Pattern Recognition 84 (2018) 251–261

Contents lists available at ScienceDirect

Pattern Recognition

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

Conditional convolution neural network enhanced random forest for facial expression recognition

Yuanyuan Liu[a], Xiaohui Yuan[b,*], Xi Gong[a], Zhong Xie[a], Fang Fang[a], Zhongwen Luo[a]

[a] Faculty of information engineering, China University of Geosciences, Wuhan, China
[b] Department of Computer Science and Engineering, University of North Texas, Denton, TX, USA

ARTICLE INFO

Article history:
Received 15 December 2017
Revised 24 June 2018
Accepted 10 July 2018
Available online 17 July 2018

Keywords:
Classification
Feature extraction
Facial expression recognition
Head pose alignment
Conditional CoNERF

ABSTRACT

In real-world applications, factors such as head pose variation, occlusion, and poor image quality make facial expression recognition (FER) an open challenge. In this paper, a novel conditional convolutional neural network enhanced random forest (CoNERF) is proposed for FER in unconstrained environment. Our method extracts robust deep salient features from saliency-guided facial patches to reduce the influence from various distortion types, such as illumination, occlusion, low image resolution, etc. A conditional CoNERF is devised to enhance decision trees with the capability of representation learning from transferred convolutional neural networks and to model facial expression of different perspectives with conditional probabilistic learning. In the learning process, we introduce a neurally connected split function (NCSF) as the node splitting strategy in the CoNERF. Experiments were conducted using public CK+, JAFFE, multi-view BU-3DEF and LFW datasets. Compared to the state-of-the-art methods, the proposed method achieved much improved performance and great robustness with an average accuracy of 94.09% on the multi-view BU-3DEF dataset, 99.02% on CK+ and JAFFE frontal facial datasets, and 60.9% on LFW dataset. In addition, in contrast to deep neural networks which require large-scale training data, conditional CoNERF performs well even when there are only a small amount of training data.

1. Introduction

Facial expression recognition (FER) has become a hot research topic of human–computer interaction. Human facial expressions provide important clues concerning human emotion and behavior. Recognizing facial expressions is crucial to applications such as digital entertainment, customer service, driver monitoring, and emotional robots [1–3]. There have been extensive studies, and methods were developed. A majority of the proposed methods were evaluated with constrained frontal FER, and their performance degenerates when dealing with cases of non-frontal and multi-view FER [4,5]. To deal with such a challenge, this paper proposes a conditional convolutional neural enhanced forests (CoNERF) to recognize facial expression under multi-view and unconstrained environment with great efficiency and robustness.

A general FER framework consists of two major steps: feature extraction and classifier construction. Extracting robust facial features and devising an effective classifier are the two key components for the unconstrained FER task. Both local facial feature and global facial feature methods have been developed for FER. The accuracy of methods based on local features relies on the detection accuracy of eyes, eyebrows, nose, and lips [6–8]; the methods based on global feature usually use texture features to recognize expressions [9–11], which is applicable for low-resolution images but lacks robustness to occlusion and illumination variance. In real-world cases, the head pose variation, partial occlusion, and low image quality make feature extraction a challenging task. Among the three degrees of freedom for the head poses, yaw motion introduces the greatest variation in facial images. With respect to classification, convolutional neural networks (CNN) has recently gained great popularity for real-life applications because of their superior performance and robustness. CNN automatically learn high-level feature representations from images [1,12–14] but demands a large training dataset and high-performance computing, e.g., GPUs [14,15]. In this paper, we propose a learning method that leverages a global deep salient representation and requires a small amount of image data. Our method aims at improving both accuracy and efficiency in multi-view FER. The workflow of our proposed conditional convolutional neural network enhanced forests (CoNERF) is shown in Fig. 1. The deep feature is extracted from salient facial patches by suppressing the influence of illumination, occlusion, and low image resolution. Yaw angle is estimated to overcome the variance among head poses. The multi-view fa-

* Corresponding author.
E-mail addresses: xiaohui.yuan@unt.edu, xyuan@cse.unt.edu (X. Yuan).

https://doi.org/10.1016/j.patcog.2018.07.016
0031-3203/

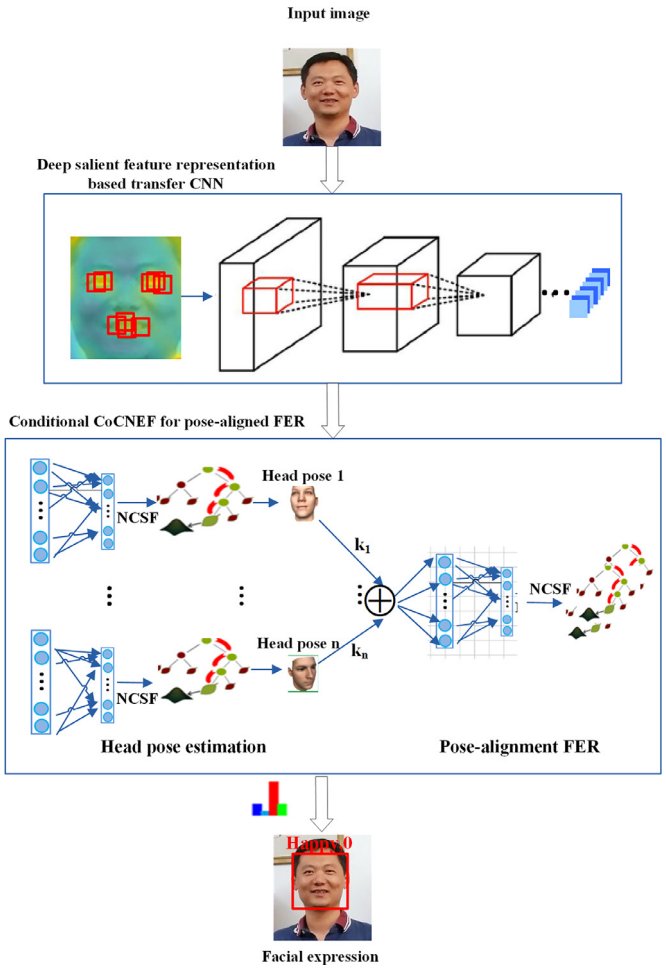

Fig. 1. An overview of our proposed conditional CoNERF for facial expression recognition.

cial expressions are estimated under the conditional probability of head pose alignment. The multi-view facial expressions are estimated under the conditional probability of head pose alignment.

Our contributions include the following:

1. a conditional convolutional neural network enhanced random forests (CoNERF) for pose-aligned facial expression recognition in an unconstrained environment, which is unified classification trees with the representation learning from deep convolution networks, by training them in an end-to-end way. Besides, we introduce a neurally connected split function (NCSF) as new split node learning in a CoNERF. The CoNERF method can achieve fast and accurate recognized results in the limited amount of image data, rather than a large amount of data required by CNN.
2. a conditional probabilistic learning method for pose alignment in multi-view facial expression recognition to suppress errors from head pose variations.
3. a robust deep salient feature representation based on saliency-guided facial patches using “visual attention” mechanisms.

The rest of this paper is organized as follows: Section 2 reviews the related work. Section 3 presents our conditional CoNERF method. Section 4 discusses the experimental results using publicly available datasets. Section 5 concludes this paper with a summary of our method.

2. Related work

Convolutional neural network. Convolutional neural network (CNN) is a type of deep learning that can learn deep generic features and classify data [16]. The design of CNN follows the discovery of visual mechanisms in living organisms, including convolutional layers, pooling layers, fully connected layers and softmax classification. Inspired by the success in face recognition and ImageNet Classification [15,17], CNNs have been applied to FER [6,13,18–21]. However, the improvement heavily relies on a large number of training sets and high-performance computing power. Moreover, recent studies reveal that a deep CNN can learn transferable features which generalize well to novel tasks for domain adaptation [12–14,22]. Bulo and Kontschieder [23] jointly tackled deep data representation and discriminative learning within randomized decision trees for semantic image labeling, which improved results and significantly compressed 70 trees compared to conventional decision trees. Yang et al. [22] proposed to extract complete and robust local regions and learned convolutional features by using CNNs by densely sampling and sparsely detecting facial points. These features are adaptive to the local regions and discriminative to the face expression.

