Gergo Gyori
IT University of Copenhagen

Outline

Pattern Recognition Letters
Volume 119, 1 March 2019, Pages 49-61
Deep spatial-temporal feature fusion for facial expression recognition in static images
Author links open overlay panelNing Sun, Li Qi, Ruizhi Huan, Jixin Liu, Guang Han

https://doi.org/10.1016/j.patrec.2017.10.022Get rights and content
Highlights
* ?
Using optical flow to represent the temporal features of FER in static image.
* ?
A MDSTFN is present to extract and fuse the temporal and spatial features for FER.

* ?
Using Average-face as a substitution for neutral-face.
* ?
Achieving recognition accuracy 98.38% on CK+, 99.17% on RafD, and 99.59% on MMI.
ABSTRACT
Traditional methods of performing facial expression recognition commonly use hand-crafted spatial features. This paper proposes a multi-channel deep neural network that learns and fuses the spatial-temporal features for recognizing facial expressions in static images. The essential idea of this method is to extract optical flow from the changes between the peak expression face image (emotional-face) and the neutral face image (neutral-face) as the temporal information of a certain facial expression, and use the gray-level image of emotional-face as the spatial information. A Multi-channel Deep Spatial-Temporal feature Fusion neural Network (MDSTFN) is presented to perform the deep spatial-temporal feature extraction and fusion from static images. Each channel of the proposed method is fine-tuned from a pre-trained deep convolutional neural networks (CNN) instead of training a new CNN from scratch. In addition, average-face is used as a substitute for neutral-face in real-world applications. Extensive experiments are conducted to evaluate the proposed method on benchmarks databases including CK+, MMI, and RaFD. The results show that the optical flow information from emotional-face and neutral-face is a useful complement to spatial feature and can effectively improve the performance of facial expression recognition from static images. Compared with state-of-the-art methods, the proposed method can achieve better recognition accuracy, with rates of 98.38% on the CK+ database, 99.17% on the RaFD database, and 99.59% on the MMI database, respectively.
* Previous article in issue
* Next article in issue
Keywords
Facial expression recognition
Deep neural network
Optical flow
Spatial-temporal feature fusion
Transfer learning
1\. Introduction
Facial expression recognition (FER) is an important aspect of face image analysis and has been a popular research topic in computer vision for decades. Recently, FER has been successfully implemented in the applications of smart video surveillance, personal robot and interactive games [1]. Facial expressions are the facial changes in response to a person's internal emotional states, intentions, or social communications. In 1978, Ekman et al. [2] reviewed psychological research on facial expressions and emotions and found six emotions to be universal. The six universal emotions are disgust, sadness, happiness, fear, anger, and surprise.
A FER system generally consists of three steps: image pre-processing, feature extraction, and expression recognition. The main purpose of image pre-processing is to crop the detected face from the image and align it to a fixed size and position based on facial landmark points. Features are then extracted from the aligned face to represent the intrinsic property of facial expressions. In the third step, a classifier is trained to recognize the facial expression.
Due to the complexities of the human face and emotional expressions, together with the effects of the illumination, perspective, and pose, effective feature

extraction is vital to FER. Feature extraction methods for FER can be divided into geometry-based and appearance-based methods. Geometry-based methods recognizes facial expressions by measuring the geometric characteristics of the face, such as distance and curvature of facial fiducial points or salient regions. The early work in this area was the Facial Action Coding System (FACS), designed by Ekman[2]. FACS encoded a facial expression in 44 facial Action Units (AUs), which defined the contraction of one or more facial muscles. A face registration algorithm such as Active Shape Model (ASM) [3] and Supervised Descent Method (SDM) [4] was then proposed for locating the facial fiducial points. According to the deformation of these fiducial points, Pantic's method [5] can determine the category of the facial expression. Geometry-based methods are intuitive and convenient, but they are sensitive to noise and disturbance and have difficulties describing some subtle changes in the face. Appearance-based methods use appearance features to represent facial expressions. These appearance features include image density, edge, texture, and more discriminative features that are obtained through a local descriptor or global space transformation such as LBP descriptor [6], PHOG descriptor [7], Gabor filter [8], and Non negative Matrix Factorization (NMF) [9].
Compared with geometry-based methods, appearance-based methods have an obvious advantage in robustness to noise and in retaining detailed information of the facial expression.
In recent years, research on deep learning has had tremendous success in many competitions and computer vision applications [10], [11], [12], [13], [14].
Deep learning methods stack a number of intermediate layers from input data to a classification layer, and can automatically learn high-level semantic features from a large amount of training data. FER methods based on CNN (Convolutional Neural Network) [15], which extract a hierarchy of nonlinear facial features by means of multi-layers of convolution and pooling, can achieve higher rates of accuracy on several facial expression benchmarks. Others deep neural network models such as Deep Belief Network (DBN) [16] and Deep Boltzmann Machine (DBM) [17] have also been successfully applied to FER.
FER methods are commonly designed to deal with two types of input: a static image and a dynamic image sequence (video). Image-oriented methods determine the class of facial expression from a single still image. This image indicates the momentary appearance of a facial expression (usually, the apex of the expression). Compared with a static image and momentary expression, video shows the temporal changes to facial appearance when an expression occurs. Besides the spatial features extracted from a single face image, video-oriented methods can represent the temporal features of expressions from dynamic transitions of between different stages of an expression rather than their corresponding static key frames. Generally, a video-oriented method can achieve a better recognition accuracy, because it can extract spatial and temporal features from dynamic image sequences. However, video-oriented methods recognize expressions from large-scale image sequences, which inevitably lead to higher computational complexity and can also introduce noise and disturbance.
In this paper, we present a FER method to simultaneously extract and fuse the temporal and spatial features from static face images. This method can not only improve the recognition performance of FER from static images, but also avoid the processing cost of large-scale image sequences. As an initial step, we assume that the face image with expression (emotional-face) implicitly contains the temporal information of the changes in facial appearance, which can be obtained by measuring the difference between emotional-face and the face image without expression (neutral-face). This kind of difference is

represented by means of the optical flow between emotional-face and neutral-face. A Multi-channel Deep Spatial-Temporal feature Fusion neural Network (MDSTFN) is proposed to extract and fuse the spatial-temporal features from the static face image. There are several deep neural network channels in MDSTFN. One channel is used to extract spatial features from gray-level image of emotional-face, and the others are utilized to extract temporal features from optical flow images. These temporal and spatial features are fused by several fusion schemes to generate the final output of recognition results. This multi-channel deep neural network architecture has been successfully applied in the field of video-oriented FER [18] and action recognition [19], [20].

