International Journal of Signal Processing, Image Processing and Pattern Recognition
Vol.8, No.12 (2015), pp.323-334
http://dx.doi.org/10.14257/ijsip.2015.8.12.30

Face Expression Recognition Based on Motion Templates and 4-layer Deep Learning Neural Network

Jianzheng Liu[1], Xiaojing Wang[1], Jucheng Yang[1]*, Chao Wu[1] and Lijun Liu[2]

[1]*College of Computer Science and Information Engineering*
Tianjin University of Science & Technology, Tianjin, China
[2]*Wuhan TipDM Intelligent Technology, No.999, Gaoxin Road, Wuhan, China*
[1]*{jz_leo,wangxiaojing, jcyang, superwoo}@tust.edu.cn,*
[2]*liu_lijun_9@hotmail.com*

Abstract

A human facial expression is the formation of facial muscle movement. In our previous research, we proposed a method of identifying facial muscle movement which based on motion templates and GentleBoost. But the method was not robust enough to recognize human expression due to insufficient learning stage. So in this paper, we proposed a new method based on motion templates and 4-layer deep learning neural network to identify human's facial expressions. We recognized Action Unit as a kind of features by using motion templates and adaboost firstly, and then the extracted features were used to feed a 4-layer deep learning neural network to recognize the facial expression. The experimental results have proved that the proposed method can solve the problem encountered in our previous research.

Keywords: *MHI; Deep Learning; Facial Expression*

1. Introduction

Affective computing is an important field which computer science will focus on in nowadays. It is allow computers to identify human feelings and emotions, and interact with people by having feelings of human. The simplest and most direct way of the research is begin with analysis of human facial expression.

Most researchers' works on affective computing are based on CV (Computer Vision) technology [1-2]. A human facial expression is the formation of facial muscle movement. To research facial expressions of human, we should start with the research of identify the specific movements of human facial muscle, and focus on the movement itself but not the face pictures. So we have been studying on how to identify this kind of movements. Our work is based on Ekman's FACS (Facial Action Coding System) [3]. Psychologists' study had demonstrated that human facial expressions corresponding to a fixed form of muscle movement which is not subject to age, gender, race, education and other factors influence.

In this paper, we proposed an approach which can recognize facial expression by identify human facial muscle movement. In our previous work, we had approached several ways to recognize Action Units defined in [3] and identify facial expressions based on BP neural network [4-6]. In this paper, we described an approach which can recognize human facial expressions by using deep learning neural network. The researches were based on our previous research. Experimental results show that the method can identify human face expression in real-time and the performance was better than we had proposed in [6].

The remainder of the paper is organized as follows: we present the relevant theories of our method in Section 2. Our method is elaborated in Section 3. In Section 4, we report on the experimental setup and results. Finally, Section 5 concludes the paper and outlines directions for future work.

ISSN: 2005-4254 IJSIP

2. Methodology

2.1. Motion Templates

We proposed an approach in [5] which can quickly and automatically identify AU (Action Unit). The proposed method was based on motion templates that can identify human facial muscle movement in real-time. Motion templates were invented in the MIT Media Lab by Bobick and Davis [7-8] and were further developed by Davis and Bradski [9-10]. The algorithm depends on generating silhouettes of the object of interest. There are a lot of methods to achieve the silhouettes; we don't describe the method we used because it isn't the main issue in this paper.

Assume that we have an object silhouette. A floating point Motion History Image [11] where new silhouette values are copied in with a floating point timestamp is updated as follows:

$$MHI_{\delta}(x,y) = \begin{cases} \tau & \text{if current silhouette at}(x,y) \\ 0 & \text{else if } MHI_{\delta} < (\tau - \delta) \end{cases} \quad (1)$$

where τ is the current time-stamp, and δ is the maximum time duration constant associated with the template. This method makes the representation independent of system speed or frame rate so that a given gesture will cover the same MHI area at different capture rates [10]. Figure 1 shows a schematic representation for a person doing shake head movement. Regardless of the number of images between the left 2 pictures, the MHI will cover the same area even if only the first and the last image were used to generate the MHI.

