Title: Multimodal learning for facial expression recognition

Outline

Figures (9)

Tables (8)

Pattern Recognition
Volume 48, Issue 10, October 2015, Pages 3191-3202
Multimodal learning for facial expression recognition
Author links open overlay panelWei Zhang a, Youmei Zhang a, Lin Ma b, Jingwei Guan a, Shijie Gong a
Show more

https://doi.org/10.1016/j.patcog.2015.04.012Get rights and content
Highlights
 * ?
Multimodal learning for facial expression recognition (FER) is proposed.
 * ?
The first attempt to do FER from the joint representation of texture and landmarks.

 * ?
The multimodal structure combines feature extraction and classification together.
 * ?
Structured regularization is used to enforce the sparsity of different modalities.

Abstract

In this paper, multimodal learning for facial expression recognition (FER) is proposed. The multimodal learning method makes the first attempt to learn the joint representation by considering the texture and landmark modality of facial images, which are complementary with each other. In order to learn the representation of each modality and the correlation and interaction between different modalities, the structured regularization (SR) is employed to enforce and learn the modality-specific sparsity and density of each modality, respectively. By introducing SR, the comprehensiveness of the facial expression is fully taken into consideration, which can not only handle the subtle expression but also perform robustly to different input of facial images. With the proposed multimodal learning network, the joint representation learning from multimodal inputs will be more suitable for FER. Experimental results on the CK+ and NVIE databases demonstrate the superiority of our proposed method.

 * Previous article in issue
 * Next article in issue

Keywords
Multimodal learning
Facial expression recognition
Texture
Landmark

1\. Introduction

Facial expression presents a rich source of affective information and thus is one of the most direct ways for us to understand the psychological state of a person. Automatic facial expression recognition (FER) is an important and challenging problem in the communities of computer vision and pattern recognition, which attracts much attention recently due to its potential applications in many areas such as human?machine interfaces [21], robotics [22], driver safety [23], communication and health-care [24].

There exist a number of FER approaches in the past years. Generally speaking, the current methods can be broadly classified into two categories based on the availability of the data for recognition. The first category can be regarded as texture-based methods [4], [5], [6], [7]. Texture modality for FER represents the facial image information, which displays face expression in pixel space. As such, texture-related features are extracted from the pixel value, which is capable of capturing detailed and subtle information of facial expression. On the other hand, the features are very sensitive to the image changes, such as luminance and masking effects. Furthermore, the texture-related features correlate very closely to each individual for FER. The other category is the landmark-based methods [2], [25], [31]. Landmark indicates face key points, the corresponding movements of which can help capture the facial expression. However, the landmark movements cannot efficiently capture the subtle changes, which may not be able to distinguish the expressions with similar landmark information.

If the texture modality (facial image) is available, facial features are extracted from the images which are further fed into classifiers for recognition. The method in [5] firstly convolves the video clip with Gabor motion energy (GME) filter in a filter bank for feature extraction. In order to make the problem close to reality, the first six frames are employed for feature extraction. Afterwards, support vector machine (SVM) is employed to train the features for expression recognition. Similarly, SVM is also employed in [4] for FER. Prior to being fed into SVM, non-negative matrix factorization (NMF) [14] is performed by minimizing a cost function. Firstly, local patches are extracted from each facial image, based on which the NMF is performed to reconstruct a sparse and part-based representation of the patches. Then SVM comes in handy to perform classification. Moreover, Yang et al. [6] proposed to represent the dynamics of facial expression for recognition. Haar-like features are employed for the sake of simplicity and effectiveness. The K-means clustering method is employed to generate the temporal pattern models of the expressions, and the Adaboost learning is employed as the classifier for FER.

If the landmark modality (face key point) is available, features can be extracted from the landmarks for FER. Similar to texture-based methods, most landmark-based methods extract handcrafted features from input landmark before performing recognition. In [25], Perveen et al. proposed to search the bounding boxes which help compute facial characteristic points (FCP). The facial animation parameters, such as the openness of eyes, width of eyes and height of eyebrows, are then evaluated via referring to the FCPs. With these animation parameters, the expression can be further recognized by employing the Gini Index [28]. More recently, Lorincz et al. [2] did a pioneering work on extracting features from the landmark in 3D space for FER. Only the landmark information is incorporated in 3D constrained local model (CLM). Such process makes the proposed FER robust against head pose variations. Additionally, they use either dynamic time warping (DTW) or global alignment (GA) kernel algorithm to deal with multi-frames considering the spatio-temporal attribute of facial expression, and the landmark is tracked by using 3D CLM. Afterwards, the Euclidean distance is calculated to build matrix, where the nearest correlation matrix is found with kernel, and the gram matrix by DTW kernel or global alignment kernel is further employed for SVM training. Finally, in order to minimize the classification error, the best parameters are searched for both kernels. With such processes, state-of-the-arts FER performance was obtained. He et al. [31] conducted spontaneous facial expression recognition based on landmarks. First, they normalized the sequences according to the pupil?s coordinates. Afterwards, they labeled landmarks on the onset and apex images manually and tracked landmarks on the whole sequences. The features depicting the point distance variation are extracted and the hidden Markov model (HMM) is employed to recognize the facial expression.

Although tremendous progresses of FER have been made in the past few decades, the problem remains with great challenges. Mostly, all previous work treats the texture or landmark modality independently, where only the texture or landmark modality is employed for FER. It has been demonstrated that each single modality is useful for FER. However, one single modality alone cannot help obtain the details of facial expression variation while avoiding the extraneous affections. The texture modality captures the detailed changes of the face information, which will be helpful for recognizing the subtle facial expression. However, external variations, such as the lightning condition and masking effect, will significantly affect the texture features, which will make the textural-based FER very sensitive. On the contrary, the landmark modality presents more robust property to the external affections. However, the landmark modality just simply outlines the shapes and contours of the face

