http://dx.doi.org/10.14801/jkiit.2018.16.11.105

A Study of Facial Organs Classification System Based on Fusion of CNN Features and Haar-CNN Features

Biao Hao*[1], Hye-Youn Lim*[2], and Dae-Seong Kang**

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government (NO. 2017R1D1A1B04030870)

Abstract

In this paper, we proposed a method for effective classification of eye, nose, and mouth of human face. Most recent image classification uses Convolutional Neural Network(CNN). However, the features extracted by CNN are not sufficient and the classification effect is not too high. We proposed a new algorithm to improve the classification effect. The proposed method can be roughly divided into three parts. First, the Haar feature extraction algorithm is used to construct the eye, nose, and mouth dataset of face. The second, the model extracts CNN features of image using AlexNet. Finally, Haar-CNN features are extracted by performing convolution after Haar feature extraction. After that, CNN features and Haar-CNN features are fused and classify images using softmax. Recognition rate using mixed features could be increased about 4% than CNN feature. Experiments have demonstrated the performance of the proposed algorithm.

요 약

본 논문에서는 사람 얼굴의 눈, 코, 입을 효과적으로 분류하는 방법을 제안한다. 최근 대부분의 이미지 분류는 CNN(Convolutional Neural Network)을 이용한다. 그러나 CNN으로 추출한 특징은 충분하지 않아 분류 효과가 낮은 경우가 있다. 분류 효과를 더 높이기 위해 새로운 알고리즘을 제안한다. 제안하는 방법은 크게 세 부분으로 나눌 수 있다. 첫 번째는 Haar 특징추출 알고리즘을 사용하여 얼굴의 눈, 코, 입 데이터셀을 구성한다. 두 번째는 CNN 구조 중 하나인 AlexNet을 사용하여 이미지의 CNN 특징을 추출한다. 마지막으로 Haar 특징 추출 뒤에 합성(Convolution) 연산을 수행하여 Haar-CNN 특징을 추출한다. 그 후 CNN 특징과 Haar-CNN을 혼합하여 Softmax를 이용해 분류한다. 혼합한 특징을 사용한 인식률은 기존의 CNN 특징 보다 약 4% 향상되었다. 실험을 통해 제안하는 방법의 성능을 증명하였다.

Keywords
AlexNet, CNN, softmax classifier, Haar-CNN, image classification

* Dong-A University Electronic Engineering
- ORCID[1]: https://orcid.org/0000-0001-8127-4908
- ORCID[2]: https://orcid.org/0000-0002-5189-461x
** Dong-A University Electronic Engineering Professor
- ORCID: https://orcid.org/0000-0003-0186-2430

· Received: Aug. 01, 2018, Revised: Oct. 24, 2018, Accepted: Oct. 27, 2018

· Corresponding Author: Dae-Seong Kang
Dept. of Dong-A University,37 NaKdong-Daero 550, beon-gil saha-gu, Busan, Korea,
Tel.: +82-51-200-7710, Email: dskang@dau.ac.kr

Ⅰ. 서 론

Image processing is a very important technology, it can be roughly divided into three parts, such as image classification[1], object detection[2], and semantic segmentation[3]. In recent years, deep learning has been one of the hottest research topics in the field of image processing. Before the rise of deep learning, machine learning is the most used algorithm in image processing. But now in the field of image processing, deep learning algorithm is the mainstream algorithm. In the deep learning and machine learning[4], neural network is very important. The neural network mimics the human brain to make the machine perform a series of operations autonomously. Perceptron is a very important algorithm before the emergence of the neural network. We can think of the neural network as a further development of the perceptron.

