International Journal of Pattern Recognition and Artificial Intelligence

HCNN: A Neural Network Model for Combining Local and Global Features towards Human-like Classification

Tielin Zhang[1]†, Yi Zeng[1,2]‡,* Bo Xu[1,2]

1. Institute of Automation, Chinese Academy of Sciences
Beijing 100190, P. R. China
2. CAS Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences
Shanghai, 200031, P. R. China
†*zhangtielin2013@ia.ac.cn*
‡*yi.zeng@ia.ac.cn*

Brain-inspired algorithms such as Convolutional Neural Network (CNN) have helped machine vision systems to achieve state-of-the-art performance for various tasks (e.g. image classification). However, CNNs mainly rely on local features (e.g. hierarchical features of points and angles from images), while important global structured features such as contour features are lost. Global understanding of natural objects is considered to be essential characteristics that the human visual system follows, and for developing human-like visual systems, the lost of consideration from this perspective may lead to inevitable failure on certain tasks. Experimental results have proved that well-trained CNN classifier cannot correctly distinguish fooling images (in which some local features from the natural images are chaotically distributed) from natural images. For example, a picture that is composed of yellow-black bars will be recognized as school bus with very high confidence by CNN. On the contrary, human visual system focuses on both the texture and contour features to form representation of images and would not mistake them. In order to solve the upper problem, we propose a neural network model, named as HOG improved CNN (HCNN), that combines local and global features towards human-like classification based on CNN and Histogram of Oriented Gradient (HOG). The experimental results on MNIST datasets and part of ImageNet datasets show that HCNN outperforms traditional CNN for object classification with fooling images, which indicates the feasibility, accuracy and potential effectiveness of HCNN for solving image classification problem.

Keywords: Convolutional Neural Network; Object Classification; Histogram of Oriented Gradient; Human-like Performance.

1. INTRODUCTION

Understanding human intelligence is a grand challenge for science, and various efforts in Artificial Intelligence (AI) try to design building blocks that capture a certain aspect of human intelligence. Recent advances have shown that the integration of fragmented building blocks may help to refine and advance the understanding on

*Author for correspondence.

the nature of human intelligence. One recent example is the efforts from DeepMind which combines CNN[1] with reinforcement learning to form a new deep reinforcement learning network (known as the Deep Q-network[2,3]). Just like human being, the new network can learn a variety of different arcade game tasks from scratch given only minimal starting information. The Deep Q-network shows the advantages on the integration of machine vision and decision making to achieve human like behavior.

Recent efforts on deep learning focuses on automatic extraction of high-level features from data, this enables their effectiveness in visual image classification, speech recognition and many other areas related to intelligent information processing. Many neural network architectures can be categorized as deep neural networks, and CNN is one of the most typical architectures which is very suitable for two-dimension image classification and object recognition. It has been considered as one of the state-of-the-art methods for object detection and recognition.[1] Features learnt from CNN are not engineered by hand, while the network architectures are still manually designed (e.g. the number of layers, and the kernel size).

Standard CNN can be separated into two main parts: the features extraction part and the classification part. On the feature extraction part (usually related to the first layers), different layers of CNN describe different scales and types of the feature representations. Neurons in CNN utilize gradients to form learned features automatically, which is very different from the handcrafted features in Scale Invariant Feature Transform (SIFT)[4] and HOG[5]. In CNN, important textual features from images can be figured out by convolving with a sliding window and forming a filter. Filters from different layers make the layer-by-layer representation of image possible. The training of CNN makes the network filters suit with the image features by comparing the predicted category with realistic classification results. On the classification part (usually related to the final layers), the classification based on the extracted features at the final layer is made according to the candidate class which owns the biggest confidence value. The procedure of classification part is very similar with the procedure of traditional three-layered (i.e. input, hidden and output layers)neural network.

The local and global features[6] have been playing an important role on image retrieval and object detection tasks respectively.[7,8] Usually the local features are the textual features, angles features and point features. On the contrary, the global features are contour features and structural features. In the processing of CNN classification, on the one hand, hierarchical feature extraction obtains main obvious local textual features (e.g. angles and point features) from the images. On the other hand, some important global features such as contour features are lost inevitably. Thus makes the final CNN classification cannot distinguish fooling images (which contains local features of the natural images, but chaotically distributed) from natural images. For example, CNN will classify the pictures of yellow-black bars as school bus with even greater classification confidence value compared to the pictures of real school bus.[9]

From the upper example, we can conclude that the mechanism for recognizing objects by CNN is still very different from human visual system, since it ignores structured global features, while human eyes focus on both the texture and contour features to form representation of images and would not mistake the school bus with yellow-black bars. So it is necessary to combine the contour feature extraction method with CNN to produce more human-like and intelligent classification results. In this paper HOG method is utilized to achieve this goal.

