturb channels clustering. Zhao et al. [23] estimated multiple 2-dimensional attention maps, they have equal spatial size of convolutional feature maps to weight. This approach is straightforward but do not take occlusion patterns into consideration.

Attention models allow for salient features to dynamically come to forefront as needed. This is especially beneficial when there is some occlusion or clutter in an image. They also help interpret the results by visualizing where the model attends to for certain tasks. Compare with existing attention models, Our approach adopts facial landmarks for region decomposition, which is straightforward and easily implemented. Meanwhile, PG-CNN adopts CNN based Patch-Gated Unit for occlusion perception, guiding the model to shift attention to informative as well as unblocked facial patches.

III. Proposed method

A. Method overview

We propose a Patch-Gated CNN (PG-CNN) for facial expression recognition with partially occlusions. To address the occlusion issue, PG-CNN is end-to-end trainable with two key schemes: region decomposition and occlusion perception.

Figure 2 illustrates the framework of the proposed PG-CNN. As can be seen in Fig. 2 , the network takes input as a facial image. The image is fed into VGG net and is represented as some feature maps. Then, PG-CNN decomposes the feature maps of the whole face to 24 sub-feature-maps for 24 local patches. Each local patch is encoded as a weighted vector of local feature by a Patch-Gated Unit (PG-Unit). PG-Unit computes the weight of each patch by an Attention Net, considering its obstructed-ness (to what extent the patch is occluded). Finally, the weighted local features are concatenated and serve as a representation of the occluded face. Three fully connected layers are followed to assign the face to one of the emotional categories. PG-CNN is optimized by minimizing the soft-max loss.

Below, we present the details of the two key schemes, region decomposition and occlusion perception, in PG-CNN.

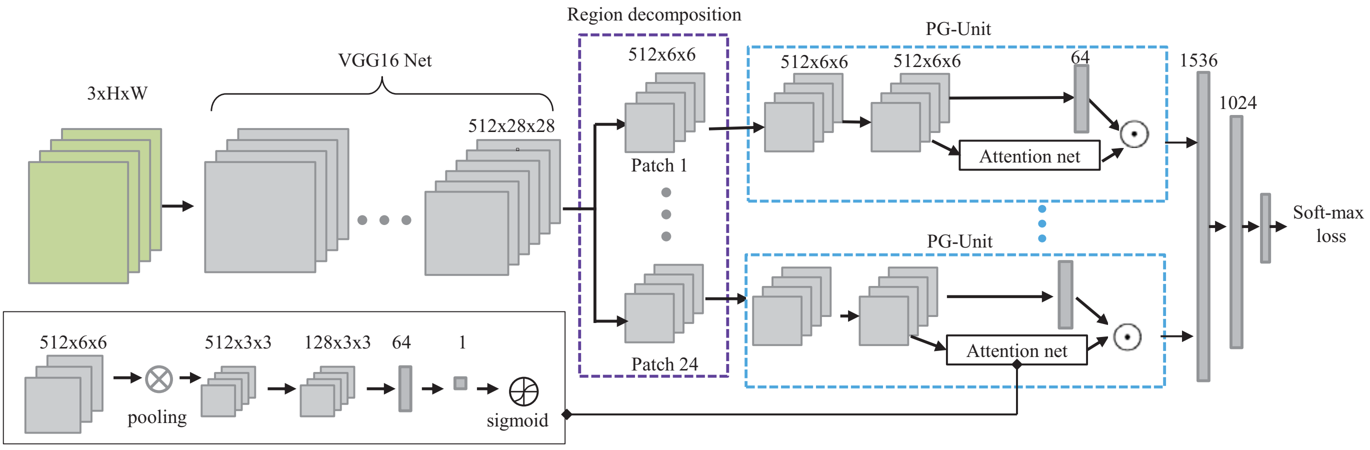


Fig. 2. Framework of the proposed PG-CNN. PG-CNN takes a facial image as input and encodes the image with VGG-16 Net. The feature maps from the last convolution layer (conv4_2 in VGG [24]) are cropped into 24 local patches through a region decomposition scheme. Each patch is then processed by a Patch-Gated Unit (PG-Unit). PG-Unit encodes a patch by a vector-shaped feature and estimates how informative the patch is through an Attention net. The soft-max loss is attached at the end. Parameters in the overall network are learned by minimizing the soft-max loss.