Image classification is the most basic algorithm in image processing. It is used in a variety of tasks. The mean of image classification is that computer automatically assign images to their respective classes. The classification task is a binary classification or a multi-class classification. The process of image classification involves two steps training system and testing system. In the training step, model extracts the features of image and form a description for every class. In the testing step, the trained model of the training step is used to classify the test images. In the image processing, features are necessary. Here are many kinds of extracted features, it can be divided into machine learning features and deep learning features. The machine learning features include Histogram of Oriented Gradient(HOG)[5], Local Binary Pattern(LBP)[6], Haar, Scale Invariant Feature Transform(SIFT)[7] and so on. The HOG constructs a feature by calculating and counting a gradient direction histogram of a local region of the image. LBP is an operator used to describe the local texture features of an image, it has significant advantages, such as rotation invariance and gray invariance. The SIFT algorithm obtains good results by finding the interest points and their descriptions about scale and orientation in a graph and performing image feature point matching. Such algorithms require artificial manipulation to extract features of the image, and the extracted features lack abstract features, and the classification results are relatively low. After that, Support Vector Machine(SVM)[8], Random Forest and Adaboost classifiers use extracted features to make a final decision of image classes. SVM construct hyperplane in high-dimensional space, which can be used for classification[9].

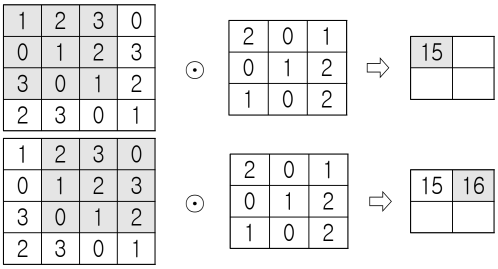
Convolutional Neural Network(CNN)[10] mimics the human nervous system and enables the computer to complete feature extraction autonomously. CNN has a huge breakthrough on ILSVR2012 image competition, it beat all others models in image classification. Now CNN is widely used in image processing. Compared with the traditional deep neural network, CNN reduced the calculated amount and the calculation complexity of the neural network. The CNN can be described as convolutional blocks and fully connection blocks. A convolutional block includes linear filters, non-linear activation function, and local pooling operations. The network of fully connection block is the traditionary neural network. CNN use the hierarchical structure of network to extract feature. Then fully connection layers use the extracted features to classify the input image. The traditional CNN includes LeNet-5[11], AlexNet, Visual Geometry Group(VGG)[12], GoogLeNet[13] and so on. LeNet-5 is the most basic CNN. Alexnet is an 8-layer neural network, VGG is a 16-layer neural network, and GoogLeNet is a 22-layer neural network. The most used networks are AlexNet and VGG.

In this paper, in order to improve the classification effect, a new algorithm is proposed. First, Haar feature is trained to obtain low-level features of input image, then CNN is used to extract Haar-CNN features based on Haar feature. Second, CNN is used to extract CNN feature from input image. Finally, the model fuse CNN features and Haar-CNN features to train softmax classifier, and realize classification task.

II. Related Algorithm

2.1 Convolutional Calculation

Weight-sharing network structure of CNN make it more similar to the biological neural network, it reduces the number of weights. CNN is used to extract more abstract features. Convolutional calculation and pooling calculation are very important in the CNN. The convolutional calculation of CNN is the same as the traditional neural network. The input is multiplied by the weights. But the weight values of the CNN are in the kernel. As shown in the Fig. 1, it is the convolutional calculation process of CNN. The result can be obtained by multiplying the input with weight. For example, the convolutional calculation result is (1*2)+(2*0)+(3*1)+(0*0)+(1*1)+(2*2)+(3*1)+(0*0)+(1*2)=15, so the first part is 15. The remaining parts get the results by the same calculation method.