Random Forest. RF is a popular method in computer vision given its capability to handle large training datasets, high generalization power and speed, and easy implementation [24–28]. It has emerged as a powerful and versatile method successful in real-time FER system, head pose estimation, facial point detection and action recognition. Sun et al. [26] employed a conditional RF for real-time body pose estimation from depth data. Fanelli et al. [29] presented a Hough forests for facial expression recognition from image sequences, which achieved a recognition rate of 76% on MMI spontaneous expression dataset. A conditional RF also has been proposed to estimate facial feature points under various head poses in [25]. Multi-class RF becomes a popular method for multi-view facial analysis in unconstrained environment owing to their robustness.

Facial expression recognition. Lots of works have existed and obtained excellent results on constrained frontal FER [7,29–33]. Comparing to frontal FER, non-frontal FER is more challenging and more applicable in real scenarios. However, only a part of work address some challenging issues in multi-view and unconstrained environment [1,6,8,34,35]. Dapogny et al. [35] proposed Pair Conditional Random Forests (PC-RF) to capture low-level expression transition patterns on the condition of head pose estimation for multi-view dynamic facial expression recognition. On the multi-view BU3D-EF dataset, the average accuracy reached 76.1%. To reduce head pose influence, Jung et al. [6] trained a jointly CNNs with facial landmarks and color images, which achieves 72.5%, and it contains three convolutional layers and two hidden layers. The higher accuracies are achieved with SIFT using the deep neural network (DNN) [1,34], which are 78.9% and 80.1% separately. Lopes et al. [4] proposed a combination of Convolutional Neural Network and special image pre-processing steps (C-CNN) to recognize six expressions under head pose at 0° and achieved an averaged accuracy of 90.96% on the BU3D-EF dataset. It is noted that head poses have large influence for FER in an unconstrained environment. How to address pose-aligned facial expression recognition with limited amount of data and unconstrained multi-view environment for improved performance is still an open problem.

3. Conditional CoNERF for facial expression recognition

The detailed steps of our proposed approach are shown in Fig. 2. The deep feature is extracted from saliency-guided facial patches by transferring CNN model to suppress the influence of illumination, occlusion, and low image resolution. CoNERF estimates

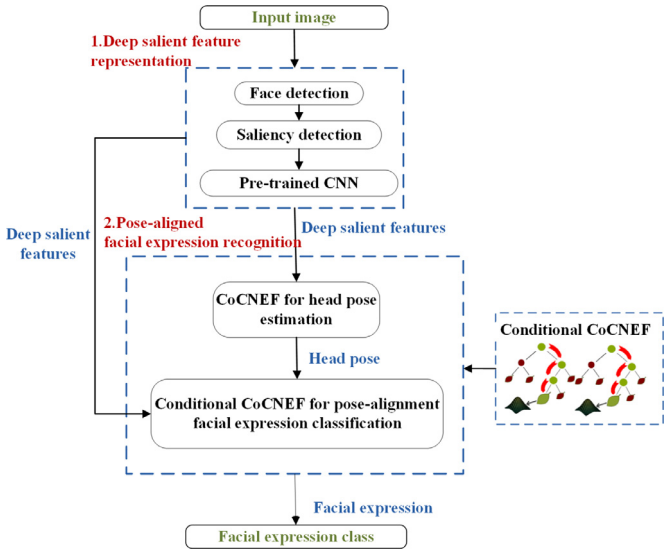

Fig. 2. The flowchart of the proposed approach for pose-aligned facial expression recognition.

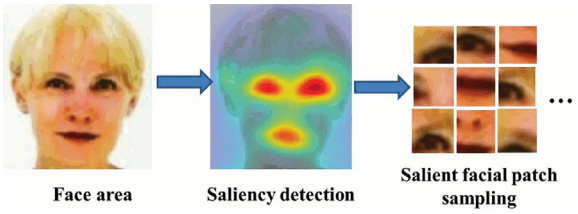

Fig. 3. The salient patch sampling in a face.

head poses with yaw motion, and the facial expressions are recognized under the conditional probability of head pose alignment.

3.1. Deep salient feature representation

We extract salience-driven deep features from facial patches with a pre-trained CNN model, i.e., VGG-face [17]. The patch sampling is shown in Fig 3. Different from randomly or densely sampled patches in the literature [29], we adopt a salience detection algorithm to aid patch selection, which highlights sparse salient regions based on image signature [36]. The image signature is computed by Discrete Cosine Transform (DCT) in the algorithm, which contains information about the foreground of an image, underlying the usefulness of the descriptor for detecting salient image regions. The salience detection obtains a map that shows the probability of salient regions in the image. In this map, the higher values represent the more informative regions as shown in Fig 3. To sample more representative facial patches, we use the method that unifies local and global salience regions by measuring the similarity between each facial patch and the other patches. Let x^i and x^n denote two patches randomly sampled from an image. The size of the patch can be sampled at the rate 50% of the salient region. $d_R(x^i, x^n)$ is the 2-norm of x^i and x^n in the map feature space, normalized to the range [0, 1]:

$$d_R(x^i, x^n) = \left\| R(x^i) - R(x^n) \right\|_2, \tag{1}$$

where $R(x^i)$ and $R(x^n)$ are the centers of patches x^i and x^n in the map feature space, respectively.

The dissimilarity between a pair of patches is computed as follows:

$$D(x^i, x^n) = \frac{d_R(x^i, x^n)}{1 + c \cdot d_p(x^i, x^n)}, \tag{2}$$

where $d_p(x^i, x^n)$ computes the Euclidean distance between the centers of patches x^i and x^n, normalized by the corresponding image dimension, width or height, to the range [0, 1]. c is a constant and we set $c = 3$ referred to [37]. A facial patch x^i is considered a salient patch if its salience s^i is significant,

$$S_{x_i} = 1 - \exp\left\{ -\frac{1}{M} \sum_{m=1}^{M} D(x^i, x^n) \right\}. \tag{3}$$

For every patch x_i, we search for the most similar M patches in the image. In our experiments, the number of the most similar patches is nine, i.e., $M = 9$. Fig. 3 illustrates a salience map and patches with large salience.

After selecting the salient patches, the multi-scale salient patches are fed to the pre-trained VGG-face network and preprocessed to the size of 224 by 224. The VGG-face architecture that is pre-trained with the LFW and YTF face datasets [17] to derive deep high-level feature representation, as shown in Fig. 4. The model includes 13 convolutional layers, 5 max-pooling layers, and 3 fully connected layers. The deep salient feature is described as:

$$y^j = \max\left(0, \sum_i x_c^i w^{i,j} + b^j\right), \tag{4}$$

where y^j is the output high-layer feature representation in the first fully connected layer, x_c^i is the convolution map of the salient facial patch x^i in the last convolutional layer, $w^{i,j}$ indicates the weight between the ith convolution map and the jth output feature, and b^j denotes the bias of the jth feature. The 4096-dimensional activation of the first fully connected layer is used as the final deep salient feature representation.

3.2. Conditional CoNERF training

3.2.1. Decision nodes

For saliency-guided facial patches, we extract a set of deep salient features P, and $P = \{y^j|\theta\}$ under each head pose θ. We propose a NCSF-f_n to reinforce the learning capability of a splitting node by deep learning representation (Fig. 5). Each output of f_n is brought in correspondence with a splitting node $d_n(P, Y|\theta)$,

$$d_n(P, Y|\theta) = \sigma(f_n(P, Y|\theta)), \tag{5}$$

where $\sigma(x) = (1 + e^{-x})^{-1}$ is the sigmoid function, Y is the parametrization of the network and n is a decision node.