Recognition techniques before CNN are very different with modern multi-layer convolution-based framework, like SIFT[4] (or PCA-SIFT[10]) and HOG[5] which focus on point features and global structural features respectively. SIFT[4] is a kind of method which can make detailed description about local features by converting each point or pixel in the image into a characteristic vector. Thus makes the images which contain the similar contents share more similar characteristic vectors than other images. These vectors of point features are robust even when some big changes happened to the images (e.g. distortion, rotation and scale transformation). Combining SIFT with the bag of words method is one of the main traditional image classification methods. Another variation of SIFT is PCA-SIFT[10] method which remains only 20 principal components of characteristic vectors of SIFT. So the SIFT method focuses more on the tiny features of points. Another commonly used method is HOG. HOG[5] is a kind of character descriptor method which ignores the features of specific gradient and edge locations, but rather detects the directions of them, thus makes the detecting results more robust.

Combining modern deep neural networks (classification method to extract local features) and earlier HOG method (generation method to extract global features) to form a new network will make the image classification more intelligent and human-like, at least not easily fooled by human-made images. In this paper, we propose a neural network model, named as HOG improved CNN (HCNN), that combines local and global features towards human-like classification performance based on CNN and HOG. Through experimental validation, we observe that HCNN could avoid some of the shortness of the current CNNs from the perspective that HCNN considers both local and global characteristics.

Based on the understanding from information theory perspective, the more types of features from images are considered, the more likely the right classes can be obtained for classification. Through experimental evaluation, combining the advantages of both texture and structure method will make the image classification more human-like and therefore more intelligent.

2. HOG improved Convolutional Neural Network (HCNN)

The overall structure of HCNN is shown in Fig. 1. Traditional CNN is constructed with two parts: The first part is feature extraction part (from layer 1 to layer N-A, in which N is the total layers of CNN and A is the number of layers for classification, for example, in Section 2.1, we will introduce two versions of CNN for two

tasks, the ImageNet CNN is with A=3, and for MNIST CNN, A=1). The functions of convolutional layers and sample layers in this part are different, and they are used for cutting down the trivial parts of images and for forming main description of input images respectively. The second part with obvious characteristics of full-connections is classification part in which the final classification function after dimension reduction (in first part) is formed.

Previous work[11] has shown that the traditional CNN does not care enough about global structures. The principle of CNN is very different with that of human visual systems which focus on both the structural characteristics (e.g. contour information) and texture characteristics (e.g. points and angles features) of images. High accuracy result of classification on the natural images on ImageNet datasets shows that the CNN (which makes the combination of the upper first and second parts) is feasible and efficient. In other words, the first part of CNN does collect some main characteristics from images (with more focus on the local texture features). Since the HCNN framework proposed in this paper tries to combine both local and global features, it is natural and feasible to keep the local texture features in the first part of CNN and utilize them in the new HCNN. In addition, global contour features are designed to be collected by HOG in the HCNN framework so that it can handle the fooling images.

2.1. *Local texture features from CNN*

We use two kinds of datasets to train different CNNs. One is with 40 classes in 1.3-million-image ILSVRC 2012 ImageNet dataset[12] which contains 1000 categories of images (Alexnet). The other is MNIST dataset[13] which contains ten categories of hand-written numbers, namely zero to nine. We name the CNNs which are trained by these two kinds of datasets ImageNet CNN and MNIST CNN respectively.

To get the trained-well ImageNet CNN features, we choose Caffe software package.[14,15] Caffe packages provide a convolutional network which is trained on 1.3 million images and has got the ILSVRC 2012 ImageNet champion. Another reason we use ImageNet CNN provided by Caffe is that it has the same structure as the former paper,[9] so that our experiments can be compared with it. The details are shown in Section 3.3 and Section 3.4.

To get the trained hand-written CNN features, we use the standard MNIST dataset[13] as training and testing data, and we use the deep learning toolbox provided by Matlab to get the trained CNN features. Detailed training structures of ImageNet CNN and MNIST CNN are shown in Fig. 2 and Fig. 3.

