Convolutional Neural Networks based Pornographic Image Classification

KaiLong Zhou,Li Zhuo,Zhen Geng,Jing Zhang,Xiao guang Li
Signal & Information Processing Laboratory
Beijing University of Technology
Beijing, China
zklkey@emails.bjut.edu.cn

***Abstract*—Considering the fact that pornographic images are flooding on the web, we propose a pornographic image recognition method based on convolutional neural network. This method can be divided into two parts: coarse detection and fine detection. Because majority of images are normal, we use coarse detecting to quickly identify the normal images with no or fewer skin-color regions and facial images. For the images which contain much more skin-color regions, they need further identification through fine detecting. At first, we trained the CNN using the strategy of pre-training mid-level features non-fixed fine-tuning, then based on the trained model, we can classify whether the image is pornographic or not. Compared with exiting methods, performance of our method is better than the state-of-the-art.**

Keywords: image classification; convolutional neural networks; pornographic image recognition;

I. INTRODUCTION

With the rapid development of the internet, obscene and pornographic images/videos can more easily spread and cause greater harm to the social stability and teenagers' mental health. Thus, how to identify pornographic images or videos automatically has become an important research subject of purifying the internet environment and promoting the network healthy development.

For the recognition of pornographic images on the internet, the recognition accuracy and speed are the two important subjects. Currently, a lot of methods for the network pornographic images recognition have been proposed. The mainstream methods are based on Content Based Image Retrieval (CBIR) [1]. This method no longer needs the participation of labor, but describes the contents of images by extracting some visual features (such as color, texture, outline, etc.), classification model can be obtained by training these features. The pornographic image recognition technology based on the content can be subdivided into three categories: the first category is the rule based on the image, to estimate whether it is pornographic according to the rule or model. Due to the complexity of pornographic image, and the unfixed body movement, it is very difficult to obtain the precisely result of recognition. In [2], the author introduced a skin color model which could filter the non-skin color area, and then according to the threshold value, if the skin color area is greater than the threshold, the image would be estimated as pornographic. Though this is the easiest method, this method would receive many wrong judgments.

The second category is based on the image retrieval technology [3, 4]. This method constructs a image database which containing a vast of the pornographic and normal image firstly. The image to be recognized is used as the query image, comparing with the images in the database; this image is recognized as the same category with the most of the retrieval results. But due to the variety of pornographic images, it is difficult to build the image database.

The third category is to consider the pornographic image recognition as the binary classification (non-pornographic or pornographic) [5-7], It describes the content of pornographic image through abstracting of low-level visual feature (such as color, texture, outline, etc.), then they adopt the machine learning method to get the classification model based on those feature vector. Finally, the trained model can identify the images. Though this method have achieved better results, the choice of feature is difficult which need the professional staffs with professional knowledge.

CNN is one of the artificial neural networks, because CNN has a very good performance in some computer vision tasks, and it currently has become the research hotspot of speech and image recognition. CNN adopts depth network structure that every level represents a feature, and the high level is the abstract of low level feature, these hierarchical features can be more effective, and it also avoids the difficulty of feature choosing. But CNN has many parameters to been learned, large amount of images are required, and the network structure directly related to the abstract level and the feature dimension need to be designed. This paper proposes a method to recognize pornographic image based on the CNN. This method is divided into coarse detection and fine detection. The coarse detection can quickly recognize the normal images with no or fewer skin color and human faces with the help of prior knowledge; while the rest of images are identified through the fine detection process. In fine detection process, we train CNN using the strategy of the pre-training mid-level features non-fixed fine-tuning, then make use of trained model to identify whether the image is pornographic or not.

The work in this paper is supported by the National Natural Science Foundation of China (No.61372149, No.61370189, No.61471013), the Importation and Development of High-Caliber Talents Project of Beijing Municipal Institutions (No.CIT&TCD20150311,No. CIT &TCD 201304036,CIT&TCD201404043), the Program for New Century Excellent Talents in University(No.NCET-11-0892),the Specialized Research Fund for the Doctoral Program of Higher Education(No.20121103110017), the Natural Science Foundation of Beijing (No.4142009), the Science and Technology Development Program of Beijing Education Committee(No. KM201410005002), Funding Project for Academic Human Resources Development in Institutions of Higher Learning Under the Jurisdiction of Beijing Municipality.

