Hand-drawn sketch recognition with a double-channel convolutional neural network

Lei Zhang

Correspondence: 201127001@cqcet.edu.cn
Digital Media College of Chongqing College of Electronic Engineering, Chongqing 401331, China

Abstract

In hand-drawn sketch recognition, the traditional deep learning method has the problems of insufficient feature extraction and low recognition rate. To solve this problem, a new algorithm based on a dual-channel convolutional neural network is proposed. Firstly, the sketch is preprocessed to get a smooth sketch. The contour of the sketch is obtained by the contour extraction algorithm. Then, the sketch and contour are used as the input image of CNN. Finally, feature fusion is carried out in the full connection layer, and the classification results are obtained by using a softmax classifier. Experimental results show that this method can effectively improve the recognition rate of a hand-drawn sketch.

Keywords: Hand-drawn sketch recognition, Multi-channel, Convolution neural network, Deep learning, Double-channel CNN

1 Introduction

As a very important field of hospital cultural image design, popular science publicity can better promote the construction of hospital culture. The high-quality hand-painted popular science images reflect the spread of high-quality popular science. The application of this method in the recognition of medical hand-painted popular science sketches can better spread health knowledge to the public.

With the popularity of portable touch devices, the application of hand sketching is more and more diversified. There are more and more researches on a freehand sketch, including sketch recognition [1–3], image retrieval based on freehand sketch [4, 5], and 3D model retrieval based on freehand sketch [6].

Hand-drawn sketch recognition is still a very challenging problem, which can be attributed to the following [7]: (1) a hand-drawn sketch has highly abstract and symbolic attributes; (2) due to the different painting level and ability of each person, the same kind of objects may have great differences in shape and abstract degree; and (3) a hand-drawn sketch lacks visual clues, color, and texture information.

The early recognition of a hand-drawn sketch mainly followed the traditional image classification model. In other words, manual features are extracted from the sketch

and sent to the classifier. General hand features include shape context feature [8], scale-invariant feature transform [9], and directional gradient histogram feature [10], but these hand features designed for natural images are not suitable for abstract and sparse hand-drawn sketches. The fusion of different local features by multi-core learning helps to improve the recognition performance, which was proved by Li et al. [11]. Fisher vector (FV) is applied to the recognition of a hand-drawn sketch, and a high recognition rate is obtained [12].

In recent years, deep learning in the field of machine learning has developed rapidly. The essence of general deep learning is a nonlinear network model with multiple hidden layers. Through the training of large-scale original data, we can extract the characteristics of the original data from the network model and predict or classify the samples. In the field of image recognition and computer vision, CNN has achieved the most remarkable results [13]. In addition, deep learning has been widely used in pedestrian detection [14], gesture recognition [15], natural language processing [16], data mining, and speech recognition. Compared with other deep neural networks such as deep belief network [17] and S-layer automatic coding [18], CNN can directly process two-dimensional images. When a two-dimensional image is converted into a single image, the spatial structure of the input data will be lost. With the development of deep learning, some deep learning models for sketch recognition have been proposed, such as VGg [19], RESNET [20], and Alex net [21]. However, these deep learning models are mainly designed for color texture natural images. Due to the lack of color and texture information in a hand-painted sketch, they are not suitable for hand-painted sketch recognition. In reference [22], the user is required to draw semantic symbols by using an explicit prompt and then click the button. In reference [23], the use of a time threshold requires users to have a clear pause after drawing semantic symbols. In addition, special graphical symbols (such as arrows) are used for grouping [24]. The constraints of these algorithms weaken the natural rendering features of the hand-drawn interface and limit the ability of rapid expression and modeling.

CNN's multi-channel mechanism is used to access different data views (such as red, green, blue channels of color images, and stereo audio tracking) [25]. By adding the input information, CNN can learn more features and improve the classification effect of the model. Therefore, in order to improve the recognition rate of sketch recognition, this paper also uses the sketch contour as the input data of CNN and proposes a dual-channel CNN.

2 Introduction

In this section, firstly, we will introduce the CNN. Secondly, we present the contour extracting method of the hand-drawn sketch. The proposed algorithm will be introduced at last.

2.1 Convolution neural network

CNN is an algorithm with less human intervention. The traditional BP neural network is used for reference in the process of weight updating. Error backpropagation is used to update parameters automatically. Due to the lack of human intervention, CNN can directly take the image as input and automatically extract image features for

recognition. The weight sharing and local sensing characteristics of CNN not only reduce the number of parameters in the network, but also work in a way similar to that of animal visual nerve cells. The recognition accuracy and efficiency of the network are greatly improved.

CNN has two typical characteristics. The first is the local connection between the two layers of neurons through the convolution nucleus, rather than the complete connection. Therefore, the convolution layer connected to the input image is a local link constructed for pixel blocks, rather than the traditional pixel-based full connection. Secondly, the weight parameters of the convolution kernel are shared in the same layer. These two features greatly reduce the number of parameters of the deep web, reduce the complexity of the model, and accelerate the training speed. This makes CNN have a great advantage in the pixel value of the processing unit. The main components of CNN include convolution layer, pool layer, activation function, full connection layer, and classifier, as shown in Fig. 1.