The ImageNet CNN structure in Fig. 2 is similar with the CNN proposed in Caffe packages.[16] The second convolutional layer takes the output of the first convolutional layer as input. The next three layers (the third, the fourth and the fifth layers) are the same types as the second layer. All the convolutional layers are filtered with different size of kernels from the first layer to the fifth layer. The kernel sizes are 11×11, 5×5, 3×3, 3×3 and 3×3 respectively. The final connection

between max pooling layer with output layers is full connection which stands for the final classification of the CNN. Before the full connection network, the layer of max pooling (i.e. green cube in Fig. 2) contains the common features which are

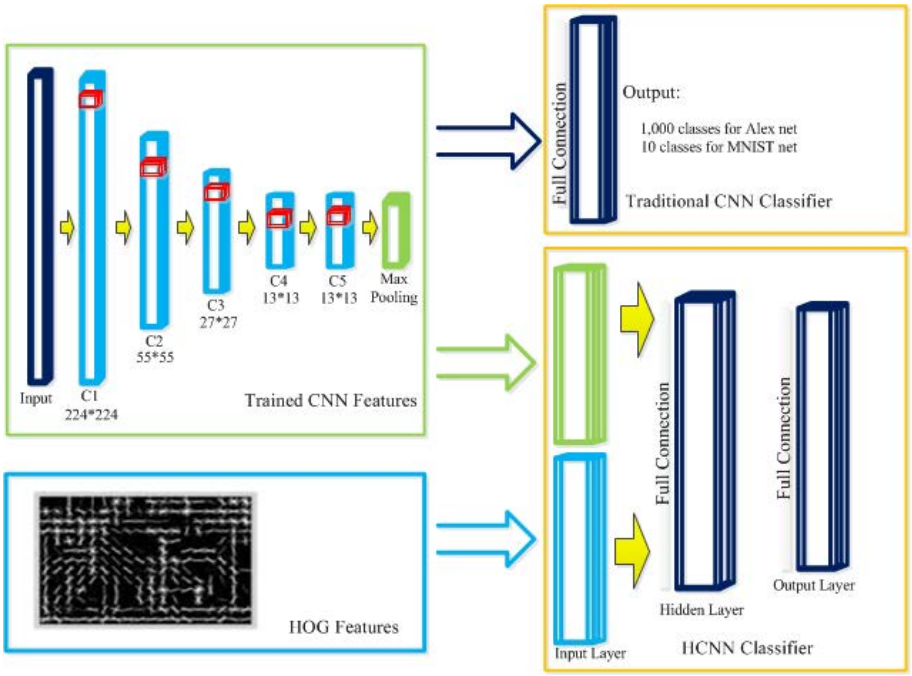

Fig. 1: The overall architecture for HCNN based on features of trained CNN and HOG. The CNN structure here is based on the ImageNet CNN.

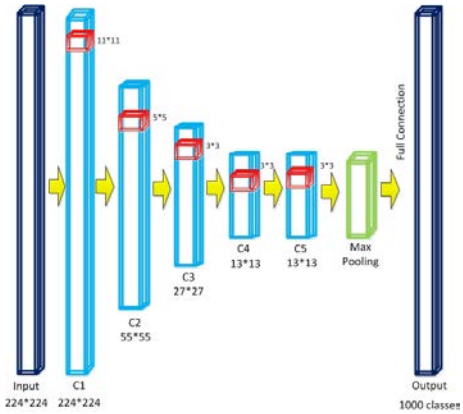

Fig. 2: The ImageNet CNN structure

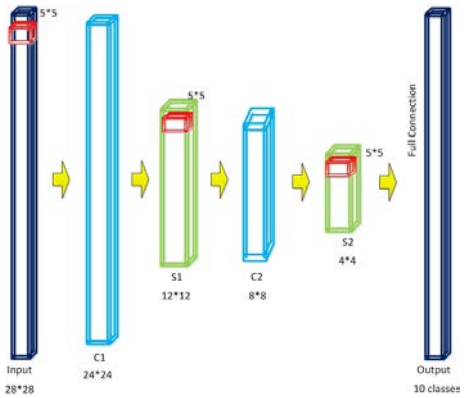

Fig. 3: The MNIST CNN structure

extracted from input images with the same type. To a certain extent, these features are a part of the feature representations of original image.

For the MNIST hand-written CNN classifier, six layers (including input and output layers) are shown in Fig. 3. The image size of the input layer is 28×28. The second layer (layer size 24×24) and fourth layer (layer size 8×8) are convolution layers, while the third layer (layer size 12×12) and the fifth layer (layer size 4×4) are sample layers. Kernel sizes for both input and convolutional layers are 5×5. Similar with ImageNet CNN, the connections between final two layers are also full connections in MNIST CNN. On the training procedure of the CNN, the standard CNN features from different categories of pictures are formed at the same time. In HCNN, both ImageNet and MNIST CNN features are saved as local texture features after training procedure(as Fig. 1 shows).

2.2. *Global structural features from HOG*

Identifying the shape and boundary characteristics quickly and robustly from images is one of the key outstanding functions for human visual system. To achieve human level performance of visual processing, a large amount of computer vision algorithms have been proposed to get the structural information (e.g. contour features) from images. Contour information enables the possibility for computational systems to make better understanding of images. How to get reliable expression of structural features from target images has been one of the most challenging topics in computer vision.

