Neurocomputing 296 (2018) 12–22

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

A visual attention based ROI detection method for facial expression recognition

Wenyun Sun[a], Haitao Zhao[b], Zhong Jin[a,*]

[a] School of Computer Science and Engineering, Nanjing University of Science and Technology, China
[b] School of Information Science and Engineering, East China University of Science and Technology, China

ARTICLE INFO

Article history:
Received 13 April 2017
Revised 10 February 2018
Accepted 10 March 2018
Available online 20 March 2018

Communicated by H. Yu

Keywords:
Facial expression recognition
Action unit
Visual attention mechanism

ABSTRACT

In this paper, an eleven-layered Convolutional Neural Network with Visual Attention is proposed for facial expression recognition. The network is composed of three components. First, local convolutional features of faces are extracted by a stack of ten convolutional layers. Second, the regions of interest are automatically determined according to these local features by the embedded attention model. Third, the local features in these regions are aggregated and used to infer the emotional label. These three components are integrated into a single network which can be trained in an end-to-end scheme. Extensive experiments on four kinds of data (namely aligned frontal faces, faces in different poses, aligned unconstrained faces, and grouped unconstrained faces) prove that the proposed method can improve the accuracy and obtain good visualization. The visualization shows that the learned regions of interest are partly consistent with the locations of emotion specific Action Units. This founding confirms the interpretation of Facial Action Coding System and Emotional Facial Action Coding System from a machine learning perspective.

1. Introduction

The research on facial expression was started by psychologists. Mehrabian et al. [1] suggested that the combined effect of simultaneous verbal, vocal and facial attitude communications is a weighted sum of their independent effects with the coefficients of 7%, 38% and 55%, respectively. The facial expression plays an important role in Human Computer Interaction (HCI), affective computing, human behavior analysis, etc. Facial Action Coding System (FACS) [2] and Emotional Facial Action Coding System (EMFACS) [3] were proposed by Ekman and Friensen. FACS and EMFACS define a set of Action Units (AUs) associated with six basic emotions including angry, disgust, fear, happy, sad and surprise. The Action Units and six basic emotions became the most commonly used expression labels for classification / detection tasks in machine learning.

In the computer vision community, deep learning based methods have become more and more popular nowadays. Kahou et al. [4] and Levi and Hassner [5] use Convolutional Neural Networks (CNNs) to solve the Facial Expression Recognition (FER) problem. CNNs are universal non-linear fitting tools for image data. In the classification problem, they learn posterior probability functions by using back-propagation (BP) algorithm. CNNs make decisions according to the learned posterior probability functions. But, human experts usually make decisions according to small local Regions Of Interests (ROIs) of faces which are more explainable. Inspired by the human behavior, the visual attention mechanism [6] can focus attentions on small regions of images. As variants of the classic CNNs, the CNNs with Visual Attention are promising methods for solving the facial expression recognition problem.

In this work,

- An eleven-layered CNN with Visual Attention is proposed for solving the facial expression recognition problem. The network extracts deep convolutional features from faces, detects the regions of interests, and uses convolutional features in these regions to infer the emotional label. Like some existing neural networks with attention mechanism, our convolutional feature extraction model, attention model and classification model are integrated into a single network which can be trained in an end-to-end scheme.
- Some state-of-the-art methods use ensembles of deep neural networks, temporal data and multimodal data to achieve superior performance. We focus our attention on facial expression recognition algorithm based on single network and single frame rather than breaking the state-of-the-art. We use controlled experiments to analyze the effects of the visual attention based ROI detector on four kinds of data, namely aligned frontal

* Corresponding author.
E-mail address: zhongjin@njust.edu.cn (Z. Jin).

https://doi.org/10.1016/j.neucom.2018.03.034
0925-2312/

faces, faces in different poses, aligned unconstrained faces, and grouped unconstrained faces. In all cases, good visualizations are obtained. The visualizations show that the learned regions of interests are partly consistent with the locations of emotion specific Action Units. This founding confirms the interpretation of FACS and EMFACS from a machine learning perspective.
- The rest of the paper is organized as follows. Section 2 reviews the related work. In Section 3, the main method is proposed. The experiments and results are presented in Section 4. Section 5 gives the conclusions.

2. Related work

2.1. Attention mechanism

The proposed method is inspired by several existing work. Xu et al. [6] have proposed a method of generating captions of images. The main idea is using visual attention mechanism to focus the network on a small region of the image. A single word is inferred from the small region and the language context. While the language context is changed over the time steps of the Recurrent Neural Network (RNN), a sequence of regions and the related words are generated.

Moreover, the attention mechanism can also be applied on the non-visual tasks. Riemer et al. [7] have proposed an attention based method for forecasting time series data. The main idea is using attention mechanism to select factors from multiple external sources. The algorithm can give explanations of decisions which may be interesting to a human analyst.

Focusing the attention on small regions of interests of a high dimensional data is a human-like behavior. But, as a universal nonlinear fitting tool, traditional neural networks accept the whole high dimensional inputs. Then, data are processed in parallel in the black box. Finally, unexplainable decisions are given. The lack of explanation is a serious disadvantage of the traditional neural networks. There are limited number of methods of looking inside of the traditional neural networks. With the help of attention mechanism, the relation of regions of interests and decisions can be learned automatically. The decisions can be explained by the main factor of the data.

2.2. Image-specific class saliency visualization

The Image-Specific Class Saliency Visualization (ISCSV) [8] is similar to the method proposed in this work. The tasks of the two pieces of work are similar. Saliency maps / attention weights are extracted from classification CNNs trained on the image-level labels. And, no additional pixel-level annotations is required. But, the methods of the two pieces of work are totally different:

- The ISCSV is a visualization method. First, the traditional CNN is trained without considering the saliency map. Then, the saliency map is extracted by solving an independent optimization problem, i.e. maximizing the class score with respect to the input image while a pair of image and label is given.
- In this work, we employ visual attention mechanism which is embedded in the proposed CNN. The attention weights is applied on the middle-level feature map rather than the original image. Besides, CNN with Visual Attention can be divided into three parts, namely convolutional part, attention mapping part and fully connected part. Since the attention mapping is a build-in part of the network, it can be trained jointly with other parts. In the test stage, the attention weights is extracted by performing a forward-propagation pass rather than solving an independent optimization problem.

Table 1
The comparison between the visualization based methods and attention based methods.

Method	Same		Different	
	Purpose	Data	Network	Locating procedure
Visualization based methods [8–10,12]	Locate regions of interests	Images +labels	CNN	Independent procedure
Attention based methods	Locate regions of interests	Images +labels	CNN +attention	Forward −propagation

2.3. Deconvolutional neural network

The Deconvolutional Neural Network [9,10] is another visualization method for CNNs. Like the ISCSV method, the Deconvolutional Neural Network solves an independent optimization problem on a pre-trained CNN. It projects features in the middle layers back to the input space, and visualizes what activate these features. It discovers locatable patterns in a variety of forms: low-level edges, mid-level edge junctions, high-level object parts and complete objects. It also can be seen as a locator for these objects.

2.4. Simplifying images

Simplifying Images is a strategy adopted for testing human visual recognition [11]. Inspired by this method, Zhou et al. [12] designed an automatic procedure to test pre-trained CNNs. First, a test image is divided into regions by edge segmentations or annotated segmentations. Then, segment that produces the smallest decrease of the correct classification score is removed. Next, the removal is repeated until the image is not correctly classified. Finally, the remaining regions contain the minimal information needed by the network to make a true prediction.

The ISCSV method, the Deconvolutional Neural Network, the Simplifying Images method and our method learn locators as by-products of solving the main classification problems. As it is already differentiated in Section 2.2, the main differences between visualization based methods [8–10,12] and attention based methods are their network architectures and the locating procedures. The visualization based methods train normal CNNs first. Then, an independent locator needs to be re-trained / re-calculated for inferring the regions of interests. The attention based methods train customized CNNs, while the attention mapping is trained as its build-in part. The comparison is summarized in Table 1.

2.5. AU-aware deep networks

Liu et al. have proposed an AU-aware Deep Network (AUDN) [13] for facial expression recognition. The network is composed of four components. First, a single convolutional layer is used for extracting an over-complete representation. Then, an AU-aware receptive fields selection procedure seeks a subset of the over-complete representation for simulating the combination of Action Units. Next, a multi-layer Restricted Boltzmann Machine (RBM) is used to learn hierarchical features. Finally, supervised logistic regression and SVM are used for classification.