2.2 Convolution layer

The convolution layer is the most important network layer in CNN feature extraction. The convolution operation is the process of obtaining a new feature map by convolution kernel and input sample image or upper output feature map under the action of the activation function. At each level, there are multiple feature maps, which represent a feature of the image. Convolution operation can be expressed as follows.

$$x_j^l = f\left(\sum_{i \in M_j} x_j^{l-1} * k_{kj}^l + b_i^l \right) \tag{1}$$

Each level in the CNN has multiple feature maps. Suppose the *jth* feature map of the layer l is x_j^l, where $f(\cdot)$ represents the activation function which will be described in more detail in the following chapters. M_j represents the input sample image or the set of all the input feature graphs, and k_l^{ij} represents the convolution kernel in the layer l, and the convolution is expressed as $*$. After the convolution operation, we need to add the bias b after the result, then the new feature graph is formed by the activation function.

The inverse error propagation algorithm is used to update CNN weight. And the first step in the update process is to calculate the gradient at each level.
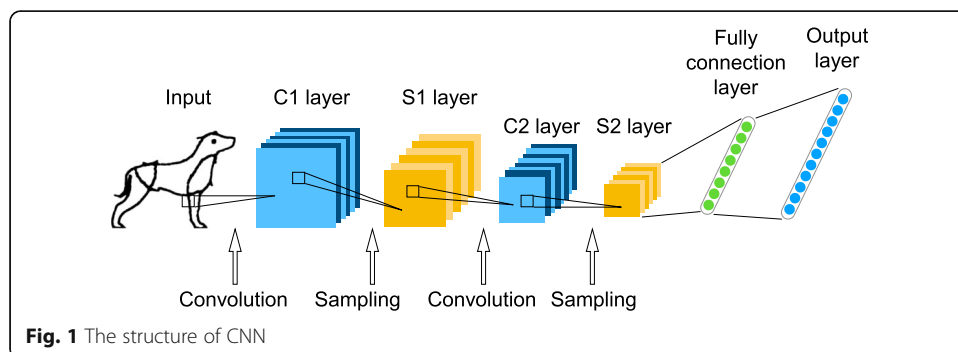

Fig. 1 The structure of CNN

2.2.1 Downsampling and pooling layer

The lower sampling layer is the process of feature extraction. By reducing the dimension of the image feature graph, it is usually called pooling. In the process of downsampling, the dimension of the upper layer of the feature graph is reduced to obtain the feature graph satisfying the one-to-one correspondence. Therefore, n output characteristic maps of the upper convolution layer are used as the input of the lower sampling layer.

After dimension reduction, n output feature graphs are obtained. The following formula represents the process of downsampling and merging.

$$x_j^l = f\left(\beta_j^l down\left(x_j^{l-1}\right) + b_j^l\right) \tag{2}$$

where $down(\cdot)$ represents the downsampling operation. The pixel value in $n \times n$ region of the input feature graph is selected to obtain a value in the output feature graph. The dimensions of the input feature graph are reduced by n times in both horizontal and vertical directions. The final value of the pixels in the output feature graph is also related to the multiplier offset β and the additional offset b. After the activation function, the final pixel value is obtained.

2.2.2 Fully connection layer and softmax classifier

The whole join layer is usually connected to the pool layer and the last layer of the classifier to fuse different features represented by multiple feature graphs.

Each neuron in the full connectivity layer is connected with all the neurons in the bottom layer and has output characteristics.

The full join layer combines all the features of the previous features and then inputs them into the softmax classifier.

The input sample image is convoluted and downsampled layer by layer to get a relatively complete feature set. These features need to be classified by a classifier to get the predictive value of the sample image category. Then, the difference between the predicted value and the actual value is obtained. The input sample image is sent back by gradient-based algorithm to train the whole neural network. Generally, the last layer of downsampling cannot be directly connected to the classifier, and the dimension transformation can only be used as the input of the classifier after one or two layers are completely connected. Softmax classifier is usually used in CNN.

Softmax classifier is suitable for multi-classification, and its prototype is a logistic regression model for binary classification. In logistic regression, assuming the sample category label is $y(y = 0$ or $y = 1)$. There are m data samples $\{(x_1, y_1), (x_1, y_1), ..., (x_m, y_m)\}$ and the input characteristic $x^{(i)} \in R^{n+1}$ of these samples. The category label of the sample is 0 or 1, that is, $y^{(i)} \in \{0, 1\}$, then its hypothetical function is shown as follows:.

$$h_g(x) = \frac{1}{1 + \exp\left(-\theta^T x\right)} \tag{3}$$

where θ is an important parameter and θ can constitute the cost function. By adjusting parameter θ to minimize the cost function, the predicted category of the input sample can be obtained.

2.2.3 The training process

There are three main ways of CNN training, namely, full supervision, full nonsupervision, and the combination of supervision and nonsupervision. This paper adopts the supervised learning method. Supervised learning is trained on neural networks in the form of supervised signals. The supervised signal is the true value of the classification in each sample. In the learning process, CNN learns and extracts the features of the input image, and gives the predictive value of the sample classification at the output end. CNN backpropagates the difference between the predicted value and the actual value to continuously adjust the network parameters. Finally, it enables all input classes on the network to make correct image samples.