which is lack of sufficient detailed information. In this case, the landmark modality cannot accurately distinguish the subtle facial expression, specifically for the two expressions with similar landmark information.
Texture and landmark modalities seem to be complementary to each other. Therefore, how to integrate the two modalities to improve the performance of the FER system remains an open question. The two modalities are of great difference, where the texture modality mostly describes the facial detailed expression, specifically the facial image content, and the landmark modality describes the positions of face key points.
Nowadays, some algorithms were proposed to address the representation learning for multiple modalities. In [8], [9], [10], [11], multimodal deep belief network (DBN) [1] is developed for learning the joint representations from the input multiple modalities. In [8], the video and audio inputs were employed to learn a bimodal DBN. In order to further discover the correlations among the two modalities, both modalities are presented during feature learning but only a single modality is used for supervised training, which means that the deep autoencoder is trained to reconstruct both modalities when given only one modality (video or audio input). In [10], the multimodal DBN is trained to learn the joint representation of the multimodal data, specifically the text and image modalities. Firstly, two DBNs are trained for image and text respectively. To form a multimodal DBN, the two trained DBNs are combined by learning a joint RBM on top of them. In [11], Deep Boltzmann Machine (DBM) is employed to train each modality. In order to form a multimodal DBM, the two trained DBMs are combined by adding an additional layer of binary hidden units on top of them. From the work in [10], [11], it is possible for the model to find representations such that some hidden units are tuned only for one modality while others are tuned only for the other modalities [8].
Besides, there exists another defect with previous methods on FER. As aforementioned, the previous FER methods can be regarded as a type of two-step methods. Firstly, handcrafted features from texture or landmark modality are extracted, which are expected to represent the expression. Subsequently, the classifiers, such as SVM or Adaboost, or employed for training on the extracted features for FER. Therefore, in such cases, features are the key components of the whole FER system. If the features can accurately depict the expression and are of great discriminations to different expressions, the classifier can recognize the expressions well. However, all the features are tuned by hand and thus can hardly ensure the classifier to distinguish the expression well. Therefore, it would be better to have feature extraction and classification assembled together to be globally optimized for FER.
In this paper, we make the first attempt to employ different modalities and assemble the feature representation and classification together for FER. Specifically, the facial texture and landmark modalities are combined together to benefit from the inherent properties of the two different modalities. A joint representation for FER is learned from the texture and landmark modalities. In order to ensure that the two modalities interact with each other for the joint representation, a structured regularization method is employed for each modality to control the connection tightness of representations. FER is then performed based on the learnt representation.
With such multimodal learning process, the proposed FER method can not only ensure the robustness of the system to time resolution of the expressions but also make the method robust against head pose variations. Additionally, the multimodal learning combines feature extraction and classification together and thus avoids the cumbersome task of features? handcrafting.
The rest of this paper is organized as follows. In Section 2, our proposed

multimodal learning method is introduced. Experimental results are given and discussed in Section 3. Finally, Section 4 concludes the paper.
2\. Multimodal FER by integrating texture and landmark
The proposed multimodal FER is introduced to jointly learn the representations from multimodal inputs, specifically the texture and landmark modalities. The texture modality is a collection of local image patches cropped from the positions indicated by the face key points, while the landmark modality depicts movements of facial key points in the face expression sequence. The data processing details are given in Section 3.2.
2.1. Multimodal learning architecture
The proposed multimodal learning architecture is illustrated in Fig. 1, which takes different numbers and types of modalities as inputs and outputs the final classification results. The proposed multimodal learning architecture not only considers each modality property but also accounts for the interactions of different modalities. The proposed multimodal learning architecture is built by stacking several layers together and feeding the hidden representation of the _k_ th layer as the input into the (k+1)th layer.
 1. Download: Download high-res image (222KB)
 2. Download: Download full-size image
Fig. 1. Multimodal learning architecture for FER.
The multimodal learning architecture in Fig. 1 can be formulated as(1)L^=?(fk(f(k?1)?f2(f1SR(x1,x2,?,xm))))Eq. (1) represents the global function of the proposed multimodal learning method. L^ is the output class label of the multimodal learning network. f1SR(·) is the function that firstly maps the visual input layer to the first hidden layer. As our method targets at a multimodal learning network, f1SR(·) is an auto-encoder (AE) with the structured regularization (SR), which enforces the modality-specific sparsity and density of each modality. As illustrated in Fig. 2(a), AE is a simple learning circuit aiming to transform inputs into outputs with the least possible amount of distortion, where _z_ _i_ is the reconstructed signal of _x_ _i_. It can be observed that AE treats each input node of different modalities equally, where the contributions of different modalities to the hidden nodes cannot be well learned. However, different modalities may contribute differently to the specific classification task, as demonstrated in Section 3. To overcome this limitation and fully exploit the contributions of different modalities, AE with SR is employed, which allows the network to distinguish different modalities for individual treatments. Fig. 2(b) illustrates the structure of AE with SR, where the connections between the visual input nodes and hidden nodes as well as the weights are learnt in a data-driven manner, which can distinguish and learn the representation from different multimodal inputs for the final classification task.
 1. Download: Download high-res image (369KB)
 2. Download: Download full-size image
Fig. 2. The structure of AE without SR (a) and with SR (b).
After the AE with SR mapping process, different modalities have been transformed to the first hidden layer, each node of which takes different modalities into consideration. AEs, f2,?,fk, are thus employed to map the feature to the final representation for the classification. By stacking several AEs, the non-linear properties are fully exploited to generate the final joint representation of the multimodal inputs (x1,x2,?,xm). Afterwards, ?(·) denoting the classifier, such as SVM, KNN, and softmax, takes the joint representation as the input to perform the final classification tasks.
The training process can be performed greedily layer by layer. This stacking architecture ensures the scalability of the learning ability. On one hand,

more layers can help improve the nonlinearity representation ability of the neural network. On the other hand, more layers will inevitably introduce more parameters, especially for the top fully connected layers. Intuitively, more parameters demand more training data to build a robust deep network and avoid the over-fitting problem. Therefore, the depth of the proposed network should be adaptively determined by the specific problem and the number of the training samples at hand.

2.1.1. Autoencoder (AE)

Each layer constituting the multi-layer learning architecture is an autoencoder (AE) shown in Fig. 2, which consists of two components, the encoder and decoder. An encoder e(·) encodes the input x?Rd to some hidden representation e(x)?Rdh, while a decoder d(·) decodes the obtained hidden representation back to a reconstructed version of _x_ , to make the reconstructed signal to be as close as possible to the input. Therefore, the encoder process can be viewed as a single mapping function f:Rd?Rdh:(2)yi=f(xi)=?(Wxi+b),where _y_ _i_ represents the encoder output and _x_ _i_ represents the input of the encoder. W?Rd×dh and _b_ are the mapping weight and encoder bias, respectively. _?_ denotes a non-linear function, which can employ sigmoid, tanh, and rectified linear unit (ReLU) function. With this non-linear mapping process, AE can present strong feature learning capabilities [12].

In order to obtain the encoder parameters, the following optimization problem needs to be solved by minimizing the reconstruction error introduced from AE:(3)minW,b,c(l(x,W,b,c)),where _c_ is the decoder bias, l(x,W,b,c) denotes the loss function to capture the reconstruction error. There are some alternatives to define the loss functions, such as the squared error or Kullback?Leibler divergence (KLD) while the feature values lie in [0,1]. Taking the squared error as the reconstruction error, l(x,W,b,c), Eq. (3) can be further represented as(4)l(x,W,b,c)=12n?i=1n?zi?xi?22,(5)yi=?(Wxi+b),(6)?i=yiyi?yi,(7)zi=W??i+c.where _y_ _i_ is the obtained hidden representations through the feedforward encoder, Eq. (7) represents the decoder with the bias _c_ , and _z_ _i_ is the reconstructed signal through performing a round of feedforward encoder and backward decoder. In order to reduce the effect of filter scale, the _L_ 2-normalization is normally performed on all hidden nodes of the encoder level as expressed in Eq. (6).