Both AUDN and our proposed network use a feature selection procedure for seeking AU-aware convolutional features, but there are two major differences as follows. First, to get benefit from the deep CNN, we select AU-aware features from the deep convolutional features rather than the shallow ones. The feature extracted by the stack of ten convolutional layers can describe more complex structures [9,10]. Second, all components in our network are derivable. The entire network can be trained in an end-to-end scheme. Training the entire network jointly is simpler and more effective

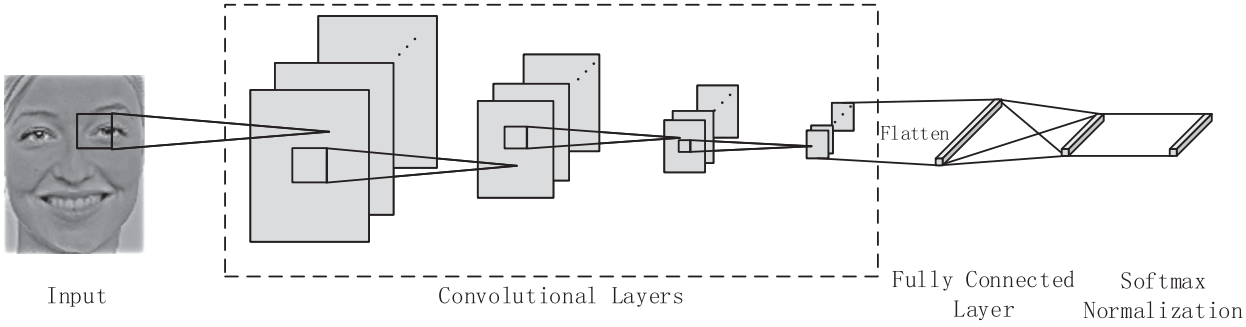


Fig. 1. The baseline CNN.

than training each component separately. End-to-end training has already become the mainstream for training deep networks.

Zhou and Shi have proposed a CNN based AU-aware feature transferring and selecting method [14]. They found that the deep CNN features pre-trained on generic images are selective to facial AUs. Based on a feature selection algorithm, these AU-aware features can boost performance on FER task.

Both [14] and our method select features, and both studies care about the relevance between features and AUs, But, the differences are obvious. The method of Zhou et al. chooses a subset of the feature maps. Our method spatially aggregates feature maps into AU-aware feature vectors by attention mapping.

2.6. AU-aware patches for facial expression recognition

Yao et al. [15] have proposed a pair-wise learning strategy to automatically seek a set of discriminative patches of a facial image. These learned local patches are consistent with the locations of expression specific Action Units. Based on the AU-aware features extracted from these patches, an SVM classifier was trained for facial expression recognition. This work won the Emotion Recognition in the Wild Challenge (EmotiW) 2015 [16].

In our work, a different way is used to find regions of interests in facial expression recognition which is smoother than those patches in [15]. We are going to propose a CNN with Visual Attention. The CNN accepts aligned facial images as its input and generate probabilities of seven emotions as its output. The visual attention mechanism embedded in the proposed CNN is similar to the one in the work of Xu et al. [6]. It aggregates the activations of the last max-pooling layer spatially to form the global feature vector. With the help of supervised data, the proposed CNN can be trained in an end-to-end scheme. More details will be given in Section 3.2.

2.7. The state-of-the-art facial expression recognition methods

The Emotion Recognition in the Wild Challenge (EmotiW) has been held for five years (2013–2017). The challenge is based on two benchmarks / datasets: the Static Facial Expressions in the Wild (SFEW) and the Acted Facial Expressions in the Wild (AFEW). There are many state-of-the-art facial expression recognition methods break the records of the two benchmarks every year. On the SFEW benchmark, Kim et al. [17] train multiple CNNs, and formed hierarchical committees of CNNs using the validation-accuracy-based exponentially-weighted average rule. Levi and Hassner [5] present a novel Local Binary Patterns (LBP) mapping and combines it with CNN classifiers. On the AFEW benchmark, Kahou et al. [4] combine CNNs, Deep Belief Networks (DBNs) and Autoencoders for different data modalities to make the frame-level and audio-level decisions, and then these decisions are together. Fan et al. [18] use 3D CNNs and the combination of CNN and Recurrent Neural Networks (CNN-RNN) to recognize facial expressions in videos. Yao et al. [19] design a novel CNN namely HoloNet to recognize facial expressions in frames, and fuse the frame-level decisions together. Hu et al. [20] present a new learning method named Supervised Scoring Ensemble and a new fusion structure. Knyazev et al. [21] transfer the knowledge learned on the large-scale face recognition task to the facial expression recognition task to improve the performance. Vielzeuf et al. [22] propose improved face descriptors based on 2D and 3D CNNs, and explore a novel hierarchical method combining features and scores.

Almost without exception, these methods use ensembles of CNNs to achieve superior performance. For the audio-video based emotion recognition task on the AFEW dataset, decisions based on temporal and multimodal data are also fused together. Although, the current state-of-the-art method [20] achieves a record-breaking test accuracy of 60.34% on AFEW benchmark. Fewer studies care about the single network performance based on a single frame. As we know, the single network denoted as "$PREPiNor, oA - \{CNNL - FC3072\}_{R1}$" in [17] achieves a test accuracy of 52.50% on SFEW benchmark. The single network denoted as "LBP1, cyclic, VGG_M-4096 - Oversampling" in [5] achieves a test accuracy of 44.73% on SFEW benchmark. They are much lower than the accuracy obtained by ensemble network when temporal and multimodal data is provided. Since the challenge can be decomposed into a set of orthogonal problems, we simplify our research objective by focusing our attention on one of the directions, design facial expression recognition algorithm based on single network and single frame.

Some methods use external data such as FER-2013 dataset [23] and CASIA Web Face dataset [24] to alleviate the small sample size problem. We use FER-2013 dataset to enlarge the training set in our experiment.

3. The proposed method

3.1. The baseline CNN

At the very beginning, inspired by the VGG network [25], we will propose a CNN for facial expression recognition as the baseline. The baseline CNN is illustrated in Fig. 1. The configuration details including the layer types and the numbers of activations / parameters are listed in Table 2. As listed in Table 2, the CNN accepts 112×144 aligned facial images as its input and generate probabilities of seven emotions as its output. The baseline CNN contains eleven trainable layers including ten 3×3 convolutional layers and one fully connected layer. The total number of trainable parameters is 360,720 which is much smaller than usual. Such a network is capable of recognizing emotions of aligned faces. But, unlike a human expert, the baseline network can not explain the decisions made by itself.

Table 2
The configuration of the baseline CNN.

Layer	Number of activations	Number of parameters
Input layer	$112 \times 144 \times 1$	0
Convolutional layer	$112 \times 144 \times 16$	$3 \times 3 \times 1 \times 16$
Convolutional layer	$112 \times 144 \times 16$	$3 \times 3 \times 16 \times 16$
Convolutional layer	$112 \times 144 \times 16$	$3 \times 3 \times 16 \times 16$
Max-pooling layer	$56 \times 72 \times 16$	0
Convolutional layer	$56 \times 72 \times 32$	$3 \times 3 \times 16 \times 32$
Convolutional layer	$56 \times 72 \times 32$	$3 \times 3 \times 32 \times 32$
Convolutional layer	$56 \times 72 \times 32$	$3 \times 3 \times 32 \times 32$
Max-pooling layer	$28 \times 36 \times 32$	0
Convolutional layer	$28 \times 36 \times 64$	$3 \times 3 \times 32 \times 64$
Convolutional layer	$28 \times 36 \times 64$	$3 \times 3 \times 64 \times 64$
Max-pooling layer	$14 \times 18 \times 64$	0
Convolutional layer	$14 \times 18 \times 128$	$3 \times 3 \times 64 \times 128$
Convolutional layer	$14 \times 18 \times 128$	$3 \times 3 \times 128 \times 128$
Max-pooling layer	$7 \times 9 \times 128$	0
Dropout layer[1]	$7 \times 9 \times 128$	0
Flatten layer[1]	8064	0
Fully connected layer[1]	7	8064×7
Soft-max normalization layer	7	0

[1] the bold texts highlight the differences between the CNNs with / without Visual Attention.

3.2. The CNN with visual attention

As illustrated in Fig. 2, the CNN is improved by adding an attention mapping between its convolutional part and the fully connected part. Let f: $\boldsymbol{A}_0 \rightarrow \boldsymbol{z}$ be an attention mapping. $\boldsymbol{A}_0 \in \Re^{W \times H \times C}$ is the unfocused convolutional feature (i.e the activation of the last max-pooling layer in Fig. 2). $\boldsymbol{z} \in \Re^{C}$ is the focused feature (i.e. the input of the fully connected layer in Fig. 2).

$$\boldsymbol{z}_i = \sum_{j=1}^{W} \sum_{k=1}^{H} \boldsymbol{B}_{j,k} (\boldsymbol{A}_0)_{j,k,i}, \tag{1}$$

$$i \in \{1, 2, \ldots, C\},$$

where $\boldsymbol{B} \in \Re^{W \times H}$ is a weight map whose elements are determined by a two layered fully connected neural network mapping g: $(\boldsymbol{A}_0)_{i,j} \rightarrow (\boldsymbol{A}_2)_{i,j}$ and a soft-max function h: $(\boldsymbol{A}_2)_{i,j} \rightarrow \boldsymbol{B}_{i,j}$.