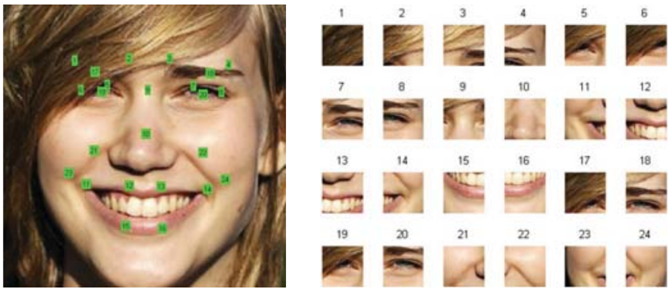


Fig. 3. Region decomposition of the face. The left figure shows the selected landmarks (green dots with numbers), around which the patches in right figure are cropped. We select 24 points in total, covering the region on or around each subject's eyebrows, eyes, nose, mouth, and cheek.

B. Region decomposition

Facial expression is distinguished in specific facial regions, because the expressions are facial activities invoked by sets of muscle motions. Localizing and encoding the expression-related parts is of benefit to recognize facial expression [12]. Additionally, dividing the face into multiple local patches helps to find the position of occlusions [13].

To find the typical facial parts that related to expression, we extract the patches according to the positions of each subject's facial landmarks. Fig. 3 shows the selection of facial patches. We first detect 68 facial landmark points by the method in [25] and then, based on the detected 68 points, we select or re-compute 24 points that cover the informative region of the face, including the two eyes, nose, mouth, cheek, and dimple. The selected patches are defined as the regions taking each of the 24 points as the center. It is noteworthy that face alignment method in [25] is robust to occlusions, which is important for precise region decomposition.

As can be seen in the overall framework (Fig. 2), the patch decomposition operation is conducted on the feature map from convolution layers rather than from the original image. This is because sharing some convolutional operations can decrease the model size and enlarge the receptive fields of subsequent neurons. Based on the $512 \times 28 \times 28$ feature maps as well as the 24 local region centers, we get a total of 24 local regions, each with a size of $512 \times 6 \times 6$.

C. Occlusion perception with PG-Unit

We embed the PG-Unit in the PG-CNN to automatically percept the blocked facial patch and pay attentions mainly to the unblocked and informative patches. The detailed structure of PG-Unit is illustrated in the blue dashed rectangle in Fig. 2. In each patch-specific PG-Unit, the cropped local feature maps are fed to two convolution layers without decreasing the spatial resolution, so as to preserve more information when learning region specific patterns. Then, the last $512 \times 6 \times 6$ feature maps are processed in two branches. The first branch encodes the input feature maps as the vector-shaped local feature. The second branch consist an attention net that estimates a scaler weight to denote the importance of the local patch. The local feature is then weighted by the computed weight.

Mathematically speaking, let us suppose $\mathbf{p}_i$ the input $512 \times 6 \times 6$ feature map of the i-th patch. Therefore, the i-th PG-Unit takes the feature map $\mathbf{p}_i$ as the input and outputs its weighted feature ϕ_i. We formulate PG-Unit as:

$$\phi_i = \mathcal{I}_i(\tilde{\mathbf{p}}_i) \odot \psi(\tilde{\mathbf{p}}_i), \tag{1}$$

where $\tilde{\mathbf{p}}_i = \tilde{\phi}(\mathbf{p}_i)$ is the last $512 \times 6 \times 6$ feature maps ahead of the two branches. PG-Unit estimates patch i's importance or 'unobstructed-ness' as $\alpha_i = \mathcal{I}_i(\tilde{\mathbf{p}}_i)$ and then uses α_i to weight the local feature $\psi_i = \psi(\tilde{\mathbf{p}}_i)$. $\psi(\tilde{\mathbf{p}}_i)$ is a vector that represents the un-weighted feature. $\odot$ denotes production. $\alpha_i = \mathcal{I}_i(\tilde{\mathbf{p}}_i)$ is a scaler that represent the patch i's importance or 'unobstructed-ness' (to what extent the patch is occluded). $\mathcal{I}(\cdot)$ means the operations in the attention net, consisting a pooling operation, one convolution operations, two inner productions, and a sigmoid activation. The sigmoid activation forces the output α_i ranges in $[0, 1]$, where 1 indicates the most salient unobstructed patch and 0 indicates the completely blocked patch.