As aforementioned, if training each modality separately and learning a joint representation (e.g. RBM) on top of them, it is possible for the model to find representations such that some hidden units are tuned only for one modality while others are tuned only for the other modalities [8]. Similarly, if we simply employ AE in Eq. (2) to map the multimodal inputs into the hidden nodes, the network is to connect all nodes of visible layer to nodes of the hidden layer, which means that all the different modality features are treated equally. Ignoring the specific properties of different modalities, AE will be trained to the form that some hidden nodes are strongly connected with some individual modality inputs while weakly connected to other modalities. As such, the correlations between different modalities cannot be well learned and represented. Therefore, to overcome this limitation, we employed the structured regularization (SR) [17], [18], which allows the network to distinguish different modalities for individual treatment. Also the modality-specific sparsity and modality-specific density of the features from different modalities are enforced and further learned. SR is employed in the layer with multimodal inputs of Fig. 1 to distinguish and learn the representation from different multimodal inputs.

2.1.2. Structured regularization (SR)
As aforementioned, the SR function is employed for AE with multimodal inputs inspired by [17], [18]. Suppose Sr,i as an K×N modality binary matrix, where _K_ denotes the numbers of modalities and _N_ indicates the number of units in corresponding modality. For SR, each modality will be used as a regularization group separately for each hidden unit, applied in a manner similar to the group regularization, compared with the traditional regularization that treats each input unit equally and ignores the relationship and correlation between different modalities. SR is defined as(8)SR(W[1])=?j=1M?k=1K(?i=1NSr,i|(Wi,j[1])P|)1/pwhere _M_ denotes the total number of hidden units. _K_ is the total number of the modalities. _N_ indicates the total number of input units in each modality. The regularization can be viewed as the summation of the corresponding Minkowski distance. For p?1, the Minkowski distance is a metric as a result of the Minkowski inequality. When p<1, the Minkowski distance violates the triangle inequality. In the limiting case of _p_ reaching infinity, the regularization will be changed to the summation of Chebyshev distance:(9)SR(W[1])=?j=1M?k=1K(maxi(Sk,i|Wi,j[1]|))which only penalizes the maximum weight from each input unit to each hidden unit. In order to prevent over-constraining, the regularization function is modified to penalize nonzero weight maxima for each modality for each hidden unit without additional penalty for larger values of these maxima. The regularization function in Eq. (9) are further modified as(10)SR(W[1])=?j=1M?k=1KB((maxi(Sk,i|Wi,j[1]|))>0)where _B_ indicates a Boolean function that takes a value of 1 if its variable is true, and 0 otherwise. The regularization function in Eq. (10) performs a direct penalty on the number of modalities used for each weight, without further constraining the weights of modes with nonzero maxima.
By integrating SR into the multimodal AE training as in [17], the objective function can be further represented as(11)?W[1]?=argminW[1]?i=1n[1]?zi[1]?xi[1]?22+?·SR(W[1])where(12)zi[1]=?j=1k[1]?j[1]Wi,j[1]where ?j[1] is the hidden node generated by the encoder of the multimodal AE, while zi[1] is signal reconstructed by the decoder from ?j[1]·n[1] is the number of the input nodes including all the modality features, and k[1] is the number of the hidden nodes of the multimodal AE. Wi,j[1] is the corresponding weights of the multimodal AE by introducing SR. _?_ is the parameter to balance the error and the regularization terms, which is experimentally set to 3× _e_ ?4 in practice.
Fig. 2(b) illustrates the structure of AE with SR to demonstrate how SR with AE works for the multimodal inputs. By integrating SR into AE, the connection between the visual input layer and the first hidden layer is learned. As Eq. (10) shows, to minimize SR(W[1]), the zero number of Wi,j[1] should be as large as possible. As such, only some effective nodes of the visual input layer get connected with those of the first hidden layer. The comparison of the structures of AE with and without SR demonstrates that the multimodal network could distinguish different modalities and learn the correlations between them automatically.
2.2. Multimodal FER
Based on the learning architecture in previous section, we propose a simple network for FER. A softmax layer is added on top of the multimodal learning architecture, which takes the learned joint representation as inputs and outputs the classification results for each facial expression. For each expression, the softmax layer will determine whether the given inputs, including the texture and landmark modalities, will result in the specific

expression or not. Consequently, the number of output nodes in the classification layer (top layer) is two. We can employ the introduced network including SR in the multimodal AE to build the network. The depth of the network depends on the problem and the number of training samples. As aforementioned, insufficient training samples will incur overfitting with high probabilities, for the specific FER case, due to the constraint of the training sample number, only one hidden layer is employed, of which the network structure is I?H?C. Fig. 3 shows the structure of our network succinctly. Take the experiment conducted on the CK+ database as an example, _C_ is defined as two to distinguish whether the inputs is the latent facial expression we aim to recognize. _I_ is defined as the size of the data from all the multimodal inputs, which is set as 4040 in this paper. _H_ denotes the size of the hidden layer nodes. As facial expressions affect the eyes and mouth significantly in each frame, the patches covering eyes and mouth are extracted and resized into 16×16 and 16×10, respectively. Afterwards, these corresponding patches will be concatenated as individual vectors, respectively. Supposing that _F_ is the number of frames imported to the network, the size of eye and mouth modalities become 256× _F_ and 160× _F_ by temporally concatenating the modality vector from each frame. Furthermore, the displacements of the landmarks provide more persuasive representation than static coordinates. Besides, we suppose that features in different directions contribute differently to FER. As a result, the displacements of the landmarks in _X_ and _Y_ directions are separated as different modalities. As there are 68 marked face key points in every frame, temporally concatenating the landmark displacement results in a vector with size of 68× _F_ for each direction. By considering the texture and landmark modalities together, the vector size is 4040, with _F_ equals to 5, which is fed into the network for further training. The output of each hidden nodes is generated by a sigmoid function ?(a)=1/(1+exp(?a)) of the weighted input:(13)hj[1]=?(?i=1IxiWi,j[1])(14)p(o|x;?)=?(?i=1Hhi[1]Wi[2])where _o_ denotes the output nodes which indicate whether the expression exists or not, and _?_ indicates all the parameters in the network, specifically W[1] and W[2].

1. Download: Download high-res image (221KB)
2. Download: Download full-size image

Fig. 3. The structure of the network.

The object of the learning network is to realize the non-linear mapping function for the FER. The inference can be realized by the following function:(15)o^=argmaxop(o|x;?)As mentioned before, each facial expression will be treated separately, for each of which we construct a network for the classification. Consequently, Eq. (15) will help distinguish whether the multimodal input are the facial expression that we aim to recognize.