978-1-5090-2179-6/16 $31.00 © 2016 IEEE
DOI 10.1109/BigMM.2016.29

II. THE METHOD BASED ON CNN

The pornographic image recognition method presented in this paper is divided into two phases: coarse detection and fine detection. First, it will do the coarse detection for the images and exclude the normal image with non-skin or fewer skin color and facial image quickly by skin color detection and face detection. For the rest of images, they will be recognized by pre-trained CNN model. The integrated framework of the method presented in this paper is shown in Figure 1. Now introduce implementation details of each part.

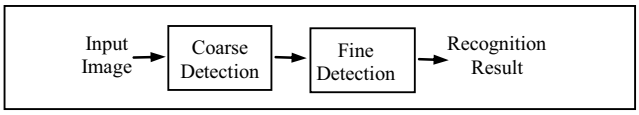


Figure 1. Diagram of the proposed method

A. Coarse Detection

The pornographic images usually contain a large exposed skin area. The feature of skin color is the most stable feature of pornographic image. So the skin color detection is able to do the preliminary recognition for the pornographic image

The coarse detection presented in this article mainly includes skin color detection and face detection. It's possible to exclude parts of images quickly and accurately with the aid of some priori knowledge. It can not only improve the integral recognition efficiency but also contribute to the improvement of recognition speed. The flow chart of the coarse detection is shown in Figure 2.

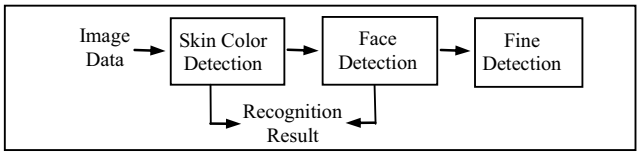
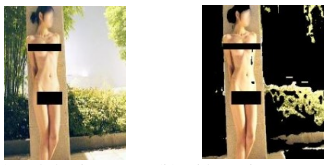


Figure 2. The flow chart of coarse detection

1) Skin-color Detection

In this paper, YCbCr color space is used for the skin-color detection, because the distribution of all races skin-color is more concentrated in YCbCr space than in other color spaces[8].

In this paper we use the same skin detection model detecting the skin color of the images by fixed threshold value [9]. The threshold value of the skin color model in YCbCr Color Space is:(Cb > 97.5 And Cb < 142.5) And (Cr > 134 And Cr < 176). Figure 3 shows the results of skin color detection. It is observed that the Skin-color Detection presented in this paper can acquire accurate detection result.



(a) Original Image (b) Skin-color Detection Result

Figure 3. Skin-color Detection Result of the Images

2) Face Detection

Face detection technology has been widely applied, and many algorithms are able to achieve good performance. In this paper, the algorithm which combines Adaboost algorithm with Haar-like feature is chosen.

For the facial images, the skin-color regions are mainly the face. But the majority of skin-color regions are body regions for the pornographic image. We can use this priori knowledge to judge the facial image. Face region (FR) can be obtained by face detection algorithm, and the result of skin color detection except the face region is body region (BR). If BR/FR>2, then it will be judged as facial image.

The coarse detection can recognize the image with no or fewer skin color and facial image quickly and accurately. The image with a mass of skin color will be judged by fine detection further. Next, the fine detection will be introduced.

B. Fine detection

In the fine detection, the pornographic image dataset is used to train the CNN model [13]. The pornographic image can be identified according to the trained model. In CNN, a small portion of the image (the local area) is the input of the lowest layer of the hierarchy, and the information is transmitted to different layers. The most significant feature of the observed data can be obtained. The structure and training method of CNN are described in detail below.