2.3 Proposed algorithm

In this part, we first introduce the method of drawing sketch outlines. Then, a dual-channel CNN is proposed.

The extraction process of sketch contour features is shown in Fig. 2. First, the input sketch is preprocessed to obtain a smooth sketch. Then, the outline of the sketch is extracted.

Due to the characteristics of hand drawing, there will inevitably be two overlapping areas, in which there are redundant unclosed curve segments. In order to get a smooth sketch outline, it is necessary to preprocess the input sketch.

The algorithm of eliminating the unclosed curve segment can be described as follows:.

(a) Scan the picture according to the direction of the line. If a point is found to belong to the curve endpoint, then turn to (b). If the whole picture still does not have the curve endpoint, then exit. The curve endpoint can be judged when one point in a 3×3 area. If there is no other point present in the eight directions at this point, which is an isolated point, then it belongs to the curve endpoint. If there is only one direction that has a point in eight directions, which is a little bit, then it is a curve endpoint. If there are three directions and more than three directions have a point in eight directions, then it is a curve endpoint. If there are two directions that have a point in the eight directions, and the two directions are adjacent, then it is a curve endpoint; otherwise, it is not.

(b) Find the endpoint of the curve and eliminate this endpoint. Then, determine whether the point adjacent to this endpoint is the curve endpoint. If it is the curve endpoint, then continue to eliminate this point and determine the next adjacent point. If it is not, then go to (a).

For the hand-drawn sketch, an adaptive tracking algorithm based on the direction of the eight-connected domain is used to extract the contour of the sketch. The original image is represented by $I(x, y)$, $C(x, y)$ represents a 2-value image of the contour. The current direction is D_i, starting from the right side and starting from the counterclockwise direction of 0,1,... 7. Select a point $point_c$ at the left of the top line of the image as the first point. In the 3×3 area at this point, there are no other points in the upper, left, and right directions. Then, $d_i = 2$ is selected and look for the contour of the sketch in a counterclockwise direction. The specific algorithm is described as follows:

Step 1. Initialization. Set C to zero and $d_i = 2$, then the direction array DI is set to $DI= \{0, 1, 2, 3, 4, 5, 6, 7, 0, 1, 2, 3, 4, 5, 6, 7\}$.
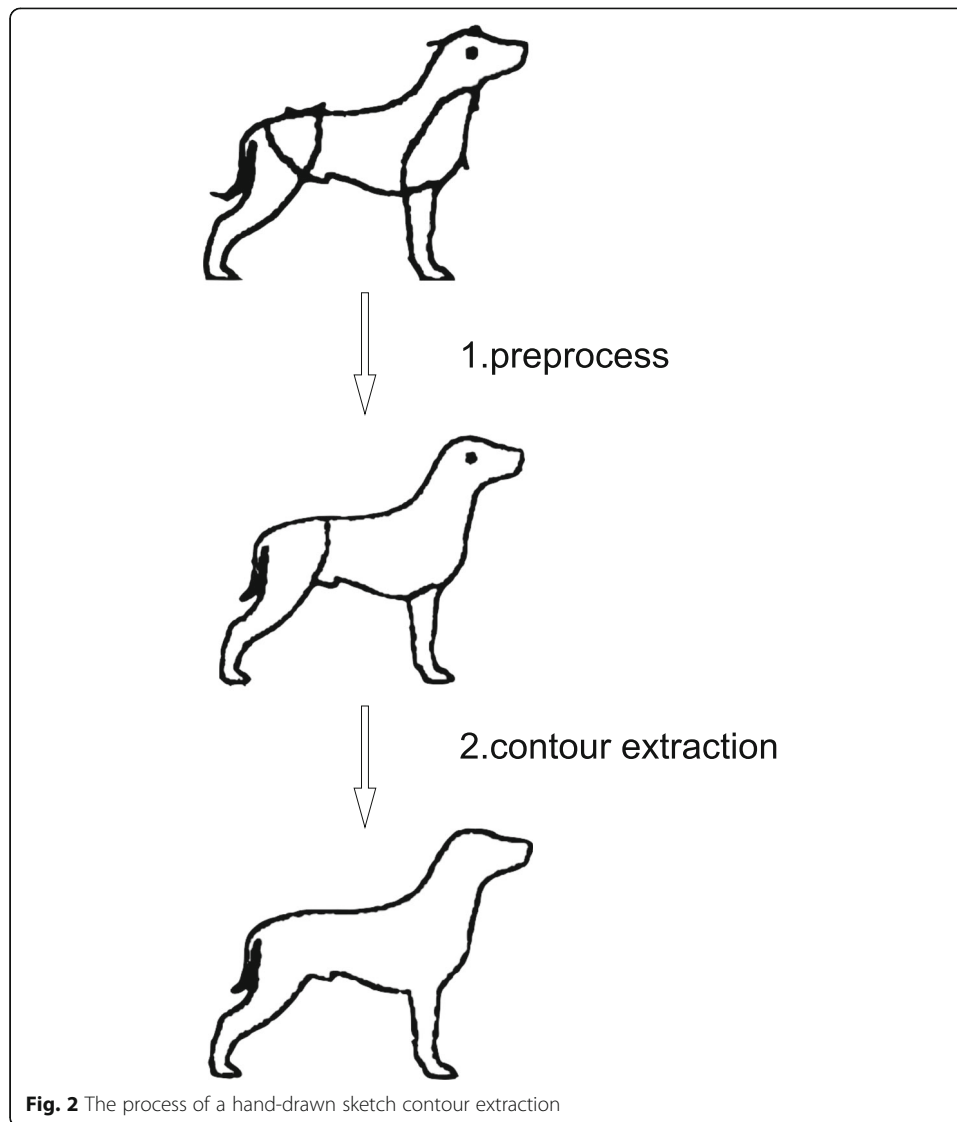
Fig. 2 The process of a hand-drawn sketch contour extraction

Step 2. Scan the original image I line by line from top to bottom, left to right, then the starting point of the contour that is $point_c$ can be obtained. The current point $point_{now}$ is initialized to $point_c$.