1	2	3	0
0	1	2	3
3	0	1	2
2	3	0	1

⊙

2	0	1
0	1	2
1	0	2

⇨

15	

1	2	3	0
0	1	2	3
3	0	1	2
2	3	0	1

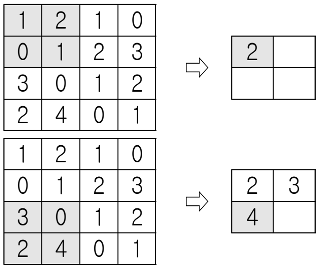
⊙

2	0	1
0	1	2
1	0	2

⇨

15	16

Fig. 1. Convolution calculation



1	2	1	0
0	1	2	3
3	0	1	2
2	4	0	1

⇨

2	

1	2	1	0
0	1	2	3
3	0	1	2
2	4	0	1

⇨

2	3
4	

Fig. 2. Max pooling operation

The appearance of pooling in CNN further reduces the training parameters of the network and further increases the speed of operation. There are two main pooling methods, average pooling and max pooling. Here we use the max poling operation. The calculation process is shown in Fig. 2, it extracts the largest pixel in the area.

2.2 Haar Feature

Haar features[14][15] are pixel-based rectangular patters, we can also call them rectangular features. Haar features consist of edge features, linear features, center features, and diagonal features. Image can be represented using these haar features. A Haar feature represents a part of an object. In the Haar feature, it has one or two white rectangular areas and one or two black rectangular areas. The value of Haar feature can be got by the sum of white rectangle pixel minus the sum of black rectangle pixel. Haar feature value reflects the change in grayscale of the image. For example, some features of the face can be described simply by rectangular features. Such as, the eye color is deeper than the cheek, the color of both sides of the nose is deeper than nose, the mouth is deeper than the surrounding color. However, rectangular features only correspond to some simple graphic structures. For example, it is more sensitive to edges and line segments. So only specific structures (horizontal, vertical, and diagonal) can be described. Fig. 3 is Haar feature template.

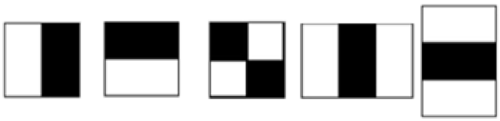


Fig. 3. Haar feature template

2.3 Integral Image

Haar feature uses an integral graph to calculate features values. The main idea of the integral image is to store the sum of the pixels in the rectangular region formed from the start point to each point of the image as an array element in the memory. When you want to calculate the pixel sum of a certain area, you can directly index the elements of the array without having to recalculate the pixel sum of the area, thereby speeding up the calculation. No matter how large the computation, integral image can use the same time to calculate it, thus greatly improving the detection speed.

The formula for calculating the feature value is that value=Sum white-Sum black. Fig. 4 is haar feature value calculation process using an integral image.

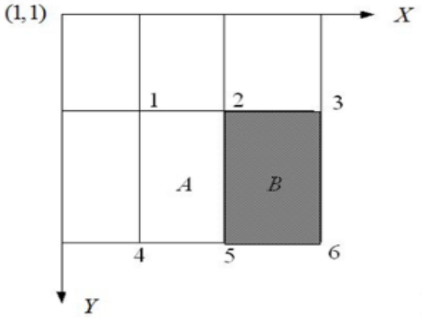


Fig. 4. Process of haar feature value calculation

Haar feature value can be computed using an integral image. For an image point A(x, y), the integral image at the location (x, y) can be defined as the sum of pixels on the top and left of (x, y). Eq. 1 is the function. $ii(x, y)$ and $i(x,y)$ are the integral image and original image respectively. $i(x',y')$ is the value of the point (x', y'). Using the recurrences of Eq. 2 and Eq. 3, the integral image can be calculated in the original image.

$$ii(x, y) = \sum_{x' \leq x, y' \leq y} i(x', y') \tag{1}$$

$$s(x, y) = s(x, y-1) + i(x, y) \tag{2}$$

$$ii(x, y) = ii(x-1, y) + s(x, y) \tag{3}$$

where $s(x,y)$ is the cumulative row sum, $s(x,-1)=0$, and $ii(-1, y)=0$. In other word, $s(x,y)$ is the integral value of the column.