The proposed method uses pre-trained deep neural network models, which are fine-tuned on a facial expression database to build MDSTFN instead of designing a new network model trained from scratch. These pre-trained deep models were trained on large-scale image databases (such as ImageNet) and have achieved the top performance in the recent years in object recognition tasks set by the ImageNet Large-Scale Visual Recognition Competition (ILSVRC). The pre-trained models are employed at an initialization stage and fine-tuned to fit the FER task on the benchmark facial expression databases. Transfer learning from pre-trained models can offer a strong initial feature extractor and make the large-scale CNN model easier to fine-tune on a limited facial expression database. It is also convenient for research as the feature extraction channels of MDSTFN can be substituted with more powerful deep models in the future. However, it is usually difficult to find the neutral-face corresponding to a certain emotional-face in real-world applications. To solve this problem, we use average-face, which is the average of a large number of facial images, to replace neutral-face for computing optical flow.

The distinctive features of this research can be summarized as follows:

* (1)
We use the optical flow extracted from the changes between emotional-face and neutral-face to represent the temporal features of a facial expression in a static image;
* (2)
An MDSTFN is present to extract and fuse the temporal and spatial features of the facial expression in a static image. This architecture not only improves the of recognition performance but also makes the proposed method easy to update;
* (3)
Using Average-face as a substitute for neutral-face makes the proposed method suitable for implementation in real-world applications;
* (4)
We conduct extensive experiments to comprehensively evaluate the proposed FER method on benchmarks including CK+, MMI, and RaFD. Results show that the optical flow is able to represent the temporal changes to facial expression in a static image. It can effectively improve the performance of FER by fusing these spatial-temporal features extracted by the proposed MDSTFN-based method.

2\. Related work

In this section, we review recent related work in FER which uses methods based on deep learning architecture. Firstly, as regards the image-oriented FER method, Ranzato et al [21] presented a DBN-based deep generative model composed of three layers to learn expression features from a face image; this worked well when a face image was disrupted by heavy occlusion. Liu et al. [16] developed a Boosted Deep Belief Network (BDBN) for FER. This method attempted to perform feature learning, feature selection, and classifier

construction in a loopy process. He et al. [17] proposed a Deep Boltzmann Machine model composed of a two-layers Boltzmann machine for FER from infrared images. The advantage of the DBN model is that it can be pre-trained in an unsupervised way, but the DBN model has difficulties to dealing with a large number of high-resolution images due to its fully-connected structure.

The essential characteristics of a CNN model are local receptive field and weight sharing, which give it a significant advantage in dealing with high-resolution data such as images and video. Khorrami et al. [22] proposed a CNN that ignored the biases of the convolutional layers; this zero-bias CNN was trained on facial expression data and achieved good performance on two expression recognition benchmarks. Burkert et al [23] proposed a novel CNN-based FER method called DeXpression. The core of their proposed CNN model comprised two Parallel Feature Extraction (FeatEx) blocks; FeatEx blocks create two parallel paths of features with different scales. Liu et al. [24] constructed an AU-inspired Deep Network (AUDN) architecture, which was inspired by Ekman's psychological theory that expressions can be decomposed into multiple facial Action Units (AUs). AUDN consisted of three sequential modules; Micro-Action-Pattern (MAP) representation learning, receptive field construction and group-wise subnetwork learning. Lopes et al. [25] presented a FER system that used a combination of CNN and specific image pre-processing steps. These pre-processing steps were used to augment the presentation order of the samples for training the deep neural network. Hamester et al. [26] proposed a two-channel deep learning architecture for a FER system. One channel was a standard CNN; the second had the same topology but its first layer weights were trained as a Convolutional AutoEncoder (CAE). The two channels were connected with a fully connected (FC) layer that generates the output for recognition.

For video or sequential image data, the correlation between consecutive frames provides discriminative information for FER. Jung et al. [27] proposed a two-channel deep neural network to recognize expressions from video data. The first channel extracted temporal appearance features from image sequences, while the second extracted temporal geometric features from temporal facial landmark points. These two channels were combined using joint fine-tuning to boost the performance of FER. Byeon et al. [28] developed a 3D-CNN to learn spatial-temporal expression information from five successive frames. The authors claimed that the 3D-CNN model can handle some degrees of shift and deformation invariance.

Compared with the aforementioned FER methods, this paper presents a MDSTFN model that extracts and fuses the spatial-temporal features for FER from a static image. We use optical flow extracted from the changes between emotional-face and neutral-face to represent the temporal features of facial expression. Based on transfer learning, the channels in MDSTFN are fine-tuned from pre-trained CNN models. This not only can solve the problem of insufficient facial expression training for large-scale deep neural networks, but also facilitates future improvement.

3\. Proposed facial expression recognition method

The proposed MDSTFN-based FER method involves the following three steps: image pre-processing, deep spatial-temporal feature learning, and feature fusion and recognition. A schematic diagram of the proposed method is shown in Fig. 1.

1. Download: Download high-res image (539KB)
2. Download: Download full-size image

Fig. 1. The architecture of the MDSTFN-based FER method.

3.1. Image pre-processing

The image pre-processing step includes two parts: face alignment and optical

flow extraction. Face alignment locates the key points of face and deforms the face image to a fixed size and position. The optical flow features are extracted from the differences between emotional-face and neutral-face. Whether the positions of face parts of two images are aligned or not can greatly affect the performance of optical flow extraction. In our work, the input face images are aligned by the ASM algorithm.

The key pre-processing of the proposed method is the extraction of optical flow. In comparison with extracting multiple optical flow from consecutive frames in video data, we use only one optical flow image computed from difference between emotional-face and neutral-face to represent the temporal changes in a static expression. This strategy can effectively capture the facial changes when a certain expression occurs, and reduce the computational cost of a frame-by-frame optical flow extraction. The optical method proposed by Brox [29] is chosen as the optical flow extractor in this paper. This optical method integrates a coarse-to-fine warping strategy and implements the non-linear optical flow constraint to yield excellent results even under a considerable amount of noise. Images of optical flow extracted by this method are shown in Fig. 6.

3.2. Multi-channel deep spatial-temporal feature fusion neural network (MDSTFN)

After the aforementioned pre-processing, we have three channels of input data. One channel is the gray-level image of emotional-face; the others two the X and Y components of optical flow extracted from emotional-face and neutral-face. In order to learn and fuse spatial-temporal features from three channels of input data, a Multi-channel Deep Spatial-Temporal feature Fusion neural Network (MDSTFN) is constructed. The architecture of MDSTFN is shown in Fig. 2. The pre-trained CNN model is used to form the feature extraction channel of MDSTFN. To fit the FER task, transfer learning is performed to fine-tune the pre-trained CNN model on facial expression databases. There are two benefits of using a pre-trained CNN model instead of training a new CNN model from scratch. Firstly, these publicly available pre-trained CNN models have strong generalized capabilities in image feature representation. Transfer-learning from a pre-trained model is an effective way to improve the performance of training since a facial expression image database is usually small. Secondly, the pre-trained CNN model can be replaced easily by a new better deep neural network model in the future.