Step 3. Add the current point $point_{now}$ to the binary image $C(x, y)$ of the contour. Search $point_{now}$ in the order of $DI[d_i]$, $DI[d_{i+1}]$, …, $DI[d_{i+7}]$, to find the next adjacent boundary point $point_{next}$. If a point in one direction is found to belong to the original image I and is not equal to the initial point $point_c$, then this point is $point_{next}$. If the direction $DI[i]$ of the next point is found, then the new search direction d_i is the next direction in the opposite direction of $DI[i]$, that is $d_i = (DI[i] + 4 + 1)$ mod8. Then, assign the value of $point_{next}$ to $point_{now}$.

Step 4. If the point $point_{now}$ coincides with the starting $point_c$, then exit. Otherwise, it should be returned to step 3.

The effect of the algorithm is shown in Fig. 2, which has good robustness. The smooth contour of the input sketch can be obtained by preprocessing the sketch.

CNN's multi-channel mechanism is used to access different data views, such as red, green, blue channels, and stereo audio tracks of color images [26]. By adding the input information, CNN can learn more features and improve the classification effect of the model. Therefore, in order to optimize the training process of a hand-drawn bone recognition, this paper proposes a dual-channel CNN. Figure 3 shows the structure of the network. The network consists of two relatively independent convolution networks. The first input of the network is the hand-drawn image, and the second input is the outline of the hand-drawn sketch [27–35].

In a dual-channel CNN, each channel contains the same number of convolution layers and parameters, but has independent weights. After the pool layer, the two channels are connected to the full connection layer and perform the full connection mapping. The two channels are connected to a fully connected hidden layer, which generates the output of the logistic regression classifier. Each channel's weight has its own update. But the final error is obtained through two output layers. So, the two output layers are like a layer that deviates from each other.

