2019 International Conference on Power, Energy, Environment and Material Science (PEEMS 2019)
ISBN: 978-1-60595-669-5

FERCaps: A Capsule-Based Method for Face Expression Recognition from Frontal Face Images

Qi-di HU[1], Qian SHU[2], Ming-ze BAI[2], Xiao-ming YAO[2,3] and Kun-xian SHU[2,*]

[1]School of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065 China

[2]Chongqing Key Laboratory on Big Data for Bio Intelligence, Chongqing University of Posts and Telecommunications, Chongqing 400065 China

[3]mProb Center, Queen Mary Hospital, University of Hong Kong, Hong Kong, China

*Corresponding author

Keywords: Facial expression recognition, convolutional neural networks, capsule, deconvolution.

Abstract. A novel method which is named FERCaps (capsule-based method for facial expression recognition) is presented in this paper. We process these images containing the faces, and segment these gray images only containing the front faces for training the model. The model framework consists three layers of convolutional neural networks (CNN), two layers of capsules, and a decoder with a fully connected layer and four deconvolution. The ReLU activation function is used to speed up the experimental model training speed in the convolution layers and constructed a decoder suitable for facial expressions using deconvolution. The original capsule is the bottom layer of the multidimensional entity. Through an iterative routing process, each raw capsule selects a corresponding high-level capsule in the upper level. Instantiated parameters can characterize facial expressions in high-level capsules. We verified the effectiveness of the model through experiments on the public benchmarking datasets JAFFE and extended Cohn-Kanade (CK+), and achieved an accuracy of 98.18% in ck+ and 88.33% in JAFFE.

Introduction

Facial expression recognition (FER) has a large number of applications in the fields of social robots, medical care, and driver fatigue monitoring[1, 2], and gradually moving from laboratory dataset testing to challenging real-world scenarios. When we have mainly frontal faces in our database, machine learning might be good solutions for FER. However, the traditional machine learning generally has four parts, including face detection, image preprocessing, feature extraction, and expression classification[3]. Among them, feature extraction and expression classification are the key points, and have important influence on FER. Scholars have done a lot of research in these two aspects, and have proposed many solutions. Feature extraction methods include Gabor wavelet transform[4, 5], Histogram of oriented gradients (HOG)[6], local binary mode (LBP)[7-9], and manifold learning[10, 11]. Expression classification method mainly has Support Vector Machine (SVM)[2, 8, 12].

With the rise of deep learning, the convolutional neural network (CNN) has made a major breakthrough in the field of image classification and recognition[13, 14], and also provided a corresponding basis and reference for the study of FER models. The end-to-end learning approach of the CNN is considered to automatically extract the best image features for classification or recognition [15] .

At present, most of FER models are trained on the CK+ database. For example, an Peak-Piloted Deep Network (PPDN) model designed by Zhao[16], experimentally on the CK+[17] database, achieved a recognition accuracy of 92.06%. CNN[13, 14, 18, 19] will discard a lot of details of the entities in the area during the pooling process. Hinton[20] pointed out that the translation invariance in CNN was not a perfectly reasonable design. It is considered that the pooled subsampling will lose

the spatial correlation in the layer-by-layer calculation process, leading to failure in the recognition task, and the capsule network might solve this problem[20,21]. The location information in the underlying capsules is "place coded" by the active capsule, and the location information of the capsule in the high level is "rate encoded" in the component of the output vector. It is also because of this shift from position coding to rate coding that high-level capsules have more degrees of freedom and characterizes more complex entities. Therefore, in order to improve the generalization ability of FER, we proposes a capsule-based model for FER, and verifies the accuracy and robustness of the model on CK+ database and JAFFE database.