We employ a Stochastic Gradient Descent (SGD) approach to minimize the risk with respect to Y:

$$Y^{(t+1)} = Y^{(t)} - \frac{\eta}{|B|} \sum_{(P,\pi)\in B} \frac{\partial L(Y, \pi; P)}{\partial Y}, \tag{6}$$

where $\eta > 0$ is the learning rate, π is facial expression label and B is a random subset (a.k.a. mini-batch) of samples. $L(Y, \pi; P)$ is the log-loss term for the training sample P, which is defined as

$$L(Y, \pi; P) = -\sum_n p(\pi|d_n, Y, P) \log(p(\pi|d_n, Y, P)), \tag{7}$$

where $p(\pi|d_n, Y, P)$ is the facial expression probability. The gradient with respect to Y is obtained by chain rule as follows:

$$\frac{\partial L(Y, \pi; P)}{\partial Y} = \sum_{n\in N} \frac{\partial L(Y, \pi; P)}{\partial f_n(P, Y|\theta)} \cdot \frac{\partial f_n(P, Y|\theta)}{\partial Y}. \tag{8}$$

Here, we have the gradient term that depends on the decision tree with the splitting child nodes as follows:

$$\frac{\partial L(Y, \pi; P)}{\partial f_n(P, Y|\theta)} = -\sum\nolimits_n (d_n^R(P, Y|\theta) + d_n^L(P, Y|\theta)), \tag{9}$$

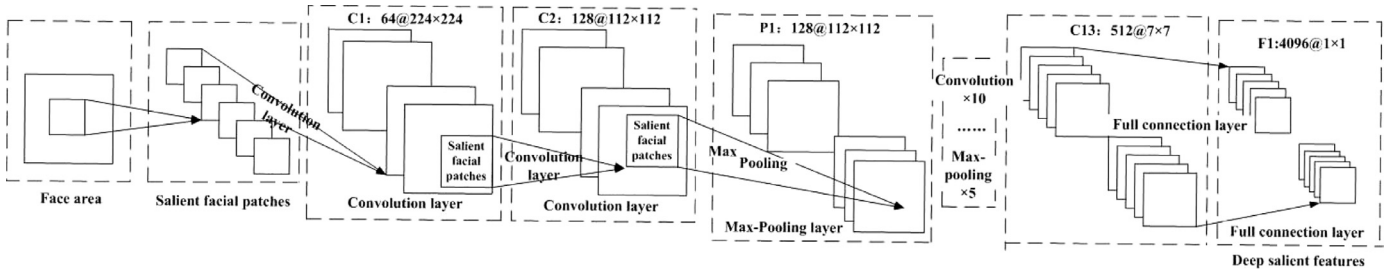

Fig. 4. The structure of pre-trained CNN network for deep salient feature representation. The trained network model includes 13 convolutional layers, 5 max-pooling layers, and 3 fully connected layers. In our work, we extract deep salient features from salient facial patches on the first fully connected layer.

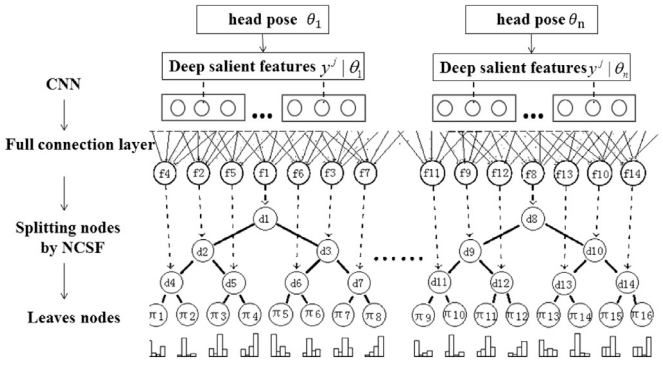

Fig. 5. The training implement of a conditional CoNERF integrated CNN and RF models. The implement includes learning splitting nodes by NCSF and generates decision life nodes.

where $d_n^R(P, Y|\theta)$ and $d_n^L(P, Y|\theta)$ denote the conditional probabilities under different pose in its left and right child nodes of the decision node, respectively. For splitting optimization, Information Gain (IG) is used to split a node into its left and right child nodes in the tree construction:

$$\tilde{\varphi} = \arg\max_{\varphi}(H(d_n) - \sum_{S\in\{N_r,N_l\}} \frac{|d_n^S|}{|d_n|}(H(d_n)), \tag{10}$$

where $\frac{|d_n^S|}{|d_n|}, S \in \{R, L\}$ is the probability between the number of feature samples in d_n^L (arriving at the left child node), set d_n^R (arriving at the right child node), and $H(d_n)$ is the entropy of d_n.

3.2.2. Leaf nodes

When *IG* is below a threshold or when a maximum depth is reached, a leaf l is created. For a leaf node in a conditional CoNERF tree, it stores the conditional probability $p(\pi|\theta, l)$. Therefore, we simplify the distribution over the facial expression class and head poses by a multivariate Gaussian Mixture Model (GMM) [38]:

$$p(\pi|\theta, l) = N(\pi|\theta; \overline{\pi|\theta}, \Sigma_l^{\pi|\theta}), \tag{11}$$

where $\overline{\pi|\theta}$ and $\Sigma_l^{\pi|\theta}$ are the mean and covariance of the facial expression probabilities given the head pose probability, respectively.

3.3. Conditional CoNERF

We train a conditional CoNERF to estimate head poses in nine yaw categories: { −90°, −60°, −45°, −30°, 0°, +30°, +45°, +60°, 90° }. In the leaves of a conditional CoNERF forest, there are nine probabilistic models of head poses. We simplify the distributions over multi-probabilities by adopting multivariate GMM as:

$$p(\theta|l) = N(\theta; \overline{\theta}, \Sigma_l^{\theta}), \tag{12}$$

where $\overline{\theta}$ and Σ_l^{θ} are the mean and covariance of head pose probabilities, respectively. While Eq. (12) models the probability for a sample P ending in a leaf l, the probability of the forest is obtained by averaging over all trees:

$$p(\theta|P) = \frac{1}{T}\sum_t p(\theta|l_t(P)), \tag{13}$$

where l_t is the corresponding leaf for a tree T_t, T is the number of trees in CoNERF. The estimated head pose is used in the conditional CoNERF for expression recognition.

The conditional CoNERF models the probability of pose-aligned facial expression, denoted with $p(\pi|P)$, which is an integration of conditional probability for all head pose, denoted with θ:

$$p(\pi|P) = \int p(\pi|\theta, P)p(\theta|P)d\theta. \tag{14}$$

To compute $p(\pi|\theta, P)$, we divide the space of head pose into disjoint subspaces, denoted with θ and probability of pose-aligned facial expression becomes

$$p(\pi|P) = \sum\nolimits_i (p(\pi|\Omega_i, P)\int p(\theta|P)d\theta). \tag{15}$$

We select T trees from the conditional CoNERF forest $F(\Omega_i)$ based on the estimated probability $p(\theta|P)$. To this end, the final facial expression probability is computed by weighted average:

$$p(\pi|P) = \frac{1}{T}\sum_i\sum_{t=1}^{k_i} p(\pi|l_{t,\Omega_i}(P)), \tag{16}$$

where l_{t,Ω_i} is the corresponding leaf for feature representation of the tree. The discrete number of trees k_i are computed:

$$k_i = T\cdot\int_{\theta\in\Omega_i} p(\theta|P)d\theta, \tag{17}$$

where $\sum_i k_i = T$.

4. Experimental results

4.1. Datasets and settings

To evaluate our approach (available at http://covis.cse.unt.edu/Demo/CoNERF/), four face expression datasets were used: frontal Cohn-Kanade (CK+) dataset [39], frontal JAFFE [40] facial expression dataset, multi-view BU-3DFE [41] dataset, and LFW [42] facial dataset. The CK+ database is a widely used benchmark for evaluating expression recognition techniques, which contains 593 image sequences across 128 subjects, which contains 6 facial expression images from neutral to peak expression. The JAFFE database contains 213 images of 7 facial expressions (6 basic facial expressions and 1 neutral) posed by 10 Japanese female models. The multi-view BU-3DFE database contains 100 people of different ethnicities, including 56 females and 44 males. Six facial expressions (anger, disgust, fear, happiness, sadness, and surprise) are elicited by various manners and head poses, and each of them includes 4 levels of intensities which yield 2400 facial expression models.

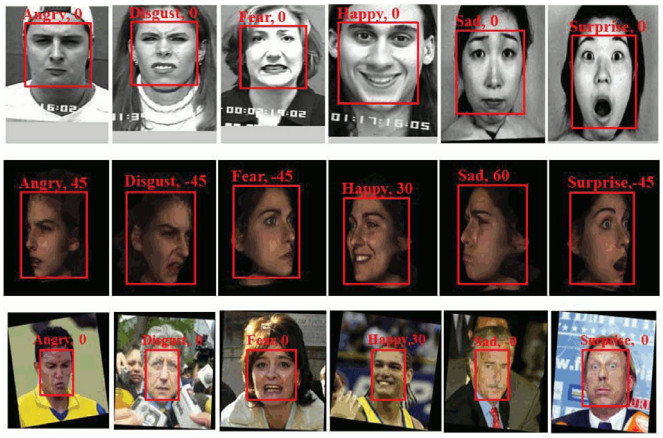

Fig. 6. Examples of recognition results images from CK+, JAFFE, BU-3DFE and LFW datasets. Top row: results using images from CK+ (the first four images) and JAFFE datasets (the last two images on the top row). Middle row: results using images from the BU-3DFE dataset. Bottom row: results using images from LFW dataset.

Table 1
Accuracy (%) of SVM classifier using different image features.