Fig. 3 The structure of double-channel CNN

Fig. 4 The samples of hand-drawn sketches
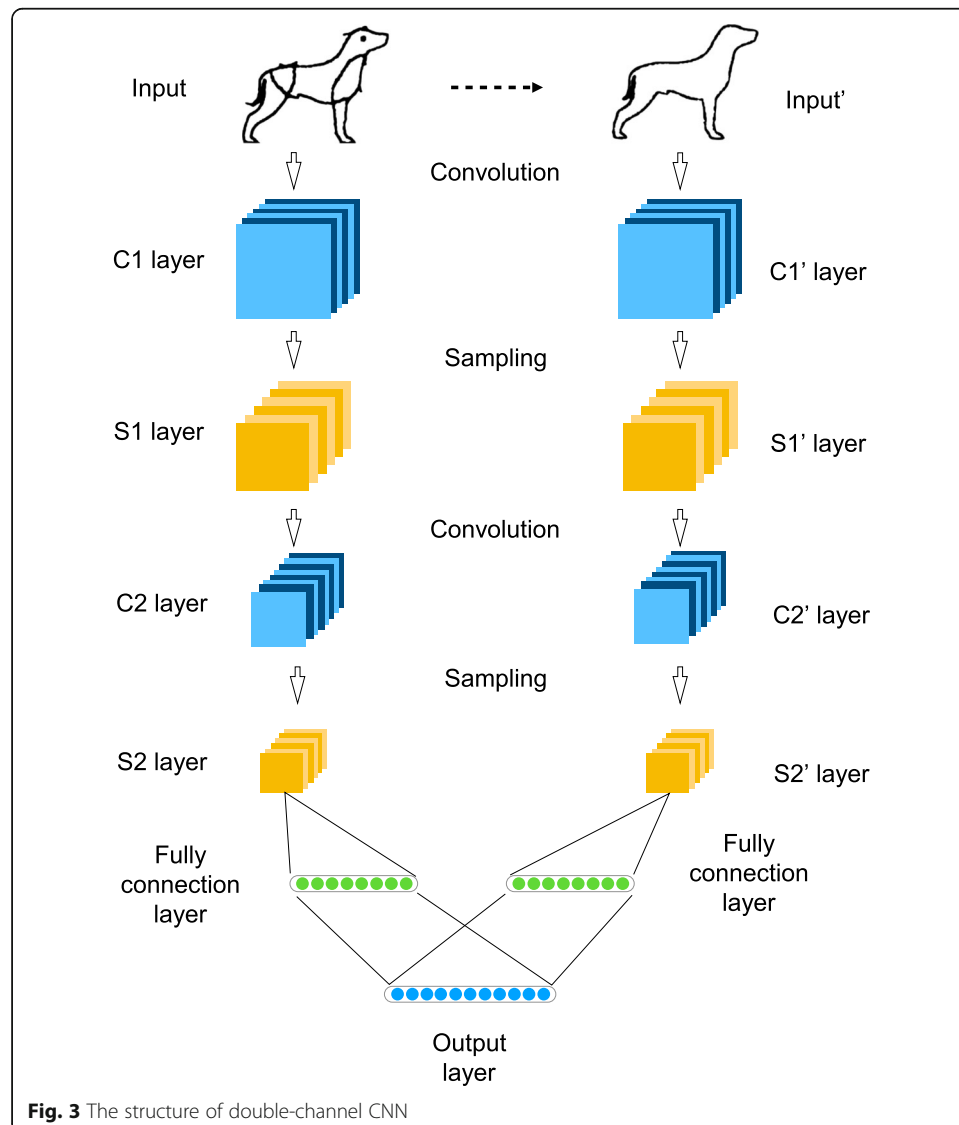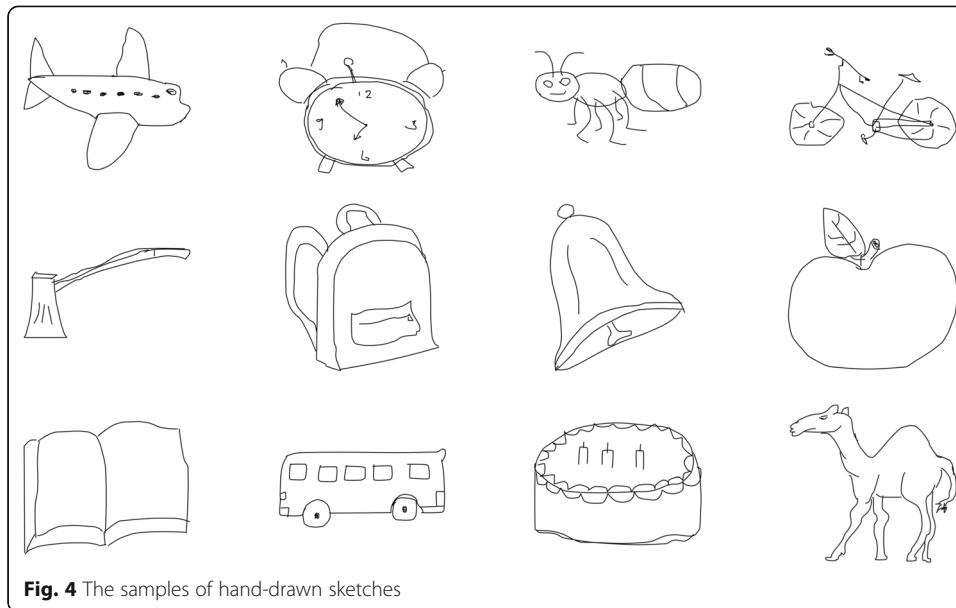
3 Results and discussion

3.1 Experimental preparation

In this experiment, the configuration of the computer is as follows. Windows 7, 3.60GHz, i7 processor, 32GB ddr, and 1024GB hard disk. The software of the experiment is Matlab 2017a.

In 2012, Eitz et al. [1] organized and collected a collection of the largest hand-sketched sketch, it contains 250 hand-drawn sketches, and each containing 80 different hand-drawn sketches. The original pixel size of the sketch is 1111×1111, as shown in Fig. 4. In the experiment, 4 fold cross-validation was used, three for training and one for testing. The evaluation index of this experiment is the recognition rate of all test samples.

3.2 Experimental results and analysis

Deep learning requires a large amount of training data, and the lack of training data tends to create an over-fit problem. In order to reduce the influence of over-fitting, this paper makes a manual expansion of the hand-drawn sketch data set used in the experiment and obtains a new amplified data set. Specific steps are as follows:

Table 1 The comparison of different methods

Method	Reference
HOG-SVM	[1]
SIFT-Fisher	[2]
MKL-SVM	[10]
FV-SP	[2]
Alex-Net	[11]
Proposed	/

Table 2 The comparison of recognition rate

| Method	Reference	Recognition rate%
HOG-SVM	[1]	56.54
SIFT-Fisher	[2]	63.26
MKL-SVM	[10]	67.34
FV-SP	[2]	70.04
Alex-Net	[11]	68.14
Proposed	/	73.24

Step 1. Dimension reduction. Reduce all the hand-painted sketch images from the original size of 1111×1111 to 256×256.

Step 2. Extract the slices. From the 256×256 diagram, select five slices of the center, upper left corner, lower left corner, upper right corner, and lower right corner, which size is set to 225×225. In the resulting five slices, the original dataset is made up of all 225×225 slices of pixel size in the center.

Step 3. Flip horizontally. Take the five slices obtained by step 2 and flip them horizontally, and five new slices are taken again. The 10 slices of each sample obtained by step 2 and step 3 constitute the amplified data set, so the data volume of the amplified data set is 10 times of the original data set.