In PG-Unit, each patch is weighted differently according to its occlusion conditions or importance. Through the end-to-end

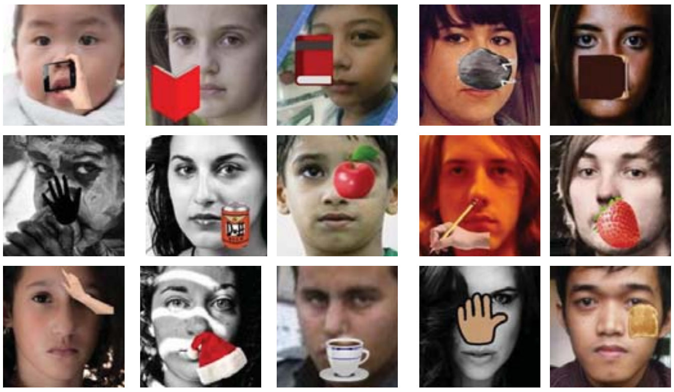


Fig. 4. Examples of the synthesized occluded facial images from RAF-DB dataset. The occluders are various in color, shape, and positions.

training of the overall PG-CNN, PG-Units can automatically learn low weights for the occluded parts and high weights for the unblocked and discriminative parts.

IV. Experiment

In this section, we present the experimental evaluations of PG-CNN. Then, we compared our method with the state-of-the-art FER methods and methods with attention mechanism. Finally, we provide an ablation analysis of the proposed PG-CNN.

A. Experimental setup

1) Datasets: We evaluated the methods on both in-the-wild datasets (RAF-DB [4] and AffectNet [5]) and in-the-lab datasets(CK+ [1], MMI [2], and Oulu-CASIA [3]). **RAF-DB** contains 30,000 facial images annotated with basic or compound expressions by 40 trained human coders. In our experiment, only images with basic emotions were used, including 12,271 images as training data and 3,068 images as test data. **AffectNet** is the largest database with annotated facial emotions. It contains about 400,000 images manually annotated for the presence of seven discrete facial expressions and the intensity of valence and arousal. We only used the ones with neutral and 6 basic emotions, containing 280,000 training samples and 3,500 test samples. **The Extended Cohn-Kanade database (CK+)** contains 593 video sequences recorded from 123 subjects. we selected the first and final frame of each sequence as neutral and target expressions, which results in 634 images. **MMI** database includes more than 30 subjects of both genders (44% female), ranging in age from 19 to 62. There are 79 sequences of each subject. Each begin and end with neutral facial expression. We extracted the neutral and peak frames from each sequence, resulting in 7348 images. **Oulu-CASIA** dataset contains six prototypic expressions from 80 people between 23 to 58 years old. We selected peak and neutral frames from sequences captured in normal illumination, which results in 9431 images.

2) Synthesis of occluded images: It seems unlikely that any reasonable sized set of training images would serve to densely probe the space of possible occlusions. We tackle the problem by manually collecting about 4k images as masks for generating occluders. These mask images were collected from search engine using more than 50 keywords, such as beer, bread, wall, hand, hair, hat, book, cabinet, computer, cup et al. All the items were selected due to their high frequency of occurence as obstructions in facial images. Since Benitez et al. [26] verified that small local occluders take no affects on current FER algorithms, we heuristically restrain occluder size S satisfying $S \in [96, 128]$, which is smaller or equal to half size of expression images. Fig. 4 shows some occluded examples derived from RAF-DB dataset. These artificial synthesised images are various in occlusion patterns and can better reflect occluder distribution in wild condition.

TABLE I
Test accuracy (%) on RAF-DB and AffectNet. (*clean*: original images. *occ.*: synthetically occluded images.)