HOG is one of the methods for structural features expression. Traditional HOG method can extract object contour information in very cluttered and illuminated backgrounds. The detected result of object features shows that HOG descriptors are better compared to other global feature detection methods, such as wavelets method.[17,18]

The contour representation from HOG method shows various advantages. Firstly, the edge and gradient structures captured by HOG are important characteristics of both local and global shapes. Secondly, the local representation of HOG is controllable, and to some extent invariant to local geometric and photometric transformations, especially when the rotation scale is much smaller than the local orientation bin size. As shown in Fig. 5 and Fig. 7, the HOG method detects the contours of objects correctly. This makes the classification of HCNN in the next step possible.

HOG descriptor significantly outperforms some other feature extraction methods and has shown great performance on contour detection. In practice, many aspects (e.g. appropriate scale of gradients, appropriate bin size of orientation) are important for achieving good HOG results. So we select the best experimental result based on the optimal structure (as shown in Fig. 4 and Fig. 6) to form HOG features in HCNN. Input images are scaled to the size of 64×64 (pixels), at the same time, different sizes of blocks and sizes of cells in images are considered. The

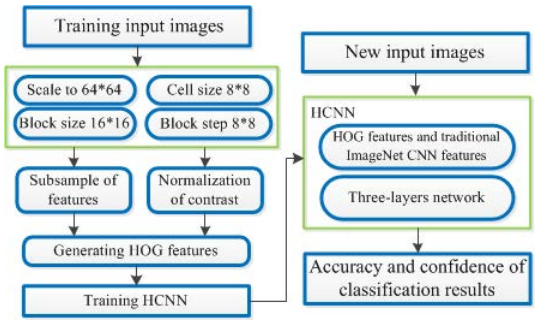

Fig. 4: HOG model for detecting contour information in ImageNet datasets.

Fig. 5: Contour information of natural images.

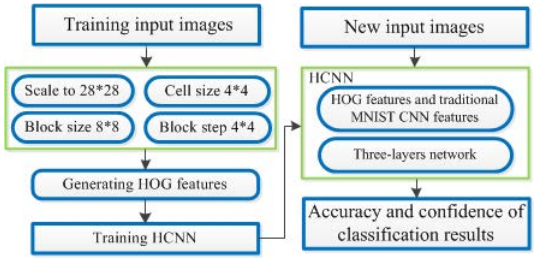

Fig. 6: HOG model for detecting contour information in MNIST datasets.

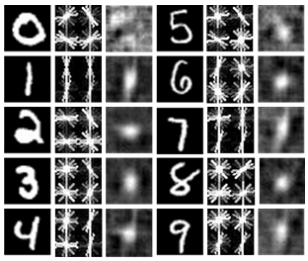

Fig. 7: Contour information of hand-written images.

features from original images are sub sampled into 196 features which are the main structural descriptions of the images. After training, the HCNN which contains both texture and structure features will be used for classifying and recognizing natural and fooling images.

2.3. *The easily fooled CNNs*

There are many differences between CNN and human visual system.[11] Even when some tiny noises are added to an image, the CNN labels this image with completely different class compared to the original label, while human can still recognize this image even when some large changes are made to it. Further, some methods[9] are created to make special images which are unrecognizable to human but still can be recognized by CNN as some categories with very high confidence (around 99.99%). These special kinds of ways to make images to cheat CNN include the evolutionary algorithm[19] and the posterior-probability-maximizing method[20,21].

In this paper, we follow the same method[9] to make different kinds of special images to cheat the traditional CNNs. The procedure and architectures are shown in Fig. 8. The evolutionary algorithm is one of the effective methods to form fooling

images from original input images. Another easier way is to add random noises on each small part of the image at every iteration time. Once the confidence of the new image is bigger than the input image, input image is updated to the new one until the confidence is high enough (here is 99.99%). Thus, for MNIST dataset, this kind of method is used to create some special images that do not belong to any of the 10 hand-written categories. Human can recognize these correctly while CNN still classifies the images as some hand-written numbers with high confidence rates. For example, MNIST CNN classifies the natural images with high accuracy (around 97.5% in MNIST datasets with only one time iteration, the test dataset is 10,000 images), but at the same time they classifies the fooling images with 99.99% confidence.

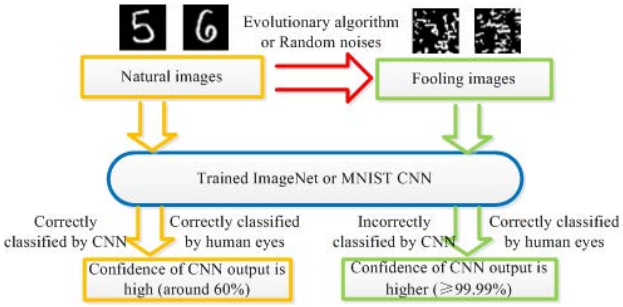

Fig. 8: The architecture of fooling images generation method.