1. Download: Download high-res image (197KB)
2. Download: Download full-size image

Fig. 2. The architecture of Multi-channel Deep Spatial-Temporal feature Fusion neural Network.

According to recent developments in deep learning, the most straightforward way of improving the performance of deep neural networks is by increasing their size. The size of representative CNN architectures has dramatically increased from AlexNet [11] with 8 layers in 2012, to VGG [12] with 19 layers, and GoogLeNet [13] with 22 layers in 2014, then to ResNet [15] with 152 layers in 2015. Deeper neural networks achieve better performance in large scale image recognition. However, a bigger size typically means a larger number of parameters, which makes the network more prone to overfitting. Especially with regard to the task of FER, the number of labeled examples in the training set is limited. It is always difficult to obtain a satisfactory performance on a large size neural network with fewer training samples. In this paper, we use a simplified version of GoogLeNet (GoogLeNetv2), which comprises the layers below the second softmax output of GoogLeNet, as the feature extraction channel in the MDSTFN. The size of the receptive field in GoogLeNetv2 is 224 ×

224. This is followed by six Inception modules and several convolution and pooling layers. The detail of GoogLeNetv2 is shown in Fig. 3; details of the Inception module are shown in Fig. 4. The bottom layers of convolution and pooling reduce significantly the dimension of input data and extract low-level features from the face image. The six Inception modules basically act as multiple convolution filters that process the same input and also do pooling at the same time. This allows the model to take advantage of multi-level feature extraction from each input.

1. Download: Download high-res image (169KB)
2. Download: Download full-size image

Fig. 3. The architecture of GoogLeNetv2.

1. Download: Download high-res image (167KB)
2. Download: Download full-size image

Fig. 4. The detail of Inception module.

The parameters of GoogLeNetv2 (as trained on the ImageNet database) are retained to initialize the training of the MDSTFN-based method. We fine-tune GoogLeNetv2 on facial expression benchmark databases to fit the FER task. Based on transfer learning like this, we can obtain a better CNN classifier on a small number of image samples while reducing the effect of overfitting.

3.3. Spatial-temporal feature fusion

We investigate three approaches to fusing temporal and spatial information from three feature extraction channels, as shown in Fig. 5. They are: score averaging fusion, Support Vector Machine (SVM) based fusion and neural network based fusion. A detailed comparison of the three fusion methods follows a in the discussion of experiments (4.4 below).

1. Download: Download high-res image (324KB)
2. Download: Download full-size image

Fig. 5. Three different fusion strategies. Red, blue, yellow, and black boxes indicate convolutional, pooling, fully-connected and softmax layers respectively. The larger green dashed line boxes are Inception module. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Score Averaging Fusion (SAF): We average the softmax output of three channels of MDSTFN to fuse the spatial-temporal feature. The expression corresponding to the max value of fusion is the recognition result. SAF is easy to compute but the improvement in performance is limited.

SVM _based Fusion (SVMF):_ After being independently fine-tuned on a facial expression image database, the output of the top FC layer of GoogLeNetv2 is concatenated in order to train a SVM-based classifier. The SVM-based classifier performs feature fusion and FER simultaneously.

Neural Network based Fusion (NNF): At first, three channels of MDSTFN are fine-tuned independently. We then adjust the model of the fine-tuned MDSTFN. The softmax layers of the three channels are removed, and a new FC layer is added which is now connected to the top FC layers of the three channels; the new FC layer is followed by a softmax layer. We then freeze the rest layers of MDSTFN and just train the top two FC layers of MDSTFN. When a new face image is input, the MDSTFN can generate directly the final recognition result.

3.4. Replacing neutral-face with average-face

In this method, the key step of the proposed method is to compute the optical flow between emotional-face and neutral-face. Generally, neutral-face images exist in most facial expression databases, but it is not easy to find the neutral-face corresponding to a certain emotional-face in real-world application. To address this problem, average-face can be used to replace neutral-face when our MDSTFN-based FER method is applied in a real-world

application. Average-face is the average of a large number of faces, which effectively smooths the appearance differences in face images caused by changes in expression and illumination. Although there is still a difference between average-face and neutral-face, it has very little influence on FER, as shown by the results of subsequent experiments (4.5 below).

4\. Experiments and discussion

In this section, we conduct extensive experiments to comprehensively evaluate the proposed FER method. The experiments are performed on three publicly facial expression benchmarks: the extended Cohn?Kanade (CK+) database [30]; the Radboud faces database (RaFD) [31]; and the MMI facial expression database [32]. These three benchmarks contain facial expression images obtained from a wide variety of subjects (i.e., as regards age, gender, and ethnicity).

4.1. Database and experimental protocols

CK+ database consists of 593 sequences from 123 subjects, which is an extended version of Cohn-Kanade (CK) database. The subjects in the database are 81%, Euro-American, 13% Afro-American, and 6% other groups with 69% female, from 18 to 50 years of age. The validating emotion labels were only assigned to 327 sequences which were found to meet criteria for one of 7 discrete emotions (anger, contempt, disgust, fear, happiness, sadness, and surprise). Each of the sequences contains images from onset (neutral frame) to peak expression (last frame). In our experiments, we firstly select seven successive the most peak images in each sequence, and the first, fourth, and seventh images are collected to be as the emotional-face. The reason why we do not use three consecutive face images is to reduce the sample correlation caused by too-similar face images. And, the first image in each sequence is used to be as the neutral-face. So we have 1083 emotional-face images and 123 neutral-face images.

MMI database holds 2885 videos and over 500 images of 88 subjects displaying various facial expressions on command, where 236 videos have basic emotion annotation (anger, disgust, fear, happiness, sadness, surprise, scream, bored, and sleepy). There is not fixed pattern of expression change in each video. Some videos begin with neutral state, and others begin with peak expression. So we manually select 5042 emotional-face images and 88 neutral-face images from these videos for the experiments.

RaFD database is a set of pictures of 67 models (including Caucasian males and females, Caucasian children, both boys and girls, and Moroccan Dutch males) displaying 8 emotional expressions, which are anger, disgust, fear, happiness, sadness, surprise, contempt, and neutral. Each emotion was shown with three different gaze directions and all pictures were taken from five camera angles simultaneously. Only the front view images are collected for training the MDSTFN-based FER method. There are 1407 emotional-face images and 201 neutral-face images. Compared with other two database, the image quality and acquisition environment of RaFD database is the best.