Methods	RAF-DB(clean/occ.)	AffectNet(clean/occ.)
VGG-16 [24]	80.96/75.26	51.11/46.48
DLP-CNN [23]	80.89/76.29	54.47/51.07
P-CNN	81.64/76.09	53.9/50.32
PG-CNN (proposed)	**83.27/78.05**	**55.33/52.47**

3) Implementation details: We implemented PG-CNN using Caffe deep learning framework [27]. We adopted VGG-16 [24] as base for PG-CNN due to its simple structure and excellent performance in object classification. We only choose the first nine convolution layers as the feature map for region decomposition then attached 24 PG-Units. The pre-trained model based on ImageNet dataset was used for initializing the model. For each dataset, Both train and test corpus are mixed with occluded images with a ratio of 1:1. We adopt a batch-based stochastic gradient descent method to optimize the model. The base learning rate was set as 0.001 and was reduced by polynomial policy with gamma of 0.1. The momentum was set as 0.9 and the weight decay was set as 0.0005. The training of models was completed on a Titan-X GPU with 12GB memory. During the training stage, we set the actual batch size as 128 and the maximum iterations as 50K. It took about 1.5 days to finish optimizing the model.

4) Evaluation metric: All the datasets are mixed with their modifications with synthesized facial occlusions with 1:1 ratio. We report FER performance on both non-occluded and occluded images. For both occluded and non-occluded FER scenarios we use the overall accuracy on seven facial expression categories(i e. six prototypical plus neutral category) as a performance metric. Both cross-dataset evaluation and 10-fold evaluation within dataset are used in our experiments.

B. Comparison with state of arts

1) Comparison with other attention models: We compare PG-CNN with DLP-CNN [23]. DLP-CNN estimates K spatial maps for attention parts generation. The hyper-parameter K is fine-tuned to the best in out experiments. Table I reports the results of PG-CNN and DLP-CNN on RAF-DB and AffectNet databases. PG-CNN outperforms DLP-CNN on non-occluded images because the patch-based model can better reflect subtle muscle motions than the model with global attention. PG-CNN exceeds DLP-CNN on occluded datasets with the help of PG-Unit, which encodes occlusion patterns in model weights and enable model attend to unblocked & distinctive patches. From

TABLE II

10-FOLD TEST ACCURACY (%) ON CK+ DATASET WITH SYNTHETIC OCCLUSIONS. (R8, R16, R24 DENOTE THE SIZE OF THE OCCLUSION AS 8×8, 16×16, 24×24. THE FULL-IMAGE SIZE IS 48×48.)

Occlusion	*PG-CNN**	PG-CNN	WLS-RF [28]	RGBT [15]
non-occlusion	*90.37*	**97.03**	94.3	94.4
R8	*89.74*	**96.58**	92.2	92.0
R16	*87.22*	**95.70**	86.4	82.0
R24	*83.91*	**92.86**	74.8	62.5
eyes occluded	*85.02*	**96.50**	87.9	88.0
mouth occluded	*82.96*	**93.92**	72.7	30.3

* denotes cross-dataset test accuracy on CK+ by the PG-CNN trained on AffectNet.

RAF-DB to AffectNet database, the performance gap becomes narrowed because significant increase in training data.

2) Comparison with other methods handling FER with occlusion: We compare PG-CNN with state-of-the-arts methods WLS-RF [28] and RGBT [15]. WLS-RF adopted multiply weighted random forests and RGBT converted a set of Gabor based part-face templates into template match distance features for FER with occlusion. We followed the same occlusion protocol of WLS-RF and RGBT and evaluated performance on model trained by AffectNet dataset.

Table II show the comparisons. The overall performance of PG-CNN is significantly better than that of WLS-RF and RGBT. Specially, PG-CNN suffers 4.30% performance degradation under random occlusion with R24 pattern, while eyes or mouth occlusion has little impact on PG-CNN. The proceeds of PG-CNN are due to PG-Unit as well as large amount of training data in AffectNet database.

PG-CNN* in Table II shows that without training on CK+ dataset, PG-CNN* can achieve comparable performance compared with WLS-RF and RGBT.