Detailed experiments are shown in Section 3.1. Traditional CNNs cannot tell the differences between natural number images and the fooling number images. To address this problem, HOG which provides global structure features will be added in the procedure of traditional CNN feature extraction and classification. The method of generating fooling images is shown in Algorithm 1.

2.4. *The process of HCNN*

One of the structural information extraction algorithms that should be considered to refine traditional CNN is HOG. As Fig. 9 shows, the new network keeps the original local CNN texture features formation part and adds global structure features from HOG.

The numbers of features of HOG and CNN are various with different kinds of images. For colorful images in ImageNet datasets, after scaled by the method in Fig. 4 and Fig. 6, the number of features from CNN is 507 (i.e. 169 dimension features for each color channel, and the color channel is 3 for colorful images) which are good texture descriptions of images and the number of HOG features is 256 (i.e. $8\times8\times4$, 8 is the block size, 4 is the number of cells in a block) which are fine contour descriptions of images. For gray-scale images in MNIST dataset, the number of

Algorithm 1 The framework of forming fooling images on MNIST dataset.

Input: A natural image as source image I_s; The trained CNN CNN_t; The threshold of the confidence C_t;

Output: The final fooling image I_f;

1: Load the image I_s and trained CNN_t. Calculate the confidence of I_s:$Conf_s$;
2: Randomly select two rectangle area in I_s (the size of rectangle is d×d pixel, d is much smaller than the size of I_s) and replace them with evolutionary or random-noise image blocks. Save as the new image I_n;
3: Calculate the confidence of I_n by CNN_t, if its confidence is higher than $Conf_s$, update I_s to I_n;
4: Repeat Step 2 and Step 3 until the new confidence is bigger than C_t;
5: Output the I_n as I_f;
6: **return** I_f;

CNN feature is 96 (i.e. 4×4×6, 4 is the size of the last convolutional layers, 6 is the map size), and the number of HOG feature is 144 (i.e. 6×6×4, 6 is the block size, 4 is the number of cells in a block).

Two kinds of features (i.e. both CNN features and HOG features) are trained in a three layer neural network in HCNN. The number of hidden nodes in the classification layers of the HCNN is 3,000. Compared to the traditional CNN, HCNN successfully distinguished the fooling images. But at the same time, it shows a little decrease on accuracy (around 0.19%, and the details are shown in Section 3.3 and Section 3.4).

3. EXPERIMENTAL RESULTS AND ANALYSIS

3.1. *The classification performance of traditional CNNs on easily fooled images*

The problems for fooling images in CNN classification have been described.[9] In the classification strategy of CNN, the value of confidence for each candidate category is calculated respectively, then a comparison among these values of confidence is made and the category which has the maximum value of confidence will be selected as the output of the classification network. Here we use confidence rate to describe

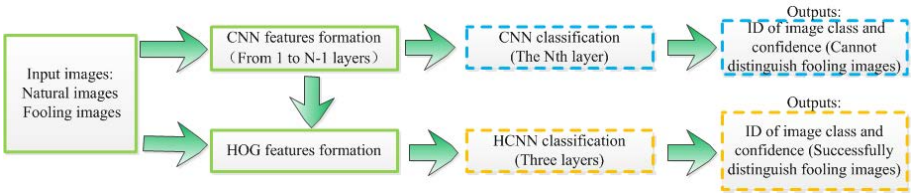

Fig. 9: The process of HCNN.

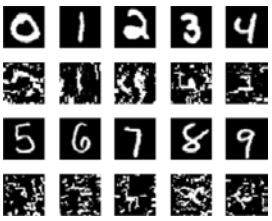

Fig. 10: MNIST Datasets and fooling images created by random noises method.

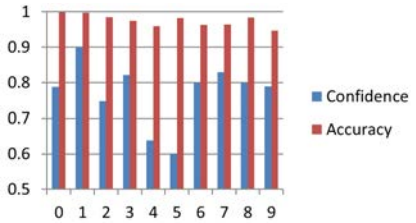

Fig. 11: The confidence rate and accuracy for natural images in MNIST dataset in traditional CNN classification.

how much the trained MNIST CNN is confident with the right classification result. For the natural images (e.g. the pictures in the first and third lines in Fig. 10), the trained MNIST CNN could recognize these kinds of hand-written number images with the accuracy of 97.5% on average (in Fig. 11). The confidence rates of the natural images are not very high (in Fig. 11). Fooling images are created by the procedure of fooling method described in Section 2.3 and Section 2.4 (e.g. the pictures in the second and fourth lines in Fig. 10 which are processed into human-unrecognized pictures from natural images). The confidence of these kinds of fooling images are very high (The confidence rate is 99.99%). Note that the structure for the trained MNIST CNN is the same one for classifying natural images and the fooling images.