Figure 1. Movement and MHI

A novel modification to the MHI gradient algorithm has an advantage in this regard – by labeling motion regions connected to the current silhouette using a downward stepping flood fill, we can identify areas of motion directly attached to parts of the object of interest [10]. We can isolate regions of the motion template MHI and determine the local motion within the region.

We can scan the MHI for current silhouette regions. When a region marked with the last time stamp is found, the region's perimeter is searched for sufficiently recent motion just outside its perimeter. When such motion is found, mark it with a downward flood fill, then remove that region, and repeat the process until all regions are found (Figure 2).

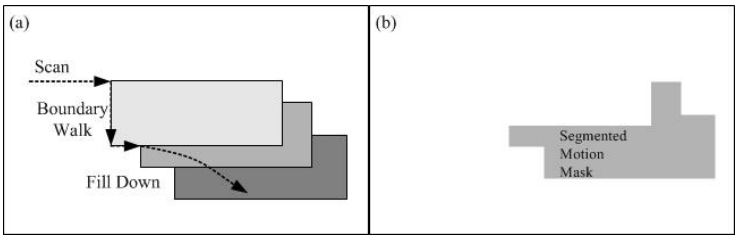

Figure 2. Motion Segmentation

Here, we illustrate our approach by showing how to train a classifier which can identify the movement of an eyebrow [5]. We identify the image region which contains an eyebrow in an image which is the every picture in a video or sequence as an ROI, and then generate MHI of the ROI image in every frame of the video firstly. After that job, for all segmentations in every MHI, we select the segmentations which corresponding to a movement of eyebrow as positive samples and set other segmentations as negative samples manually. Then these samples are normalized to a unified size. Finally, we train a classifier based on adaboost by using these samples. The trained classifier can identify the movement quickly and accurately.

2.2. Deep Learning Neural Network

Deep learning neural network is a new area of Machine Learning research. In 2006, Hinton presented a way to reduce the dimensionality of data with neural networks [12]. In the paper, Hinton described a way to train a multilayer neural network which can convert high-dimensional data to low-dimensional codes. He used gradient descent to fine-tuning the weights in such "autoencoder" networks. Usually, autoencoders use three or more layers, including a visible layer, a number of hidden layers, and an output layer. Hinton built a four-layer network in which each layer was an RBM (restricted Boltzmann machine). The network could then encode a picture and reduce its dimension from 784 to 30, with a performance superior to a PCA (principal components analysis).

RBMs were invented by Paul Smolensky [13] and have gain popularity after Hinton published the famous paper [12]. RBMs are probabilistic graphical models that can be interpreted as stochastic neural networks. They have attracted much attention as building blocks for the multi-layer learning systems called deep belief networks, and variants and extensions of RBMs have found applications in a wide range of pattern recognition tasks [14].

An RBM is an MRF associated with a bipartite undirected graph, shown in Figure 3. It consists of m visible units and n hidden units. $W = \{w_{nm}\}$ is the weight between the visible layer and the hidden layer. The visible units constitute the first layer and correspond to the components of an observation (*e.g.,* one visible unit for each sample value of a digital signal). The hidden units model dependencies between the components of observations (*e.g.,* dependencies between the sample values in the signal) and can be viewed as non-linear feature detectors [15]. In the RBMs network graph, each unit is connected to all the units in the other layer, with no connections between units in the same layer.

An RBM can be trained by adjusting its weight matrix (W) such that the probability distribution the RBM represents fits the training data as well as possible. Although training an RBM is computationally demanding, many researchers have presented different ways to solve the problem [12, 14]. However, that problem is not the focus of this paper, and will not be discussed here.