The two layered fully connected neural network g can be defined as

$$(\boldsymbol{A}_1)_{i,j,k} = tanh\left(\sum_{l=1}^{C} (\boldsymbol{W}_1)_{l,k} (\boldsymbol{A}_0)_{i,j,l} + (\boldsymbol{b}_1)_k \right), \tag{2}$$

$$i \in \{1, 2, \ldots, W\}, j \in \{1, 2, \ldots, H\}, k \in \{1, 2, \ldots, C_1\},$$

$$(\boldsymbol{A}_2)_{i,j} = tanh\left(\sum_{k=1}^{C_1} (\boldsymbol{W}_2)_k (\boldsymbol{A}_1)_{i,j,k} + b_2 \right), \tag{3}$$

$$i \in \{1, 2, \ldots, W\}, j \in \{1, 2, \ldots, H\},$$

where $\boldsymbol{W}_1$, $\boldsymbol{b}_1$, $\boldsymbol{W}_2$, $\boldsymbol{b}_2$ is the parameters of the fully connected neural network g. As illustrated in Fig. 2, the convolutional feature tensor $\boldsymbol{A}_0$ is spatially divided into vectors. Each vector denotes the feature at a specific location. The neural network takes one of these vectors as its input and generates an importance scalar as its output. The neural network evaluate the importance of the input vector according to its content rather than its location.

Furthermore, these scalars of importance are normalized by

$$\boldsymbol{B}_{i,j} = \frac{exp(\beta (\boldsymbol{A}_2)_{i,j})}{\sum_{k=1}^{W} \sum_{l=1}^{H} exp(\beta (\boldsymbol{A}_2)_{k,l})}, \tag{4}$$

$$i \in \{1, 2, \ldots, W\}, j \in \{1, 2, \ldots, H\}.$$

This equation encourages the sparsity of the weight map $\boldsymbol{B}$ which is controlled by β. Thus, Eq. (1) can select fewer regions of interests and produces the focused feature. In this work, we let $\beta = 1$, and let Eq. (4) be a standard soft-max function, which seems suitable for most applications.

Alternatively, extra constraints can be applied on the weight map for some specific tasks. Such as a symmetric constraint is usually used for processing aligned faces. Formally, the symmetric weight map can be defined as

$$\boldsymbol{B}_{i,j} = \frac{exp(\frac{1}{2}\beta (\boldsymbol{A}_2)_{i,j} + \frac{1}{2}\beta (\boldsymbol{A}_2)_{W+1-i,j})}{\sum_{k=1}^{W} \sum_{l=1}^{H} exp(\beta (\boldsymbol{A}_2)_{k,l})}, \tag{5}$$

$$i \in \{1, 2, \ldots, W\}, j \in \{1, 2, \ldots, H\}.$$

The symmetric weight map Eq. (5) is a drop-in replacement of the free symmetric weight map Eq. (4) in our framework. It can generate symmetric ROIs, which human can clearly recognize and understand. But it also has some disadvantage. First, it must be used on the aligned front faces. Second, it even brings some negative effects on the performance. In most cases, people should use Eq. (4) rather than Eq. (5). Comparisons will be shown in Section 4.3.

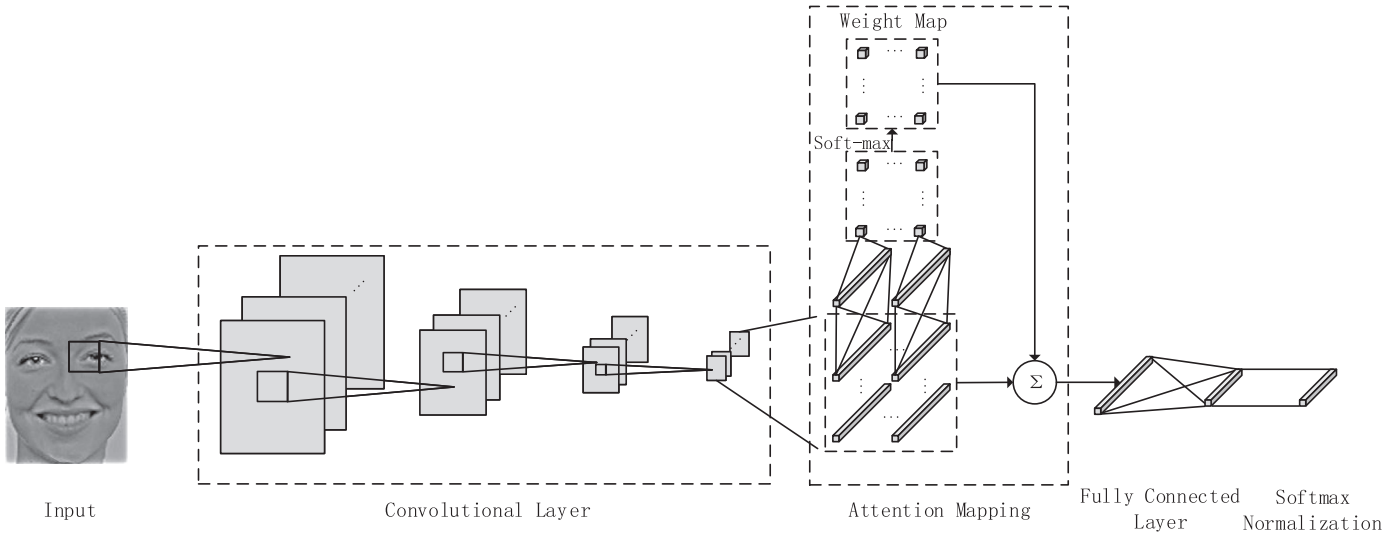


Fig. 2. The CNN with visual attention.

Table 3
The configuration of the CNN with visual attention

Layer	Number of activations	Number of parameters
Input layer	$112 \times 144 \times 1$	0
Convolutional layer	$112 \times 144 \times 16$	$3 \times 3 \times 1 \times 16$
Convolutional layer	$112 \times 144 \times 16$	$3 \times 3 \times 16 \times 16$
Convolutional layer	$112 \times 144 \times 16$	$3 \times 3 \times 16 \times 16$
Max-pooling layer	$56 \times 72 \times 16$	0
Convolutional layer	$56 \times 72 \times 32$	$3 \times 3 \times 16 \times 32$
Convolutional layer	$56 \times 72 \times 32$	$3 \times 3 \times 32 \times 32$
Convolutional layer	$56 \times 72 \times 32$	$3 \times 3 \times 32 \times 32$
Max-pooling layer	$28 \times 36 \times 32$	0
Convolutional layer	$28 \times 36 \times 64$	$3 \times 3 \times 32 \times 64$
Convolutional layer	$28 \times 36 \times 64$	$3 \times 3 \times 64 \times 64$
Max-pooling layer	$14 \times 18 \times 64$	0
Convolutional layer	$14 \times 18 \times 128$	$3 \times 3 \times 64 \times 128$
Convolutional layer	$14 \times 18 \times 128$	$3 \times 3 \times 128 \times 128$
Max-pooling layer	$7 \times 9 \times 128$	0
Visual attention mechanism	128	$128 \times 66 + 66 \times 1$
Dropout layer	128	0
Fully connected layer	7	128×7
Soft-max normalization layer	7	0

[1] the bold texts highlight the differences between the CNNs with / without visual attention.

The configuration details of the improved CNN are listed in Table 3. The convolutional parts of the two networks described in Table 2 and Table 3 are the same. The main difference between the two networks is the attention mapping part in Table 3 which aggregates $7 \times 9 \times 128$ dimensional activation together to form an activation of 128 dimensions.

Similarly to the proposed attention mapping, the global average pooling method [26] also averages every activation map of the last max-pooling layer into a scalar value. But, it does not use a trainable weight map. It uses an inefficient fixed uniform weight map in which every element is fixed at $1/WH$. In the facial expression recognition task, the local features extracted at background, hair, etc. have no correlation with the emotional labels. The global feature averaged by these meaningless local features is inefficient. The global average pooling method does not suitable for such task. By contrast, the advantage of using a trainable weight map is obvious. The feature aggregated from the region of interest is more informative than the feature in global average pooling method.