Related Work

Convolutional Neural Network

CNN performs well in understanding low-level or high-level features in images, but loses spatial information about the image at the pooling level[22]. In the convolutional layer[23], the image matrix is operated through a convolution kernel to obtain a feature map. In existing convolutional networks, a convolutional layer can be considered a feature extractor, and the features learned by using the 1st to 3rd layers can be regarded as the basic features[22]. In existing convolutional networks, a convolutional layer can be considered a feature extractor, and the features learned by using the 1st to 3rd layers can be regarded as the basic features[22]. We design a FERCaps-2 model with 2 layers of convolution, and a FERCaps-3 model with 3 layers of convolution 3. The two FER models were validated on the CK+ database and JAFFE database, and the performance of different convolutional layers on the face recognition model was tested.

Capsule Definition and Computing

A capsule's neuron activity is used to represent various attributes that appear in a particular entity, and to represent different types of instance parameters[20]. For example, the parameters include direction, position, size, deformation, texture, and color, etc. And one special property is the existence of the instantiated entity in the image. The modulus of the parameter vector of the instance can be used to represent the probability of existence of the entity, while different direction parameters of the vector are used to represent different attributes of the entity. Because the modulus of the final capsule is to represent the probability of existence of the entity, and in order to increase its non-linear expression ability, the original vector is compressed with a nonlinear function. This function makes the direction of the vector unchanged, while ensuring that the modulus length of all vectors is also compressed to a length below 1 and the short vector is compressed to almost zero.

$$v_j = \frac{\left\|s_j\right\|^2}{1+\left\|s_j\right\|^2}\frac{s_j}{\left\|s_j\right\|} \tag{1}$$

where v_j is the output vector of capsule j and s_j is its full input. The first layer of the capsule is obtained by convolution of the last layer. Then, the output u_i of the capsule is multiplied by the weight matrix W_{ij} to obtain the prediction vector $u_{j|i}$, and then the weighted sum of the prediction vector $u_{j|i}$, is obtained to obtain a capsule s_j.

$$s_j = \sum_i c_{ij} u_{j|i}, u_{j|i} = W_{ij} u_i \tag{2}$$

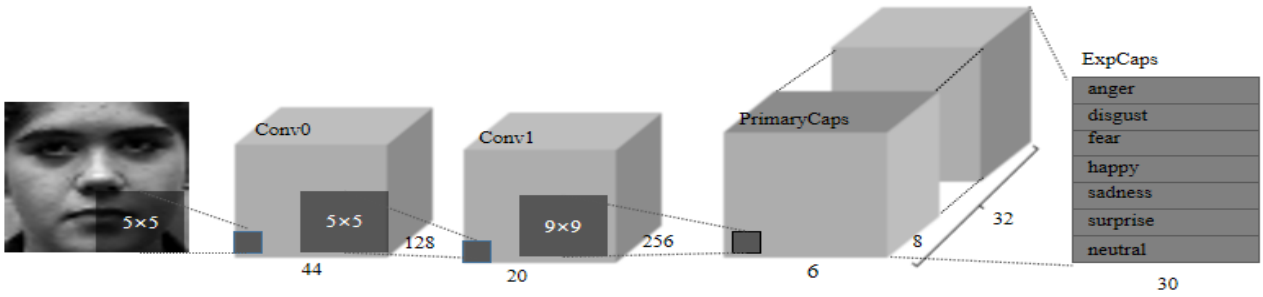

Figure 1. FERCaps-3 Model: Capsule-based facial expression recognition. The activation vector modulus of each capsule in the ExpCaps layer gives an instance of each class.