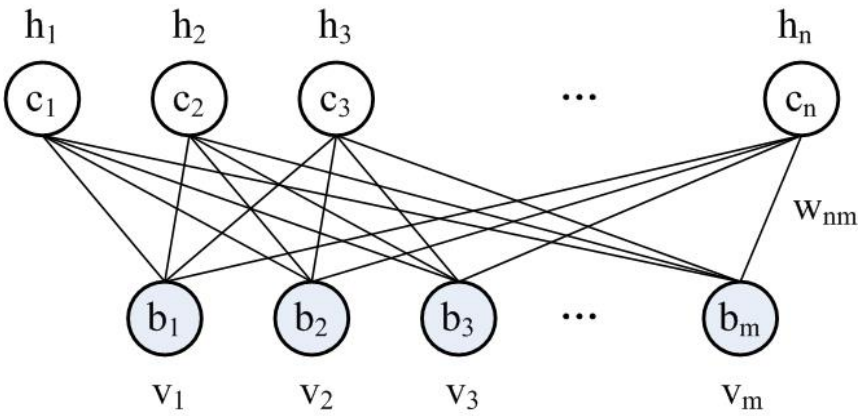

Figure 3. Architecture of an RBM

As a kind of unsupervised learning algorithms, autoencoder systems can make use of unlabeled data. Obviously, this kind of data is more abundant than labeled data. These algorithms are usually used in multi-class classification problems. A popular way to train such network is as following: Firstly, the whole network was pretraining using unsupervised learning except the last layer, just as the method presented by Hinton in [12]. The trained part of the network can be seen as a kind of auto feature extractor which converts high-dimensional to low-dimensional codes. The converted codes are used to feed the last layer of the network. The last layer can be seen as a softmax regression, which is a generalized form of logistic regression and will be trained in supervised learning with the labels of the training data. Finally, as the method Hinton used, combine each part of the network and train it using backpropagation algorithm.

3. The Proposed Approach

We designed a real-time system which can recognize angry, disgust, fear, happy, sadness and surprise. Figure 4 is the general diagram of our approach.

3.1. Input of the Network

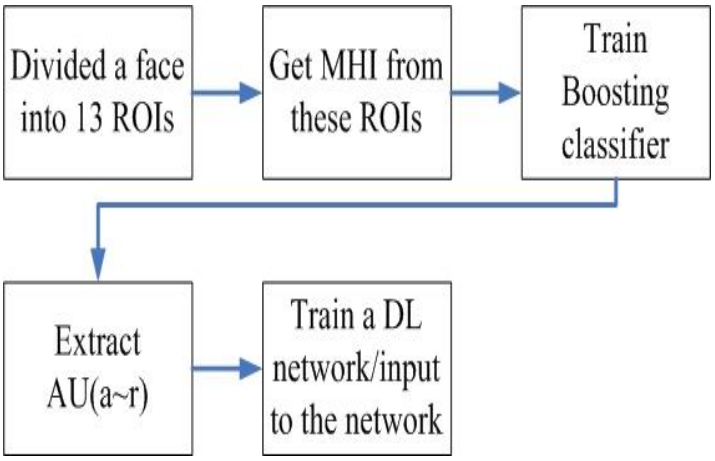

Figure 4. Outline of our Approach

Table 1. Region of Interest on Face

Name	Region
Left inside corner	x: W/2, y: H/4, w: W/4, h: H/4
Right inside corner	x: W/4, y: H/4, w: W/4, h: H/4
Left eyebrow	x: W/2, y: H/8, w: W/4, h: H/4
Right eyebrow	x: W/4, y: H/8, w: W/4, h: H/4
Nostril	x: W*2/7, y: H/2, w: W*3/7, h: H/4
Left mouth corner	x: W*7/15, y: H*2/3, w: W/3, h: H/3
Righ mouth corner	x: W/5, y: H*2/3, w: W/3, h: H/3
Left nasolabial fold	x: W/2, y: H*5/8, w: W*3/8, h: H*3/8
Right nasolabial fold	x: W/8, y: H*5/8, w: W*3/8, h: H*3/8
Upper lip	x: W/4, y: H*5/8, w: W/2, h: H/4
Lower lip and jaw	x: 0, y: H*1, w: W*1, h: H/4
Glabella	x: W*3/8, y: H/8, w: W/4, h: H/4
Nose	x: W*3/8, y: H/3, w: W/4, h: H/3