Features	CK+	JAFFE	BU-3DFE	LFW
Deep salient feature	**97.57**	**95.46**	**76.42**	**44.21**
CONV.13 from VGG-face	93.28	91.57	69.85	31.65
FC.1 from VGG-face	96.72	93.25	76.14	35.46
FC.2 from VGG-face	96.55	93.2	70.47	34.73
FC.3 from VGG-face	95.17	92.59	65.5	34.26
SIFT	78.66	65.38	51.02	30.83
HOG	75.83	70.39	62.53	30.57
SIFT+HOG	71.35	61.92	48.91	27.33
Distances between 21 points	86.25	83.68	60.37	40.84

These models are described by both 3D geometrical shapes and color textures with 83 Feature Points identified on each model. The LFW dataset consists of 5749 individual face images. The images were collected in the wild and varied in expressions, poses, lighting conditions, resolutions, occlusions, make-ups, etc. In our experiments, we labeled six facial expressions of 5000 images following the scheme in [39].

Examples of facial expression recognition of CK+, BU-3DFE, JAFFE, and LFW datasets are shown in Fig. 6. A conditional CoNERF model for each head pose was trained with 1086 images from CK+ dataset, 134 images from JAFFE dataset, 1722 images from the BU-3DFE dataset and 2000 images from LFW dataset. In our evaluation, we used 368 images from CK+ dataset, 49 images from JAFFE dataset, 574 images from the BU-3DFE dataset, and 500 images from LFW dataset. We used the Caffe framework [1] [16] for implementing CNN and deep salient feature representation. The important training parameters in the experiments include learning rate (0.01), epochs (5000), splitting interactive times (1000) and trees' depth(15).

4.2. Analysis of image features

To understand the influence of features on the accuracy, we conducted experiments with deep features and two popular classical features (SIFT and HOG) as well as the combination of them. Table 1 lists the average accuracy with eight single features and the combination features on four expression datasets. The single features include our proposed deep salient feature, the convolution feature from the 13th layer of VGG-face (CONV.13), features from the first, second, and third fully connected layers of VGG-face, denoted with FC.1, FC.2, and FC.3, respectively, SIFT description, HOG feature, and distances between 21 facial points. The experimental results demonstrate that the proposed deep salient feature improve the average accuracy in all cases.

[1] http://caffe.berkeleyvision.org/.

Table 2
Accuracy (%) of the conditional CoNERF using different image features.

Features	CK+	JAFFE	BU-3DFE	LFW
Deep salient feature	**99.02**	**98.13**	**94.09**	**60.09**
CONV.13 from VGG-face	95.27	93.34	75.36	37.19
FC.1 from VGG-face	98.52	96.14	86.85	45.4
FC.2 from VGG-face	98.39	94.86	83.34	42.27
FC.3 from VGG-face	96.21	94.73	79.28	40.33
SIFT	83.46	70.36	78.9	35.62
HOG	89.78	89.7	74.0	36.53
SIFT + HOG	93.48	91.75	80.93	42.11
Distances between 21 points	95.21	95.30	80.52	55.53

Table 3
Confusion matrix of facial expression recognition of CK+ and JAFFE datasets.

	Anger	Disgust	Fear	Happy	Sadness	Surprise
Anger	**1**	0	0	0	0	0
Disgust	0.028	**0.958**	0	0	0	0.014
Fear	0	0	**1**	0	0	0
Happy	0	0	0	**1**	0	0
Sadness	0	0	0	0	**1**	0
Surprise	0	0	0	0.009	0	**0.991**

Table 4
Accuracy (%) and standard deviations (STD) using different methods on CK+ and JAFFE frontal facial datasets.

Methods	# of Expression	Accuracy	STD.
T-DCN [13]	6+neutral	80.49	0.9
HF [29]	6	87.1	0.7
PCRF [43]	6	96.4	1.1
AU-DNN [20]	6	92.05	0.7
M-SVM [44]	6	93.6	0.8
JFDNN [6]	6	97.3	1.2
CNN [18]	6	97.8	1.3
C-CNN [4]	6	91.64	2.5
Conditional CoNERF	6	99.02	0.5

Table 2 presents the average accuracy of the conditional CoNERF using different image features and it is clear that the proposed deep salient feature obtained the best performance. With CK+, JAFFE, and BU-3DFE datasets, the accuracies are in the range of mid to high end of 90%. On the multi-view BU-3DFE dataset, the average accuracy significantly improved using the deep salient feature. On the challenging LFW dataset, the recognition rate reaches 60.09% using the deep salient features, which improves 8.2% with respect to the second best result. The conditional CoNERF exhibits much improved performance in comparison to the SVM classifier (see Table 1 for results of SVM), especially in the multi-view BU-3DFE and LFW datasets. It is evident that the deep salient feature provides a better description of the multi-view facial expression images.

4.3. Experiments with CK+ and JAFFE frontal facial datasets

Table 3 shows the confusion matrix of the expression recognition with CK+ and JAFFE frontal facial datasets. The accuracies are all above 95% with the average accuracy of 99.02% for the frontal faces.

In comparison with the state-of-the-art facial expression recognition methods, Table 11 lists the average accuracy and standard deviation (STD) on CK+ and JAFFE frontal facial datasets using Transfer learning from deep convolutional networks (T-

Table 5
Confusion matrix of head pose estimation on BU-3DFE dataset.

	−90°	−60°	−45°	−30°	0°	30°	45°	60°	90°
−90°	**1**	0	0	0	0	0	0	0	0
−60°	0.009	**0.988**	0.002	0	0	0	0	0	0.001
−45°	0	0.016	**0.981**	0.001	0	0.002	0	0	0
−30°	0.002	0	0.005	**0.993**	0	0	0	0	0
0°	1	0	0	0	**0.997**	0.002	0	0	0.001
30°	1	0	0	0	0	**0.991**	0.007	0.002	0
45°	1	0	0	0	0	0.01	**0.977**	0.013	0
60°	1	0	0	0	0	0	0.014	**0.984**	0.002
90°	1	0	0	0	0	0	0	0.002	**0.998**

DCN) [13], Hough Forests (HF) [6], Pairwise conditional random forests (PCRF) [43], Au-aware deep networks (AU-DNN) [20], Multi-class Support vector machine (M-SVM) [44], Joint fine-tuning in deep neural networks (JFDNN) [6], CNN [18], Combined-CNN (C-CNN) [4] and our CoNERF. Xu et al. [13] proposed a facial expression recognition model based on transfer features from deep convolutional networks for six expressions and one neutral expression. The average accuracy achieved 80.49% and STD is 0.9 on the frontal facial datasets. Fanelli et al. [29] presented a Hough forests for facial expression recognition from image sequences, which achieved an average recognition rate of 87.1% on six expressions recognition. Dapogny et al. [43] proposed to learn Random Forest from heterogeneous derivative features upon pairs of images and achieved the average accuracy of 96.4% in the front facial expression datasets. Au-aware deep networks [20] constructed a deep architecture for six expression recognition by elaborately utilizing the prior knowledge that the appearance variations caused by expression can be decomposed into a batch of local facial Action Units. The final recognition results can be obtained 92.05% of average accuracy and 0.7% of STD. Jung et al. [6] proposed a jointly deep network based on two different models , such as temporal appearance features and temporal geometry features, which obtained an average accuracy of 97.3% and STD of 1.2%. Mollahosseini et al. [18] uses a deep CNN architecture to address six expression recognition and achieved the average accuracy of 97.8%. Lopes et al. [4] used a combination of CNNs and special image pre-processing steps to achieve the average accuracy of 98.8% on CK+ dataset and 84.48% on JAFFE datasets, whose average accuracy can reach 91.64% with STD of 2.5% on the two frontal datasets. Our CoNERF method outperforms other methods with an average accuracy of 99.02% in the front facial expression recognition on the CK+ and JAFFE datasets. The lowest STD is 0.5% using our proposed method (Table 4).

4.4. Experiments with multi-view BU-3DFE dataset

When dealing with multi-view images, the head pose is estimated for correction of pose-induced inconsistency in expression recognition. A 4-fold cross-validation was conducted.

4.4.1. Head pose estimation

For head pose estimation, we use the same settings as the facial expression recognition. Each image in the BU-3DFE dataset is automatically annotated with one out of the nine head pose labels in the yaw rotation ({ −90°, −60°, −45°, -30°, 0°, +30°, +60°, +75°, 90° }). We train a CoNERF of 50 neural trees using 15,498 head pose images. Table 5 shows the confusion matrix of head pose estimation on BU-3DFE dataset. The CoNERF estimated 9 head pose classes in the horizontal direction and achieved the average accuracy of 98.99%. Examples of the estimated head pose are shown in Fig. 6. Our method aligned head poses for expression recognition.