4. Experiments

4.1. Dataset

The Radboud Faces Database (RaFD) [27], the FER-2013 dataset [23] and the Static Facial Expressions in the Wild dataset 2.0 (SFEW) [28] are used in the experiment for evaluating the proposed method. The RaFD contains constrained facial images of 67 subjects. For each subject, faces were recorded in different poses and gaze directions. The samples belonging to the contempt class are ignored. The RaFD contains 7035 (67 subjects $\times$ 5 poses $\times$ 7 emotions $\times$ 3 gaze directions) 681 $\times$ 1024 images. These images are divided into five folds. Subjects in these folds are not overlapped (fold #1: subject 1–15, fold #2: subject 16–29, fold #3: subject 30–43, fold #4: subject 44–57, fold #5: subject 58–73).

The SFEW dataset 2.0 contains 1766 unconstrained facial images extracted from movies. These images are divided into training set, validation set, and test set. Subjects in these sets are not overlapped. The training set and the validation set are labeled with six basic expressions and the neutral class We report the performance on the validation set since only the labels of test samples are not public.

Some researchers involved in EmotiW used FER-2013 as external training set. The FER-2013 dataset contains 35887 unconstrained facial images labeled with six basic expressions and the neutral class in real-world conditions. All these samples are used as the external training data to improve the performance of unconstrained faces.

The HAPpy PEople Images (HAPPEI) dataset [29] is also used for evaluating the fully convolutional [30] version of the proposed method. The HAPPEI dataset is designed for group happiness intensity analysis. It contains 4886 samples downloaded from Flickr and manually with group level mood intensities. We select four images (see the first row of Fig. 7) in the HAPPEI dataset for visualizing the learned attention model in a fully convolutional scheme.

4.2. Preprocessing and data augmentation

As illustrated in Fig. 3, the RaFD is preprocessed into two sets namely RaFD-FRONT and RaFD-POSE. The RaFD-FRONT contains only the aligned frontal faces. Tight bounding boxes of the faces in RaFD are detected using the Histogram of Oriented Gradients (HOG) feature combined with a linear classifier, an image pyramid, and sliding window detection scheme [31]. The landmarks are detected using a regression based method [32]. 3D shapes of the faces are estimated from the detected landmarks [33]. Faces are aligned to a pre-defined 3D facial geometry. The RaFD-FRONT contains 1407 (67 subjects $\times$ 1 poses $\times$ 7 emotions $\times$ 3 gazes) 125 $\times$ 160 images. The RaFD-POSE contains faces in different poses. Faces in this set are cropped by a fixed bounding box. They are not precisely aligned. The RaFD-POSE contains 7035 (67 subjects $\times$ 5 poses $\times$ 7 emotions $\times$ 3 gazes) 284 $\times$ 284 images. Only the intensity channels of images in RaFD-FRONT and RaFD-POSE are kept. The intensity channels are illumination normalized by the isotropic diffusion method [34]. To enrich the training set, images are augmented by small translations and horizontal reflections. We extract random 112 $\times$ 144 / 272 $\times$ 272 patches (and their horizontal reflections) from the 125 $\times$ 160 / 284 $\times$ 284 images and training networks on these extracted patches. During test, the network makes predictions by extracting patches in the center.

The training and validation set of the SFEW is used. As illustrated in Fig. 4, faces in the SFEW dataset are detected, aligned, and illumination normalized by using similar pre-processing steps. For simplicity, samples are rejected if no face bounding box is detected. Otherwise, only the biggest box is kept. Finally, the training and validation sets have 865 and 400 samples respectively since some samples are rejected by the face detector.

The images in HAPPEI dataset are resized to fit the scale of RaFD-FRONT and RaFD-POSE. Similarly, the isotropic diffusion method is employed for illumination normalization in a fixed-sized sliding window.

4.3. Facial expression recognition on RaFD-FRONT dataset

The baseline CNN (see Fig. 1 and Table 2) was trained and validated on RaFD-FRONT using a 5-fold cross-validation scheme. The cross entropy loss and the Adam Optimizer [35] was employed. After 10,000 epoches, the network achieved a validation accuracy of 97.3% $\pm$ 1.2% which is a good result among the recent work. The baseline network is capable of recognizing emotions of faces in RaFD-FRONT. After the baseline CNN was modified by adding an attention mapping between its convolutional part and fully connected part (see Fig. 2 and Table 3), the proposed CNN with Visual Attention achieved a good accuracy of 95.2% $\pm$ 2.0% (use free weight map Eq. (4)) and 94.8% $\pm$ 0.8% (use symmetric weight map Eq. (5)).

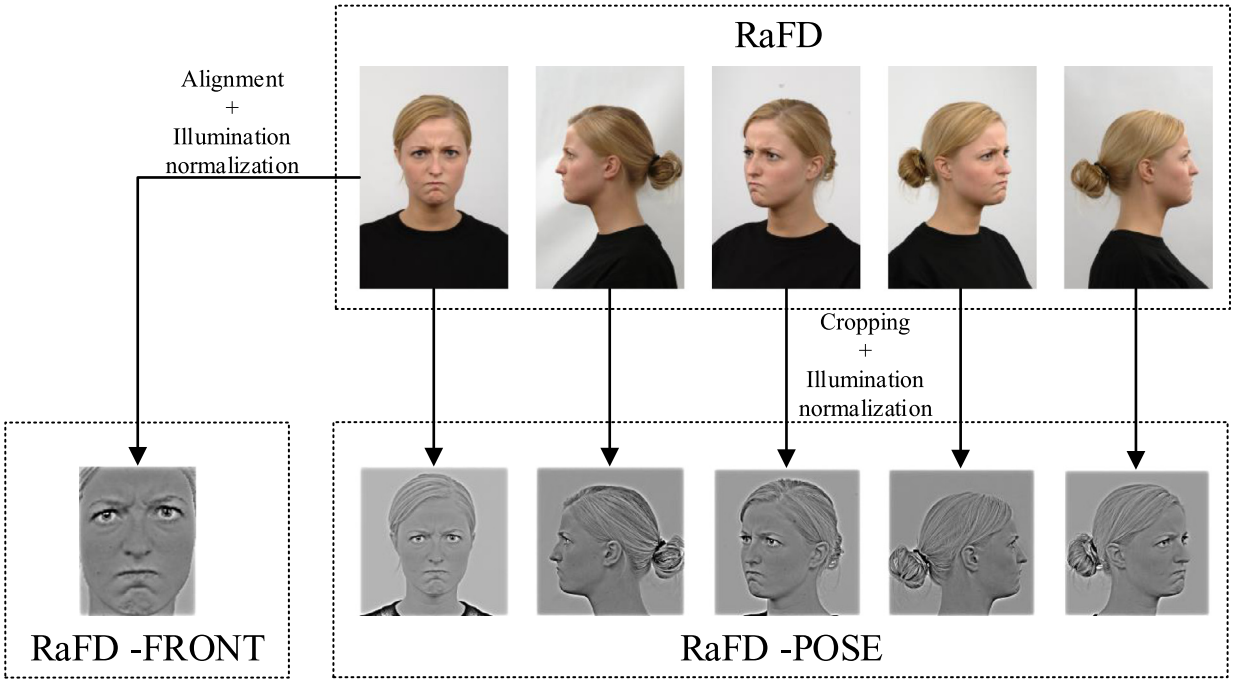


Fig. 3. The RaFD and the preprocessed data.

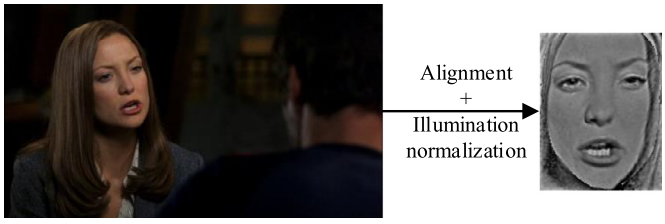


Fig. 4. The SFEW dataset and the preprocessed data.

Besides the evaluation of the performances, the input images and weight maps $\boldsymbol{B}$ (projected back to the input space) are averaged and visualized for each emotion. As illustrated in Fig. 5, the proposed CNNs with Visual Attention using free weight map Eq. (4) and symmetric weight map Eq. (5) are compared. In Fig. 5(a), without the symmetric constraint, the attention mechanism prefers to choose only one side of the face since the other side contains redundant information for spatial aggregating. in Fig. 5(b), the symmetric weight maps for each type of the emotions are also demonstrated. The visual attention mechanism using symmetric weight map discovers the symmetric ROIs, which human can clearly recognize and understand. These regions are partly consistent with the locations of expression specific Action Units listed in Table 4. Our new founding confirms the interpretation of FACS and EMFACS from a machine learning perspective.

Further more, the global average pooling method (fixed uniform weight map) is also evaluated for comparison. All the experiment results are summarised in Table 5. These compared methods share the same meta-parameters in their convolutional parts and fully connected parts. The baseline CNN uses $7 \times 9 \times 128$ dimensional features for the final classification. The CNN with Visual Attention / global average pooling achieve comparable cross validation accuracies, when the dimension of their features is extremely compressed (from $7 \times 9 \times 128$ to 128). Although the proposed method does not achieve the best performance, it is the only one which can give an explainable result among the compared methods.