In the Fig. 4, the feature value of the rectangular feature is that the pixel value of the area A minus the pixel value of the area B using an integral image. The pixel value of area A is ii(5)+ii(1)-ii(2)-ii(4). The pixel value of area B is ii(6)+ii(2)-ii(5)-ii(3). So the feature value of this rectangular feature is equal to ii(5)+ii(1)-ii(2)-ii(4)-[ii(6)+ii(2)-ii(5)-ii(3)]=[Ii(5)-ii(4)]+[ii(3)-ii(2)]-[ii(2)-ii(1)]-[ii(6)-ii(5)]. The feature values of a rectangular feature are only related to the endpoint of integral graph of the rectangular feature. But it has nothing to do with the coordinates of the image.

2.4 AlexNet Neural Network

CNN is one of the representative network structure in deep learning, it is widely used in image processing. The emergence of AlexNet is the beginning of CNN’s rise. The Alexnet Convolutional Neural Network achieved the first place in the imagenet contest in 2012. So it becomes a research hotspot in image processing, more and more scholars and experts research on CNN. Fig. 5 is the structure of AlexNet.

The number of output of the first CNN layer Conv_1 is 98, and the output feature map size is 55*55. The number of output of the second CNN layer Conv_2 is 256, and the output feature map size is 27*27.

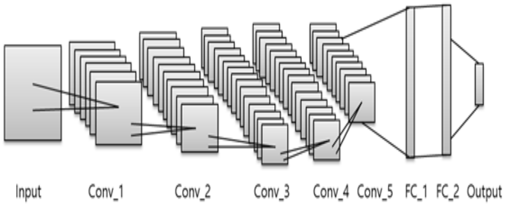


Fig. 5. Structure of AlexNet

The number of output of the third CNN layer Conv_3 is 384, and the size of the output feature map is 13*13. The number of output of the fourth CNN layer Conv_4 is 384, and the size of the output feature map is 13*13. The number of output of fifth CNN layer Conv_5 is 256, and the size of the output feature map is 13*13. The number of output of the first fully connection layer FC_1 is 4096, and the size of the input feature map is 6*6. The number of output of the second fully connection layer FC_2 is 4096. The number of output of the last layer is image classes.

III. Proposed Algorithm

The proposed algorithm is a fusion algorithm based on Haar and CNN algorithm. In the CNN part of this paper, we select AlexNet, in classifier part, we choose the softmax classifier. Fig. 6 is the operation process of proposed algorithm.

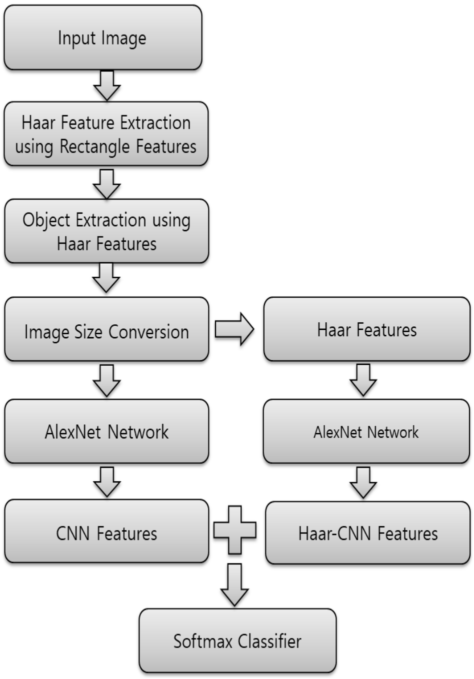


Fig. 6. Framework of proposed algorithm

Now, there are many datasets for image processing, but no dataset about eye, nose, and mouth. So here we need to use Haar technology to extract the eye, nose and mouth images to form a dataset for training. Firstly, the existing LFW face dataset is selected, and the Haar algorithm is used to detect three target images of the eye, nose, and mouth from the face. Then they are extracted to form an eye, nose, and mouth image dataset. Second, because the fully connection network has requirements for the size of the input image, the scale of the image is needed to convert to 227*227 here. Third, the CNN features and Haar features of the image are extracted separately. Fourth, in the extracted Haar features, the convolutional operation is processed to extract the Haar-CNN features of the image. Fifth, a new feature set is constructed using the CNN feature and the Haar-CNN, and then the new feature set is inputted into the softmax classifier for classification. In this experiment the used CNN model is the basic AlexNet, and the model fuses features at the 5th layer of the network.