3) Cross database evaluation: We evaluated the generalization ability of PG-CNN under the cross-dataset evaluation protocol. In our experiments, PG-CNN was trained on RAF-DB or AffectNet dataset and evaluated on CK+, MMI, Oulu-CASIA dataset with or without synthetic occlusions. Table III shows the results compared with other FER methods. Among the compared experiments, [29] adopted an inception based CNN and provided the average cross-database recognition accuracy. [30] and [31] reported the highest cross-database results, which were both trained on MMI and evaluated on CK+ or vice versa. PG-CNN(A) exceeds [29]–[31] by at least 40.7% and 3.02% on CK+ and MMI dataset respectively. It suggests that PG-CNN(A) can generalize better than PG-CNN(R) due to a larger amount of training data.

C. Ablation analysis

We conducted ablation analysis to figure out how PG-CNN boosts performance on FER with occlusion task.

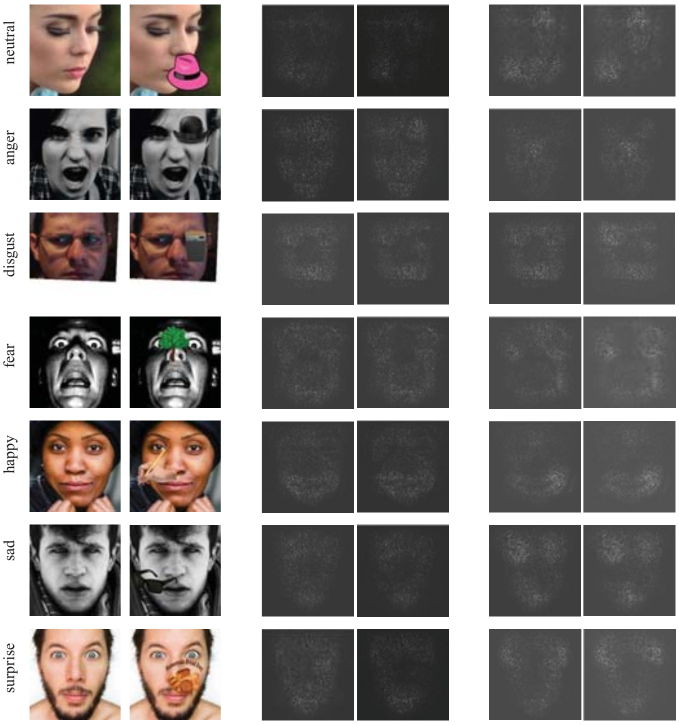
1) CNN VS P-CNN: We compared VGG-16 and P-CNN (PG-CNN without PG-Unit) to verify benefit of region decomposition. As listed in Table I, P-CNN exceeds VGG-16 on both original and occluded images. The promotions of P-CNN suggest that globally encoded representation has fallen behind in reflecting subtle muscle motions compared with locally learned patterns.

TABLE III

CROSS DATASET EVALUATION (ACCURACY%) ON IN-THE-LAB DATASETS (*clean*: ORIGINAL IMAGES. *occ.*: SYNTHESIZED OCCLUDED IMAGES.)

method	CK+(clean/occ.)	MMI(clean/occ.)	Oulu-CASIA(clean/occ.)
[29]	64.2 / –	55.6 / –	– / –
[30]	60.8 / –	60.3 / –	– / –
[31]	61.2 / –	66.9 / –	– / –
P-CNN(R)	79.81 / 76.02	57.02 / 53.70	49.83 / 46.98
P-CNN(A)	89.27 / 85.33	66.94 / 61.26	54.77 / 51.05
PG-CNN(R)	80.28 / 79.49	55.61 / 53.44	50.04 / 47.15
PG-CNN(A)	**90.38 / 86.27**	**68.92 / 63.94**	**57.93 / 54.18**

A denotes models trained on AffectNet dataset.
R denotes models trained on RAF-DB dataset.