1) The Structure Of CNN

CNN has demonstrated outstanding performance in many computer vision tasks,(e.g. image classification [16,17], object detection [20]), compared with traditional hand-engineered representations, CNN has a sophisticated structure. It is mainly composed of three types of layers: convolution layer, pooling layer, the whole connection layer. The convolution layer is closely followed by the next pooling layer, at last the fully connected layer will obtain the image feature. These layers are organized according to figure 4.

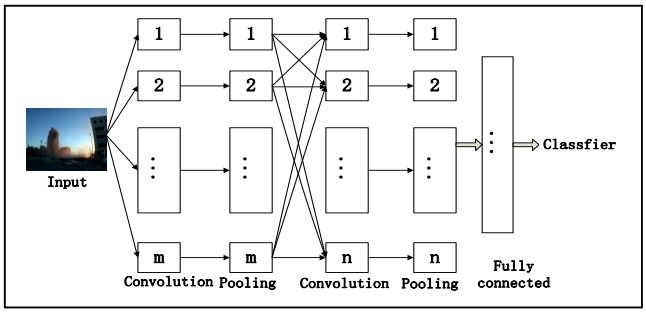


Figure 4. CNN overall structure

a) Convolution layer

CNN with fewer parameters is mainly due to the existence of convolution layer which is composed by multiple feature maps. Each feature map is composed of three stages: first, the input layer is multiplied by convolution kernels which parameters need trained, second, the value add with a bias, finally the result is input to the excitation function [14, 15]. Each feature map corresponds to a convolution kernel.

Convolution features can be gained by formula (1). y_n^l represents the l layer of the n-th feature maps. $\omega_{m,n}^l$ represents convolution kernel when extracting features of layer l. b_n^l is the bias. v_n^l means characteristic pattern set linked to layer l.

$$y_n^l = f_l(\sum_{m \in v_n^l} y_m^{l-1} \otimes \omega_{m,n}^l + b_n^l) \quad (1)$$

b) Pooling layer

Pooling layer has the same number feature maps with the former convolution layer. It divides the input into disjoint zones, and gains output by using given pooling method against each zone. Then adds offset and outputs to the excitation function [13, 14]. The pooling layer can make features more robust to resist some deformation.

$$y_n^l = f_l(z_n^{l-1} \otimes \omega_n^l + b_n^l) \quad (2)$$

z_n^{l-1} is the value after fixed window of 1-1 convolution layer features according to the pooling algorithm（average-pooling, max-pooling). ω_n^l is map weight and b_n^l is the offset.

c) Fully connected layer

Fully connected layer often consists of sigmodal neure or RBF neure. In general, the last layers of network are fully connected layer. Decrease the dimension of features. Neure number and input image types of the last layer are the same.

$$y_n^l = f_l(\sum_{m=1}^{N_{l-1}} y_m^{l-1} \omega_{m,n}^l + b_n^l) \quad (3)$$

Sigmodal neure of output layer uses formula (3) for calculation. N_l is the number of neure in output layer. The m characteristic features in 1-1 layer. y_m^{l-1} is the weight of the m characteristic pattern in the former layer joint with n neure in layer l.

The network structure used by many image recognition Imagenet tasks is shown in Figure 5. Such network structure has presented much superior performance [16]. The input of the network needs to make normalization processing to the image. The paper sets the size of image as 227×227. The network includes 8 layers: 5 convolution layers and 3 fully connected layer. Each convolution is followed by ReLU, pooling and Normalization. The former two fully connected layers are formed by inner product and ReLU. And dropout strategy is used to improve performance of neural network. The last fully connected layer only contains inner product. The number of output neuron is the same with that of recognition subject types.

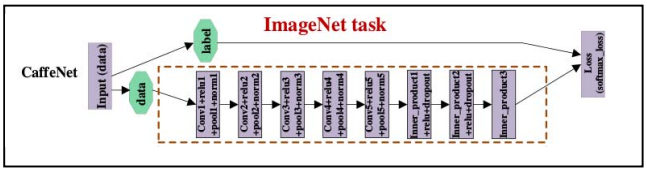


Figure 5. The structure of CaffeNet

2) Convolutional neural network training

CNN is a supervised training method, which can learn relation between output and input of many samples. The specific training includes forward propagation and backward propagation stage. The forward propagation: training images input to network, by network model parameters, the final output can be calculated. Back propagation: according to the error between output value and given label, error back-propagation algorithm (SGD and ADAGRAD) correct network parameters and make the network training again.