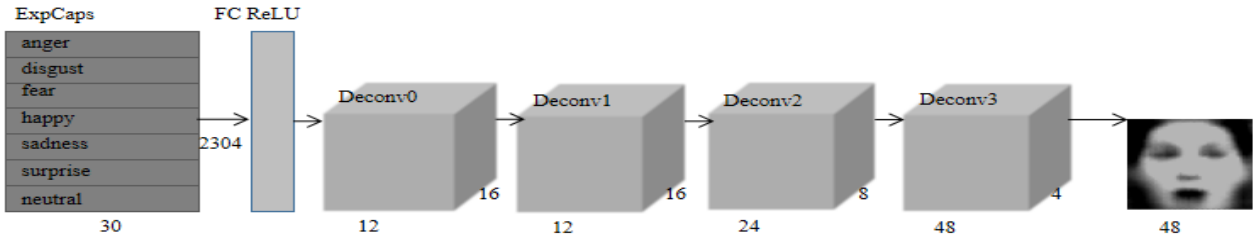

Figure 2. FERCaps-3 Decoder: Reconstruct the decoding structure of the facial expression image, the input of the structure is ExpCaps.

where the c_{ij} are coupling coefcients that are determined by the iterative dynamic routing process[20].

Capsule Definition and Computing

The first appearance of the concept of deconvolution was in Zeiler's 2010 paper[24]. With the successful application of deconvolution in neural network visualization, many other works are also adopting deconvolution. Deconvolution is also known as tansposed cnvolution, factional frided cnvolution[22], and so on. With the successful application of deconvolution in neural network visualization, many other works are also adopting deconvolution. Deconvolution is also known as tansposed cnvolution, factional frided cnvolution[22], and so on. We can produce images by sending features to a deconvolution structure. In order to use the generated image for regularization, it is naturally necessary to restore the extracted feature map to the same size as the original image.

Methodology

Capsule Network for Facial Expression Recognition

For efficiently facial expression recognition, we propose two models of FERCaps according to numbers of CNN layers,FERCaps-2 and FERCaps-3, of which the FERCaps-3 model is shown in Figure. 1. The FERCaps-2 model structure is roughly the same as FERCaps-3. We will describe the two model structures in detail below.

FERCaps-2 has two layers of convolutional layers. The first layer has 256, 9×9 convolution kernels with a stride of 2. The second layer of PrimaryCaps is a convolutional capsule layer. The original 256 channel convolutional layer is considered to be 32 channels. Each channel has 8 convolutional units. The convolution kernel is 9×9 and the stride is 2.

FERCaps-3 has 3 layers of convolutional layers. Conv0 has 128, 5×5 convolution kernels with a stride of 1. Conv1 has 256, 5×5 convolution kernels with a stride of 2. The last two layers are the same as FERCaps-2. We use dynamic routing[20] between the primary capsule layer and the expression capsule layer (ExpCaps), with 3 routing iterations.

In order to realize facial expression, we use the modulus of the vector to indicate whether the entity represented by the capsule exists. We use the Margin loss commonly used in SVM [8,12] for optimization, and it was defined by,

$$L_c = T_c \max\left(0, m^+ - \|v_c\|\right)^2 + \lambda\left(1 - T_c\right)\max\left(0, \|v_c\| - m^-\right)^2 \quad (3)$$

c is the expression category. T_c is the indicator function of the classification. where $T_c = 1$ iff the expression category c is present. Penalty false negative $m^+ = 0.9$, penalty false positive $m^- = 0.1$, $\lambda = 1$.

Decoder Network

We input ExpCaps into the decoder network as shown in fig. 2. The decoder network consists of a fully connected layer and four deconvolution layers, with the activation function Sigmoid applied at the last layer, and the ReLU at the others. In this decoder model, the first layer is a fully connected layer that converts the expression capsule into 2304 outputs, which are then converted into $12 \times 12 \times 16$ feature maps. Layers 2~5 are deconvolution layers. In the beginning, we used the fully connected layer to build the decoder, but the results were not ideal. Then we used deconvolution to build a new decoder to make the model performance superior and stable.

Experiments and Results

This experiment was done on the Tensorflow[25] deep learning framework. The training was accomplished by the Adam optimizer [26] and its default parameters.