Table 6 lists the comparison results of our CoNERF method, CNN [45], GSRRR [34], and SIFT-CNN [1]. The CNN [45] uses AlexNet architecture, which contains three convolutional layers and two fully connected layers. The filter size is 5 by 5. The input images are rescaled to 224 by 224. The average accuracy of CNN is 69.61% with STD of 0.9. The GSRRR algorithm [34] based on sparse SIFT features obtains an accuracy of 87.36%. The average accuracy using the improved SIFT-CNN proposed in [1] is 92.26%. Our CoNERF method achieves the average accuracy of 98.99%, which is competitive to the methods above. The STD of 0.5 of our method demonstrates the robustness of the proposed CoNERF for head pose estimation.

Table 6
Accuracy (%) and STD of head pose estimation using different methods on BU-3DFE dataset.

Methods	Features	Poses	Accuracy	STD.
CNN [45]	Image	9	69.61	0.9
GSRRR [34]	Sparse SIFT	9	87.36	0.8
SIFT-CNN [1]	SIFT	9	92.26	0.7
CoNERF	Deep transfer feature	9	98.99	0.5

4.4.2. Pose-aligned facial expression recognition

Table 7 lists the confusion matrices under different head poses from the BU-3DFE dataset. The average accuracy of expression recognition is 94.09% under overall head poses. The highest accuracy is 96.424% of sadness followed by that of surprise, happiness, and anger, which are above 95%. The lowest accuracy is 87.95% for disgust.

The accuracy reached 95.98% under head pose at 0°. The lowest accuracy is 87.8% under head pose at 90°, which is partly caused by a great degree of self-occlusion. Nevertheless, it is demonstrated that our proposed method achieved robust results under large head pose motion.

The average accuracy of our CoNERF method is compared with that of CNN, Local binary patterns based SVM(LBPs-SVM) [9], PCRF [35], JFDNN [6], Coupled gaussian process regression (CGPR) [8], Group sparse reduced-rank regression (GSRRR) [34], Deep neural network-driven SIFT feature (SIFT-CNN) [1] and C-CNN [4] in Table 8. The CNN in our experiment contains three convolutional layers followed by three max-pooling layers and two fully connected layers. Each filter is of size 5×5 and there are 32, 64, and 128 such filters in the first three layers, respectively. The numbers of the hidden nodes in two fully connected layers are 1024 and 512. The input images are rescaled to 224 by 224.

The accuracy of the CNN on BU-3DFE dataset is 68.9% as presented in Table 8. The accuracy of multi-class SVM with LBP and LGBP in [9] is 71.1%. Dapogny et al. [35] proposed pair conditional random forests to capture low-level expression transition patterns on the condition of head pose estimation for multi-view dynamic facial expression recognition. On the pose variances BU-3DFE dataset, the average accuracy reaches 76.1%. JFDNN achieves 72.5% which contains three convolution layers and two hidden layers, where the filters in the three convolution layers are in size 5×5, the numbers of the hidden nodes are set to be 100 and 600.

Table 7
Confusion matrix of facial expression recognition using the BU-3DFE dataset under various head poses.

	Head pose 0%						Head pose 30%					
	Anger	Disgust	Fear	Happy	Sadness	Surprise	Anger	Disgust	Fear	Happy	Sadness	Surprise
Anger	**0.977**	0	0.02	0	0.003	0	**0.977**	0	0.02	0	0.003	0
Disgust	0.08	**0.904**	0	0.006	0.01	0	0.08	**0.904**	0	0.006	0.01	0
Fear	0	0.06	**0.901**	0.039	0	0	0	0.06	**0.901**	0.039	0	0
Happy	0	0.002	0.011	**0.986**	0	0.001	0	0.002	0.011	**0.986**	0	0.001
Sadness	0.02	0.03	0	0	**0.95**	0	0.02	0.03	0	0	**0.95**	0
Surprise	0	0	0.01	0.000	0	**0.99**	0	0	0.01	0.000	0	**0.99**
	Head pose 45%						Head pose 60%					
Anger	**0.959**	0.03	0	0	0.011	0	**0.889**	0.101	0.01	0	0	0
Disgust	0	**0.94**	0.01	0.01	0	0.02	0.03	**0.95**	0	0.02	0	0
Fear	0	0.0	**0.863**	0.063	0	0	0.011	0.01	**0.874**	0.053	0	0.052
Happy	0	0	0.011	**0.989**	0	0	0	0.011	0.011	**0.978**	0	0
Sadness	0.034	0.023	0.012	0	**0.931**	0	0.046	0	0	0	**0.954**	0
Surprise	0	0.01	0	0.02	0.01	**0.96**	0.01	0.01	0.01	0.01	0.01	**0.95**
	Head pose 90%						Overall					
Anger	**0.889**	0.03	0	0.03	0.03	0.021	**0.955**	0.015	0.005	0.005	0.019	0.011
Disgust	0.03	**0.85**	0.08	0.02	0.01	0.01	0.037	**0.933**	0.009	0.007	0.001	0.013
Fear	0.063	0.042	**0.789**	0.084	0.01	0.012	0.012	0.036	**0.88**	0.037	0.009	0.026
Happy	0.011	0	0.043	**0.935**	0	0.011	0	0.004	0.014	**0.963**	0.002	0.017
Sadness	0.011	0.022	0.012	0.012	**0.943**	0	0.02	0.01	0.005	0	**0.965**	0
Surprise	0	0.02	0.08	0.03	0	**0.87**	0.006	0.011	0.013	0.014	0.003	**0.952**

Table 8
Accuracy (%) and STD using different methods on multi-view BU-3DFE dataset.

Methods	Features	Poses	Accuracy
CNN	Image	9	68.9 (1.5)
LBPs-SVM [9]	LBP and LGBP	7	71.1 (1.2)
PCRF [35]	Heterogeneity	5	76.1 (1.0)
JFDNN [6]	Image and landmarks	5	72.5 (1.3)
CGPR [8]	facial landmarks	5	76.5 (0.8)
GSRRR [34]	Sparse SIFT	9	78.9 (1.0)
SIFT-CNN [1]	SIFT	9	80.1 (0.8)
C-CNN [4]	Intensity	1(0°)	90.96 (1.0)
Conditional CoNERF	Deep salient feature	9	94.09 (0.6)

Table 9
Average accuracy (%) of pose-aligned FER under each head pose on the BU-3DFE dataset.

Methods	0°	30°	45°	60°	90°
Conditional CoNERF	95.08	93.55	95.3	93.21	87.8
GSRRR [34]	78.9	80.1	80.1	78.4	77.0
SIFT-CNN [1]	79.7	80.7	81.0	80.5	79.51

The higher accuracies are achieved with SIFT using GSRRR [34] and SIFT-CNN [1], which are 78.9% and 80.1%, respectively. Our method achieved 94.09%, which is highly competitive to the aforementioned methods. The standard deviation of 0.6% using our method demonstrates the robustness of our proposed method. In addition, Lopes et al. [4] used intensity features to recognize six expressions with head pose at 0° and achieved an averaged accuracy of 90.96%. In contrast, our method resulted in an average accuracy at 95.98% for the same case.

Table 9 lists the average accuracy under five head poses of our method, GSRRR [34] and SIFT-CNN [1]. The accuracy of our method is significantly greater than that of the other two methods in terms of different views. The highest accuracies achieved by methods [1,34] are 81.0% and 80.1% at 45°, respectively. The highest accuracy of our method is achieved under the head pose 0°, which is are 95.08%. And the lowest accuracies appear both under the head pose 90° because of occlusion and facial deformation.

Table 10
Confusion matrix of head pose estimation on LFE dataset.

	−90°	−45°	0°	45°	90°
−90°	**0.895**	0.099	0.006	0	0
−45°	0.01	**0.864**	0.126	0	0
0°	0.001	0.036	**0.903**	0.058	0.003
45°	0	0	0.045	**0.863**	0.092
90°	0	0	0.018	0.107	**0.875**

Table 11
Accuracies (%) and STDs of head pose estimation using different methods on LFW dataset.