In this paper, we firstly randomly initialize network parameters and uses SGD to make network training. Then, we also use a pre-trained model (caffenet) as mid-level image feature representation [19], re-used on our dataset and train network parameters of the last two layers and gain the final classification model [18].

III. EXPERIMENTAL RESULTS AND ANALYSIS

In order to verify the property of the pornographic image recognition method presented in this paper, we conduct a contrast experiment between the proposed method and existing various machine learning method. Now there is a lack of standard recognition dataset of pornographic image in the world, so this paper adopts the same pornographic image dataset with [21]. There are 19000 images in total, including 8000 pornographic images and 11000 normal images(including animals, media of communication, plants, normal characters, foods, natural scenery and architectures and so on). The Figure 6 shows some images of the dataset.

Figure 6. Parts of image dataset

We have selected randomly from the dataset that have 5000 pornographic images and 6000 normal images (11000 images in total) as the training set, the remaining is test set. All the tests in the paper are set up by caffe source code under the Ubuntu system. The system configuration is Intel(R) Core (TM) i5-3210M CPU 2.5GHz with 4G internal storage and NVIDIA GTX980 graphics card. The stochastic gradient descent algorithm is chosen to train network in this paper, the batch size of training parameter is 200, momentum is 0.9, weight attenuation is 0.0005.

The Figure 7 shows the visualized results of the first layer connection weights. CNN is able to abstract the image features

layer by layer. This first layer will execute a convoluted extraction for the marginal information of the images while abstracting higher layer features afterwards.

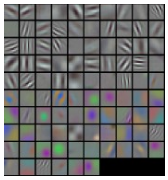


Figure 7. The first layer connection weights of the training network

We compare with the method in [21, 22], and the same database is adopted. In [21], it has proposed a pornographic image recognition method based on the ORB feature. ORB descriptors are extracted and represented using the bag-of-(visual-) words histogram based on BOW model. Next, ORB feature is combined with 72-dimensional HSV color feature of the whole image to construct the feature vector. At last, SVM is chosen to train the feature vectors to get the classification model. In [22], local feature algorithms(ORB, SIFT, SURF, ORB, BRISK, FREAK, KAZE) were analyzed in the performance of pornographic image recognition.

TABLE I. THE COMPARISON OF RECOGNITION RATE AND TIME OF DIFFERENT PORNOGRAPHIC IMAGE

Proposed method	Recognition precision	Time(ms)
SURF+HSV	94.13%	2391
ORB+HSV	93.03%	769
our method	97.20%	84

The table 1 shows respectively the comparison result of the two methods' time and recognition precision in pornographic image recognition. We use two training method for the CNN, finally the fine-tuning achieves the better results. From this table, the CNN method wins a higher recognition rate comparing to other machine learning method. For the time, the results of the CNN based on GPU are superior than and ORB and SURF feature method.

IV. CONCLUSION

In this paper, we propose a pornographic image recognition method based on CNN. Coarse detection makes contribution to identify the non-pornographic images with no or fewer skin-color regions and facial images quickly. The rest of unrecognized images need further identification--fine detection. Then we train CNN using the strategy of the pre-training mid-level features non-fixed fine-tuning, and make use of trained model to identify whether the image is pornographic or not. Compare to other machine learning method, the method in this paper has a higher recognition precision.

REFERENCES

[1] Sui L, Zhang J, Zhuo L, Yang Y C. Research on pornographic images recognition method based on visual words in a compressed domain. IET Image Processing, 2012, 6(1): 87-93

[2] H. Yin, X. Xu, L. Ye, Big skin regions detection for adult image identification, in: Digital Media and Digital Content Management (DMDCM), 2011 Workshop on, IEEE, 2011, pp. 242–247.