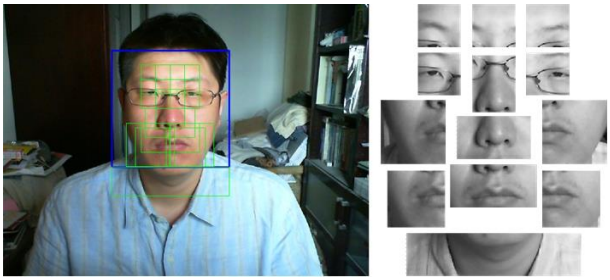

Figure 5. Divide a Face into 13 ROIs

There many ways to get ROIs in an image based on object detection. For an example, the method proposed in [16] is a very popular algorithm. But in our method, it is according to previous knowledge that we divided a human face image into 13 ROIs (Table 1 and Figure 5). In Table 1, x & y is the coordinates of ROI in the face area, w is the width of ROI, h is the high of ROI, W is the width of face area, H is the high of face area. In fact, the accurate positions of ROIs are not necessary for our method. Based on previous knowledge, our method can work more simply and lower computational cost. For every ROI, one or more boosting classifiers will be trained to recognize a kind of muscle movement; each classifier corresponded to an Action Unit defined in FACS or a form of facial muscle movement. In order to distinguish with FACS, we named the output of these classifier as AUa~AUr, see Table 2.

Table 2. The Definition of AUa~AUr (AU1~AU27 is Defined in FACS [1])

classifier	ROI	Explain
AUa	Left eyebrow	AU1
AUb	Right eyebrow	AU1
AUc	Left eye	AU5,6,7
AUd	Right eye	AU5,6,7
AUe	Glabella	AU4
AUf	Nose	AU9
AUg	Left nasolabial fold	deeper
AUh	Right nasolabial fold	deeper
AUi	Left mouth corner	Corresponding to a simle
AUj	Left mouth corner	Corresponding to a sadness
AUk	Left mouth corner	Corresponding to a surprise
AUl	Right mouth corner	Corresponding to a simle
AUm	Right mouth corner	Corresponding to a sadness
AUn	Right mouth corner	Corresponding to a surprise
AUo	Upper lip	AU10
AUp	Upper lip	Corresponding to a sadness
AUq	Upper lip	Corresponding to a surprise
AUr	Lower lip and jaw	AU26,27

Boosting classifier is a binary classifier, but its output is a floating point value. Actually, the value is a superposition of many weak classifiers' output. Through experimental observation, we found that, Boosting classifiers' output values can reflect the similarity of test sample and training samples. That is, test samples with a high positive output are more similar to the positive training samples than those samples which have a lower positive output. According to this phenomenon, we used the boosting classifiers as a kind of feature extractors. A classifier can output a feature when feeding MHI to it. The output value can be seen as a feature corresponding to the movement in ROI. Therefore, we defined AUa~AUr as the boosting classifiers' output value and used it to feed our deep learning neural network.

3.2. Structure of Network

After repeated tests, we constructed a 4-layer deep learning neural network which had 18 visible input units, 4 layers and 7 output units. Each layer of the network was an RBM. Figure 6 shows the structure of our network. Each input corresponds to a boosting classifier's output (AUa~AUr) which is a floating point value, and the 7 outputs correspond to the six basic facial expressions and non-expression. In the training phase, the output unit was set to 1 which corresponds to the expression represented by the input data (AUa~r), and other six output units were set to 0. In the predict phase, the expression corresponds to the unit which has the maximum output value reveals the recognition result.

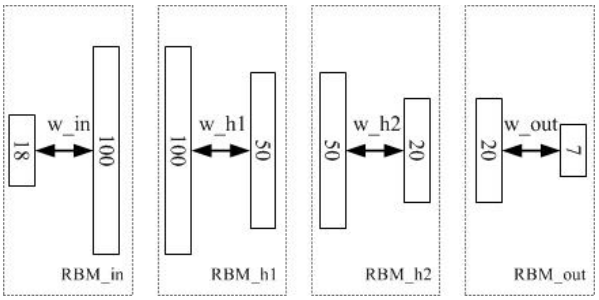

Figure 6. Architecture of the DL Neural Network

4. Results

4.1. Training Set

In order to demonstrate the improve of our method presented in this paper, we tested our method on the same database which was reported in [6]. Most of the current expression databases were made of static images or image sequences, few databases contain video expression library. So we trained a deep learning neural network on the database which captured in our laboratory. In this database there are 10 real-color videos in nearly frontal view from 10 participants, male and female in 50-50%, being 25 to 60 years of age. The videos were captured in the scale of 640X480 pixels at 30 fps (frame per second). In each section, the participants were asked to make 6 basic facial expressions clearly, each expression was required to show twice (it is 5 times in the current version database). The 10 video clips were randomly divided into two groups; each group contained 5 videos, named test set 1 and test set 2. In the training phase, one group of the two set was used as the training set, the other group as the test set. We hand-picked the best two continuous 10 frames which can representative one expression in training set videos and extracted AUa~AUr from every frame. That is, there are 140 samples in a training set; each expression has 20 samples and 20 non-expression samples.