This observation indicates that the traditional CNNs are very easily to be fooled by the fooling images.

3.2. *Comparative study among HCNN and other algorithms*

In order to show that in HCNN, the global structural information will contribute more than traditional point features (e.g. features obtained by SIFT and PCA-SIFT), we replace the HOG method in HCNN with point features method (e.g. SIFT and PCA-SIFT), and make comparison with these different methods, including SIFT-CNN, PCA-SIFT-CNN, HCNN and traditional CNN. The structure of the final three layers for classification in HCNN are not changed. The comparison results are shown in Fig. 12. All the experiments are made on a computer with Intel (i5-3470) processor with one unit of GPU (GT 640) and 16G RAM running the Linux operating system (Ubuntu 14.04). The results show that the HCNN, SIFT-CNN and PCA-SIFT-CNN methods are better than traditional CNN method from the perspectives of judging fooling images and scale-rotation flexibility. But HCNN and CNN show more advantages on judging natural images, illumination and execution time. Compared with HCNN, both SIFT-CNN and PCA-SIFT-CNN are mainly focusing on the point features, which causes the fact that they cannot judge fooling images well. The HCNN method shows more accuracy than other methods on

fooling images classification. Since the CNN features also contain some points and self-learnt angle features, HOG mainly extracts global structural features which will avoid the shortness of CNN features extraction. Combining the texture and structure features, the HCNN classification method will be more human-like and be sensitive with fooling images. Further detailed comparisons and experiments are shown in Section 3.3 and Section 3.4. These comparative studies indicate why we use HOG method to form additional structure information.

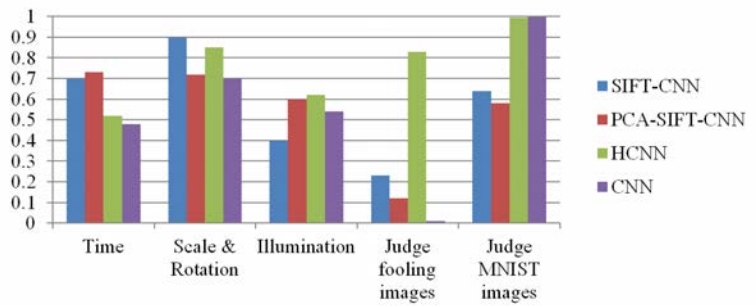

Fig. 12: Comparison results for different types of models. (The values in y axis for time is the execution time after normalization, and the base value of time is 0.1 second. The y axis for others are the accuracy of image classification.)

3.3. *Reduction of confidence on classifying fooling images through HCNN*

As Section 2.4 shows, additional structure information from HOG and original CNN texture information are combined. The classification result of HCNN (in Table 1) shows that the confidence of classification for the fooling images decreased. Some numbers are even classified incorrectly (e.g. number 5). The new HCNN shows much lower confidence properties about fooling images than the traditional CNNs, and on the other hand, HCNN obtains nearly the same accuracy on natural images with the traditional CNN.

3.4. *Retraining HCNN by adding a new class of fooling images*

(1) Experimental results on MNIST datasets

For MNIST datasets, besides the original ten classes, a new class which is composed of fooling images is added to the training datasets. Previous work[9] has used the CNN to retrain the new class. And the result was that the new retrained CNN could not classify the fooling images correctly. In this paper, we use new built HCNN to retrain the new fooling class, and the result is shown in Table 2. The result is that HCNN can classify both natural and fooling images correctly, which shows the

Table 1: The confidence decreased after adding structural information on MNIST. The bold numbers of each lines are top two confidence results.

Class	0	1	2	3	4	5	6	7	8	9
0	**0.62**	0.01	0.04	0	0.02	0.01	**0.5**	0.01	0	0.01
1	0	**0.7**	0	0.01	0.1	0.05	0.01	**0.25**	0	0.01
2	0.01	0	**0.49**	**0.33**	0	0	0.01	0	0	0
3	0	0	0	**0.34**	0.01	0	**0.3**	0	0	0
4	0	**0.1**	0	0.01	**0.693**	0.01	0.02	0.02	0	0
5	0	0	0	0	0.02	**0.29**	**0.41**	0	0	0.01
6	0.01	0	**0.4**	0	0	0	**0.59**	0	0	0
7	0.01	0.01	0	0.06	0	0	0.04	**0.79**	**0.2**	0.03
8	0.01	0	0	0	**0.47**	0.01	0	0	**0.49**	0
9	0.001	0	0	0	0	**0.3**	0	0.05	0	**0.52**

effectiveness and feasibility of the proposed method. The most important point is that HCNN does not decrease the classification result of natural images very much. For 10,000 hand-written images in MNIST test datasets, the accuracy is 99.56% (just a little decrease compared with the accuracy of original MNIST CNN which is 99.75%).