Table 4
The relation between emotions and Action Units [43].

Emotions	Action units	ROI
Angry	4(Brow Lowerer)+5B(Slight Upper Lid Raiser)+ 7(Lid Tightener)+23(Lip Tightener)	Brow+Lid+Lip
Disgust	9(Nose Wrinkler)+15(Lip Corner Depressor)+ 16(Lower Lip Depressor)	Nose+Lip
Fear	1(Inner Brow Raiser)+2(Outer Brow Raiser)+ 4(Brow Lowerer)+5(Upper Lid Raiser)+ 7(Lid Tightener)+20(Lip Stretcher)+26(Jaw Drop)	Brow+Lid+Lip+Jaw
Happy	6(Cheek Raiser)+12(Lip Corner Puller)	Check+Lip
Sad	1(Inner Brow Raiser)+4(Brow Lowerer)+ 15(Lip Corner Depressor)	Brow+Lip
Surprise	1(Inner Brow Raiser)+2(Outer Brow Raiser)+ 5B(Upper Lid Raiser)+26(Jaw Drop)	Brow+Lid+Jaw

The performances reported in the related work are also compared in Table 5. The Local Binary Pattern (LBP) [36–38], Gabor [39] and Local phase quantization (LPQ) [40] are commonly used hand-crafted features for facial expression recognition. The Down Sampling (DS) features are down-sampled 12×10 facial patches. Four SVM classifiers are trained on these features separately. The result shows that our CNNs are superior to these conventional methods. The Facial Expression Generic Elastic Model (FE-GEM) [41] is a novel 3D reconstruction method for aligned faces. It can reconstruct the depth map from a single 2D face. Four SVM classifiers are trained on features extracted from both the original faces and the depth maps estimated by FE-GEM. The combination of Gabor+FE-GEM+SVM achieves the state-of-the-art performances. The only limitation is that the FE-GEM model needs to be trained on an external 3D facial dataset [42]. The performances of our CNNs are very close to the state-of-the-art method. CNNs and the FE-GEM are orthogonal techniques. CNNs can accept both the original faces and the depth maps estimated by FE-GEM as the multi-channel inputs. Such networks may get better results. The combinations of our CNNs and the FE-GEM method are beyond the scope of this paper.

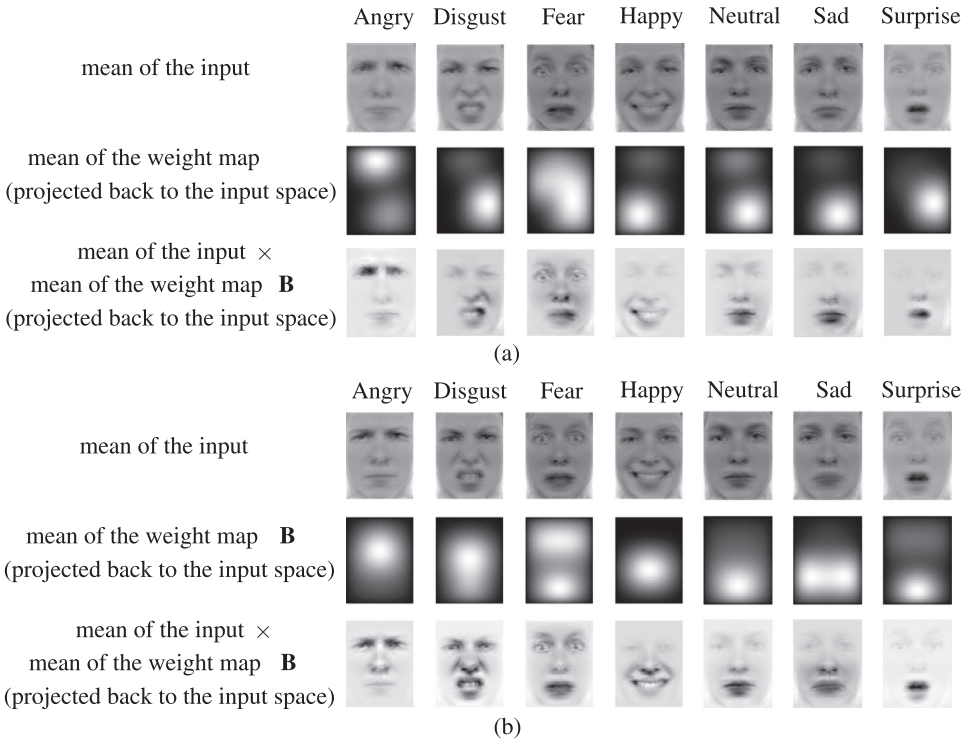


Fig. 5. The visualization of visual attentions for aligned frontal faces.
(a) use free weight map Eq. (4), (b) use symmetric weight map Eq. (5).

Table 5
The accuracies of the classifiers.

Dataset	Method	Validation accuracy (%)
RAFD-FRONT	The baseline CNN	97.3 ± 1.2
	The CNN with Visual Attention (use Eq. (4))	95.2 ± 2.0
	The CNN with Visual Attention (use Eq. (5))	94.8 ± 0.8
	The CNN with global average pooling	96.3 ± 1.7
	LBP+SVM [36,37]	86.5
	Gabor+SVM [38]	83.1
	LPQ+SVM [39]	84.8
	DS+SVM [40]	79.0
	LBP+FE-GEM+SVM [41][a]	94.5
	Gabor+FE-GEM+SVM [41][a]	98.1
	LPQ+FE-GEM+SVM [41][a]	94.4
	DS+FE-GEM+SVM [41][a]	90.8
RAFD-POSE	The baseline CNN	87.2 ± 2.8
	The CNN with Visual Attention (use Eq. (4))	93.1 ± 0.6
	The CNN with global average pooling	92.8 ± 1.7
	Human recognition rate [27]	82.0
SFEW	The baseline CNN	38.5
	The CNN with Visual Attention (use Eq. (4))	40.0
	The CNN with global average pooling	32.8
	The baseline CNN[a]	46.3
	The CNN with Visual Attention (use Eq. (4))[a]	48.3
	The CNN with global average pooling[a]	42.8
	Kim's CNN[b]	52.5
	Levi's LBP+CNN[c]	44.7
	PHOG+LPQ+SVM [27]	36.0

[a] Trained on external dataset.
[b] The winner of EmotiW-2015-SFEW, the network is referred as "$PREPiNor, oA - \{CNNL - FC3072\}_{R1}$" in [17].
[c] The network is referred as "LBP1, cyclic, VGG_M-4096 - Oversampling" in [5].

4.4. Facial expression recognition on RaFD-POSE dataset

In this section, our networks are evaluated on the unaligned RaFD-POSE. This task is more challenging than that in Section 4.3. For the reason that the image size of the RaFD-POSE is different from the one of RaFD-FRONT, slight modifications should be made based on the network described in Tables 2 and 3. The dimension of the input layer is enlarged from 112 × 144 to 272 × 272. The dimensions of the activation maps of the convolutional layers and the dimensions of the hidden fully connected layers are enlarged correspondingly. The other meta-parameters remain the same as those listed in Tables 2 and 3.