We train the proposed method using two kinds of experimental protocol. The first is subject-independent. Using this protocol, we construct ten person-independent subsets by sampling in ID ascending order with a step size equal to 10, which means that if image of subject X is in one group, no image of X will be in any other group). The second protocol is random-division. Using this protocol, samples are randomly partitioned into ten groups of equal size without constraint of person independence between groups. With both training protocols, ten-fold cross-validation is used to test the recognition accuracy of the proposed method. The experimental protocol are specified in the following experiments.

The steps of facial alignment and optical flow extraction are developed in the

OpenCV and C++environment, and the MDSTFN is trained using Caffe Python API. All experiments run on an image processing workstation with an Intel Xeon 2.4 GHz 8 core CPU, 128GB memory, and two NVIDIA Titan X GPUs.

4.2. Evaluation of different channel combinations

The primary idea of the proposed method is to extract optical flow between emotional-face and neutral-face as the information of temporal change, and to fuse this with complementary information from the gray-level emotional-face to improve the performance of FER. In order to verify the effectiveness of this idea, samples of six universal emotional-face and the corresponding directional components of optical flow are extracted from the CK+, RaFD, and MMI databases, as shown in Fig. 6. It can clearly be seen that there is a unique pattern for each expression although these face images are captured from subjects of different age, gender, and ethnicity, and even though some images have facial accessories such as glass and scarves. For example, in the expression of surprise, the significant change of face is opening one's eyes and mouth widely. The corresponding Y-component of optical flow has a strong response in the areas of the eyes and mouth. There are similar results in other expressions. Compared with the X-component of optical flow, the Y-component of optical flow looks more discriminative for FER. The main result is that the change in appearance caused by facial expression change in the Y direction is stronger than the one in the X direction.

1. Download: Download high-res image (573KB)
2. Download: Download full-size image

Fig. 6. Optical flow extracted from three databases cross different expressions.

We showed above that optical flow extracted between emotional-face and neutral-face can represent the temporal features of a facial expression. In the next step, we conduct experiments to assess the performance of the MDSTFN-based FER method with different channel combinations on the CK+, RaFD, and MMI databases. For the MDSTFN-based FER method with a single channel, we use the softmax output of the channel to recognize the facial expression. For MDSTFN with two or three channels, the NNF method is used to fuse the spatial-temporal features and generate the final recognition result. These experiments are conducted based on the random-division protocol. The recognition accuracy is shown in Table 1.

Table 1. The recognition accuracy of the MDSTFN-based FER method with different channel combinations on three benchmark databases. The letter G, Y, and X in the second row means gray-level emotional-face, Y component of optical flow, and X component of optical flow, respectively.

Database	Channel						
Empty Cell	G	X	Y	G + X	G + Y	X + Y	G + X + Y
CK+	89.60%	66.52%	65.06%	96.65%	96.89%	85.27%	**98.38** %
RaFD	93.19%	45.27%	75.88%	98.33%	**99.17%**	82.37%	98.75%
MMI	85.09%	82.10%	94.62%	99.40%	99.50%	99.08%	**99.59%**

The proposed method is able to achieve an accuracy of 66.52% and 65.09% using only the Y or X component of optical flow respectively on the CK+ databases. Obviously, the recognition accuracy improves greatly when fusing the spatial-temporal feature: 98.38% for G+X+Y up from 89.6% for G on the CK+ database; 99.17% for G + Y up from 93.19% for G on the RaFD database; and 99.59% for G + X + Y up from 85.09% for G on the MMI database. The recognition accuracy shows once again that the optical flow extracted between emotional-face and neutral-face plays an effective complementary role to gray-level emotional-face. Fusing the spatial-temporal features extracted in this way can significantly

improve the performance of FER. From Table 1, it can be seen that the results achieved by Y or Y + G are generally better than the results achieved by X or X + G. This confirms that the Y-component of optical flow is more discriminative for FER than the X-component of optical flow, as also demonstrated in Fig. 6.

We show the confusion matrixes of the proposed method using channel G or using channel G + X + Y on three benchmark facial expression database in Fig. 7. It can be clearly seen that those expressions hardly distinguished by the proposed method only using gray-level emotion-face are effectively classified by the proposed method combining gray-level emotion-face and optical flow information, such as anger vs. fear in the CK+ and MMI database. The rest 15 confusion matrixes of our method using various channel combinations are shown in Appendix.

1. Download: Download high-res image (446KB)
2. Download: Download full-size image

Fig. 7. Confusion matrixes of the proposed method using various channel combination on CK+, RaFD, and MMI database, which is shown in the first, second and third row, respectively. The first column is the results by the proposed method using one feature channel of gray-level emotional-face. The second column is the results by the proposed method using three feature channel including gray-level emotional-face, Y-component, and X-component of optical flow.

4.3. Evaluation of different pre-trained CNN models

We train the MDSTFN-based FER method using four publicly available CNN models; namely, AlexNet and three versions of GoogLeNet. The best model in this comparison is chosen as the model for feature extraction in the proposed method. The three versions of GoogLeNet are GoogLeNetv1 (the layers below the first softmax output of GoogLeNet); GoogLeNetv2 (the layers below the second softmax output of GoogLeNet); and the full GoogLeNet. Other CNN models like VGG and ResNet are not tested in the experiment because of the lack of pre-trained models. The experiments are conducted using the random-division protocol.

Table 2 shows the accuracy of the proposed FER method with the four CNN models. Overall, the results of the three GoogLeNet versions are better than that of AlexNet under almost all kinds of feature extraction combinations. It shows that the deeper architecture and Inception module provide more powerful capabilities in non-linear feature representation. GoogLeNetv2 (the layers below the second softmax output of GoogLeNet), achieves the best recognition accuracy in comparison with GoogLeNetv1 and GoogLeNet. This shows that the model complexity of GoogLeNetv2 is most suited for the sample size of the three facial expression databases (about 1 k to 5 k). Larger scale model like GoogLeNet always results in overfitting while smaller scale model like GoogLeNetv1 leads to under-fitting.

Table 2. The recognition accuracy of the MDSTFN-based FER method with different models on three benchmark databases.