The proposed algorithm in this paper is compared with some other popular sketch recognition methods, such as HOG-SVM [1], SIFT-Fisher [2], MKL-SVM [10], FV-SP [2], and Alex Net\cite [11]. The experimental results are shown in Table 1. Compared with traditional non-deep learning methods, HOG-SVM, SIFT-Fisher, MKL-SVM, and FV-SP, the recognition rates of the proposed algorithm are 16.1, 9.98, 5.9, and 3.2, respectively. The results show that the depth learning method has a stronger feature and nonlinear expression than the non-depth learning method. Compared with the depth

Fig. 5 The comparison of training time

**Fig. 6** The comparison of recognition rate
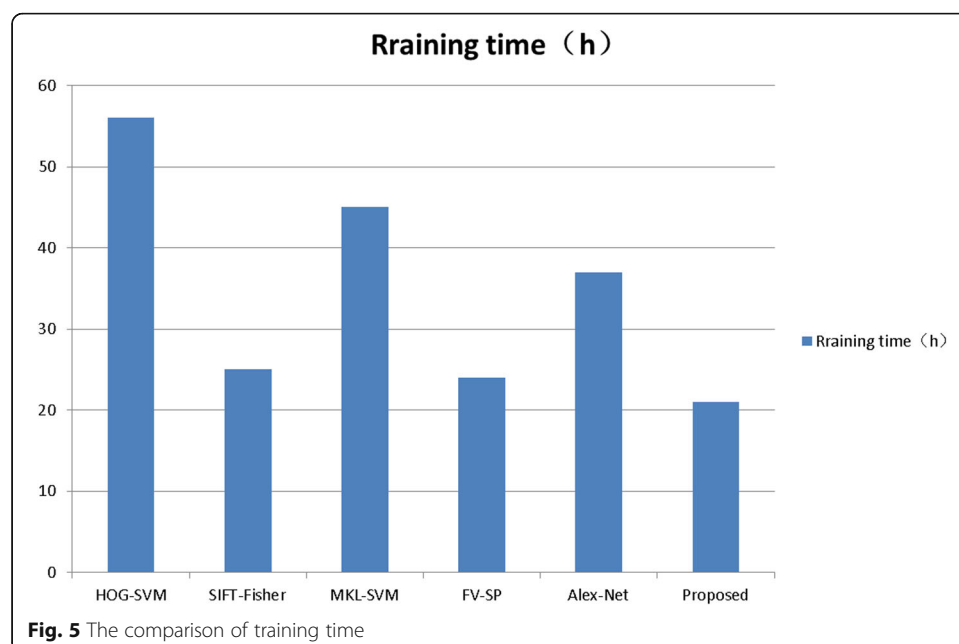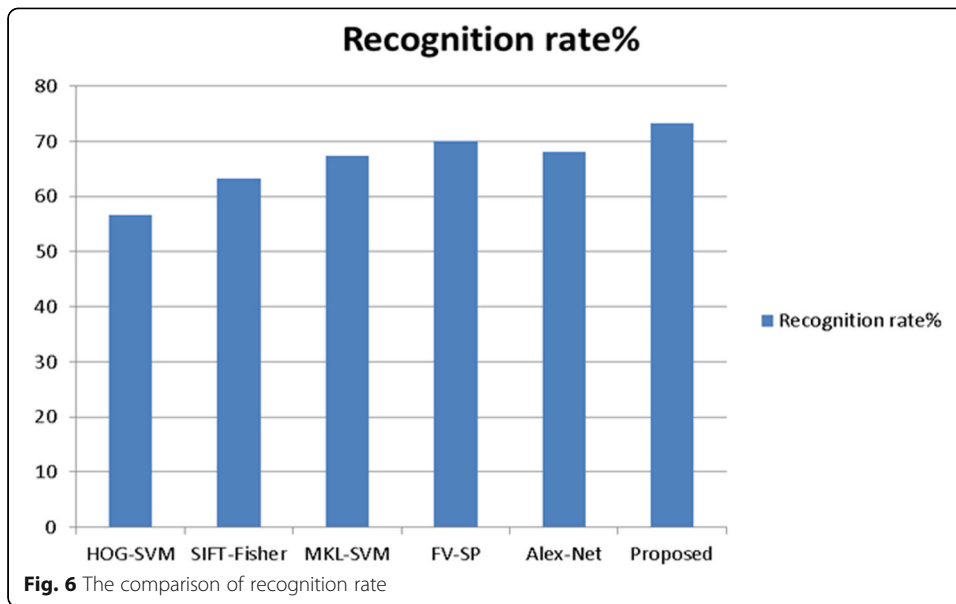
of the classical learning method Alex-Net, the accuracy rate is improved by 5.1. The results show that the proposed algorithm, namely a double-channel CNN, can help improve the recognition rate of hand-drawn sketches (Table 2).

The results are shown in Fig. 1 to Fig. 5, and the comparison between our method and other methods on the COAD dataset (Figs. 6 and 7).