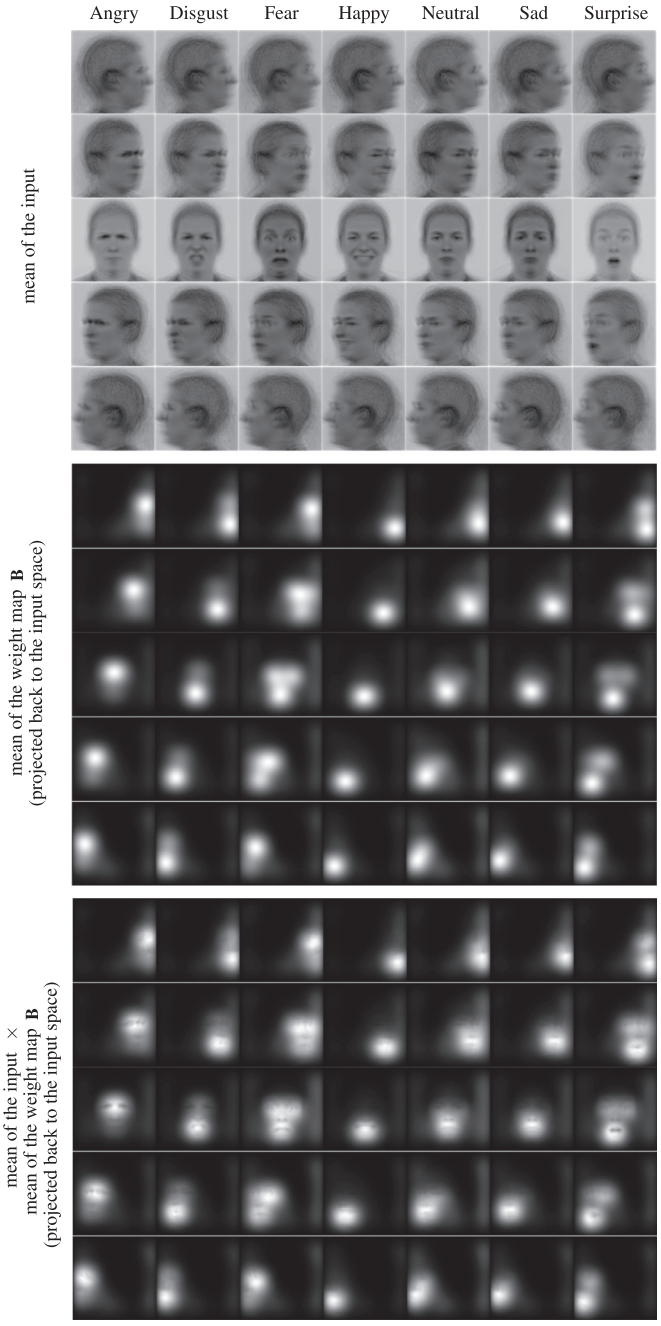


Fig. 6. The visualization of visual attentions for faces in different poses.

Both CNNs with / without visual attention mechanism were trained and validated on RaFD-POSE using a 5-fold cross-validation scheme. The cross entropy loss and the Adam Optimizer [35] was employed. After 130 epoches, the proposed network achieved good performance as listed in Table 5. Compared with the classification task on aligned frontal faces in Section 4.3, the visual attention mechanism is more important in this experiment. It may play a role of landmark locator for unaligned faces. Therefore, the cross validation accuracy increases 5.9% (from 87.2 to 93.1%) with the help of the visual attention mechanism.

As illustrated in Fig. 6, the input images and weight maps $\boldsymbol{B}$ (projected back to the input space) are averaged and visualized per emotion per pose. Similar to the visualization in Fig. 5, the regions of interests for each type of the emotions have been successfully discovered. Furthermore, these regions are pose invariant. In each column of Fig. 6, the region of interest always locates the corresponding Action Units, while the face is rotating. Suppose that we have ideally collected sufficient facial images in continuous yaw / pitch angles. The difference between two adjacent images can be seen as a kind of small deformation. The continuous mapping: *input* $\boldsymbol{X} \rightarrow \boldsymbol{A}_0 \rightarrow \boldsymbol{B}$ may be trained to give smooth and correct attention weights while the facial image deforms continually. In practice, even the training data RaFD-POSE have only five poses, the continuous mapping still seems to be well trained.

Further more, the global average pooling method (fixed uniform weight map) is also evaluated for comparison. As listed in Table 5, the baseline CNN is not suitable for unaligned faces. It achieves the lowest cross validation accuracy of 87.2%. The CNN with global average pooling achieves a better accuracy of 92.8%. By contrast, the proposed CNN with Visual Attention (use Eq. (4)) achieves the best accuracy of 93.1%, even the dimension of the feature is extremely compressed (from $17 \times 17 \times 128$ to 128). Our CNNs surpass the human recognition rate which is reported by the creator of the dataset [27].

The attention model has different effects on different datasets. The accuracy on the RaFD-POSE dataset increases 5.9% (from 87.2 to 93.1%) by inducing the attention model. Compared to it, the attention model does not improve the accuracy on the RaFD-FRONT dataset (from 97.3 to 95.2%). A suitable explanation is that the attention model has its two sides. On the positive side, it has a similar function to a face detector or an AU detector. It detects the ROI, amplifies the signals in the ROI and suppresses the background noises. On the negative side, the attention model is a learning system. The ROI predicting error has bad effects on the subsequent processing. In Figs. 5 and6, the ROI in the aligned / unaligned facial expression recognition task is different. In unaligned facial expression recognition tasks, the positive side dominates the result since the attention model makes the feature pose invariant. The ROIs of the unaligned faces are generally smaller and more variable than those of the aligned faces. In aligned facial expression recognition tasks, the ROI is stable. The fully connected layer of the neural network can learn connecting weights which tells the importance of different local regions, reduce the requirement of the attention based ROI detection. The negative side dominates the result. In a word, we should make the best use of the advantages and bypass the disadvantages according to the above analysis.

The confusion matrices evaluated on the first fold of the RaFD-FRONT dataset and the RaFD-POSE dataset are illustrated in Table 6. The confusion matrices show that the networks separate each class from others well, except for the surprise-fear and neutral-sad pairs which are naturally hard to distinguish.

4.5. Facial expression recognition on SFEW dataset

The proposed method is also evaluated on the unconstrained SFEW dataset and FER-2013 dataset. The learning method is similar to that in Section 4.3. The cross entropy loss and the Adam Optimizer [35] was employed to train the baseline CNN, the CNN with Visual Attention, and the CNN with global average pooling on the preprocessed SFEW dataset. The validation accuracies at epoch 300(trained with FER-2013) / 10,000(trained without FER-2013) are listed in Table 5. The visual attention model improves the the performance of the baseline CNN by 1.5% (from 38.5 to 40.0%). Moreover, by using the external training data the performance is further improved by 8.3% (from 40.0 to 48.3%). Compared to it, the global average pooling makes the performance worse since it is very sensitive to background noises.

Some state-of-the-art methods use ensembles of deep neural networks to achieve superior performance [4,5,17–22]. This paper focuses on facial expression recognition algorithm based on single network and single frame. For the sake of fairness, only single

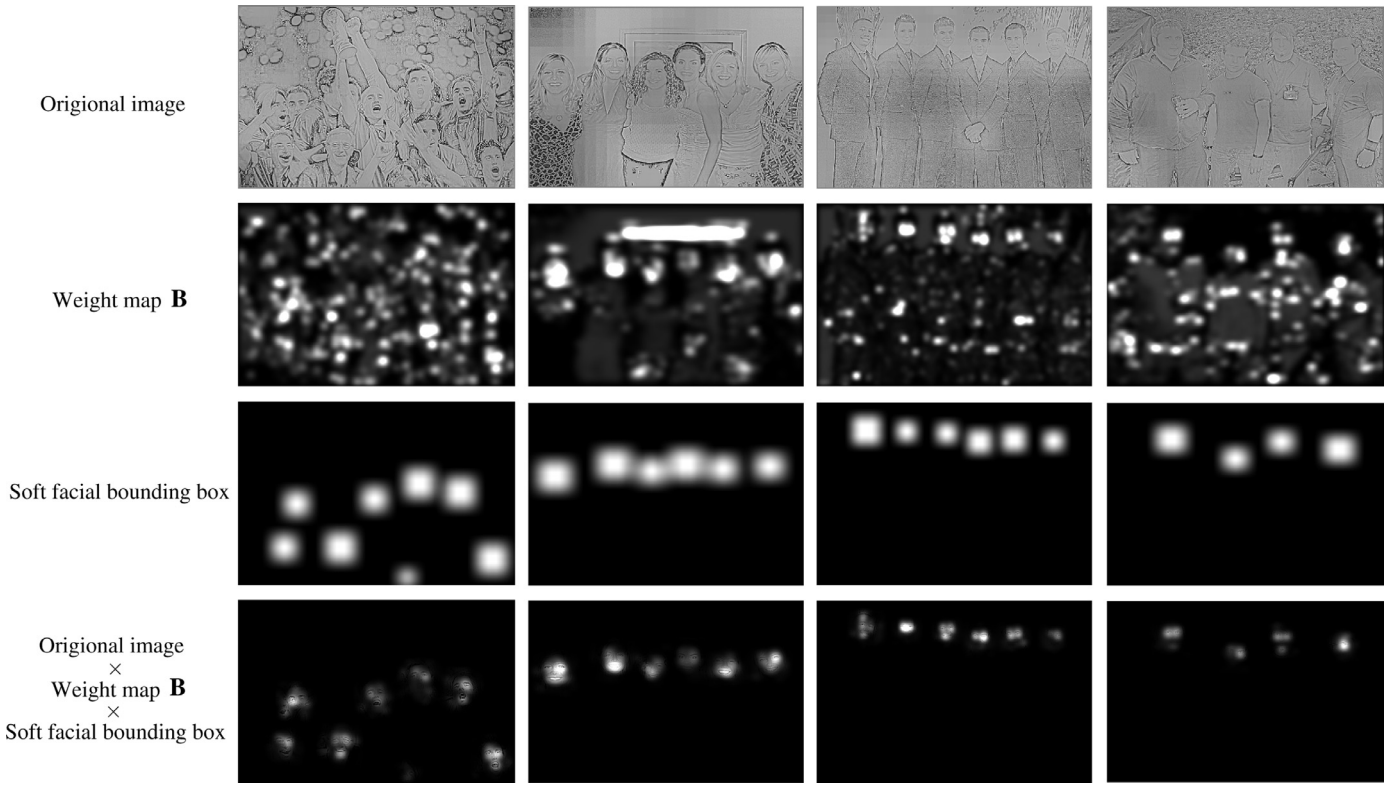


Fig. 7. The visualization of visual attentions on the HAPPEI dataset.