Methods	Features	Poses	Accuracy	STD
CNN [45]	Image	5	65.25	1.2
GSRRR [34]	Sparse SIFT	9	77.2	0.9
SIFT-CNN [1]	SIFT	5	83.4	0.9
CoNERF	Deep saliency-guided feature	5	88	0.7

4.5. Experimental results of in-the-wild LFW facial dataset

4.5.1. Pose estimation

The images from LFW dataset vary in expressions, poses, lighting conditions, resolutions, occlusions, make-ups, etc. We categorized the images using 5 yaw angles (−90°, −45°, 0°, +45°, 90°) to represent left, partial left, front, partial right, and right. Table 10 shows the confusion matrix of head pose estimation. The CoNERF estimated five head pose classes in the horizontal direction and achieved an average accuracy of 88%. Examples of the estimation results are shown in Fig. 6.

Table 11 lists the average accuracy of CNN [45], GSRRR [34], SIFT-CNN [1] and our method using the LFW dataset. The parameters used in this experiment is the same as the one reported in Section 4.4.1. It is clear that our proposed method yielded the highest average accuracy with the smallest STD. This is consistent with the results produced with BU-3DFE dataset.

4.5.2. Expression recognition

Table 12 shows the confusion matrix of expression recognition. The average accuracy achieved 60.9%. The highest accuracy is 85.24% of happiness followed by surprise and sad, which are more

Table 12
Confusion matrix of posed-aligned FER on LFW dataset.

	Anger	Disgust	Fear	Happy	Sadness	Surprise
Anger	**0.559**	0.202	0.166	0	0.009	0.006
Disgust	0.137	**0.505**	0.087	0.02	0.111	0.139
Fear	0.055	0.106	**0.595**	0.016	0.184	0.004
Happy	0.002	0.16	0.01	**0.852**	0.004	0.012
Sadness	0.146	0.142	0.115	0	**0.572**	0.025
Surprise	0.019	0.121	0.156	0.013	0.036	**0.655**

Table 13
Average accuracy (%) of facial expression recognition under different head poses on the LFW dataset.

Methods	0°	45°	90°
Conditional CoNERF	64.85	59.4	56.51
PCRF [43]	52.1	42.9	40.2
SVM based on Gabor features	45.4	36.57	42.6
RF based on Gabor features	46.8	46.23	44.7

Table 14
Average accuracy of different methods trained with small data sets.

Convolutional models	Number of images.		Accuracy(%)	STD.
	# Training	# Testing		
ResNet-50 [46]	1086	368	82.6	0.8
AlexNet [45]	1086	368	78.96	0.9
T-DCN [13]	1086	368	80.49	0.7
JFDNN [6]	1086	368	95.3	1.0
Conditional CoNERF	1086	368	99.02	0.5

than 65.48%. The lowest accuracy is 50.53% of disgust. Relatively low accuracies appear between these expressions because the LFW dataset introduced more noises and spontaneous expressions than other datasets.

Table 13 shows the recognition accuracy using our method, PCRF [43], SVM and RF methods. The features used for SVM and RF include Gabor filter responses with eight rotation angles and five phase shifts. It is shown that the accuracy of our method is greater than the other methods. The highest accuracy of these methods is achieved under the head pose 0°, which are 64.85% using our method, 52.1% using PCRF, 45.4% using SVM, and 46.8% using RF method. And the lowest accuracies appear both under the head pose 90°. On this challenging dataset, our method outperformed the other methods under each head pose.

4.6. Training with small datasets

To evaluate our method in small training datasets, we compared the accuracy of five methods trained with the same small dataset, including conditional CoNERF with deep salient feature, AlexNet [45], ResNet50 [46], T-DCN [13], and JFDNN [6]. The weights of AlexNet, ResNet50, T-DCN, and JFDNN are refined from the existing models instead of the randomly initialized. The experiments were conducted with CK+ facial expression, which includes 1086 images for training and images for testing. A five-fold cross-validation was conducted. Table 14 lists the average accuracy of these methods. In our experiments, AlexNet [45] architecture includes five convolutional layers and three fully connected layers. The average accuracy achieved by AlexNet was 78.96%. Shallow residual network model-ResNet50 [46] yielded an average accuracy of 82.6%. JFDNN [6] resulted in an average accuracy of 95.3%, which integrates two joint convolutional neural network models. Our proposed conditional CoNERF achieved an average accuracy of 99.02% and a Standard Deviation at 0.5. It is demonstrated that the conditional CoNERF exhibits the best performance with a small amount of training data.

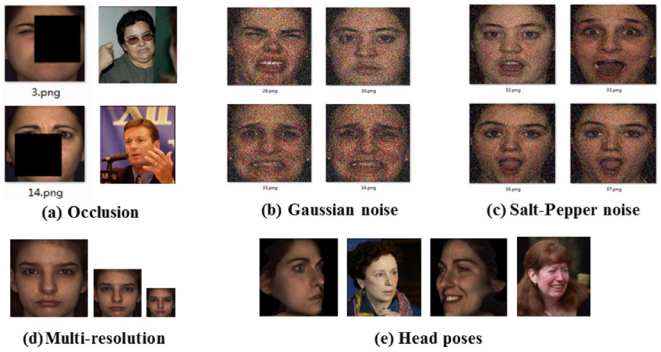

Fig. 7. Examples of distortions images in our experiments.

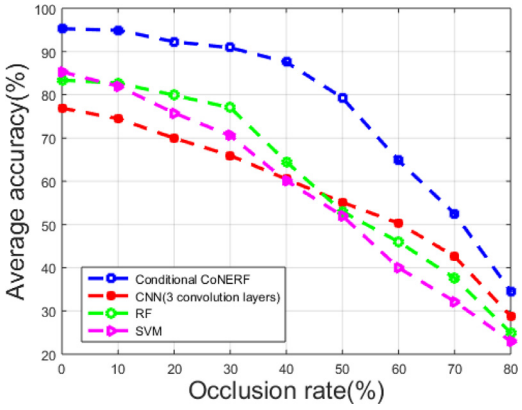

Fig. 8. Average accuracy with different degrees of occlusion.

4.7. Evaluation of image distortion

This section presents our evaluations of the impact of image distortions, which include occlusion, additive noise, scale variation, and pose variation. Fig. 7 depicts a few examples of the distorted images. We compared the proposed method with the CNN, SVM, and RF. The CNN in our experiments contains three convolutional layers followed by three max-pooling layers and two fully connected layers. The feature used in RF and SVM is HOG description.

4.7.1. Analysis of occlusion

In addition to the natural occlusions in the images, we also generated synthetic partial occlusion (see Fig. 7(a) for an example). The occlusion rate of our synthetic images is in the range of 20% to 80% and the occlusion is randomly placed on the face. Fig. 8 illustrates the average accuracy of the conditional CoNERF, CNN, RF, and SVM at different occlusion rates. The horizontal axis represents the occlusion rates up to 80% of the face. The vertical axis represents the recognition accuracy. With 60% of occlusion, our method achieved an average accuracy above 63%. The accuracy of our method is consistently higher than that of the others for all occlusion rate. The accuracy of our proposed method degrades gracefully when the occlusion rate is less than 50% and this trend accelerates beyond that point. As the portion of occluded face increases, it becomes harder for recognition. It is foreseeable that with total occlusion, the recognition accuracy reaches zero. It is interesting that when occlusion rate is moderate (at or below 40%), CNN performs worse than RF and SVM; whereas our proposed method improves the performance by more than 10%. Our experimental results demonstrate that the proposed method exhibits greater performance in the case of partial occlusion.

Table 15
Average accuracy (%) of noisy image at four different noise levels.

Methods	Gaussian noise				Salt & Pepper noise			
	0.05	0.1	0.15	0,2	0.05	0.1	0.15	0,2
CNN(3conv+2fc)	68.68	63.78	61.53	57.84	69.18	66.82	63.20	58.65
SVM	72.93	70.4	68.17	58.32	77.35	71.2	64.23	59.5
RF	74.3	73.4	69.5	59.61	79.21	75.9	70.74	66.32
Conditional CoNERF	**84.52**	**80.8**	**78.55**	**74.29**	**91.52**	**87.52**	**85.19**	**81.20**

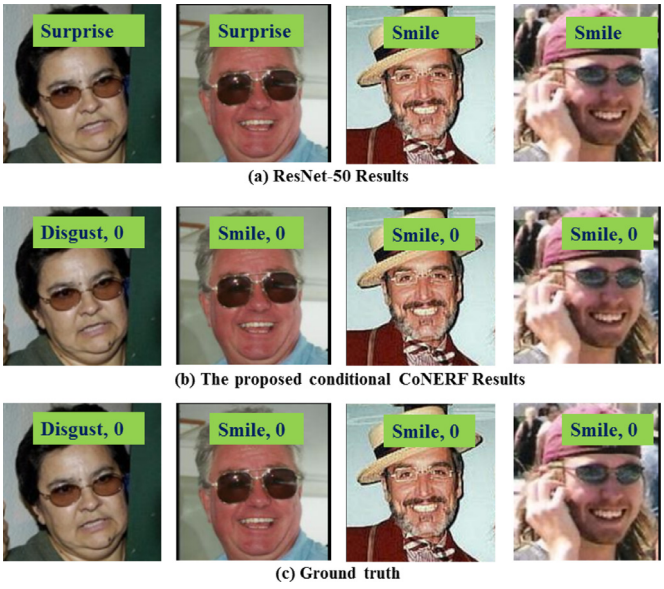

Fig. 9. Results of ResNet-50 (first row) and conditional CoNERF (second row). The bottom row shows the ground truth.