In order to make the inference, we need to obtain the parameters of the constructed network, specifically the parameters of the two layers, respectively. For the parameters W[1] in the multimodal layer, AE is first pretrained to obtain the initialized parameters. Specifically, the parameters of multimodal layer are firstly pretrained, which simply learns features from unlabeled data automatically aiming to transform inputs into outputs with the least possible amount of distortion. With the process of pre-training, the constructed network can effectively avoid the risk of trapping in poor local optima. After the pre-training process, the fine-tuning process needs to be further performed to make the network more suitable for FER. Thereby, a log-likelihood function is employed as the object function for further training the parameters W[2] in the softmax layers and fine-tuning the parameters W[1] in the multimodal layer:(16)???=argmax??t=12logP(L^=L|x;?)??SR(W[1])where _L_ represents the label of the inputs and L^ represents the outputs of the network. For the parameter training, traditional back-prorogation (BP) [26] is employed to fine-tune parameters of the constructed deep network. This algorithm is first proposed by Rumelhart and McCelland, the essence of which is to minimize the mean squared error between actual output and desired output based on gradient descent. BP algorithm is especially powerful because it can extract regular knowledge from input data and memory on the weights in the network automatically [17]. Simultaneously, it can improve generalization performance of the learning system, which is fabulous when used in FER.

Algorithm 1

Multimodal learning for facial expression recognition.

TRAINING :

Input: {Xtrain,Ltrain}

Output: ?,L^train

1:| Initialize W[1] and W[2] randomly;

2:| Pretraining: W[1] is pretrained based on ?W[1]?=argminW[1]?i=1n[1]?zi[1]?xi[1]|22+?·SR(W[1]) to learn the connections between the visual input layer and the contributions of different modalities to the hidden nodes on the benefit of SR;

3:| Finetuning: _?_ is updated according to ???=argmax??t=12logP(L^=L|x;?)??SR(W[1]) to strengthen the recognition capability of the network;

4:| Record _?_ for the multimodal learning network.

TESTING :

Input: _?_ , _X_ _test_

Output: L^test

1:| Generate the output class labels L^test of _X_ _test_ based on _?_ according to Eqs. (13), (14);

2:| Output the labels of facial expressions L^test.

Furthermore, in order to prevent over-fitting in training neural network, drop-out is introduced. Typically the outputs of neurons are set to zero with a probability of _p_ in the training stage and multiplied with 1?p in the test stage. By randomly masking out the neurons, dropout is an efficient approximation of training many different networks with shared weights. In our experiments, we applied the dropout to all the layers and the probability is set as _p_ =0.2.

We summarize our proposed multimodal learning for FER as in Algorithm 1. _X_ _train_ is the training sample which contains both textures and landmarks of facial expression. And _L_ _train_ denotes its corresponding labels. Based on the training samples, the parameters _?_ of the multimodal learning network, specifically W[1] and W[2] are trained and learned. For testing, when imported the testing sample _X_ _test_ to the trained network, the output class label L^test is generated based on the learned parameters _?_.

3\. Experimental results

In order to evaluate the effectiveness of the proposed method, Cohn?Kanade Extended Dataset (CK+) [15] and the natural visible and infrared facial expression (NVIE) database [30] are employed for experimental results. Firstly, the detailed information of the database are introduced. Afterwards, we will present how to process the input data to obtain the multimodal inputs for the proposed multimodal FER, including training and testing. Finally, experimental results are provided to demonstrate the effectiveness of the proposed multimodal method, as well as the performance comparison of the

multimodal inputs and unimodal input.
3.1. Database
The Cohn?Kanade Extended Dataset (CK+) [15] is built by Kanade et al., which is developed for automated facial analysis and has been widely used for testing the performance of FER algorithms. In this dataset, the facial behaviors of 210 adults are recorded using two hardware synchronized Panasonic AG-7500 cameras and participants are 18?50 years of age. There are posed and non-posed expressions concurrently in the dataset. The facial expression dynamics of sequences in the CK+ dataset starts from neutral expression and ends on the apex of the expression. Since we need data with labels for training and testing, only posed expressions with explicit labels are selected. There are totally 123 subjects with 593 frontal image sequences in our input data, where 327 sequences are annotated with the emotion labels (1=anger, 2=contempt, 3=disgust, 4=fear, 5=happy, 6=sad and 7=surprise). Each frame in the sequence is digitized into either 640×490 or 640×480 pixel arrays with 8-bit gray scale or 24-bit color values, and 68 face key points are detected by AAM [24] for each frame, which are regarded as the facial landmark. In this paper, six emotions are selected for FER testing and the inventory of each expression used in this experiment is shown in Table 1. When imported to the recognition system, the samples of the certain expression are set as positive with the rest as negative. Obviously, the positive samples of expression ?Fear? and ?Sad? are less than others. As the luminance information is more important for FER, the color frame is converted into gray ones to only preserve the luminance components before further processing.
Table 1. The number of expressions.

Emotion	Anger	Disgust	Fear	Happy	Sad	Surprise
Number	45	59	25	69	28	83

The natural visible and infrared facial expression (NVIE) database is newly developed for expression analysis. This database includes two sub-databases, that are posed database consisting of apex images and spontaneous database containing images and landmarks from onset to apex images. As the posed database with only apex frame could not meet our requirement, we did not take the posed one into consideration. For the spontaneous database, the facial images were recorded by DZ-GX25M camera with resolution 704×480 under three different conditions: illumination from left, front and right. There are 105 subjects under front illumination, 111 subjects under left illumination and 112 subjects under right illumination, respectively. A total of 28 landmarks are located and tracked on each image. Different from the CK+ database, the labels of samples in the NVIE database are assigned values from 0 to 2 to every expression. The larger the value, the more likely the sample belongs to that expression.
3.2. Data processing
As aforementioned, the databases contain both texture and landmark modalities for each facial image. These two modalities reflect different properties of the facial expression, which should be considered together for FER. As introduced in [13], the preprocessing of the data is critical to learning process. In the following, the texture and landmark modalities of the facial image will be first processed, respectively, before being fed into the multimodal FER system, as illustrated in Fig. 4.
1. Download: Download high-res image (214KB)
2. Download: Download full-size image

Fig. 4. The structure of our approach.
3.2.1. Texture modality

The definition of facial expression is based on the action unit (AU), which is relevant to the brows, eyes, bridge of the nose and mouth. As a result, the image patches are extracted around eyes and mouth from one frame, where the patches around eyes should cover the brows as well as bridge of the nose. These extracted image patches contain the most pivotal facial features related to expressions. As Fig. 5 shows, the green points are landmarks on the face, while the red border are the trim lines. We clip the patches according to the landmarks on the bridge and the tip of the nose. In order to cover the whole subject, for the CK+ database, the size of the eye and mouth patches are defined as 100×100 and 160×100, respectively. After that, these patches are further downsampled by ten times for dimension reduction, which can further reduce the parameter number and the computational complexity for training and testing. Finally, the image patch is concatenated into row vector before further normalization. The resulting vector size of the eye and mouth is 16×16=256 and 16×10=160, respectively. For the NVIE database, we clip 40×40 eye-patch and 40×60 mouth-patch. Afterwards, the patches are further downsampled to 20×20 and 20×30. After concatenating them together, the final vector size is 256×2+160=672 for the CK+ database and 400×2+600=1400 for the NVIE database, which represents the input of the textural modality for one frame.

1. Download: Download high-res image (138KB)
2. Download: Download full-size image

Fig. 5. The schematic diagram of eye and mouth patch extraction. (For interpretation of the references to color in this figure, the reader is referred to the web version of this paper.)