Database	Model	Channel
Empty Cell| Empty Cell| G + X| G + Y| X + Y| G + X + Y
CK+| AlexNet| 79.56%| 87.56%| 88.89%| 88.89%
GoogLeNet| 93.99%| 95.08%| 74.5%| 93.44%
GoogLeNetv1| 91.26%| 92.90%| 70.33%| 91.80%
GoogLeNetv2| 96.65%| 96.89%| 85.27%| **98.38%**
RaFD| AlexNet| 84.36%| 96.42%| 90.11%| 96.9%
GoogLeNet| 98.75%| 97.08%| 78.75%| 97.92%

| GoogLeNetv1| 94.58%| 97.92%| 77.80%| 97.08%
| GoogLeNetv2| 98.33%| **99.17%**| 82.37%| 98.75%
MMI| AlexNet| 93.23%| 92.90%| 92.71%| 93.82%
| GoogLeNet| 96.56%| 94.32%| 89.77%| 96.43%
| GoogLeNetv1| 92.33%| 93.91%| 90.8%| 93.0%
| GoogLeNetv2| 99.41%| 99.50%| 98.97%| **99.59%**
4.4. Evaluation of different fusion strategies
We test the proposed method with different fusion strategies: SAF, SVMF, and NNF (see Section 3.3 above). The experiments are conducted using the random-division protocol. Table 3 shows the recognition results on three benchmark databases. NNF is the best fusion method compared with SVMF and SAF. The accuracy achieved by the FER method with NNF under different channel combinations is about 0.5% to 6% higher than the one of method with SVMF, and is about 10% higher than the one of method with SAF. SAF is a decision-level fusion that simply uses the average of the softmax output of the feature extraction channel as the fusion result. It is computationally efficient, but the improvement in recognition performance by SAF fusion is quite limited. The input of SVMF comes from the concatenation of the outputs of multi-channels. The SVM-based classifier can learn further from the relationships among the concatenated features, but the concatenation is a kind of rough fusion strategy and hard to achieve optimal performance. NNF on the other hand is treated as a multi-input fully-connected neural network. Instead of simply averaging or concatenation, the output of MDSTFN is learning from the three-layers of the neural network. Hence the feature fusion ability of NNF is more powerful than that of SAF and SVMF.
Table 3. The recognition accuracy of the MDSTFN-based FER method with fusion strategies on three benchmark databases.
Database	Fusion strategy	Channel
Empty Cell| Empty Cell| G + X| G + Y| X + Y| G + X + Y
CK+| SAF| 91.07%| 92.55%| 80.76%| 93.28%
SVMF| 95.17%| 95.08%| 80.83%| 95.58%
NNF| 96.65%| 96.89%| 85.27%| **98.38%**
RaFD| SAF| 92.43%| 95.66%| 77.6%| 95.63%
SVMF| 97.05%| 97.54%| 78.54%| 97.15%
NNF| 98.33%| **99.17%**| 82.37%| 98.75%
MMI| SAF| 93.5%| 97.63%| 97.04%| 98.96%
SVMF| 98.74%| 98.08%| 95.35%| 99.11%
NNF| 99.41%| 99.50%| 98.97%| **99.59** %
4.5. Evaluation of the proposed method using average-face
Average-face is the alternative to neutral-face when the proposed method is used in real-world applications. We evaluate the impact of using average-face on the proposed method. Two kinds of average-face images, which are average-face of man and average-face of woman, are generated by averaging all male or female face images in the three benchmark databases, respectively. As Fig. 8 shows, average-face of man and average-face of woman are well aligned and shows a neutral state.
1. Download: Download high-res image (92KB)
2. Download: Download full-size image
Fig. 8. Average-face of man and average-face of woman.
The results of the proposed method using average-face compared to neutral-face are shown in Table 4. The experiments are conducted using the random-division protocol. Table 4 shows that the proposed method can still extract the temporal feature effectively from the optical flow between emotional-face and

average-face. The proposed method using average-face can also achieve satisfactory recognition results, such as 97.08% on the CK+ database, 98.34% on the RaFD database, and 99.38% on the MMI database. Additionally, the proposed method using average-face is able to achieve a result close to the proposed method using neutral-face. There is only a slight decline in accuracy when we use average-face instead of neutral-face, such as 1.3% (from 98.38% to 97.08%) on the CK+ database, 0.83%(from 99.38% to 99.17%) on the RaFD database, and 0.21% (from 99.59% to 99.38%) on the MMI database. The main cause of performance degradation in the proposed method using average-face is that average-face cannot replicate the neutral-face corresponding to a certain image of emotional-face. We believe that the recognition accuracy of the proposed method using average-face can be further improved when more kinds of average-face with a wider range of age and ethnicity are used.

Table 4. The recognition accuracy of the MDSTFN-based FER method using average-face and neutral-face on three benchmark databases. The letter A and N in the second column means average-face and neutral-face, respectively.

Database	Empty Cell	Channel					
Empty Cell	Empty Cell	G	X	Y	G + X	G + Y	X + Y	G + X + Y
CK+	A	89.60%	51.18%	59.19%	96.54%	96.90%	83.97%	**97.08%**
	N	89.60%	66.52%	65.06%	96.65%	96.89%	85.27%	**98.38%**
RaFD	A	93.19%	42.17%	73.33%	97.01%	**98.34%**	80.63%	97.64%
	N	93.19%	45.27%	75.88%	98.33%	**99.17%**	82.37%	98.75%
MMI	A	85.09%	82.12%	86.42%	99.41%	**99.38%**	99.42%	99.27%
	N	85.09%	82.03%	94.57%	99.41%	99.50%	98.97%	**99.59%**

4.6. Evaluation of cross-database performance

The cross-database experiment is a good evaluation of generalization performance of the proposed method in real environments. In these experiments, one database is used for evaluation and the other databases are used to train the network. The experiments are conducted using the random-division protocol. Table 5 shows the results of the cross-database experiments. As can be seen, the recognition accuracy has decreased considerably. It shows that the differences between training database and test database have a negative impact on the accuracy of the proposed method. Also, the recognition accuracies of the model trained by samples without the RaFD database is higher than those trained by samples with the RaFD database. This shows that the generalization of the model trained by a database with limited constraints is better than that of the model trained by a database captured under ideal environments.

Table 5. Recognition results of the proposed method in cross-database experiments.

Training database	Test database	Channel		
Empty Cell	Empty Cell	G + X	C + Y	G + X + Y
CK+,RaFD	MMI	59.27%	61.67%	67.18%
CK+,MMI	RaFD	77.53%	82.21%	86.80%
MMI, RaFD	CK+	66.22%	68.57%	75.13%

4.7. Comparison with state-of-the-art

In this section, we compare the performance of the proposed method with state-of-the-art FER methods on the CK+, RaFD, and MMI benchmark databases. These recently reported FER methods are all based on the architecture of deep neural networks. FER methods based on shallow learning are not tested here, because the superiority of the deep learning method has been proven in a lot of object recognition tasks.