Table 6
The confusion matrices.

(a) The baseline CNN method validated on RaFD-FRONT (%)

	A	D	F	H	N	Sa	Su
A	100.0	0.0	0.0	0.0	0.0	0.0	0.0
D	0.0	100.0	0.0	0.0	0.0	0.0	0.0
F	0.0	0.0	97.4	0.0	0.0	0.0	2.6
H	0.0	0.0	0.0	100.0	0.0	0.0	0.0
N	0.0	0.0	0.0	0.0	100.0	0.0	0.0
Sa	0.0	0.0	0.0	0.0	5.1	94.9	0.0
Su	0.0	0.0	0.0	0.0	0.0	0.0	100.0

(b) The CNN with Visual Attention (use Eq. (4)) method validated on RaFD-FRONT (%)

	A	D	F	H	N	Sa	Su
A	94.9	0.0	0.0	0.0	0.0	5.1	0.0
D	0.0	100.0	0.0	0.0	0.0	0.0	0.0
F	0.0	0.0	89.7	0.0	0.0	2.6	7.7
H	0.0	0.0	0.0	100.0	0.0	0.0	0.0
N	0.0	0.0	0.0	0.0	100.0	0.0	0.0
Sa	2.6	0.0	0.0	0.0	12.8	84.6	0.0
Su	0.0	0.0	2.6	0.0	0.0	0.0	97.4

(c) The baseline CNN method validated on RaFD-POSE (%)

	A	D	F	H	N	Sa	Su
A	86.2	1.0	2.1	0.0	7.2	3.6	0.0
D	4.1	90.3	2.1	1.5	1.0	1.0	0.0
F	0.0	0.0	81.5	0.0	7.2	0.5	10.8
H	0.5	0.5	0.0	99.0	0.0	0.0	0.0
N	6.2	0.0	0.5	0.5	83.6	8.7	0.5
Sa	10.3	0.0	5.1	0.0	20.5	62.6	1.5
Su	0.0	0.0	5.6	0.0	0.0	0.0	94.4

(d) The CNN with Visual Attention (use Eq. (4)) method validated on RaFD-POSE (%)

	A	D	F	H	N	Sa	Su
A	97.4	0.5	0.0	2.1	0.0	0.0	0.0
D	4.1	92.8	0.5	2.1	0.0	0.0	0.5
F	0.0	1.0	97.4	0.0	0.0	0.0	1.5
H	0.0	0.0	0.0	100.0	0.0	0.0	0.0
N	2.1	0.0	3.1	2.1	87.2	5.6	0.0
Sa	2.6	0.5	5.6	0.0	3.1	88.2	0.0
Su	0.0	0.0	6.7	0.0	0.0	0.0	93.3

networks are compared in Table 5. The proposed CNN with Visual Attention outperforms the baseline CNN, the CNN with global average pooling, and the PHOG+LPQ+SVM method proposed by the creator of the SFEW dataset [16,28]. Also, our method is close to the state-of-the-art single networks [5,17]. The Kim's CNN [17] is a member of the ensemble network which won the EmotiW-2015-SFEW challenge.

The unconstrained facial expression recognition task is more challenging than the constrained one. Although the faces are aligned, they have various of dirty backgrounds out of the ROI. The features extracted from the dirty background will reduce the accuracies. Theoretically, CNN feature can be invariant to these backgrounds. But due to the small sample size limitation, calculating the relative importance of the pixels (i.e. the attention map) is more practical. The performance can be improved by aggregating the features in the ROI.

4.6. Visualizing the learned attention model on HAPPEI dataset

Our last experiment provides a preliminary evaluation of the proposed network on grouped unconstrained faces. We directly extend the proposed network to its fully convolutional version [30] for processing images of arbitrary size efficiently. As illustrated in Fig. 2, the convolutional part calculates convolutional feature maps for arbitrary-sized images. The attention mapping part generates variable-sized attention weights, and aggregates variable-sized convolutional feature maps into a fixed-length 1D vector. Although

Table 7
Comparing of feature aggregating methods.

Method	Receptive field	Needs learning
Max-pooling	Small local patches	No
Stacked max-pooling	Large local patches	No
Global average pooling	The whole image	No
Visual attention mechanism	The region of interest of the whole image	Yes

the dimensions of the input and activation of these layers are variable, the numbers of parameters are always fixed.

Our Fully Convolutional Neural Network (FCNN) is trained on the RaFD-POSE dataset and visualized on the HAPPEI dataset [29]. More exactly, only four test images in the HAPPEI dataset are selected for visualization (see the first row of Fig. 7). Each image is a group photo containing several faces with various backgrounds. Our network was trained on the constrained RaFD-POSE. It has not seen such complicated test images before. The inferred attention weights are illustrated in the second row of Fig. 7. Our network makes many false positives in background areas for the reason that the background model is never learned. We simply used soft facial bounding boxes detected by an existing face detector [31] to suppress these false positives. To our surprise, the suppressed attention weight maps look good. As shown in the bottom row of Fig. 7, the regions of interests of the faces are similar to those in Figs. 5 and6. It shows that the proposed network can also be extended to its fully convolutional version for grouped faces.

4.7. Further analysis of the visual attention mechanism

The visual attention mechanism extremely compress the high-level feature from a $7 \times 9 \times 128$ / $17 \times 17 \times 128$ dimensional tensor to a 128 dimensional vector without loosing classification accuracy. The elements of the 128 dimensional vector denote the existences of factors in the region of interest. Table 7 compares three commonly used feature aggregating methods namely max-pooling, global average pooling [26] and visual attention mechanism. The differences of these methods are obvious.

- Max-pooling is often paired with convolutions. A CNN often contains a stack of convolution layers and max-pooling layers. Stacked max-pooling layers are designed to reduce the spatial dimensions of the activation map and increase the area of the receptive field of the elements in the deep activation map. For the reason that the max operator is sensitive to noise, max-pooling can not be used globally.
- Global average pooling sums out the feature map spatially, thus the network is more robust to spatial translations of the input. The receptive field of global average pooling is the whole image.
- Visual attention mechanism is very close to global average pooling. It sums out the feature map spatially based on attention weights $\boldsymbol{B}$. The weights are learned from the supervised data and tell the importance of the regions. The receptive field of visual attention mechanism is the region of interest of the whole image. Visual attention mechanism is more reasonable than global average pooling.

In facial expression recognition task, humans always pay their attentions around eyes and mouths and ignore the hairs, noses, ears, backgrounds, etc. The behavior of humans is a kind of explanation for the proposed CNN with Visual Attention. Furthermore, experts defined the Action Units precisely according the muscular movements in different regions. Then they defined six emotions according to six combinations of Action Units. The existences of Action Units are latent factors in facial expression recognition. Aggregating features using the learned attention weights can be viewed as a kind of rough detection for latent factors.

5. Conclusion

A CNN with Visual Attention is proposed for solving the facial expression recognition problem. Some empirical studies show that the proposed network is able to learn regions of interests which are partly consistent with the locations of expression specific Action Units. Our founding confirms the interpretation of FACS and EMFACS from a machine learning perspective.

Learning meaningful latent factors for specific tasks is an interesting topic. It may be a good way to understand the internal mechanism of neural networks. In this work, the regions of interests for facial expression recognition are located by the proposed CNN with Visual Attention. According to the interpretation of FACS and EMFACS, aggregating features in these regions is efficient. Furthermore, the latent factors should not be limited to the Action Units studied in this paper. Such as common properties of a person, local textures, temporal motions, etc. are important clues for expression recognition. They should be pay attention to in the further work.

Acknowledgments

This work is partially supported by National Natural Science Foundation of China under Grant nos. 61373063, 61375007, 61233011, 91420201, 61472187 and by National Basic Research Program of China under Grant No. 2014CB349303.

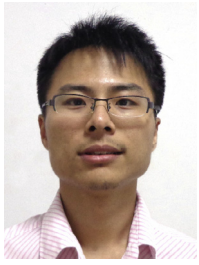
Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.neucom.2018.03.034.