3.2.2. Landmark modality

The generation of facial expression is a dynamic process. Therefore, for the landmark modality, movements of the landmarks between the current frame and the previous one in video flow provide more insightful representation of facial expression than static landmarks. Additionally, we are not sure whether the head positions in the images from different people remain unchanged or not. As a result, we calculate the different value between current frame and previous one as the movements of landmarks. Assuming that Xt+1i and Yt+1i are the _i_ th _X_ and _Y_ coordinates in the current frame, respectively, _X_ _i_ _t_ and _Y_ _i_ _t_ are the _i_ th _X_ and _Y_ coordinates in the previous frame, the landmark movements can be calculated as(17)?Xti=Xt+1i?Xti?Yti=Yt+1i?YtiNote that the first frame of the input sequence has no previous frame for reference, which only serves as reference and is excluded from the landmark modality for FER. After obtaining the movements from each frame, the movements are concatenated as the input of the landmark modality, which results in the size of the landmark input modality as 68×2=136 for the CK+ database and 28×2=56 for the NVIE database.

3.2.3. Modality-specific normalization

As the extracted vectors are from two different modalities, a normalization method under the incentive of [4], [27] is employed to make the network robust to illumination and contrast variations. In addition, the normalization process is vital to the network training. The procedure of normalization could be summarized as follows. Firstly, the mean value of the texture and landmark from one frame is obtained. Then the difference between real and the mean value is calculated to remove the individual difference for texture and landmark modalities. Finally, the standard deviation is divided to make the data to be normally distributed. Supposing _P_ _j_ as the _j_ th pixel value of stretched patch (texture modality) or the _j_ th coordinate of landmark modality in the row vector, _J_ as the number of pixels in one patch or the number of landmark modality in one frame, the normalized result P^j of the input data is obtained by(18)?=?j=1JPjJ?=?j=1J[Pj??]2P^j=Pj???+Cwhere _C_ is a constant avoiding the numerator be divided by zero. The normalization makes the network robust to illumination and contrast variation as demonstrated by [20]. Fig. 6 shows the intensity of input data before and after normalization in histogram. It can be observed that the values of the input data before normalization are mostly around 1. After pre-processing, the input data subjects to normal distribution approximately, which tends to be more suitable for network training [29].

1. Download: Download high-res image (139KB)
2. Download: Download full-size image

Fig. 6. Histogram of intensity of input data before and after normalization. (a) Before Normalization (b) After Normalization.

3.2.4. Temporal cascading

The importance of facial dynamics in FER has been established in many vision experiments [16], [19]. As stated in [5], facial dynamics is about motion among frames, rather than static patterns. Additionally, the inputs fed to the network are a row vector conventionally, which makes that the cascading the multi-frame in the same video together becomes an essential work. After integrating the texture and landmark modalities from different frames in the same sequence as one row, the multimodal input data for the network is prepared. The corresponding input data and labels can be obtained for further training. As long as the training is finished, the feedforward network with learned parameters can be employed to recognize facial expressions from texture and landmark modalities.

3.3. Multimodal learning FER results

The experimental settings are as follows. For each facial expression in the two databases, 2/3 of the whole samples are randomly selected to form the training set, with the rest as testing samples. The network was trained and tested for five times, with the average experimental results as the network?s performance. The receiver operating characteristic (ROC) curves are employed as the criterion to evaluate the performance, which is more general and reliable than recognition accuracy [6] for evaluating the FER system. The _X_ -coordinate of the ROC curve is FP/N, where _N_ represents the number of negative samples and _FP_ (false positive) as the number of samples incorrectly labeled as belonging to the positive class. Analogically, the _Y_ -coordinate is TP/P, where _P_ indicates the positive samples and _TP_ (true positive) presents the number of samples correctly labeled as belonging to _P_. To draw a complete ROC curve, the threshold value ranges from 0 to 1 with 0.01 as the step size to obtain the curve. As aforementioned, there are two units in the output layer. The value of only one unit is employed for ROC curve generation. If the value is larger than the threshold, the unit is set to 1, and 0 otherwise. The area under the ROC curve (AUC) is digital representation of the performance. Obviously, the larger the AUC, the better is the classifier.

3.3.1. Comparison to prior study on FER

In order to efficiently assess the performance of our algorithm, we compare it with existing state-of-the-arts FER algorithms. We first use the first six frames of every labeled sequence as inputs. As the first frame only serves as reference, only the texture and landmark modalities from the rest five frames are extracted for training and testing. The performance is compared with the recent work done by Lorincz et al. [2], which achieved start-of-the-art performance. The corresponding results are illustrated in Table 2. It is noteworthy here that the experimental results from [2] are employed for

performance comparison. Fig. 7 shows the ROC curves of six expressions.
Table 2. Comparison to prior study on FER (first six frames).

Method	Angary	Disgust	Fear	Happy	Sad	Surprise	Average
Wu [5]	0.829	0.677	0.667	0.877	0.784	0.879	0.786
Long [3]	0.774	0.711	0.692	0.894	0.848	0.891	0.802
Jeni [4]	0.817	0.908	0.774	0.938	0.865	0.886	0.865
DTW [2]	0.873	0.893	0.793	0.892	0.843	0.909	0.867
GA [2]	0.921	0.905	0.887	0.910	0.871	0.930	0.904
Proposed algorithm	0.948	0.929	0.890	0.916	0.903	0.930	0.919

1. Download: Download high-res image (434KB)
2. Download: Download full-size image

Fig. 7. ROC curves of six different emotions.

Obviously, the performance of algorithms in [3], [5] is inferior to other methods. Although the dynamic characteristics of facial expression have been considered and a spatiotemporal GME filter is employed, the texture modality is only adopted for FER in [5]. Long et al. [3] proved that learning spatiotemporal filters with ICA works better than spatiotemporal Gabor features. Yet the final result relies largely on the handcrafted features. Jeni et al. [4] and Lorincz et al. [2] yield satisfactory results. Jeni et al. [4] removed personal mean texture manually. Only selected portions of the face image are employed, where the overall change of the face is neglected. Conversely, Lorincz et al. [2] used only the landmarks and neglected the texture one, with some important details of face missing.

With the first six frames as inputs, our algorithm produced better results than the existing algorithms. Since texture describes the face details and landmark outlines the shapes and contours, the proposed method integrates them together to exploit the complementarity of them. However, through the observation of the CK+ database, we find that the expression process is incomplete in the first six frames. In order to improve the recognition performance, we also take six frames which contain the first, the middle four and the last frame of sequence as inputs. The experimental results displayed in Table 3 demonstrate that the substantial change of expression reduce the recognition difficulty and can generate better recognition results.

Table 3. Comparison to prior study on FER (First?Last).