4.2. Experimental Results

Like the works in [6], our test is also divided into two stages. There were 2 tests in the first stage. In test 1, the whole system was trained by using test set 1 and tested on test set 2. In test 2, we swapped the train samples and test samples. On the second stage, we used both test set 1 and 2 as training set to train a deep learning neural network which had the same structure with was trained in the first stage. We tested the trained network with another video clips captured ourselves and Cohn-Kanade expression database [17] by feeding them to the network. In this time, a set of videos named test set 3 were captured from peoples who were not in test set 1 & 2. Like the capture policy reported in [6], we didn't ask the subjects to act expressions in accordance with a prior requirement. The only rule was that he must show expressions clearly, and recorded the expressions he had shown after the video capturing. Our method can identify expressions from a video or image sequence. Generally speaking, expressions will be reflected in the continuous frames in a video. That means the output of our system is recognition result which corresponding with the last continuous several frames. In order to determine whether the recognition result is correct or not, we made the following defined: For a ten frames image sequence, if the expression is identified correctly in the last frame, we regarded it as a CORRECT result; if there is a wrong expression (does not include an expression was identified as non-expression, unless in the continuous frames which an expression was shown, each frame was identified as non-expression) identify result in the end image of a sequence, we regarded it as an ERROR.

The recognition results are shown in Table 3, Table 4 and Table 5. In the tables, row heads mean the labels of the input subjects and the column heads mean the recognition results. Compare with our previous work, correctly identify rate in the first stage was raised from 86.5% [6] to 90.8%, a total of 240 expressions were identified incorrectly 22 times. In the second state, there were 23 happy expressions, 17 angry, 11 sadness, 25 surprise 10 disgust and 12 fear expressions in all video clips. We have achieved identification rates of 87.8%, a total of 98 expressions were identified incorrectly 12 times. We select 100 image sequences from the Cohn-Kanade database, which include 4 expressions: happy, angry, sadness and surprise. We selected 25 image sequences for each expression. Ten image sequences were corresponding to each expression. We have achieved identification rates of 91%.

5. Discussion and Conclusion

We had introduced a method which can identify Action Unit in [5]. But for various reasons, the method was not robust enough to identify expressions. In in this paper, we present the approach that we used a number of boosting classifiers' floating-point output value as the input of deep learning neural network can solve the problsem.

Table 3. Results of the First Stage

						Test1/Test2
Expression	Hap.	Ang.	Sad.	Surp.	Disg.	Fear
Happy	20/20	0/0	0/0	0/0	0/0	0/0
Angry	0/0	17/18	2/0	0/0	1/2	0/0
Sadness	0/0	2/2	16/17	0/0	1/0	1/1
Surprise	1/2	0/0	0/0	19/18	0/0	0/0
Disgust	0/0	1/1	0/1	0/0	18/18	1/0
Fear	0/0	0/0	1/0	0/2	0/0	19/18

Table 4. Results of the Second Stage

Expression	Hap.	Ang.	Sad.	Surp.	Disg.	Fear
Happy	23	0	0	0	0	0
Angry	0	14	0	0	2	1
Sadness	0	0	9	0	0	2
Surprise	2	0	0	23	0	0
Disgust	0	2	0	0	8	0
Fear	0	0	1	2	0	9

Table 5. Results on Cohn-Kanade Database

	Training set: test set 1+2; Test set: Cohn-Kanade			
Expression	Hap.	Ang.	Sad.	Surp.
Happy	23	0	0	2
Angry	0	22	2	1
Sadness	0	4	21	0
Surprise	0	0	0	25