Although they use the same databases, these ten state-of-the-art methods are

tested using two kinds of experimental protocol. The results reported in the literature [33], [23], [36] are obtained using the random-division protocol, while the results reported in the literature [24], [15], [18], [35], [25], [22], [34] are achieved using the subject-independent protocol. To perform a valid comparison, we test the proposed method using both protocols. The comparison results using the subject-independent protocol are listed in Table 6, and the results based on the random-division protocol are listed in Table 7.

Table 6. Recognition accuracy comparison with state-of-the-art methods based on the subject-independent protocol.

Method	Database
Empty Cell| CK+| RaFD| MMI
Liu et al. [24]| 93.70%|	73.85%
Mollahosseini et al. [15]| 93.20%|	77.60%
Jung et al. [18]| 97.25%|	70.24%
Liu et al. [36].| 92.05%|	74.76%
Lopes et al. [25]| 96.76%|
Khorrami et al. [22]| **98.30%**|
Zhou et al. [34]| 97.50%| 97.75%
Our method (G + Y + A)| 95.93%| 97.02%| 89.79%
Our method (G + X + Y + A)| 96.48%| 97.40%| 90.21%
Our method (G + Y + N)| 96.26%| 98.13%| 91.03%
Our method (G + X + Y + N)| 97.28%| **98.17%**| **91.46%**

Table 7. Recognition accuracy comparison with state-of-the-art methods based on the random-division protocol.

Method	Database
Empty Cell| CK+| RaFD| MMI
Majumder et al. [33].| 98.95%|	97.55%
Burkert et al. [23]| **99.60** %|	98.63%
Ali et al. [35]|	93.75%
Our method (G + Y + A)| 96.90%| 98.34%| 99.38%
Our method (G + X + Y + A)| 97.08%| 97.64%| 99.27%
Our method (G + Y + N)| 96.89%| **99.17%**| 99.50%
Our method (G + X + Y + N)| 98.38%| 98.75%| **99.59%**

As the results in Tables 6 and 7 show, the accuracy of almost all versions of the proposed method (using average-face or neutral-face, random-division or subject-independent protocol, and including various models and parameters such as different feature channels) is better than the results of other deep learning based methods on three benchmark facial expression databases. It is worth noting that the experimental protocol of the results reported by the literatures listed in Tables 6 and 7 is not exactly the same as that of our method as there is no an accepted experimental standard for FER. Besides, most researchers do not describe the experimental parameters in detail. Especially in the experiments on the CK+ and MMI databases, researchers often collect the training images from videos according to their own principles. Although the proposed method does not achieve the highest recognition accuracy on CK+ database, we can safely draw the conclusion that our MDSTFN-based method is a successful deep spatial-temporal feature extraction architecture for FER and can achieve a better recognition performance compared with state-of-the-art.

5\. Conclusion

This paper presented a deep neural network architecture with multi-channels to

extract and fuse the spatial-temporal features of static image for FER. The optical flow computed from the changes between emotional-face and neutral-face is used to represent the temporal changes of expression, while the gray-level image of emotional-face is used to provide the spatial information of expression. The feature extraction channels of the MDSTFN (Multi-channel Deep Spatial-Temporal feature Fusion neural Network) are fine-tuned from a pre-trained CNN model. This transfer learning scheme can not only obtain a better feature extractor for FER on a small number of image samples but can also reduce the risk of overfitting. Three kinds of strategies were investigated to fuse the temporal and spatial features obtained by multiple feature channels. On three benchmark databases (CK+, RaFD, and MMI), extensive experiments were conducted to evaluate the proposed method under various parameters such as channel combination, fusion strategy, cross-database, and pre-trained CNN models. The results show that the MDSTFN-based method is a feasible deep spatial-temporal feature extraction architecture for facial expression recognition. Replacing neutral-face with average-face improves the practicality of the proposed method, while the results of a comparison show that the MDSTFN-based method can achieve better accuracy than state-of-the-art methods.

Acknowledgments

This work was supported by the National Nature Science Foundation of China (61471206 and 61401220), Natural Science Foundation of Jiangsu province (BK20141428 and BK20140884) and Science Foundation of Ministry of Education-China Mobile Communications Corporation(MCM20150504).

Appendix

Figs. A1?A5

1. Download: Download high-res image (275KB)
2. Download: Download full-size image

Fig. A1. Confusion matrix of the proposed method using X on three database.

1. Download: Download high-res image (264KB)
2. Download: Download full-size image

Fig. A2. Confusion matrix of the proposed method using Y on three database.

1. Download: Download high-res image (243KB)
2. Download: Download full-size image

Fig. A3. Confusion matrix of the proposed method using G + X on three database.

1. Download: Download high-res image (249KB)
2. Download: Download full-size image

Fig. A4. Confusion matrix of the proposed method using G + Y on three database.

1. Download: Download high-res image (263KB)
2. Download: Download full-size image

Fig. A5. Confusion matrix of the proposed method using X + Y on three database.

Special issue articlesRecommended articles

References

1. [1]
A. Corneanu C, M. Oliu, F. Cohn J, _et al._
Survey on RGB, 3D, thermal, and multimodal approaches for facial expression recognition: history, trends, and affect-related applications
IEEE Trans. Pattern Anal. Mach. Intell., 38 (8) (2016), pp. 1548-1568
Google Scholar
2. [2]
P. Ekman, V. Friesen W

Facial action coding system: a technique for the measurement of facial movement
Rivista Di Psichiatria, 47 (2) (1978), pp. 126-138
Google Scholar
3. [3]
F. Cootes T, J. Taylor C, H. Cooper D, _et al._
Active shape models-their training and application
Comput. Vision Image Understanding, 61 (1) (1995), pp. 38-59
Google Scholar
4. [4]
Xiong X, De l T F. Supervised descent method and its applications to face alignment. 2013, 9(4):532?539.
Google Scholar
5. [5]
M. Pantic, I. Patras
Dynamics of facial expression: recognition of facial actions and their temporal segments from face profile image sequences
Syst. Man Cybern. Part B Cybern. IEEE Trans., 36 (2) (2006), pp. 433-449
View in ScopusGoogle Scholar
6. [6]
C. Shan, S. Gong, W. Mcowan P
Facial expression recognition based on local binary patterns: a comprehensive study
Image Vision Comput., 27 (6) (2009), pp. 803-816
View PDFView articleView in ScopusGoogle Scholar
7. [7]
A. Dhall, A. Asthana, R. Goecke, _et al._
Emotion recognition using PHOG and LPQ features
IEEE International Conference on Automatic Face & Gesture Recognition and Workshops, IEEE (2011), pp. 878-883
CrossrefView in ScopusGoogle Scholar
8. [8]
W. Liu, C. Song, Y. Wang, _et al._
Facial expression recognition based on Gabor features and sparse representation
International Conference on Control, Automation, Robotics and Vision (2012), pp. 1402-1406
View in ScopusGoogle Scholar
9. [9]
J. Zhou, S. Zhang, H. Mei, _et al._
A method of facial expression recognition based on Gabor and NMF
Pattern Recognit. Image Anal., 26 (1) (2016), pp. 119-124
Google Scholar
10. [10]
O. Russakovsky, J. Deng, H. Su, _et al._
ImageNet large scale visual recognition challenge
Int. J. Comput. Vision, 115 (3) (2015), pp. 211-252
CrossrefGoogle Scholar
11. [11]
A. Krizhevsky, I. Sutskever, G.E. Hinton
ImageNet classification with deep convolutional neural networks
International Conference on Neural Information Processing Systems, Curran Associates Inc (2012), pp. 1097-1105
Google Scholar