(a) Expression images and corresponding occluded version

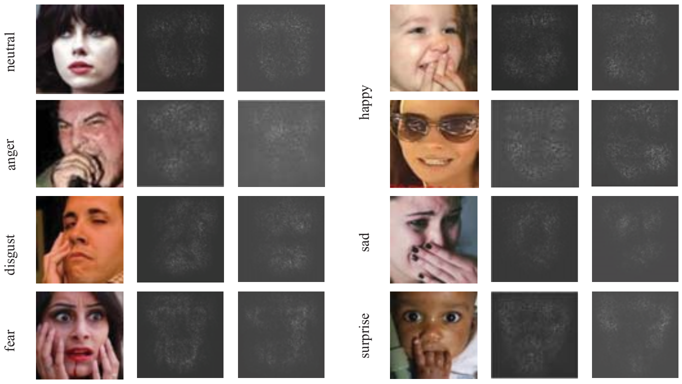
(b) Attention maps of (a), derived from P-CNN. Left: maps of original images. Right: maps of occluded images

(c) Attention maps of (a), derived from PG-CNN. Left: maps of original images. Right: maps of occluded images

Fig. 5. Attention maps of several test images and their modifications with artificial facial occlusions. Each image denotes one of seven basic expressions. A deep white corresponds to high attention and a deep dark to no attention at all. Better viewed in color and zoom in.

2) P-CNN VS PG-CNN: We compared P-CNN and PG-CNN to verify benefit of PG-Unit. As displayed in Table I, total improvements of PG-CNN on RAF-DB and AffectNet datasets are 1.99%, 2.58% and 3.65%, 4.27% respectively. This is because PG-Unit enables the model to attend to most related local patches, and shift attention to other related local parts when original ones are occluded. Similar performance improvements can be found in Table III, where PG-CNN outperforms P-CNN on nearly all datasets except for MMI.

We visualized the attention map of PG-CNN and P-CNN using the method in [32]. Simonyan et al. [32] derives a attention map by computing the gradient of the class score with respect to the input image. As can be seen in Fig. 5, P-CNN relies on almost the whole face region, while PG-CNN



(a) Image (b) Attention maps of (a), derived from P-CNN (c) Attention maps of (a), derived from PG-CNN (d) Image (e) Attention maps of (d), derived from P-CNN (f) Attention maps of (d), derived from PG-CNN

Fig. 6. Attention maps of several test images results of test images with real occlusion in RAF-DB. Better viewed in color and zoom in.

attends to local discriminative patches. This can decrease the probability that an occluder take effect in PG-CNN. Moreover, PG-CNN responses weakly to an occluder and shifts attention from the occluded patch(e.g., right eye in the subfigures for *disgust*) to other related but unobstructed one(e.g., left eye). Fig. 6 displays images with real occluders picked from test set in RAF-DB and AffectNet corpus. PG-CNN performs as consistently as on artificial occluders. Take angry category for instance, we observe only PG-CNN attends to subjects' nose, which is a strongly discriminative patch for anger.

V. Conclusion

This work presents a Patch-Gated CNN for facial expression recognition under occlusion. PG-CNN consists of region decomposition, Patch Gated Unit for robust facial expression recognition. Experiments under intra and cross database evaluation protocols demonstrated PG-CNN outperforms other state-of-the-art methods. Ablation analyses show PG-CNN is capable of shifting attention from occluded patch to other related ones. For future work, we will study how to generate attention parts in face without landmarks, as PG-CNN relies on robust face detection and facial landmark localization modules.

Acknowledgment

This work was partially supported by National Key R&D Program of China under contracts NO.2017YFB1002802, Natural Science Foundation of China (grants 61702481 and 61702486), and External Cooperation Program of CAS (grant GJHZ1843). Thanks also to Xin Liu for discussions.

References

[1] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in *CVPRW, 2010*.

[2] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," in *ICME, 2005*.

[3] G. Zhao, X. Huang, M. Taini, S. Z. Li, and M. PietikäInen, "Facial expression recognition from near-infrared videos," *Image & Vision Computing*, vol. 29, no. 9, pp. 607–619, 2011.

[4] S. Li, W. Deng, and J. Du, "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," in *CVPR, 2017*.

[5] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild," *arXiv preprint arXiv:1708.03985*, 2017.

[6] J. Wright, A. Y. Yang, A. Ganesh, S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE TPAMI*, vol. 31, no. 2, pp. 210–227, 2009.