Haar feature is a feature that reflects the change of image gray level. The Haar algorithm extracts the information it considers important from the original image. And then CNN calculation is performed on the acquired image to extract the Haar-CNN feature. The CNN model extracts features directly from the original image and uses the extracted features for classification. However, the effect of classification is not very good, so it is supplemented with Haar-CNN features to improve the classification effect.

3.1 Softmax classifier

In a neural network, the most used classifiers are SVM and softmax classifiers. In this paper, softmax classifier is selected to carry out the classification task. Eq. 4 is the softmax function.

$$f(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{N} e^{x_j}} \tag{4}$$

x_i is the first i dimension output, N is the number of classes. The $f(x_i)$ is the probability of class i. When predicting an image, the class of input image is that it has the largest probability. Eq. 5 is the function.

$$F(x_i) = \max_{i \epsilon (1,N)} f(x_i) \tag{5}$$

Ⅳ. Experimental Results

In order to evaluate the performance of the proposed algorithm, the LFW face database is used. The dataset has 13233 person face images. But we did not use it all, some of them are used. Here we use the dataset to get an eye, nose, and mouth dataset.

Table 1 is the experimental environment. Table 2 is the introduction of database.

Table 1. Experimental environment

CPU	Intel(R) Core(TM) i7-4770@ 3.4GHz
GPU	NVIDIA GeForce 980TI
Memory	8GB
OS	Windows 10
Development Tools	Pycharm
HDD	930GB

Table 2. Introduction of database

Database	Image Size	Number of image	Train Image	Test Image
LFW Datebase	250*250	13233	2000	150
Eye, Nose, Mouth Dataset	227*227	—	6000	450

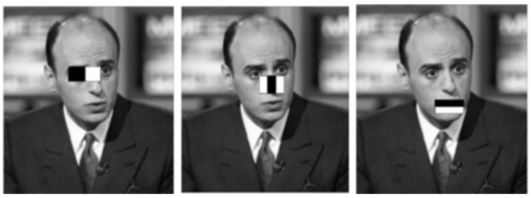


Fig. 7. Process of Haar calculation

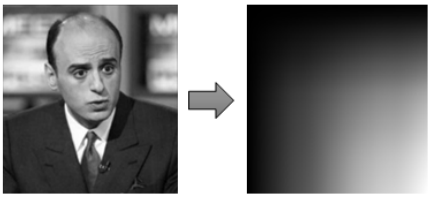


Fig. 8. Integral image of input image

This paper use fusion features of CNN and Haar-CNN features to train softmax classifier, finally trained model is used to realize the classification task. In order to evaluate the effect of the proposed algorithm, classification accuracies of three different algorithms are compared. The three algorithms are proposed algorithm, CNN algorithm, and Haar algorithm. In Haar algorithm, an integral image is used to calculate the image pixel value. Fig. 7 is the process of Haar calculation. Fig. 8 is the integral image of an image.

Table 3. Description of different algorithms

Features	Description of features
Haar features	Description of the grayscale variation. Description of edge features.
CNN features	CNN features extraction. Abstract features extraction.
Proposed algorithm features	The CNN features extraction. Haar-CNN features extraction. Obtain the new features by fusing features.

Table 3 is the description of different features, we can know the advantage of the proposed algorithm features. Fig. 9 is the display of the image's performance ability of different features.