Table 2: HCNN confidence result after retraining a new class for fooling images.

	0	1	2	3	4	5	6	7	8	9	New
0	0	0.01	0.04	0	0.08	0.05	0.01	0	0	0.02	**0.78**
1	0.01	0	0.07	0.09	0.06	0.08	0.01	0	0	0.08	**0.87**
2	0.01	0.01	0.03	0.03	0	0.1	0	0	0	0.1	**0.45**
3	0.06	0.01	0.1	0.07	0	0.08	0	0	0.01	0.03	**0.64**
4	0.03	0.02	0.07	0.07	0.04	0.01	0.03	0.09	0	0.01	**0.34**
5	0.01	0.06	0.09	0.02	0.03	0.01	0.05	0	0.03	0.07	**0.87**
6	0	0.07	0.06	0.08	0.03	0.01	0.01	0	0.01	0	**0.53**
7	0.01	0.02	0.08	0.08	0	0.01	0.06	0	0	0.09	**0.86**
8	0.05	0.05	0.02	0.07	0.03	0.01	0.01	0	0	0	**0.72**
9	0.01	0.04	0.04	0.08	0	0.01	0.01	0	0.08	0	**0.85**

(2) Experimental results on ImageNet datasets

For ImageNet datasets, we select 40 main pictures as the testing images (in Fig. 13), just the same with the paper which creates the fooling images.[9] After training the 1,000 categories of ImageNet images, CNN features from traditional CNN are saved. HOG structural features are extracted from the selected 40 items from ImageNet datasets which show the best contour features of that image (in Fig. 14). Through combination of both the traditional CNN features and HOG features, the new ImageNet HCNN is trained.

To test if the new ImageNet HCNN could recognize the fooling images, special fooling images in Fig. 15 is created. As Fig. 16 shows, the classified confidence of

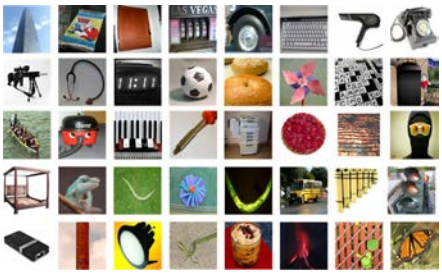

Fig. 13: The selected 40 source images from ImageNet datasets.

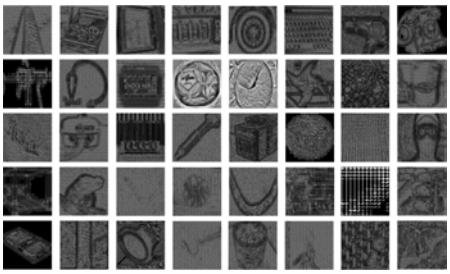

Fig. 14: HOG features of the 40 selected source images from ImageNet datasets.

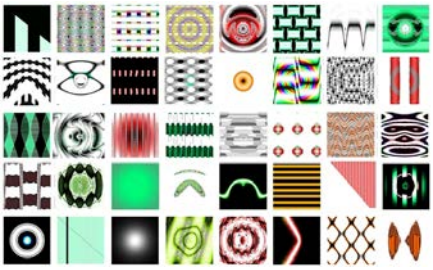

Fig. 15: The 40 fooling images of Fig. 13 from ImageNet datasets.

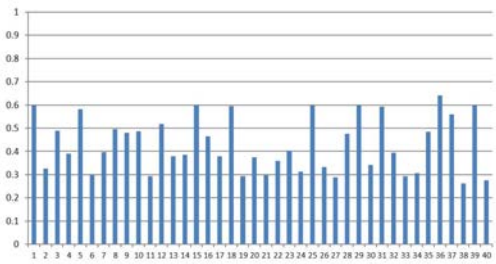

Fig. 16: The confidence values of 40 fooling images are decreased in HCNN.

the HCNN to fooling images is decreased (i.e. the confidence of the original CNN is 99.99%). For most of the testing images, the confidence is decreased more than 50%. It indicates that HCNN could reduce the confidence for fooling images significantly. Combining structural information with the traditional CNN classification will greatly improve the ability of judging fooling images.