12. [12]
Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint 2014, arXiv:1409.1556.
Google Scholar
13. [13]
C. Szegedy, W. Liu, Y. Jia, _et al._
Going deeper with convolutions
Computer Vision and Pattern Recognition, IEEE (2014), pp. 1-9
Google Scholar
14. [14]
K. He, X. Zhang, S. Ren, _et al._
Deep residual learning for image recognition
Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016), pp. 770-778
Google Scholar
15. [15]
A. Mollahosseini, D. Chan, M.H. Mahoor
Going deeper in facial expression recognition using deep neural networks
Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on, IEEE (2016), pp. 1-10
Google Scholar
16. [16]
P. Liu, S. Han, Z. Meng, _et al._
Facial expression recognition via a boosted deep belief network
IEEE Conference on Computer Vision and Pattern Recognition, IEEE Computer Society (2014), pp. 1805-1812
View in ScopusGoogle Scholar
17. [17]
S. He, S. Wang, W. Lan, _et al._
Facial expression recognition using deep Boltzmann Machine from thermal infrared images
Affective Computing and Intelligent Interaction, IEEE (2013), pp. 239-244
CrossrefView in ScopusGoogle Scholar
18. [18]
H. Jung, S. Lee, J. Yim, _et al._
Joint fine-tuning in deep neural networks for facial expression recognition
Proceedings of the IEEE International Conference on Computer Vision (2015), pp. 2983-2991
View in ScopusGoogle Scholar
19. [19]
K. Simonyan, A. Zisserman
Two-stream convolutional networks for action recognition in videos
Advances in neural information processing systems (2014), pp. 568-576
Google Scholar
20. [20]
A. Karpathy, G. Toderici, S. Shetty, _et al._
Large-scale video classification with convolutional neural networks
Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (2014), pp. 1725-1732
View in ScopusGoogle Scholar
21. [21]
M. Ranzato, J. Susskind, V. Mnih, _et al._
On deep generative models with applications to recognition
The IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2011,

Colorado Springs, Co, Usa, IEEE Xplore (2011), pp. 2857-2864
20-25 June
CrossrefView in ScopusGoogle Scholar
22. [22]
P. Khorrami, T. Paine, T. Huang
Do deep neural networks learn facial action units when doing expression recognition?
Proceedings of the IEEE International Conference on Computer Vision Workshops (2015), pp. 19-27
View in ScopusGoogle Scholar
23. [23]
Burkert P, Trier F, Afzal M Z, et al. Dexpression: deep convolutional neural network for expression recognition. arXiv preprint 2015, arXiv:1509.05371.
Google Scholar
24. [24]
M. Liu, S. Li, S. Shan, _et al._
AU-inspired Deep networks for facial expression feature learning
Neurocomputing, 159 (C) (2015), pp. 126-136
View PDFView articleView in ScopusGoogle Scholar
25. [25]
A.T. Lopes, E. de Aguiar, A.F. De Souza, _et al._
Facial expression recognition with convolutional neural networks: coping with few data and the training sample order
Pattern Recognit., 61 (2017), pp. 610-628
View PDFView articleView in ScopusGoogle Scholar
26. [26]
D. Hamester, P. Barros, S. Wermter
Face expression recognition with a 2-channel convolutional neural network
Neural Networks (IJCNN), 2015 International Joint Conference on, IEEE (2015), pp. 1-8
View in ScopusGoogle Scholar
27. [27]
H. Jung, S. Lee, J. Yim, _et al._
Joint fine-tuning in deep neural networks for facial expression recognition
IEEE International Conference on Computer Vision, IEEE (2015), pp. 2983-2991
View in ScopusGoogle Scholar
28. [28]
Y.H. Byeon, K.C. Kwak
Facial expression recognition using 3d convolutional neural network
Int. J. Adv. Comput. Sci. Appl., 5 (12) (2014)
Google Scholar
29. [29]
T. Brox, A. Bruhn, N. Papenberg, _et al._
High accuracy optical flow estimation based on a theory for warping
European conference on computer vision, Berlin Heidelberg, Springer (2004), pp. 25-36
CrossrefView in ScopusGoogle Scholar
30. [30]
P. Lucey, F. Cohn J, T. Kanade, _et al._
The extended cohn-kanade dataset (ck+): a complete dataset for action unit and emotion-specified expression
Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on, IEEE (2010), pp. 94-101
View in ScopusGoogle Scholar

31. [31]
Oliver Langner, Ron Dotsch, Gijsbert Bijlstra, _et al._
Presentation and validation of the radboud faces database
Cognition Emotion, 24 (8) (2010), pp. 1377-1388
CrossrefView in ScopusGoogle Scholar
32. [32]
M. Pantic, M. Valstar, R. Rademaker, _et al._
Web-based database for facial expression analysis
IEEE International Conference on Multimedia and Expo, IEEE Xplore (2005), p. 5
Google Scholar
33. [33]
A. Majumder, L. Behera, K. Subramanian V
Automatic facial expression recognition system using deep network-based data fusion
IEEE Trans. Cybern., 99 (2016), pp. 1-12
Google Scholar
34. [34]
Y. Zhou, B. Shi
Action unit selective feature maps in deep networks for facial expression recognition
Submitted IJCNN (2017)
Google Scholar
35. [35]
G. Ali, A. Iqbal M, S. Choi T
Boosted NNE collections for multicultural facial expression recognition
Pattern Recognit., 55 (2016) (2016), pp. 14-27
View PDFView articleView in ScopusGoogle Scholar
36. [36]
M. Liu, S. Li, S. Shan, _et al._
Au-aware deep networks for facial expression recognition
Automatic face and gesture recognition (FG), 2013 10th IEEE International Conference and Workshops on, IEEE (2013), pp. 1-6
View in ScopusGoogle Scholar
Cited by (92)