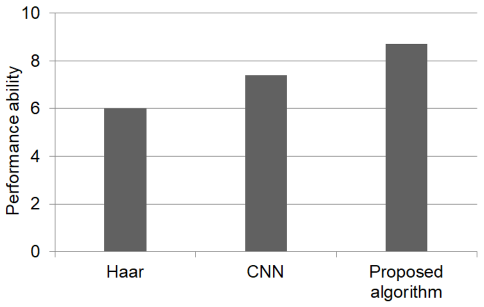


Fig. 9. Performance ability of different features

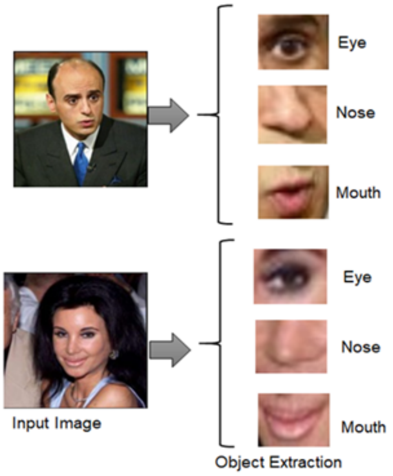


Fig. 10. Object extraction

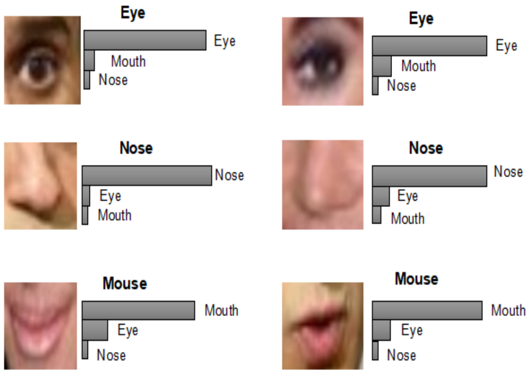


Fig. 11. Effect of classification using softmax

Fig. 10 is the process of object extraction. Fig. 11 is the prediction probability of the image using the softmax classifier. If the probability of which class is the highest, then the image is determined to be this class.

Table 4 is the classification accuracy comparison of different algorithm. The effect of the proposed algorithm is better than others.

Table 4. Accuracy comparison of different algorithms

Evaluation index	Mouth accuracy	Eye accuracy	Nose accuracy
Haar(%)	90.0	89.3	88.7
CNN(%)	92.5	93.3	94.0
Proposed algorithm(%)	97.3	96.0	96.6

Here is a comparative experiment, which selects different numbers of images as the training set, and the number of images in the test set does not change. Here are face image, the model use face images to get eye, nose, and mouth dataset. Detecting the classification ability of various features. Table 5 is the classification ability of different features.

Table 5. Classification accuracy for different number of training images

Average accuracy of different algorithm	Number of training images			
	500	1000	1500	2000
Haar(%)	81.9	85.3	87.4	89.3
CNN(%)	86.2	87.9	90.7	93.5
Proposed algorithm(%)	92.6	94.1	95.3	96.7

Ⅴ. Conclusions

This paper proposed a classification model of eye, mouth, and nose based on the CNN feature and Haar feature. Firstly, the mouth, eye, nose images dataset is obtained. Secondly, the fusion features are extracted, it composes of two parts. one is that AlexNet network is used to extract CNN features, the other is that Haar algorithm is used to obtain the Haar images, and then AlexNet is used to extract Haar-CNN features based on Haar images. Finally, a softmax classifier is used to classify images. The recognition rate can be increased by about 4% by using fusion features than using CNN features. There have 3 kinds of features used in this study, they are CNN, Haar, and Haar-CNN. Compare with the CNN feature, the Haar features are lower-level features. Here Haar features are used as input of AlexNet to extract abstract Haar-CNN features. AlexNet also is used to extract CNN features of images, and then model fuse features based on CNN and Haar-CNN features. In the experiment, the number of features of images is increased, so the classification accuracy of images is improved. We can also consider the proposed algorithm from another angle. The input images of AlexNet can be divided into two parts one is the original image, the other one is the Haar feature images of the original image. The recognition rate is also improved, because of the expansion of the training dataset.