Facial Expression Database

In order to evaluate the FER model, we use the reference public database JAFFE[27] and CK+[17] of the facial expression. The JAFFE database has a total of 213 images, consisting of 7 expressions of 10 Japanese women (angry, fear, happy, neutral, sad, disgusted, surprise).Figure 3 shows this series of pictures.

Figure 3. Examples of facial expression samples in the JAFFE database.

This database was extended on the basis of Cohn-Kanade Dataset and was released in 2010. The number of sequences is increased by 22% and the number of subjects by 27%.Figure 3 shows this series of pictures.

Figure 4. Examples of facial expression samples in the CK+ database.

Comparison of the Proposed Model and the Optimal Method

We would record the experimental results in Table 1. The performance of the FERCaps-2 model is slightly insufficient. The FERCaps-3 model achieved an accuracy of 88.33% on the JAAFE database, which is 7.33% higher than the state-of-the-art SVM (RBF) + Boosted-LBP. The experimental results on the CK+ database are recorded in Table 2. The FERCaps-2 model achieved an accuracy of 94.85% on CK+. The FERCaps-3 model achieved an accuracy of 98.18% on the CK+ database, which is 0.88% higher than state-of-the-art PPDN[16]. The results in both databases show that the FERCaps-3

model performs best. And the performance improvement on JAFFE is obvious. Our model used only three layers of convolution to get better results, surpassed the state-of-the-art PPDN.

Table 1. Comparative evaluation.

Method	JAFFE Accuracy(%)	CK+ Accuracy(%)
SVM(linear+Boosted-LBP)[8]	81.0	91.4
I2CNN [28]	75.28	96.2
Aligned crop + LSTM[29]	/	97.2
PPDN[16]	/	97.3
FERCaps-2	78.33	94.85
FERCaps-3	88.33	98.18

Conclusion

This paper proposes a capsule-based facial expression recognition model. And we use deconvolution to form a decoder suitable for facial expression recognition.Experimental results show that we are the best. The three main contributions for performance improvement are carefully designed FERCaps model, designed a deconvolution decoder, and verifies that the 3-layer convolution is optimal for this model.

Acknowledgement

This study was financially supported by Chongqing Natural Science Foundation (cstc2018cyjAX0225), the Special Project of National Science and Technology Cooperation (2014DFB30010), National Natural Science Foundation of China (61501071)

References

[1] B. Fasel and J. J. P. R. Luettin, “Automatic facial expression analysis: a survey,” vol. 36, no. 1, pp. 259-275, 1999.

[2] M. Takalkar, M. Xu, Q. Wu, Z. J. M. T. Chaczko, and Applications, “A survey: facial micro-expression recognition,” vol. 77, no. 15, pp. 1-25, 2018.

[3] K. H. Cheung, A. Kong, J. You, Q. Li, D. Zhang, and P. Bhattacharya, “A new approach to appearance-based face recognition,” in IEEE International Conference on Systems, 2005.

[4] W. Gu, C. Xiang, Y. V. Venkatesh, D. Huang, and H. J. P. R. Lin, “Facial expression recognition using radial encoding of local Gabor features and classifier synthesis,” vol. 45, no. 1, pp. 80-91, 2012.

[5] M. J. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, “Coding facial expressions with Gabor wavelets,” in IEEE International Conference on Automatic Face & Gesture Recognition, 2002.

[6] X. Wang, J. Chao, L. Wei, H. Min, L. Xu, and F. Ren, “Feature fusion of HOG and WLD for facial expression recognition,” in IEEE/SICE International Symposium on System Integration, 2014.

[7] X. Feng, M. Pietikäinen, A. J. P. R. Hadid, and I. Analysis, “Facial expression recognition based on local binary patterns,” vol. 17, no. 4, pp. 592-598, 2007.

[8] C. Shan, S. Gong, P. W. J. I. Mcowan, and V. Computing, “Facial expression recognition based on Local Binary Patterns: A comprehensive study,” vol. 27, no. 6, pp. 803-816, 2009.