Our method is superior to the other methods in training time, recognition accuracy, and energy consumption

Fig. 7 The comparison of energy consumption
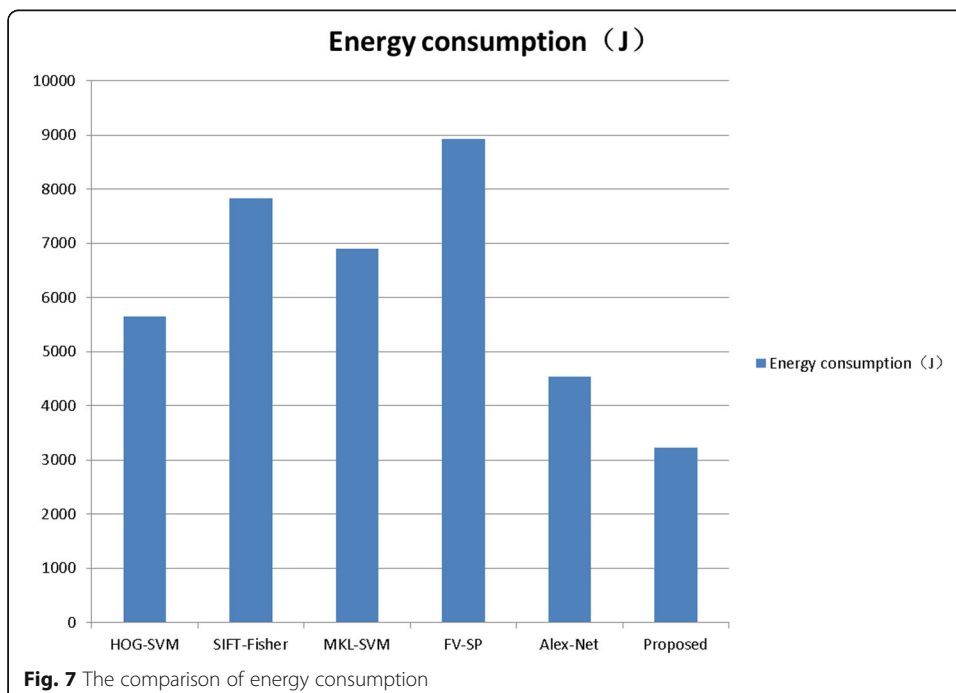
4 Conclusion and future work

In order to improve the recognition rate of a hand-drawn sketch recognition, a hand-drawn sketch recognition algorithm based on a dual-channel convolution neural network is proposed. Firstly, the sketch is preprocessed to extract the contour information. Secondly, sketch and contour are used as two input channels of a convolutional neural network. Finally, a softmax classifier is used for feature fusion in the full connection layer to get the classification results. The experimental results show that the proposed method achieves a higher recognition rate than the existing mainstream sketch recognition methods.

Our future work is to improve the recognition accuracy of handwriting input.

Author's contributions
Lei ZHANG, as the primary contributor, completed the analysis, experiments, and paper writing. The author read and approved the final manuscript.

Authors's information
Lei Zhang, Associate Professor, Chongqing Vocational College of Electronic Engineering.
Master of Arts, Chongqing University.
Main research direction: graphic image and art design research

Declarations

Ethics approval and consent to participate
Not applicable

Consent for publication
Not applicable

Competing interests
The authors declare that they have no competing interests.

References

1. M. Eitz, J. Hays, M. Alexa, How do humans sketch objects. *ACM Trans. Graph.* **31**(4) (2012)
2. P. Zhao, Y. Liu, H. Liu, S. Yao, A sketch recognition method based on deep convolutional-recurrent neural network. *Journal of Computer-Aided Design & Computer Graphics* **30**(2), 217–224 (2018)
3. O. Seddati, S. Dupont, S. Mahmoudi, in *Content-Based Multimedia Indexing (CBMI), 13th International Workshop on.* Deepsketch: deep convolutional neural networks for sketch recognition and similarity search (*IEEE*, 2015), pp. 1–6
4. S. Liang, Z. Sun, Sketch retrieval and relevance feedback with biased SVM classification. *Pattern Recogn. Lett.* **29**(12), 1733–1741 (2008)
5. M. Eitz, K. Hildebrand, T. Boubekeur, Sketch-based image retrieval: benchmark and bag-of-features descriptors. *IEEE Trans. Vis. Comput. Graph.* **17**(11), 1624–1636 (2011)
6. B. Li, Y. Lu, C. Li, *SHREC'14 track: extended large scale sketch-based 3D shape retrieval, Eurographics workshop on 3D object retrieval* (2014)
7. Yu Q, Yang Y X, and Song Y Z, Sketch-a-net that beats humans, http://arxiv.org/abs/1501.07873v3, 2017
8. G. Carneiro, A. Jepson, D, Pruning local feature correspondences using shape context, *Proceedings of the 17th International Conference on Pattern Recognition. Los Alamitos: IEEE Computer Society Press* **3**, 16–19 (2004)
9. D.G. Lowe, Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **60**(2), 91–110 (2004)