Method	Anger	Disgust	Fear	Happy	Sad	Surprise	Average
Yang [6]	0.973	0.941	0.916	0.991	0.978	0.998	0.966
Long [3]	0.933	0.988	0.964	0.993	0.991	0.999	0.978
Jeni [4]	0.989	0.998	0.977	0.998	0.994	0.994	0.992
DTW [2]	0.991	0.994	0.987	0.999	0.995	0.996	0.994
GA [2]	0.986	0.993	0.986	1.000	0.984	0.997	0.991
Proposed algorithm	0.995	0.999	0.967	0.999	1.000	1.000	0.993

Furthermore, Fig. 7 indicates that the performance of recognition on emotions ?Fear? and ?Sad? is inferior than the others, because AUC under the ROC curves of these two expression is less. The reason may be attributed to the extreme lack of training samples of these two emotions referring to Table 1. Hence. the equilibrium, correctness and scale of the dataset are crucial for training a successful neural network.

3.3.2. Unimodality vs. multimodality

The core idea of this paper is to address the integration of the texture and landmark modalities for FER. Therefore, it is necessary to compare the performance with unimodality and multimodality, respectively. Table 4 shows the corresponding experimental results, where the texture and landmark modalities are extracted from the integral multimodality dataset. The average performance of recognition results prove that multimodality is more reliable than unimodality for FER. It can be observed that the texture modality alone as the input data performs worse than the landmark modality. This is probably because the texture modality only covers portions of the face while landmarks can outline the shape and contour of the whole face. Moreover, the detailed change of face information that presented by texture and the global change of face represented by landmark can be viewed as complementary to each other. Consequently, when combined together, they yield the best FER results.

Table 4. Comparison of the algorithms with unimodality and multimodality.

Inputs	Anger	Disgust	Fear	Happy	Sad	Surprise	Average
Texture	0.770	0.790	0.584	0.921	0.577	0.877	0.753
Landmark	0.906	0.893	0.803	0.924	0.703	0.910	0.856
Multimodality	0.948	0.929	0.890	0.916	0.903	0.930	0.919

However, it seems that the recognition of ?Happy? is improved by integrating both the texture and landmark modalities. It is easy to recognize ?Happy? in this dataset. The texture and landmark modality alone already performs well. However, by integrating them together, the network will be much larger, which requires more training data. In this case, lack of training data can somewhat lead to overfitting, which results in the performance degradation.

Another issue is that the modality number may affect the FER performance. As shown in Table 5, FER is first performed with two modalities as inputs, where the left eye, right eye, and mouth are combined together as the texture modality, and the X-displacement and Y-displacement are combined together as landmark modality. The performance result is illustrated in the second row of Table 5. Furthermore, these components can be treated separately, which are regarded as five different modalities and fed into the recognition network. The corresponding results are illustrated in the third row of Table 5. It can be observed that treating these five modalities separately will help produce better results. Moreover, it can be concluded that our proposed multimodal learning network is scalable to different numbers of modalities. As such, by introducing more related modalities, the FER results will be improved further.

Table 5. Comparison of the algorithms with different numbers of modalities.

Inputs	Anger	Disgust	Fear	Happy	Sad	Surprise	Average
(Left-eye + Right-eye + mouth) + (X-displacement + Y-displacement)	0.923	0.890	0.746	0.923	0.870	0.927	0.879
Left-eye + Right-eye + mouth + X-displacement + Y-displacement	0.948	0.929	0.890	0.916	0.903	0.930	0.919

3.3.3. Comparison of the algorithms with and without pretraining

Table 6 displays the comparison of algorithms with and without pretraining. It is demonstrated that, once pretraining is added to the network, the performance is improved by six percent. BP is based on local gradient descent, and starts usually at some random initial points, which may cause poor local optima. If pretraining is employed to initialize the parameters, the network could be fine-tuned on the pretrained parameters. In this case, the parameters of the network will avoid the risk of getting stuck at local optima. Table 6 illustrates the FER results with and without pretraining, which demonstrate the necessity of pretraining.

Table 6. Comparison of the algorithms with and without AE.

Method	Anger	Disgust	Fear	Happy	Sad	Surprise	Average
BP	0.907	0.915	0.708	0.909	0.764	0.895	0.850

BP + pretraining	0.948	0.929	0.890	0.916	0.903	0.930	0.919

3.3.4. One hidden layer vs. multiple hidden layers

The number of the hidden layers _H_ and the number of the units _U_ in each layer are the hyper-parameters of the network, which is very important for the network performance. In order to find the best parameters _H_ and _U_ , we test various combinations of them on the facial expression ?Anger?. Fig. 8(a) reveals the regularity as follows. Once the number of hidden layer is set to more than one, 50 units tend to give the best performance. More units will result in more parameters to be learned. However, the training samples cannot afford a network with too many parameters, which map the inputs from visible layer to hidden layer. On the other hand, if the number of hidden units is too small, it is hard to represent the 4040 input nodes. Our intention is to build a deep network to learn the nonlinear property of the texture and landmark modalities for FER. However, the lack of training data cannot afford a deep architecture for FER. It can be observed that Fig. 8(b) shows the network with one hidden layer and 100 hidden units performed the best. As the number of hidden layer increased, the performance on FER will be degraded. Hence, as illustrated in Fig. 3, the settings of one hidden layer and 100 hidden units are adopted for FER in this paper.

1. Download: Download high-res image (464KB)
2. Download: Download full-size image

Fig. 8. The recognition result with respect to the number of (a) hidden layers, (b) hidden units.

3.3.5. The proposed method vs. other classifiers

As aforementioned, we integrate texture and landmark modalities together as the input and use multimodal learning method to perform the FER. To demonstrate that the proposed method indeed performs better than other algorithms, we compare the method to other two classifiers, specifically SVM and KNN. We first use the same row vector with 4040 units as the input to the classifiers. Experimental results demonstrate that SVM performs better than KNN. However, both results are not satisfactory. To further prove that AE with SR can not only integrate texture and landmark together but also automatically extract the meaningful features for FER, we import the learned feature of the hidden layer into SVM and KNN, respectively. As shown in Table 7, after performing the multimodal feature learning, the performances of both SVM and KNN are significantly improved. We can conclude that the multimodal learning can effectively learn the representation from the multimodal inputs.

Table 7. Comparison of the recognition results using proposed method and other classifiers.

Method	Anger	Disgust	Fear	Happy	Sad	Surprise	Average
KNN using the row vector	0.309	0.493	0.383	0.412	0.434	0.355	0.400
SVM using the row vector	0.892	0.843	0.496	0.836	0.756	0.887	0.785
KNN using the hidden layer units	0.863	0.786	0.636	0.842	0.787	0.865	0.797
SVM using the hidden layer units	0.901	0.876	0.696	0.879	0.832	0.910	0.849
Proposed algorithm	0.948	0.929	0.890	0.916	0.903	0.930	0.919

3.3.6. Experiments on spontaneous database

There are great differences between posed and spontaneous facial expression. The former is acted intentionally, while the latter is displayed unconsciously by subjects. The posed expressions are captured by asking subjects to perform different expressions in front of a camera, which are usually exaggerated. The spontaneous ones are more natural and different from the posed one both in appearance and timing. The recognition of spontaneous seems to have more profound theoretical and practical significances. However, its expression recognition is thus harder. In this experiment, ?Happy?, ?Fear? and ?Disgust? are selected as samples to conduct the three-class classification. Fig. 9 illustrates the comparison results of the proposed method and He?s method on the NVIE database. For ?Disgust?, ?Fear? and ?Happy?, the recognition accuracy is measured by the ratio of the correctly recognized specific expression over the total number of specific expression samples. For ?Total accuracy?, the recognition accuracy is calculated by the ratio of all correctly recognized samples over all the total number of the samples. It can be observed that for the comparison of the spontaneous FER, the proposed method performs better than He?s method. For ?Total accuracy?, 10% accuracy improvement is obtained by our proposed multimodal learning method.