[9] G. Zhao, M. J. I. T. o. P. A. Pietikainen, and M. Intelligence, “Dynamic Texture Recognition Using Local Binary Patterns with an Application to Facial Expressions,” vol. 29, no. 6, pp. 915-928, 2007.

[10]H. Murase and S. K. J. I. J. o. C. V. Nayar, “Visual learning and recognition of 3-d objects from appearance,” vol. 14, no. 1, pp. 5-24, 1995.

[11]M. Belkin, P. Niyogi, and V. J. J. o. M. L. R. Sindhwani, “Manifold Regularization: A Geometric Framework for Learning from Labeled and Unlabeled Examples,” vol. 7, no. 1, pp. 2399-2434, 2006.

[12]X. Wu, W. Zuo, Y. Zhu, L. J. I. T. o. N. N. Lin, and L. Systems, “F-SVM: Combination of Feature Transformation and SVM Learning via Convex Relaxation,” vol. PP, no. 99, pp. 1-15, 2018.

[13]A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” in International Conference on Neural Information Processing Systems, 2012.

[14]S. Lawrence, C. L. Giles, A. C. Tsoi, and A. D. J. I. T. O. N. N. Back, “Face recognition: a convolutional neural-network approach,” vol. 8, no. 1, pp. 98-113, 1997.

[15]J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, “How transferable are features in deep neural networks?,” 2014.

[16]X. Zhao, X. Liang, L. Liu, L. Teng, and S. Yan, “Peak-Piloted Deep Network for Facial Expression Recognition,” in European Conference on Computer Vision, 2016.

[17]P. Lucey, J. F. Cohn, T. Kanade, and J. Saragih, “The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression,” in Computer Vision & Pattern Recognition Workshops, 2010.

[18]V. Badrinarayanan, A. Kendall, R. J. I. T. o. P. A. Cipolla, and M. Intelligence, “SegNet: A Deep Convolutional Encoder-Decoder Architecture for Scene Segmentation,” vol. PP, no. 99, pp. 1-1, 2017.

[19]H. Li, L. Zhe, X. Shen, J. Brandt, and H. Gang, “A convolutional neural network cascade for face detection,” in Computer Vision & Pattern Recognition, 2015.

[20]S. Sabour, N. Frosst, and G. E. Hinton, “Dynamic Routing Between Capsules,” 2017.

[21]Y. Lecun, L. Bottou, Y. Bengio, and P. J. P. O. T. I. Haffner, “Gradient-based learning applied to document recognition,” vol. 86, no. 11, pp. 2278-2324, 1998.

[22]M. D. Zeiler and R. Fergus, “Visualizing and Understanding Convolutional Networks,” 2013.

[23]M. Oquab, L. Bottou, I. Laptev, and J. Sivic, “Learning and Transferring Mid-level Image Representations Using Convolutional Neural Networks,” in Computer Vision & Pattern Recognition, 2014.

[24]M. D. Zeiler, D. Krishnan, G. W. Taylor, and R. Fergus, “Deconvolutional networks,” in Computer Vision & Pattern Recognition, 2010.

[25]M. Abadi et al., “TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems,” 2016.

[26]D. P. Kingma and J. J. C. S. Ba, “Adam: A Method for Stochastic Optimization,” 2014.

[27]M. J. Lyons, J. Budynek, S. J. P. A. Akamatsu, and M. I. I. T. on, “Automatic classification of single facial images,” vol. 21, no. 12, pp. 1357-1362, 1999.

[28]C. Zhang, P. Wang, Chen, Joni-Kristian, K. J. J. O. S. Engineering, and Electronics, “Identity-aware convolutional neural networks for facial expression recognition,” vol. 28, no. 4, pp. 784-792, 2017.

[29]P. Rodriguez et al., “Deep Pain: Exploiting Long Short-Term Memory Networks for Facial Expression Classification,” vol. PP, no. 99, pp. 1-11, 2017.