[3] J.-L. Shih, C.-H. Lee, C.-S. Yang, An adult image identification system employing image retrieval technique, Pattern Recognition Letters 28 (16)(2007) 2367–2374.

[4] B.-b. Liu, J. Su, Z.-m. Lu, Z. Li, Pornographic images detection based on cbir and skin analysis, in: International Conference on Semantics, Knowl-edge and Grid (SKG), vol. 0, 2008, pp. 487–488.

[5] B. Choi, B. Chung, J. Ryou, Adult image detection using bayesian decision rule weighted by svm probability, in: Computer Sciences and Convergence Information Technology, 2009. ICCIT'09. Fourth International Conference on, IEEE, 2009, pp. 659–662.

[6] H. Zhu, S. Zhou, J. Wang, Z. Yin, An algorithm of pornographic image detection, in: Image and Graphics, 2007. ICIG 2007. Fourth International Conference on, IEEE, 2007, pp. 801–804.

[7] A. P. Lopes, S. E. de Avila, A. N. Peixoto, R. S. Oliveira, A. d. A. Ara′ujo, A bag-of-features approach based on hue-sift descriptor for nude detection, in Proceedings of the 17th European Signal Processing Conference, Glasgow, Scotland, Citeseer, 2009, pp. 1552–1556.

[8] D. Chai and A. Bouzerdoum, "A Bayesian Approach to Skin Color Classification in YCbCr Color Space," TENCON 2000. Proceedings, IEEE, Kuala Lumpur Malaysia,Vol. 2, pp. 421-424, Sept. 2000.

[9] Kovac J, Peer P, Solina F. Human skin color clustering for face detection[M]. IEEE,2003.

[10] Viola P, Jones M. Rapid object detection using a boosted cascade of simple features , Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on. IEEE, 2001, 1: I-511-I-518 vol. 1.

[11] Viola P, Jones M J. Robust real-time face detection. International journal of computer vision, 2004, 57(2): 137-154.]

[12] R. Lienhart, A Kuranov, and V. Pisarevsky, "Empirical Analysis of Detection Cascades of Boosted Classifiers for Rapid Object Detection," Technical report, MRL,Intel Labs, 2002.

[13] L. Cun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Handwritten digit recognition with a back-propagation network," in Advances in Neural Info rmation Processing Systems. Morgan Kaufmann, 1990, pp. 396–404.

[14] S. L. Phung and A. Bouzerdoum, "A pyramidal neural network for visual pattern recognition," IEEE Transactions on Neural Networks, vol. 18, no. 2, pp. 329–343, 2007.

[15] "Matlab library for convolutional neural networks," ICT Research Institute, Visual and Audio Signal Processing Laboratory, University of Wollongong, Tech. Rep., 2009.

[16] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, 2012, pp. 1097–1105.

[17] K. Simonyan, A. Zisserman, Very deep convolutional networks for large scale image recognition, arXiv preprint arXiv:1409.1556.

[18] M. Oquab, L. Bottou, I. Laptev, J. Sivic, Learning and transferring mid-level image representations using convolutional neural networks, in: CVPR,2014.

[19] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick,S. Guadarrama, T. Darrell, Caffe: Convolutional architecture for fast feature embedding, arXiv preprint arXiv:1408.5093.

[20] S. S. Farfade, M. Saberian, L.-J. Li, Multi-view face detection using deep convolutional neural networks, arXiv preprint arXiv:1502.02766.

[21] Zhuo Li, Geng Zhen, Zhang Jing, Li XiaoGuang. ORB feature based web pornographic image recognition. Neurocomputing, 2015.4.2

[22] Geng Zhen, Zhuo Li, Zhang Jing, Li Xiaoguang. A comparative study of local feature extraction algorithms for web pornographic image recognition. The 2015 International Conference on Progress in Informatics and Computing (PIC-2015). Nanjing, China, December 18-20, 2015. ,unpublished.