Figure 7 shows the phenomenon that when AU classifiers gave an error result, the network's output was still correct. In Figure 7 A, some AU classifiers (AUi, AUj, AUl, AUm, *etc.*) identified it as smile or sad. When we combined all of the classifiers' output and fed them to the network. The system's output was happy. In Figure 7 B, AU classifiers identified it as smile and sad, the network's output was sadness. Especially in Figure 7 C, AU classifiers identified the image as sad; the network gave a correct result: surprise. The AU classifiers can be seen as a kind of feature extractors. Its output can reflect the motion in ROI. But it is a local feature. We separated the face into 13 ROIs and trained 18 classifiers. The combined feature which was consisting of the 18 outputs can reflect every muscle movement in each face ROI. We used this kind of feature to recognize facial expressions. When change the network from BP network to deep learning network, the approach we proposed in this paper get a better result.

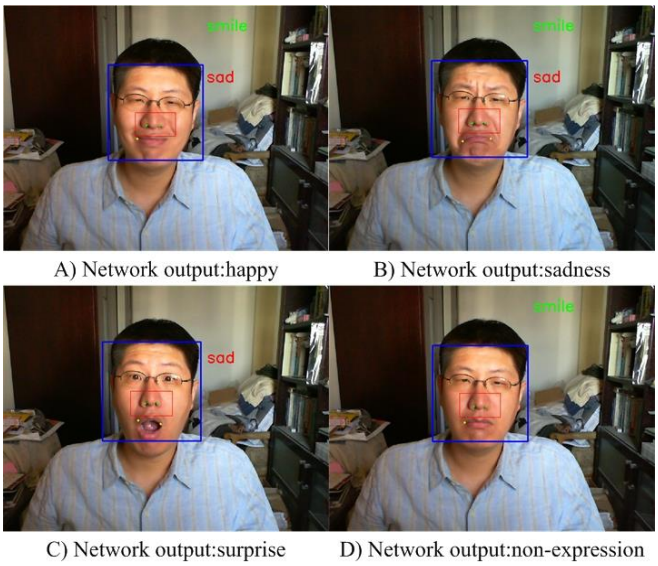

Figure 7. Network Identification Results

In this paper we present a robust, highly accurate method for identify expressions in video based DL neural network and some our previous work. We test our method and had achieved relatively good results, and solve some of the problems encountered in previous research.

In future work we will investigate try to measure the degree of an identified expression. Also we plan to identify the composition of a complex expression.

Acknowledgments

The research reported here was supported by the National Natural Science Foundation of China under Grant 61402332, and partly funded by the Tianjin High School Science & Technology Fund Planning Project: 20130802, and the young academic team construction projects of the 'twelve five' integrated investment planning in Tianjin University of Science and Technology.

This paper is a revised and expanded version of a paper entitled "Expression Recognition Based on 4-layer DeepLearning Neural Network" presented at International Symposium on Information Technology Convergence, Oct. 30-31, 2014, Jeonju, Korea.