* ### Emotion recognition and artificial intelligence: A systematic review (2014?2023) and research recommendations
2024, Information Fusion
Show abstract
Emotion recognition is the ability to precisely infer human emotions from numerous sources and modalities using questionnaires, physical signals, and physiological signals. Recently, emotion recognition has gained attention because of its diverse application areas, like affective computing, healthcare, human?robot interactions, and market research. This paper provides a comprehensive and systematic review of emotion recognition techniques of the current decade. The paper includes emotion recognition using physical and physiological signals. Physical signals involve speech and facial expression, while physiological signals include electroencephalogram, electrocardiogram, galvanic skin response, and eye tracking. The paper provides an introduction to various emotion models, stimuli used for emotion elicitation, and the background of existing automated emotion recognition systems. This paper covers comprehensive searching and scanning of well-known datasets followed by design criteria for review. After a thorough analysis and discussion, we selected 142 journal articles using PRISMA guidelines. The review provides a detailed analysis of existing studies and available datasets of emotion

recognition. Our review analysis also presented potential challenges in the existing literature and directions for future research.

* ### A systematic review on affective computing: emotion models, databases, and recent advances
2022, Information Fusion

Citation Excerpt :

Note that since facial images or videos suffer from a varied range of backgrounds, illuminations, and head poses, it is essential to employ pre-processing techniques (e.g., face alignment [225], face normalization [226], and pose normalization [227]) to align and normalize semantic information of face region. In this sub-section, we distinguish FER methods (shown in Fig. 4) via the point of whether the features are hand-crafted features based ML models [228] or high-level features based on DL-based models [229]. Table 9 provides an overview of representative FER methods.

Show abstract

Affective computing conjoins the research topics of emotion recognition and sentiment analysis, and can be realized with unimodal or multimodal data, consisting primarily of physical information (e.g., text, audio, and visual) and physiological signals (e.g., EEG and ECG). Physical-based affect recognition caters to more researchers due to the availability of multiple public databases, but it is challenging to reveal one's inner emotion hidden purposefully from facial expressions, audio tones, body gestures, etc. Physiological signals can generate more precise and reliable emotional results; yet, the difficulty in acquiring these signals hinders their practical application. Besides, by fusing physical information and physiological signals, useful features of emotional states can be obtained to enhance the performance of affective computing models. While existing reviews focus on one specific aspect of affective computing, we provide a systematical survey of important components: emotion models, databases, and recent advances. Firstly, we introduce two typical emotion models followed by five kinds of commonly used databases for affective computing. Next, we survey and taxonomize state-of-the-art unimodal affect recognition and multimodal affective analysis in terms of their detailed architectures and performances. Finally, we discuss some critical aspects of affective computing and its applications and conclude this review by pointing out some of the most promising future directions, such as the establishment of benchmark database and fusion strategies. The overarching goal of this systematic review is to help academic and industrial researchers understand the recent advances as well as new developments in this fast-paced, high-impact domain.

* ### Facial expression recognition based on a multi-task global-local network
2020, Pattern Recognition Letters

Show abstract

Facial expression recognition plays an important role in intelligent human-computer interaction. The clues for understanding facial expressions lie not in global facial appearance, but also in local informative dynamics among different but confusing expressions. In this paper, we design a multi-task learning framework for global-local representation of facial expressions. First, a shared shallow module is designed to learn information from local regions and the global image. Then we construct a part-based module, which processes critical local regions including the eyes, the nose, and the mouth to extract local informative dynamics related to facial expressions. A global face module is proposed to extract global appearance features related to expressions. The proposed network extracts both local-global and spatio-temporal information for a discriminative and robust representation of facial expressions. Through properly fusing these modules into a system, we have

achieved competitive results on the CK+ and Oulu-CASIA databases.
 * ### Deep Facial Expression Recognition: A Survey
2022, IEEE Transactions on Affective Computing
 * ### FER-net: facial expression recognition using deep neural net
2021, Neural Computing and Applications
 * ### Fine-Grained Facial Expression Recognition in the Wild
2021, IEEE Transactions on Information Forensics and Security
View all citing articles on Scopus
View Abstract

Part of special issue
Deep Learning for Pattern Recognition
Edited by
Zhaoxiang Zhang, Shiguang Shan, Yi Fang, Ling Shao
Download full issue
Other articles from this issue
 * ### Deep learning for sensor-based activity recognition: A survey
1 March 2019
Jindong Wang, ?, Lisha Hu
View PDF
 * ### Boosting deep attribute learning via support vector regression for fast moving crowd counting
1 March 2019
Xinlei Wei, ?, Lingfei Ye
View PDF
 * ### Learning domain-invariant feature for robust depth-image-based 3D shape retrieval
1 March 2019
Jing Zhu, ?, Yi Fang
View PDF
View more articles
Recommended articles
No articles found.
Article Metrics
Citations
 * Citation Indexes: 92
Captures
 * Readers: 81
View details

You may choose not to allow some types of cookies. However, blocking some types may impact your experience of our site and the services we are able to offer. See the different category headings below to find out more or change your settings.
Allow all
Manage Consent Preferences
Strictly Necessary Cookies
Always active
These cookies are necessary for the website to function and cannot be switched off in our systems. They are usually only set in response to actions made by you which amount to a request for services, such as setting your privacy preferences, logging in or filling in forms. You can set your browser to block or alert you about these cookies, but some parts of the site will not then work. These cookies do not store any personally identifiable information.
Cookie Details List?
Functional Cookies
Functional Cookies
These cookies enable the website to provide enhanced functionality and personalisation. They may be set by us or by third party providers whose services we have added to our pages. If you do not allow these cookies then some or all of these services may not function properly.
Cookie Details List?
Performance Cookies
Performance Cookies
These cookies allow us to count visits and traffic sources so we can measure and improve the performance of our site. They help us to know which pages are the most and least popular and see how visitors move around the site.
Cookie Details List?
Targeting Cookies
Targeting Cookies
These cookies may be set through our site by our advertising partners. They may be used by those companies to build a profile of your interests and show you relevant adverts on other sites. If you do not allow these cookies, you will experience less targeted advertising.
Cookie Details List?
Back Button
Cookie List
Search Icon
Filter Icon
Clear
checkbox label label
Apply Cancel
Consent Leg.Interest
checkbox label label
checkbox label label
checkbox label label
Confirm my choices