1. Download: Download high-res image (155KB)
2. Download: Download full-size image

Fig. 9. Comparison with the method [31] on the NVIE database.

4\. Conclusion

In this paper, we presented a multimodal FER algorithm, where the texture and landmark are integrated together to boost the FER performance. In order to avoid handcrafted features, which are cumbersome and time-consuming, the joint representation for FER is learned from the built neural network. By incorporating SR into AE, the proposed network can not only distinguish each modality but also learn the correlation and interaction between the texture and landmark modalities, which are complementary to each other. Various experimental results and comparisons have demonstrated the superiority of the proposed method over the existing ones.

Conflict of interest

None declared.

Acknowledgment

This work was supported by the NSFC Grant nos. 61203253 and 61233014, Research Found of Outstanding Young Scientist Award of Shandong Province (BS2013DX023), Independent Innovation Foundation of Shandong University (IIFSDU) 2013TB004, and Program of Key Lab of ICSP MOE China 2013000002.

Special issue articlesRecommended articles

References

1. [1]
G.E. Hinton, R.R. Salakhutdinov
Reducing the dimensionality of data with neural networks
Science, 313 (2006), pp. 504-507
CrossrefGoogle Scholar
2. [2]
A. Lorincz, L.A. Jeni, Emotional expression classification using time-series kernels, in: IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 889?895.
Google Scholar
3. [3]
F. Long, T. Wu, J.R. Movellan, M.S. Bartlett, G. Littlewort
Learning spatiotemporal features by using independent component analysis with application to facial expression recognition
Neurocomputing, 93 (2012), pp. 126-132
View PDFView articleView in ScopusGoogle Scholar
4. [4]
L.A. Jeni, J.M. Girard, J.F. Cohn, F.D.L. Torre, Continuous AU intensity estimation using localized, sparse facial feature space, in: IEEE

International Conference and Workshops on Automatic Face and Gesture Recognition, 2013, pp. 1?7.
Google Scholar
5. [5]
T. Wu, M.S. Bartlett, J.R. Movellan, Facial expression recognition using Gabor motion energy filters, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2010, pp. 42?47.
Google Scholar
6. [6]
P. Yang, Q. Liu, D.N. Metaxas, Facial expression recognition using encoded dynamic features, in: IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp. 1?8.
Google Scholar
7. [7]
M. Liu, S. Shan, R. Wang, X. Chen, Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 1749?1756.
Google Scholar
8. [8]
J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, A.Y. Ng, Multimodal Deep Learning, in: International Conference on Machine Learning, 2011, pp. 689?696.
Google Scholar
9. [9]
H. Hu, B. Liu, B. Wang, M. Liu, X. Wang, Multimodal DBN for predicting high-quality answers in cQA portals, in: Association for Computational Linguistics, 2013.
Google Scholar
10. [10]
N. Srivastava, R. Salakhutdinov, Learning representations for multimodal data with deep belief nets, in: International Conference on Machine Learning, 2012.
Google Scholar
11. [11]
N. Srivastava, R. Salakhutdinov, Multimodal learning with deep boltzmann machines, in: Neural Information Processing Systems, 2013, pp. 2222?2230.
Google Scholar
12. [12]
M. Chen, Z. Xu, K. Weinberger, Marginalized denoising autoencoders for domain adaptation, in: International Conference on Machine Learning, 2012.
Google Scholar
13. [13]
A. Krizhevsky, G.E. Hinton, Learning Multiple Layers of Features From Tiny Images, Technical Report, University of Toronto, 2009.
Google Scholar
14. [14]
Y. Wanf, Y. Zhang
Non-negative matrix factorization: a comprehensive review
IEEE Trans. Knowl. Data Eng., 29 (2013), pp. 1336-1353
Google Scholar
15. [15]
T. Kanade, J.F. Cohn, Y. Tian, Comprehensive database for facial expression analysis, in: IEEE International Conference on Automatic Face and Gesture Recognition, 2000, pp. 46?53.
Google Scholar
16. [16]

G. Zhao, M. Pietickaine
Dynamic texture recognition using local binary patterns with an application to facial expression
IEEE Trans. Pattern Anal. Mach. Intell., 29 (2007), pp. 915-928
View in ScopusGoogle Scholar
17. [17]
I. Lenz, H. Lee, A. Saxena
Deep learning for detecting robtic grasps
Int. J. Robotics Res. (2014)
Google Scholar
18. [18]
A. Jalali, P. Ravikumar, S. Sanghavi, C. Ruan
A dirty model for multi-task learning
Neural Inf. Process. Syst. (2010), pp. 964-972
Google Scholar
19. [19]
M. Black, Y. Yacoob
Recognizing facial expressions in image sequences using local parameterized models of image motion
Int. J. Comput. Vis., 25 (1997), pp. 23-48
View in ScopusGoogle Scholar
20. [20]
Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, in: Proceedings of the IEEE, 1998, pp. 2278?2324.
Google Scholar
21. [21]
M. S. Bartlett, G. Littlewort, I. Fasel, J.R. Movellan, Real time face detection and facial expression recognition: development and applications to human computer interaction, in: IEEE Conference on Computer Vision and Pattern Recognition, 2003, p. 53.
Google Scholar
22. [22]
K. Berns, J. Hirth, Control of Facial Expressions of the Humanoid Robot Head ROMAN, in: IEEE International Conference on Intelligent Robots and Systems, 2006, pp. 3119?3124.
Google Scholar
23. [23]
E. Vural, M. Cetin, A. Ercil, G. Littlewort, M. Bartlett, J. Movellan, Automated drowsiness detection for improved driver safety-comprehensive databases for facial expression analysis, in: International Conference on Automotive Technologies, 2008.
Google Scholar
24. [24]
P. Lucey, J. Cohn, S. Lucey, I. Matthews, S. Sridharan, K. Prkachin, Automatically detecting pain using facial actions, in: IEEE International Conference on Affective Computing and Intelligent Interaction, 2009, pp. 1?8.
Google Scholar
25. [25]
N. Perveen, S. Gupta, K. Verma, Facial expression recognition using facial characteristic points and Gini index, in: IEEE Students Conference on Engineering and Systems, 2012, pp. 1?6.
Google Scholar
26. [26]
L. Deng, Three classes of deep learning architectures and their applications:

a tutorial survey, in: APSIPA Transactions on Signal and Information
Processing, 2012.
Google Scholar
27. [27]
L. Kang, P. Ye, Y. Li, D. Doermann, Convolutional neural networks for no-
reference image quality assessment, in: IEEE Conference on Computer Vision and
Pattern Recognition, 2014.
Google Scholar
28. [28]
C. Strobl, A.L. Boulesteix, T. Augustin
Unbiased split selection for classification trees based on the Gini index
Comput. Stat. Data Anal., 52 (2007), pp. 483-501
View PDFView articleView in ScopusGoogle Scholar
29. [29]
D. Ciresan, U. Meier, J. Masci, J. Schmidhuber
Multi-column deep neural network for traffic sign classification
Neural Netw., 32 (2012), pp. 333-338
View PDFView articleView in ScopusGoogle Scholar
30. [30]
S. Wang, Z. Liu, S. Lv, _et al._
A natural visible and infrared facial expression database for expression
recognition and emotion inference
IEEE Trans. Multimed., 12 (2010), pp. 682-691
View in ScopusGoogle Scholar
31. [31]
S. He, S. Wang, Y. Lv. Spontaneous facial expression recognition based on
feature point tracking, in: IEEE 2011 Sixth International Conference on Image
and Graphics (ICIG), 2011, pp. 760?765.
Google Scholar
Cited by (97)
* ### Facial expression recognition with Convolutional Neural Networks: Coping with few data and the training sample order
2017, Pattern Recognition
Show abstract
Facial expression recognition has been an active research area in the past 10 years, with growing application areas including avatar animation, neuromarketing and sociable robots. The recognition of facial expressions is not an easy problem for machine learning methods, since people can vary significantly in the way they show their expressions. Even images of the same person in the same facial expression can vary in brightness, background and pose, and these variations are emphasized if considering different subjects (because of variations in shape, ethnicity among others). Although facial expression recognition is very studied in the literature, few works perform fair evaluation avoiding mixing subjects while training and testing the proposed algorithms. Hence, facial expression recognition is still a challenging problem in computer vision. In this work, we propose a simple solution for facial expression recognition that uses a combination of Convolutional Neural Network and specific image pre-processing steps. Convolutional Neural Networks achieve better accuracy with big data. However, there are no publicly available datasets with sufficient data for facial expression recognition with deep architectures. Therefore, to tackle the problem, we apply some pre-processing techniques to extract only expression specific features from a face image and explore the presentation order of the samples during training. The experiments employed to evaluate our technique

were carried out using three largely used public databases (CK+, JAFFE and BU-3DFE). A study of the impact of each image pre-processing operation in the accuracy rate is presented. The proposed method: achieves competitive results when compared with other facial expression recognition methods ? 96.76% of accuracy in the CK+ database ? it is fast to train, and it allows for real time facial expression recognition with standard computers.

* ### Learning structure of stereoscopic image for no-reference quality assessment with convolutional neural network
2016, Pattern Recognition
Show abstract
In this paper, we propose to learn the structures of stereoscopic image based on convolutional neural network (CNN) for no-reference quality assessment. Taking image patches from the stereoscopic images as inputs, the proposed CNN can learn the local structures which are sensitive to human perception and representative for perceptual quality evaluation. By stacking multiple convolution and max-pooling layers together, the learned structures in lower convolution layers can be composed and convolved to higher levels to form a fixed-length representation. Multilayer perceptron (MLP) is further employed to summarize the learned representation to a final value to indicate the perceptual quality of the stereo image patch pair. With different inputs, two different CNNs are designed, namely one-column CNN with only the image patch from the difference image as input, and three-column CNN with the image patches from left-view image, right-view image, and difference image as the input. The CNN parameters for stereoscopic images are learned and transferred based on the large number of 2D natural images. With the evaluation on public LIVE phase-I, LIVE phase-II, and IVC stereoscopic image databases, the proposed no-reference metric achieves the state-of-the-art performance for quality assessment of stereoscopic images, and is even competitive to existing full-reference quality metrics.
* ### Intelligent facial emotion recognition using moth-firefly optimization
2016, Knowledge-Based Systems
Show abstract
In this research, we propose a facial expression recognition system with a variant of evolutionary firefly algorithm for feature optimization. First of all, a modified Local Binary Pattern descriptor is proposed to produce an initial discriminative face representation. A variant of the firefly algorithm is proposed to perform feature optimization. The proposed evolutionary firefly algorithm exploits the spiral search behaviour of moths and attractiveness search actions of fireflies to mitigate premature convergence of the Levy-flight firefly algorithm (LFA) and the moth-flame optimization (MFO) algorithm. Specifically, it employs the logarithmic spiral search capability of the moths to increase local exploitation of the fireflies, whereas in comparison with the flames in MFO, the fireflies not only represent the best solutions identified by the moths but also act as the search agents guided by the attractiveness function to increase global exploration. Simulated Annealing embedded with Levy flights is also used to increase exploitation of the most promising solution. Diverse single and ensemble classifiers are implemented for the recognition of seven expressions. Evaluated with frontal-view images extracted from CK+, JAFFE, and MMI, and 45-degree multi-view and 90-degree side-view images from BU-3DFE and MMI, respectively, our system achieves a superior performance, and outperforms other state-of-the-art feature optimization methods and related facial expression recognition models by a significant margin.
* ### Deep Neural Networks for wireless localization in indoor and outdoor environments

2016, Neurocomputing
Show abstract
In this paper, we propose a wireless positioning method based on Deep Learning. To deal with the variant and unpredictable wireless signals, the positioning is casted in a four-layer Deep Neural Network (DNN) structure pre-trained by Stacked Denoising Autoencoder (SDA) that is capable of learning reliable features from a large set of noisy samples and avoids hand-engineering. Also, to maintain the temporal coherence, a Hidden Markov Model (HMM)-based fine localizer is introduced to smooth the initial positioning estimate obtained by the DNN-based coarse localizer. The data required for the experiments is collected from the real world in different periods to meet the actual environment. Experimental results indicate that the proposed system leads to substantial improvement on localization accuracy in coping with the turbulent wireless signals.
 * ### Facial Expression Recognition Using Weighted Mixture Deep Neural Network Based on Double-Channel Facial Images
2017, IEEE Access
 * ### Ensemble of Deep Neural Networks with Probability-Based Fusion for Facial Expression Recognition
2017, Cognitive Computation
View all citing articles on Scopus
View Abstract

Part of special issue
Discriminative Feature Learning from Big Data for Visual Recognition
Edited by
Zhuolin Jiang, Ling Shao, Zhe Lin, Haibin Ling, Fatih Porikli, Pavan Turaga
Download full issue
Other articles from this issue
 * ### Self-taught learning of a deep invariant representation for visual tracking via temporal slowness principle
October 2015
Jason Kuen, ?, Chin Poo Lee
View PDF
 * ### CRF learning with CNN features for image segmentation
October 2015
Fayao Liu, ?, Chunhua Shen
View PDF
 * ### Deep feature learning with relative distance comparison for person re-identification
October 2015
Shengyong Ding, ?, Hongyang Chao
View PDF
View more articles
Recommended articles
No articles found.

Apply Cancel
Consent Leg.Interest
checkbox label label
checkbox label label
checkbox label label
Confirm my choices