Reference

[1] J. Wang, X. Ma, J. Sun, Z. Zhao, Y. Zhu, Puzzlement Detection from Facial Expression Using Active Appearance Models and Support Vector Machines, International Journal of Signal Processing, Image Processing and Pattern Recognition, Vol.7, No.5, **(2014)**, pp.349-360.
[2] K. Lin, W. Cheng, J. Li, Facial Expression Recognition Based on Geometric Features and Geodesic Distance, International Journal of Signal Processing, Image Processing and Pattern Recognition, Vol.7, No.1, **(2014)**, pp.323-330.
[3] P. Ekman, W. Friesen, J. Hager, Facial Action Coding System, Research Nexus, a subsidiary of Network Information Research Corporation, Made in the United States of America, **(2002)**, ISBN 0-931835-01-1.
[4] J. Liu, Z. Zhao, Fully automatic and quickly facial feature point detection based on LK algorithm, Networked Computing and Advanced Information Management (NCM), 2011 7th International Conference on, Korea, **(2011)**, pp.190-194.
[5] J. Liu, Z. Zhao, M. Li, C. Liu, Action unit recognition based on motion templates and GentleBoost, Networked Computing and Advanced Information Management (NCM), 2011 7th International Conference on, Korea, **(2011)**, pp.195-199.
[6] C. Liu, J. Liu, Expression Recognition Based on BP Neural Network, Computer Science and Electronics Engineering (ICCSEE), 2012 International Conference on. IEEE, **(2012)**, 1: 89-93.
[7] A. Bobick, J. Davis, Real-time recognition of activity using temporal templates, IEEE Workshop on Applications of Computer Vision, **(1996)**, pp. 39-42.
[8] J. Davis and A. Bobick, The representation and recognition of human movement using temporal templates. Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on. IEEE, **(1997)**: 928-934.
[9] J. Davis, G. Bradski, Real-time motion template gradients using Intel CVLib, ICCV Workshop on Framerate Vision, **(1999)**.
[10] G. Bradski, J. Davis, Motion segmentation and pose recognition with motion history gradients, IEEE Workshop on Applications of Computer Vision, **(2000)**.
[11] J. Davis, Recognizing movement using motion histograms, MIT Media Lab Technical Report #487, **(1999)**.
[12] G. E. Hinton, R. R. Salakhutdinov, Reducing the dimensionality of data with neural networks, Science 313 (5786) 504–507 **(2006)**.
[13] P. Smolensky, Information processing in dynamical systems: Foundations of harmony theory, **(1986)**: 194.
[14] A. Fischer, C. Igel, Training restricted Boltzmann machines: An introduction, Pattern Recognition 47 (1) 25–39 **(2014)**.
[15] G. E. Hinton, Boltzmann machine, Scholarpedia, 2, 1668, **(2007)**.
[16] P. Viola, Paul, M. J. Jones, Robust real-time face detection, International journal of computer vision 57.2: 137-154 **(2004)**.
[17] Kanade, Takeo, Jeffrey F. Cohn, and Yingli Tian, Comprehensive database for facial expression analysis, Automatic Face and Gesture Recognition, Proceedings. Fourth IEEE International Conference on. IEEE, **(2000)**.

Authors

Jianzheng Liu, he received the Ph.D. degree from Tianjin University, China in 2012. Currently he is a lecturer with the College of Computer Science and Information Engineering, Tianjin University of Science & Technology, Tianjin, China. His research interests include affective computing, biometric systems, signal and image processing, pattern recognition.

Xiaojing Wang, College student in College of Computer Science and Information Engineering, Tianjin University of Science and Technology, Tianjin, P.R.China. Her main research interests include artificial intelligence and machine learning.

Jucheng Yang is a full professor in College of Computer Science and Information Engineering, Tianjin University of Science and Technology, Tianjin, P.R.China. He is a Specially-appointed Professor of Tianjin City and Haihe Scholar. He received his B.S. degree from South-Central University for Nationalities, China in 2002, MS and PhD degrees from Chonbuk National University, Republic of Korea in 2004 and 2008. He did his postdoctoral work at the Advanced Graduate Education Center of Jeonbuk for Electronics and Information Technology-BK21 (AGECJEIT-BK21) in Chonbuk National University, too. He has published over 80 papers in related international journals and conferences, such as IEEE Trans. on HMS, IEEE Systems Journal, Expert Systems with Applications and so on. He has served as editor of five books in biometrics, and as reviewer or editor for international journals such as IEEE Transactions on Information Forensics & Security, IEEE Trans. on Circuits and Systems for Video Technology, IEEE Communications Magazine, and as the guster editor of Journal of Network and Computer Applications. He is the general chiar of CCBR'15, ISITC2015, and the publicity chair of ICMeCG'10-12. And he is the program committee member of many conferences such as JCeSBI'10, IMPRESS'11 and CCBR'13, CCBR'14. He owns 7 patents in biometrics. His research interests include image processing, biometrics, pattern recognition, and neural networks. E-mail: jcyang@tust.edu.cn.

Chao Wu, he received the Ph.D. degree from Northwestern Polytechnical University, China in 2007. Currently he is a lecturer with the College of Computer Science and Information Engineering, Tianjin University of Science & Technology, Tianjin, China. His research interests include machine learning, face recognition.

Lijun Liu, she received her Ph.D. degree from Chonbuk National University, South Korea in 2009 and did her post-doc researches in STFX University, Canada. Currently she is a senior researcher in Wuhan TipDM Intelligent Technology, Wuhan, China. Her research interests include Internet of Things, Big Data, Pervasive Computing, Intelligent Control, MIMO, *etc.*