4. DISCUSSION

This paper attempts to avoid the shortness of traditional CNNs that they only focus on the texture features and relatively lack of global structural information, which causes the fact that they cannot distinguish fooling images from natural images. The experimental results show that the proposed HCNN is more similar with the human classification results. The main reason is that HCNN focuses on more structural information than the traditional CNNs. However, even with the significantly decreased confidence of the fooling images, HCNN still makes the accuracy of classification a little decrease. In the future, more feasible and reasonable features should be taken into account for designing new networks in order to optimize both confidence and accuracy.

At the same time, for the 1,000 classes of natural images, the intrinsic property

of inner classes is not considered too much. For example, an apple after a little modification should be more likely to be considered as a pear instead of a book, because apple is closer to pear than book on semantic relations. The relationship between two kinds of semantic classes should also contribute to the classification results.

The existing deep neural networks have shown their stellar results on object detection and recognition. In fact, these kinds of networks are "classification model", which are very different from the network of human visual system which should be a "generative model". Adding features from HOG to trained CNN is just like making the combination of features from the upper two types of models. Thus the classification procedure and results will be more similar to human visual system.

5. CONCLUSION

In this paper, we firstly make some introduction about modern CNNs and traditional computer vision methods like SIFT and HOG which are related to our work. According to the shortness of CNNs which cannot recognize the fooling images, a new neural network named HCNN is proposed which combines the features from two parts: texture features in traditional CNNs and structural features in HOG method. The new network is much more sensitive to the fooling images, even with just a little decrease on the classification accuracy of natural images. Experiments on MNIST and ImageNet datasets are conducted for comparative studies which show the efficiency and accuracy of the proposed new method.

HCNN obtains approximately similar results for recognizing natural images and much better performance for recognizing fooling images. It can be considered as an attempt towards the goal of realizing human-level image classification.

In our future work, we will try and test HCNN on some specific application domains where CNN is applicable but still remains some limitations, while HCNN can reduce some of them. In addition, HCNN should be refined to keep the same or higher accuracy for natural images compared with traditional CNN.

6. ACKNOWLEDGEMENTS

This paper is supported by research grants from the Strategic Priority Research Program of the Chinese Academy of Sciences, and Beijing Municipality of Science and Technology.

References

1. Geoffrey E Hinton. Learning multiple layers of representation. *Trends in cognitive sciences*, 11(10):428–434, 2007.
2. Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

3. Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.
4. David G Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.
5. Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2005)*, volume 1, pages 886–893. IEEE, 2005.
6. Lawrence M Ward. Determinants of attention to local and global features of visual forms. *Journal of Experimental Psychology: Human Perception and Performance*, 8(4):562, 1982.
7. Kevin Murphy, Antonio Torralba, Daniel Eaton, and William Freeman. Object detection and localization using local and global features. In *Toward Category-Level Object Recognition*, pages 382–400. Springer, 2006.
8. Chi-Ren Shyu, CE Brodley, AC Kak, Akio Kosaka, A Aisen, and L Broderick. Local versus global features for content-based image retrieval. In *Proceedings of the IEEE Workshop on Content-Based Access of Image and Video Libraries*, pages 30–34. IEEE, 1998.
9. Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015)*, pages 427–436, 2015.
10. Yan Ke and Rahul Sukthankar. Pca-sift: A more distinctive representation for local image descriptors. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2004)*, volume 2, pages II–506. IEEE, 2004.
11. Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.
12. Jia Deng, Wei Dong, Richard Socher, Lijia Li, Kai Li, and Feifei Li. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2009)*, pages 248–255. IEEE, 2009.
13. Yann LeCun and Corinna Cortes. The mnist database of handwritten digits, 1998.
14. Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 2014 ACM International Conference on Multimedia*, pages 675–678. ACM, 2014.
15. Pedro Felzenszwalb, David McAllester, and Deva Ramanan. A discriminatively trained, multiscale, deformable part model. In *Proceedings of The 2008 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2008)*, pages 1–8. IEEE, 2008.
16. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.
17. Anuj Mohan, Constantine Papageorgiou, and Tomaso Poggio. Example-based object detection in images by components. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(4):349–361, 2001.
18. Paul Viola, Michael J Jones, and Daniel Snow. Detecting pedestrians using patterns of motion and appearance. In *Proceedings of the 9th IEEE International Conference*

on Computer Vision (ICCV 2003), pages 734–741. IEEE, 2003.
19. Dario Floreano and Claudio Mattiussi. *Bio-inspired artificial intelligence: theories, methods, and technologies*. MIT press, 2008.
20. Dumitru Erhan, Yoshua Bengio, Aaron Courville, and Pascal Vincent. Visualizing higher-layer features of a deep network. Technical Report 1341, University of Montreal, June 2009.
21. Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.