[7] E. Osherov and M. Lindenbaum, "Increasing cnn robustness to occlusions by reducing filter support," in *CVPR, 2017*.

[8] M. Ranzato, J. Susskind, V. Mnih, and G. E. Hinton, "On deep generative models with applications to recognition," in *CVPR, 2011*.

[9] X. Mao, Y. Xue, Z. Li, K. Huang, and S. Lv, "Robust facial expression recognition based on rpca and adaboost," in *WIAMIS, 2009*.

[10] I. Kotsia, I. Buciu, and I. Pitas, "An analysis of facial expression recognition under partial facial image occlusion," *Image and Vision Computing*, vol. 26, no. 7, pp. 1052–1067, 2008.

[11] A. M. Martinez, "Recognizing imprecisely localized, partially occluded, and expression variant faces from a single sample per class," *IEEE TPAMI*, vol. 24, no. 6, pp. 748–763, 2002.

[12] L. Zhong, Q. Liu, P. Yang, B. Liu, J. Huang, and D. N. Metaxas, "Learning active facial patches for expression analysis," in *CVPR, 2012*.

[13] A. Dapogny, K. Bailly, and S. Dubuisson, "Confidence-weighted local expression predictions for occlusion handling in expression recognition and action unit detection," *IJCV*, pp. 1–17, 2017.

[14] W. Li, F. Abitahi, and Z. Zhu, "Action unit detection with region adaptation, multi-labeling learning and optimal temporal fusing," in *CVPR, 2017*.

[15] L. Zhang, D. Tjondronegoro, and V. Chandran, "Random gabor based templates for facial expression recognition in images with facial occlusion," *Neurocomputing*, vol. 145, pp. 451–464, 2014.

[16] R. Min, A. Hadid, and J. L. Dugelay, "Improving the recognition of faces occluded by facial accessories," in *Automatic Face & Gesture Recognition and Workshops, 2011*.

[17] X. Huang, G. Zhao, W. Zheng, and M. Pietikinen, "Towards a dynamic expression recognition system under facial occlusion," *Pattern Recognition Letters*, vol. 33, no. 16, pp. 2181–2191, 2012.

[18] H. Towner and M. Slater, "Reconstruction and recognition of occluded facial expressions using pca," in *ACII, 2007*.

[19] Y. Deng, D. Li, X. Xie, K. Lam, and Q. Dai, "Partially occluded face completion and recognition," in *ICIP, 2009*.

[20] H. Zheng, J. Fu, T. Mei, and J. Luo, "Learning multi-attention convolutional neural network for fine-grained image recognition," in *ICCV, 2017*.

[21] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *ICML, 2015*.

[22] C. Zhu, Y. Zhao, S. Huang, K. Tu, and Y. Ma, "Structured attentions for visual question answering," in *CVPR, 2017*.

[23] L. Zhao and W. Zhuang, Jingdong, "Part-aligned network for person re-identification," in *ICCV, 2017*.

[24] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[25] J. Zhang, M. Kan, S. Shan, and X. Chen, "Occlusion-free face alignment: deep regression networks coupled with de-corrupt autoencoders," in *CVPR, 2016*.

[26] C. F. Benitez-Quiroz, R. Srinivasan, Q. Feng, Y. Wang, and A. M. Martinez, "Emotionet challenge: Recognition of facial expressions of emotion in the wild," *arXiv preprint arXiv:1703.01210*, 2017.

[27] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *ACM Multimedia, 2014*.

[28] A. Dapogny, K. Bailly, and S. Dubuisson, "Confidence-weighted local expression predictions for occlusion handling in expression recognition and action unit detection," *IJCV*, pp. 1–17, 2016.

[29] A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," in *WACV, 2016*.

[30] C. Mayer, M. Eggers, and B. Radig, "Cross-database evaluation for facial expression recognition," *Pattern recognition and image analysis*, vol. 24, no. 1, pp. 124–132, 2014.

[31] X. Zhang, M. H. Mahoor, and S. M. Mavadati, "Facial expression recognition using {1} _ {p}-norm mkl multiclass-svm," *Machine Vision and Applications*, vol. 26, no. 4, pp. 467–483, 2015.

[32] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2013.