10. X. Zheng, H. Tan, Z. Ma, Performance comparison of improved HOG, Gabor and LBP. *Journal of Computer-Aided Design & Computer Graphics* **24**(6), 787–792 (2012)
11. Y. Li, T.M. Hospedales, Y.Z. Song, Free-hand sketch recognition by multi-kernel feature learning. *Comput. Vis. Image Underst.* **137**, 1–11 (2015)
12. R.G. Schneider, T. Tuytelaars, Sketch classification and classification-driven analysis using Fisher vectors. *ACM Trans. Graph.* **33**(6) (2014)
13. L.W. Jin, Z.Y. Zhong, Z. Yang, Applications of deep learning for handwritten Chinese character recognition: a review. *Acta Automat. Sin.* **42**(8), 1125–1141 (2016)
14. V. John, S. Mita, Z. Liu, in *Proceedings of the 14th IAPR International Conference on Machine Vision Applications on IEEE*. Pedestrian detection in thermal images using adaptive fuzzy C-means clustering and convolutional neural networks (2015), pp. 246–249
15. J. Cai, J.Y. Cai, X.D. Liao, Preliminary study on hand gesture recognition based on convolutional neural network. *Computer Systems & Applications* **24**(4), 113–117 (2015)
16. Y. Goldberg, Neural network methods for natural language processing. *Synthesis Lectures on Human Language Technologies* **10**(1), 1–309 (2017)
17. G.E. Hinton, R.R. Salakhutdinov, Reducing the dimensionality of data with neural networks. *Science* **313**(5786), 504–507 (2006)
18. M.A. Ranzato, C. Poultney, S. Chopra, in *Proceedings of the 2007 Advances in Neural Information Processing Systems*. Efficient learning of sparse representations with an energy-based model (*MIT Press, USA*, 2007), pp. 1137–1144
19. Simonyan K, and Zisserman A, Very deep convolutional networks for large-scale image recognition, http:// arxiv.org/a bs/1409.1556v6, 1th, March, 2017
20. He K M, Zhang X Y, and Ren S Q, Deep residual learning for image recognition, http://arxiv.org/abs/1512.03385v1, 1th, March, 2017
21. A. Krizhevsky, I. Sutskever, G.E. Hinton, in *Proceedings of the 25th International Conference on Neural Information Processing Systems*. ImageNet classification with deep convolutional neural networks (*MIT Press, Cambridge*, 2012), pp. 1097–1105
22. T. Kurtoglu, T.F. Stahovich, in *Proc. of AAAI Spring Symposium on Sketch Understanding*. Interpreting schematic sketches using physical reasoning (*AAAI Press, Palo Alto, USA*, 2002), pp. 78–85
23. M. Fonseca, C. Pimentel, J. Jorge, in *Proc. of AAAI Spring Symposium on Sketch Understanding*. CALI: an online scribble recognizer for calligraphic interfaces (*AAAI Press, Palo Alto, USA*, 2002), pp. 51–58
24. M.G. Leslie, B.K. Levent, F.S. Thomas, Combining geometry and domain knowledge to interpret hand-drawn diagrams. *Comput. Graph.* **29**(4), 547–562 (2005)
25. P. Barros, S. Magg, C. Weber, in *International Conference on Artificial Neural Networks*. A multichannel convolutional neural network for hand posture recognition (Springer, Cham, 2014), pp. 403–410
26. Dumoulin V, and Visin F, A guide to convolution arithmetic for deep learning, https://arxiv.org/pdf/1603.07285.pdf, 23th, October, 2017
27. W. Wei, X.L. Yang, B. Zhou, J. Feng, P.Y. Shen, Combined energy minimization for image reconstruction from few views. Math. Probl. Eng. **2012**, 154630. https://doi.org/10.1155/2012/154630,2012
28. W. Wei, B. Zhou, D. Polap, M. Wozniak, A regional adaptive variational PDE model for computed tomography image reconstruction. Pattern Recogn. **92**, 64–81 (2019). https://doi.org/10.1016/j.patcog.2019.03.009
29. Wei Wei; Poap Dawid; Li Xiaohua; Wozniak Marcin; Liu Junzhe, Study on remote sensing image vegetation classification method based on decision tree classifier, Proceedings of the 2018 IEEE Symposium Series on Computational Intelligence, SSCI 2018, p 2292-2297, July 2, 2018, ISBN-13: 9781538692769; DOI: 10.1109/SSCI.2018.8628721; Article number: 8628721;Accession number: 20191106631413
30. C. Pan, H. Jian, G. Jianxing, C. Cheng, Teach machine to learn: hand-drawn multi-symbol sketch recognition in one-shot. APPLIED INTELLIGENCE, MAR **50**(7), 2239–2251 (2020)
31. Xia X, Marcin W, Fan X, Damasevicius R., Li Y. Multi-sink distributed power control algorithm for Cyber-physical-systems in coal mine tunnels. Computer Networks.Vol.161, pp.210-219, https://doi.org/10.1016/j.comnet.2019.04.017, 2019.
32. H. Song, W. Li, P. Shen, A. Vasilakos, Gradient-driven parking navigation using a continuous information potential field based on wireless sensor network. Inf. Sci. **408**(2), 100–114 (2017)
33. Q. Xu, L. Wang, X.H. Hei, P. Shen, W. Shi, L. Shan, GI/Geom/1 queue based on communication model for mesh networks. Int. J. Commun. Syst. **27**(11), 3013–3029 (2014)
34. X. Fan, H. Song, X. Fan, J. Yang, Imperfect information dynamic Stackelberg game based resource allocation using hidden Markov for cloud computing. IEEE Trans. Serv. Comput. **11**(1), 78–89 (2016)
35. J. Su, H. Song, H. Wang, X. Fan, Cdma-based anti-collision algorithm for EPCglobal c1 Gen2 systems. Telecommun. Syst. **67**(3), 1–9 (2018)

Publisher's Note