References

[1] A. Mehrabian, S.R. Ferris, Inference of attitudes from nonverbal communication in two channels, J. Consult. Psychol. 31 (3) (1967) 248.
[2] P. Ekman, E. Friensen, Facial Action Coding System(FACS): Manual, Palo Alto: Consulting Psychologists Press, 1978.
[3] W. Friensen, P. Ekman, EMFACS-7: emotional facial action coding system, Technical Report, University of California at San Francisico, 1983.
[4] S.E. Kahou, C. Pal, X. Bouthillier, P. Froumenty, Ç. Gülçehre, R. Memisevic, P. Vincent, A. Courville, Y. Bengio, R.C. Ferrari, et al., Combining modality specific deep neural networks for emotion recognition in video, in: Proceedings of the Fifteenth ACM on International Conference on Multimodal Interaction, ACM, 2013, pp. 543–550.
[5] G. Levi, T. Hassner, Emotion recognition in the wild via convolutional neural networks and mapped binary patterns, in: Proceedings of the ACM on International Conference on Multimodal Interaction, ACM, 2015, pp. 503–510.
[6] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, Y. Bengio, Show, attend and tell: Neural image caption generation with visual attention, Proceedings of the Thirty-second International Conference on Machine Learning, 2015, pp. 2048–2057.
[7] M. Riemer, A. Vempaty, F.P. Calmon, F.F. Heath III, R. Hull, E. Khabiri, Correcting forecasts with multifactor neural attention, in: Proceedings of the Thirty-third International Conference on Machine Learning, 2016, pp. 3010–3019.
[8] K. Simonyan, A. Vedaldi, A. Zisserman, Deep inside convolutional networks: Visualising image classification models and saliency maps, arXiv:1312.6034 (2013).
[9] M.D. Zeiler, D. Krishnan, G.W. Taylor, R. Fergus, Deconvolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2010, pp. 2528–2535.
[10] M.D. Zeiler, G.W. Taylor, R. Fergus, Adaptive deconvolutional networks for mid and high level feature learning, in: Proceedings of the International Conference on Computer Vision, IEEE, 2011, pp. 2018–2025.
[11] I. Biederman, Visual Object Recognition, Vol. 2, MIT Press Cambridge, 1995.
[12] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Object detectors emerge in deep scene cnns, arXiv:1412.6856 (2014).
[13] M. Liu, S. Li, S. Shan, X. Chen, Au-aware deep networks for facial expression recognition, in: Proceedings of the Tenth IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, IEEE, 2013, pp. 1–6.
[14] Y. Zhou, B.E. Shi, Action unit selective feature maps in deep networks for facial expression recognition, in: Proceedings of the International Joint Conference on Neural Networks, IEEE, 2017, pp. 2031–2038.

[15] A. Yao, J. Shao, N. Ma, Y. Chen, Capturing au-aware facial features and their latent relations for emotion recognition in the wild, in: Proceedings of the ACM on International Conference on Multimodal Interaction, ACM, 2015, pp. 451–458.
[16] A. Dhall, O. Ramana Murthy, R. Goecke, J. Joshi, T. Gedeon, Video and image based emotion recognition challenges in the wild: Emotiw 2015, in: Proceedings of the ACM on International Conference on Multimodal Interaction, ACM, 2015, pp. 423–426.
[17] B.-K. Kim, H. Lee, J. Roh, S.-Y. Lee, Hierarchical committee of deep CNNS with exponentially-weighted decision fusion for static facial expression recognition, in: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, ACM, 2015, pp. 427–434.
[18] Y. Fan, X. Lu, D. Li, Y. Liu, Video-based emotion recognition using CNN-RNN and c3d hybrid networks, in: Proceedings of the Eighteenth ACM International Conference on Multimodal Interaction, ACM, 2016, pp. 445–450.
[19] A. Yao, D. Cai, P. Hu, S. Wang, L. Sha, Y. Chen, Holonet: towards robust emotion recognition in the wild, in: Proceedings of the Eighteenth ACM International Conference on Multimodal Interaction, ACM, 2016, pp. 472–478.
[20] P. Hu, D. Cai, S. Wang, A. Yao, Y. Chen, Learning supervised scoring ensemble for emotion recognition in the wild, in: Proceedings of the Nineteenth ACM International Conference on Multimodal Interaction, ACM, 2017, pp. 553–560.
[21] B. Knyazev, R. Shvetsov, N. Efremova, A. Kuharenko, Convolutional neural networks pretrained on large face recognition datasets for emotion classification from video, arXiv:1711.04598 (2017).
[22] V. Vielzeuf, S. Pateux, F. Jurie, Temporal multimodal fusion for video emotion classification in the wild, in: Proceedings of the Nineteenth ACM International Conference on Multimodal Interaction, ACM, 2017, pp. 569–576.
[23] P. Carrier, A. Courville, I. Goodfellow, M. Mirza, Y. Bengio, FER-2013 face database, Technical Report, Technical report, 1365, Université de Montréal, 2013.
[24] CASIA, Casia webface database, 2015. http://www.cbsr.ia.ac.cn/english/CASIA-WebFace-Database.html.
[25] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv:1409.1556 (2014).
[26] M. Lin, Q. Chen, S. Yan, Network in network, arXiv:1312.4400 (2013).
[27] O. Langner, R. Dotsch, G. Bijlstra, D.H. Wigboldus, S.T. Hawk, A. van Knippenberg, Presentation and validation of the Radboud faces database, Cognit. Emot. 24 (8) (2010) 1377–1388.
[28] A. Dhall, R. Goecke, S. Lucey, T. Gedeon, Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark, in: Ptoceedings of the IEEE International Conference on Computer Vision Workshops (ICCV Workshops), IEEE, 2011, pp. 2106–2112.
[29] A. Dhall, R. Goecke, T. Gedeon, Automatic group happiness intensity analysis, IEEE Trans. Affect. Comput. 6 (1) (2015) 13–26.
[30] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3431–3440.
[31] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: Proceedings of the IEEE IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), Vol. 1, IEEE, 2005, pp. 886–893.
[32] V. Kazemi, J. Sullivan, One millisecond face alignment with an ensemble of regression trees, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1867–1874.
[33] T. Hassner, S. Harel, E. Paz, R. Enbar, Effective face frontalization in unconstrained images, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 4295–4304.
[34] R. Gross, V. Brajovic, An image preprocessing algorithm for illumination invariant face recognition, in: Proceedings of the International Conference on Audio-and Video-Based Biometric Person Authentication, Springer, 2003, pp. 10–18.
[35] D. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv:1412.6980 (2014).
[36] C. Shan, S. Gong, P.W. McOwan, Facial expression recognition based on local binary patterns: a comprehensive study, Image Vis. Comput. 27 (6) (2009) 803–816.
[37] S. Moore, R. Bowden, Local binary patterns for multi-view facial expression recognition, Comput. Vis. Image Underst. 115 (4) (2011) 541–558.
[38] C. Liu, H. Wechsler, A gabor feature classifier for face recognition, in: Proceedings of the Eighth IEEE International Conference on Computer Vision, Vol. 2, IEEE, 2001, pp. 270–275.
[39] V. Ojansivu, J. Heikkilä, Blur insensitive texture classification using local phase quantization, in: Proceedings of the International Conference on Image and Signal Processing, Springer, 2008, pp. 236–243.
[40] J. Wright, A.Y. Yang, A. Ganesh, S.S. Sastry, Y. Ma, Robust face recognition via sparse representation, IEEE Trans. Pattern Anal. Mach. Intell. 31 (2) (2009) 210–227.
[41] A. Moeini, H. Moeini, Multimodal facial expression recognition based on 3d face reconstruction from 2d images, in: Proceedings of the Face and Facial Expression Recognition from Real World Videos, Springer, 2015, pp. 46–57.
[42] A. Savran, N. Alyüz, H. Dibeklioğlu, O. Çeliktutan, B. Gökberk, B. Sankur, L. Akarun, Bosphorus database for 3d face analysis, in: Proceedings of the European Workshop on Biometrics and Identity Management, Springer, 2008, pp. 47–56.
[43] Arcadian, Facial action coding system, 2016. https://en.wikipedia.org/wiki/Facial_Action_Coding_System.



Wenyun Sun received the B.S. degree in Computer Science and Technology and the M.S. degree in Pattern Recognition and Intelligent System from Jiangsu University of Science and Technology, Zhenjiang, China in 2009 and 2012, respectively. He is currently pursuing the Ph.D. degree in Pattern Recognition and Intelligent Systems at Nanjing University of Science and Technology, Nanjing, China.

Haitao Zhao received his Ph.D. degree in pattern recognition and intelligent system from Nanjing University of Science and Technology, Nanjing, China in 2003. Now he is a professor at East China University of Science and Technology, Shanghai, China. His current interests are in the areas of pattern recognition, machine learning and computer vision.

Zhong Jin received the B.S. degree in mathematics, M.S. degree in applied mathematics and the Ph.D. degree in pattern recognition and intelligent system from Nanjing University of Science and Technology, Nanjing, China in 1982, 1984 and 1999, respectively. His current interests are in the areas of pattern recognition and face recognition.