Fusion technology is a useful technology to improve the network effect. In the future, we will use fusion technology in various algorithms to improve the performance of the algorithm.

References

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks", Communications of the ACM, Vol. 60, No. 6, pp. 84-90, Jun. 2017.

[2] P. B. Felzenszwalb, R. B. Girshick, and D. M. Allester, "Object Detection with Discriminatively Trained Part-Based Models", IEEE, Vol. 32, pp. 1627-1645, Sep. 2010.

[3] J. Long, E. Shelhamer, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation", IEEE Conference on Computer vision & Pattern Recognition, Vol. 79, No. 10, pp. 3431-3440, Jun. 2015.

[4] S. H. Park, S. Y. Ihm, and Y. H. Park, "A study on Application of Machine Learning Algorithm to Visitor Marketing in Sports Stadium", Journal of DCS, Vol. 19, No. 1, pp. 27-33, Dec. 2018.

[5] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection", IEEE Computer Vision and Pattern Recognition(CVPR), Vol. 1, No. 12, pp. 886-893, Jun. 2005.

[6] Ojala, Timo, Matti Pietikainen, and David Harwood, "A comparative study of texture measures with classification based on featured distributions", Pattern recognition, Vol. 29, No.1, pp. 51-59, Jan. 1996.

[7] A. Hassan, "Scale Invariant Feature Transform Evaluation in Small Dataset", International Conference on Sciences and Techniques of Automatic Control and Computer Engineering(STA), pp. 368-372, Dec. 2015.

[8] D. K. Srivastava and L. Bhambhu, "Data Classification Using Support Vector Machine", Journal of Theoretical and Applied Information Technology, pp. 1-7, Feb. 2010.

[9] S. J. Kim and J. J. Lee, "A study on Face Recognition using Support Vector Machine", JIIBC, Vol. 16, No. 6, pp. 183-190, 2016.

[10] H. W. Jang and S. H. Jung, "Training Artificial Neural Networks and Convolutional Neural Networks using WFSO Algorithm", Journal of DCS, Vol. 18, No. 5, pp. 969-967, Aug. 2017.

[11] Y. Lecun, L. D. Jackel, and C. Cortes, et. all, "Learning Algorithm for Classification: A Comparison on Handwritten Digit Recognition", Neural Network the Statistical mechanics Perspective, pp. 261-276, 1995.

[12] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition", ICLR, pp. 1-14, Apr. 2015.

[13] C. Szegedy, W. Liu, Y. Q. Jia, P. Sermanet, S. Rees, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions", IEEE Conference on Computer Vision & Pattern Recognition(CVPR), pp. 1-19, Sep. 2014.

[14] P. Viola and M. Jones, "Rapid Object Detection

using a Boosted Cascade of Simple Features", IEEE Company Society, Vol. 1, No. 2, pp. 511, Dec. 2001.

[15] L. Rainer and M. Jochen, "An Extended Set of Haar-like Features for Rapid Object Dectection", International Conference on Image Processing, Vol. 1, pp. 900-903, Sep. 2002.

Authors

Biao Hao

2015 : B.S., Electronic Information Engineering, Hebei University of Science and Technology.

2016 ~ 2018 : M.S., Electrical Engineering, Dong-A University.

Research interests : signal processing and pattern recognition.

Hye-Youn Lim

2005 : B.S., Electronic Engineering, DaeGu University.

2008 : M.S., Electrical Engineering, Dong-A University.

2013 : Ph.D., Electric Engineering, Dong-A University.

2013 ~ present : Assistant professor, Dong-A University.

Research interests : Computer Vision, Tracking.

Dae-Seong Kang

1994 : Ph.D., Electric Engineering, Texas A&M University.

1995 ~ present : Professor, Dong-A University.

Research interests : image processing and compression.