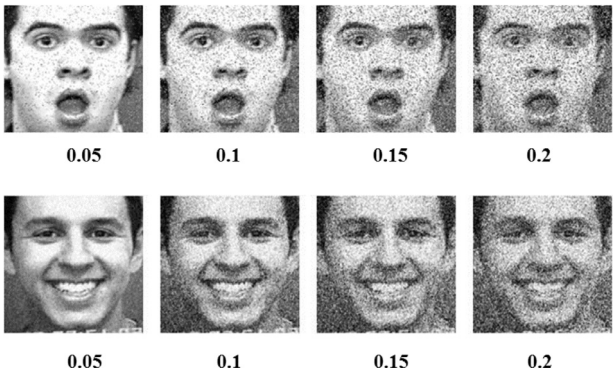

Fig. 10. Images with Gaussian (first row) and salt-pepper noise (scend row) amount of 0.05, 0.1, 0.15, 0.2.

In real-world cases, people wear attachments including sunglasses, masks, and hats. LFW dataset (see Table 11) collects face images with occlusion. Fig. 9 shows the FER results of ResNet-50 [46] and our conditional CoNERF on some challenging cases from LFW dataset. It is evidential that our method improves the robustness of FER to partial occlusions under varied situations.

4.7.2. Analysis of noise

To evaluate the influence of noise, we created distorted images by adding Gaussian and Salt & Pepper noises. We randomly selected 1000 images from the CK+ dataset and created distorted images. The noise magnitudes are 0.05, 0.1, 0.15, and 0.2. Fig. 10 illustrates examples of the distorted images. A four-fold cross-validation was conducted. Table 15 presents the average accuracy.

Table 16
Average accuracy with images at different resolutions.

Methods	100%	50%	25%	STD.
SVM	85.31	73.52	58.19	5.18
RF	83.4	77,65	65.43	4.73
CNN	68.9	69.44	67.83	0.74
Conditional CoNERF	**95.29**	**94.89**	**94.96**	0.24

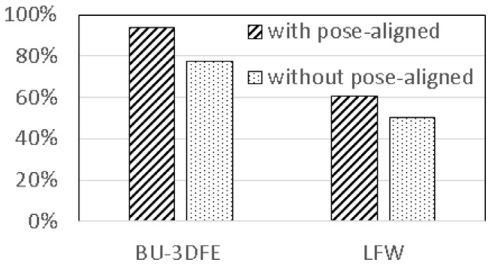

Fig. 11. Average accuracy with and without pose-aligned conditional probability.

The average accuracy declines with the increment of noise levels for all methods and our proposed method achieved the best performance under various additive noisy conditions. In contrast to the second best results, conditional CoNERF improves the accuracy by a minimum of 10.08% for Gaussian noise and 15.31% for Salt & pepper noise.

4.7.3. Scale analysis

To evaluate the impact of image resolution, we resized the images to a half and a quarter of their original size. Table 16 presents the average accuracy of our proposed method and the other three methods as well as the standard deviations across all scales. The accuracy of our method is above 94% for all cases, and the results are quite consistent compared to the other methods as demonstrated with the STD. The results of RF and SVM vary greatly with the change of image scale.

4.7.4. Head pose alignment analysis

Fig. 11 shows the results with and without pose-aligned conditional probability. As depicted in this figure, our method that employs pose-aligned conditional probability outperformed the other without pose-aligned conditional probability on both LFW and BU-3DFE datasets. The improvement is about 21%. It is clear that pose alignment is an important factor of expression recognition.

4.8. Time complexity

Table 17 reports the average time of facial expression recognition using the BU-3DFE dataset. The experiments were conducted on a PC with Intel Core i7-6700 CPU at 4 GHz, 32 GB memory, and NVIDA GeForce GTX 1080. The programs were implemented with MATLAB. Both random forest and SVM were executed on CPU and Conditional CoNERF and CNN were executed with support of GPUs. Comparing Conditional CoNERF and CNN, the training time

Table 17
Computation time of training and testing of different methods. The training times are in second and the testing times are in micro-second.

Methods	Using CPU			Using GPU	
	Conditional CoNERF	RF	SVM	Conditional CoNERF	CNN
Training Time	–	6540 s	808 s	1620 s	16920 s
Testing Time	135 ms	128 ms	378 ms	113 ms	160 ms

of Conditional CoNERF is less than one tenth of the time used by CNN and the testing time of Conditional CoNERF is also much less than that of CNN, improved by 30%. The training of SVM is most efficient among all methods, despite that SVM was trained with support of CPU only. Yet, its testing time is much inferior to the others, almost triple the time of Conditional CoNERF with support of CPU only. In all cases, Conditional CoNERF with support of GPU exhibits the greatest efficiency and its training efficiency is much superior to Random Forest (using CPU only) and CNN (using GPU).

5. Conclusion

This paper describes a novel conditional CoNERF method for pose-aligned facial expression recognition in multi-view and the unconstrained environment. Robust deep salient features are extracted from saliency-guided facial patches using transfer CNN model and the conditional CoNERF unifies random trees with the representation learning from deep convolutional neural networks. A neurally connected split function is introduced to CoNERF to split node learning. Our method performs well due to transferring a pre-trained CNN to fast decision node splitting in a Random Forest. The experimental results demonstrated that our method was of great robustness and efficiency in various poses, occlusions, and noise conditions.

Experiments were conducted using public CK+, JAFFE, BU-3DEF and LFW datasets. Our results demonstrated that the proposed deep salient feature outperformed the other popular image features. Compared to the state-of-the-art methods, the conditional CoNERF achieved improved performance and great robustness with an average accuracy of 94.09% on the multi-view BU-3DEF dataset, 99.02% on CK+ and JAFFE frontal facial datasets, and 60.9% on in-the-wild LFW dataset. The average time for performing a pose-aligned FER is about 113 ms. The average accuracies of head pose estimation on BU-3DEF and LFW datasets are about 98.9% and 89.7%. Our method achieved the great performance for head pose-varied facial expression recognition with a limit number of training examples, which is a significant advantage in contrast to deep neural networks.

In future, we plan to explore real-time CoNERF model for spontaneous facial expression recognition in videos. As demonstrated in our efficiency analysis, the prediction time (i.e., testing time) is in the range of one tenth of a second. To conduct video analysis in real-time, it is necessary to reduce the prediction time to less than 50 ms per frame. A possible solution is to employ an attention model to search for the peak frame in a video and to leverage the spatial-temporal expression features in the frame sequence to improve efficiency.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 61602429), China Postdoctoral Science Foundation (No. 2016M592406), and Research Funds of CUG from the Colleges Basic Research and Operation of MOE (No. 26420160055).

References

[1] T. Zhang, W. Zheng, Z. Cui, Y. Zong, J. Yan, K. Yan, A deep neural network–driven feature learning method for multi-view facial expression recognition, IEEE Trans. Multimed. 18 (12) (2016) 2528–2536.
[2] X. Yuan, L. Xie, M. Abouelenien, A regularized ensemble framework of deep learning for cancer detection from multi-class, imbalanced training data, Pattern Recognit. 77 (2018) 160–172.
[3] B. Fang, Q. Zhang, H. Wang, X. Yuan, Personality driven task allocation for emotional robot team, Int. J. Mach. Learn. Cybern. (2018) inpress.
[4] A.T. Lopes, E. de Aguiar, A.F. De Souza, T. Oliveira-Santos, Facial expression recognition with convolutional neural networks: coping with few data and the training sample order, Pattern Recognit. 61 (2017) 610–628.
[5] E. Hamouda, O. Ouda, X. Yuan, T. Hamza, Optimizing discriminability of globally Binarized face templates, Arab. J. Sci. Eng. 41 (8) (2016) 2837–2846.
[6] H. Jung, S. Lee, J. Yim, S. Park, J. Kim, Joint fine-tuning in deep neural networks for facial expression recognition, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 2983–2991.
[7] A. Tawari, M.M. Trivedi, Face expression recognition by cross modal data association, IEEE Trans. Multimed. 15 (7) (2013) 1543–1552.
[8] O. Rudovic, I. Patras, M. Pantic, Coupled gaussian process regression for pose-invariant facial expression recognition, in: Proceedings of the Computer Vision–ECCV (2010) 350–363.
[9] S. Moore, R. Bowden, Local binary patterns for multi-view facial expression recognition, Comput. Vis. Image Underst. 115 (4) (2011) 541–558.
[10] X. Yuan, M. Abouelenien, M. Elhoseny, Quantum Computing:An Environment for Intelligent Large Scale Real Application, Springer International Publishing, pp. 433–448.
[11] C. Shan, Smile detection by boosting pixel differences, IEEE Trans. Image Process. 21 (1) (2012) 431–436.
[12] H.V. Nguyen, H.T. Ho, V.M. Patel, R. Chellappa, Dash-n: joint hierarchical domain adaptation and feature learning, IEEE Trans. Image Process. 24 (12) (2015) 5479–5491.
[13] M. Xu, W. Cheng, Q. Zhao, L. Ma, F. Xu, Facial expression recognition based on transfer learning from deep convolutional networks, in: Proceedings of the Eleventh International Conference on Natural Computation (ICNC), IEEE, 2015, pp. 702–708.
[14] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. ng, E. Tzeng, T. Darrell, Decaf: a deep convolutional activation feature for generic visual recognition., in: Proceedings of the ICML, 32, 2014, pp. 647–655.
[15] R. Girshick, Fast R-CNN, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1440–1448.
[16] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, T. Darrell, Caffe: convolutional architecture for fast feature embedding, in: Proceedings of the Second ACM International Conference on Multimedia, ACM, 2014, pp. 675–678.
[17] O.M. Parkhi, A. Vedaldi, A. Zisserman, Deep face recognition., in: Proceedings of the BMVC, 1, 2015, p. 6.
[18] A. Mollahosseini, D. Chan, M.H. Mahoor, Going deeper in facial expression recognition using deep neural networks, in: Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, 2016, pp. 1–10.
[19] M. Liu, S. Li, S. Shan, R. Wang, X. Chen, Deeply learning deformable facial action parts model for dynamic expression analysis, in: Proceedings of the Asian Conference on Computer Vision, Springer, 2014, pp. 143–157.
[20] 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 (FG), IEEE, 2013, pp. 1–6.
[21] O. Gupta, D. Raviv, R. Raskar, Illumination invariants in deep video expression recognition, Pattern Recognit. 76 (2018) 25–35.
[22] M. Yang, X. Wang, G. Zeng, L. Shen, Joint and collaborative representation with local adaptive convolution feature for face recognition with single sample per person, Pattern Recognit. 66 (2017) 117–128.
[23] S.R. Bulo, P. Kontschieder, Neural decision forests for semantic image labeling, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 81–88.
[24] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.
[25] M. Dantone, J. Gall, G. Fanelli, L. Van Gool, Real-time facial feature detection using conditional regression forests, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012, pp. 2578–2585.
[26] M. Sun, P. Kohli, J. Shotton, Conditional regression forests for human pose estimation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012, pp. 3394–3401.

[27] A. Criminisi, J. Shotton, E. Konukoglu, Decision forests for classification, regression, density estimation, manifold learning and semi-supervised learning, Microsoft Research Cambridge, Technical Report MSR-TR-2011-114 5(6) (2011) 12.
[28] G. Fanelli, M. Dantone, J. Gall, A. Fossati, L. Van Gool, Random forests for real time 3d face analysis, Int. J. Comput. Vis. 101 (3) (2013) 437–458.
[29] G. Fanelli, A. Yao, P.-L. Noel, J. Gall, L. Van Gool, Hough forest-based facial expression recognition from video sequences, in: Proceedings of the European Conference on Computer Vision, Springer, 2010, pp. 195–206.
[30] 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.
[31] H. Kim, M.-K. Sohn, D.-J. Kim, S.-H. Lee, Kernel locality-constrained sparse coding for head pose estimation, IET Comput. Vis. 10 (8) (2016) 828–835.
[32] A. Ito, X. Wang, M. Suzuki, S. Makino, Smile and laughter recognition using speech processing and face recognition from conversation video, in: Proceedings of the International Conference on Cyberworlds, IEEE, 2005, p. 8.
[33] S.E. Kahou, P. Froumenty, C. Pal, Facial expression analysis based on high dimensional binary features, in: Proceedings of the European Conference on Computer Vision, Springer, 2014, pp. 135–147.
[34] W. Zheng, Multi-view facial expression recognition based on group sparse reduced-rank regression, IEEE Trans. Affect. Comput. 5 (1) (2014) 71–85.
[35] A. Dapogny, K. Bailly, S. Dubuisson, Dynamic pose-robust facial expression recognition by multi-view pairwise conditional random forests, arXiv:1607.06250 (2016).
[36] X. Hou, J. Harel, C. Koch, Image signature: highlighting sparse salient regions, IEEE Trans. Pattern Anal. Mach. Intell. 34 (1) (2012) 194–201.
[37] S. Goferman, L. Zelnik-Manor, A. Tal, Context-aware saliency detection, IEEE Trans. Pattern Anal. Mach. Intell. 34 (10) (2012) 1915–1926.
[38] Y. Liu, J. Chen, Z. Su, Z. Luo, N. Luo, L. Liu, K. Zhang, Robust head pose estimation using Dirichlet-tree distribution enhanced random forests, Neurocomputing 173 (2016) 42–53.
[39] P. Lucey, J.F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, I. Matthews, The extended Cohn–Kanade dataset (ck+): a complete dataset for action unit and emotion-specified expression, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, IEEE, 2010, pp. 94–101.
[40] M.J. Lyons, J. Budynek, S. Akamatsu, Automatic classification of single facial images, IEEE Trans. Pattern Anal. Mach. Intell. 21 (12) (1999) 1357–1362.
[41] L. Yin, X. Wei, Y. Sun, J. Wang, M.J. Rosato, A 3d facial expression database for facial behavior research, in: Proceedings of the International Conference on Automatic Face and Gesture Recognition, IEEE, 2006, pp. 211–216.
[42] G.B. Huang, M. Ramesh, T. Berg, E. Learned-Miller, Labeled faces in the wild: a database for studying face recognition in unconstrained environments, Technical Report 07–49, University of Massachusetts, Amherst, 2007.
[43] A. Dapogny, K. Bailly, S. Dubuisson, Pairwise conditional random forests for facial expression recognition, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 3783–3791.
[44] X. Zhang, M.H. Mahoor, S.M. Mavadati, Facial expression recognition using {l}_{p}-norm MKL multiclass-SVM, Mach. Vis. Appl. 26 (4) (2015) 467–483.
[45] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks, in: Proceedings of the Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.
[46] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016) 770–778.

Yuanyuan Liu received B.E. degree from NanChang University, NanChang, China, in 2005, M.E. degree from Huazhong University of Science and Technology, Wuhan, China, in 2007, and Ph.D. degree from Central China Normal University. She is currently a lecturer in China University of Geosciences (Wuhan). Her research interests include image processing, computer vision and pattern recognition.

Xiaohui Yuan received a B.S. degree in Electrical Engineering from the Hefei University of Technology, Hefei, China in 1996 and a Ph.D. degree in Computer Science from the Tulane University, New Orleans, USA in 2004. He is currently an Associate Professor at the Department of Computer Science and Engineering at the University of North Texas and a Visiting Professor at the China University of Geosciences, Wuhan, China. His research interests include computer vision, data mining, machine learning, and artificial intelligence. Dr. Yuan is a recipient of Ralph E. Powe Junior Faculty Enhancement award in 2008 and the Air Force Summer Faculty Fellowship in 2011, 2012, and 2013. He is a senior member of IEEE.

Fang Fang received a B.S. degree in Computer Science and Technology in 1998 and a Ph.D. degree in Management Science and Engineering in 2012 from the China University of Geosciences, Wuhan, China. She is currently an Associate Professor at the College of Information Engineering at the China University of Geosciences. Her research interests include spatial data mining, machine learning, and GIS applications. She is a member of CCF and IEEE.