BOOTSTRAPPING DEEP FEATURE HIERARCHY FOR PORNOGRAPHIC IMAGE RECOGNITION

Kai Li[1], Junliang Xing[1], Bing Li[1], Weiming Hu[1,2]

[1]National Laboratory of Pattern Recognition
[2]CAS Center for Excellence in Brain Science and Intelligence Technology
Institute of Automation, Chinese Academy of Sciences, Beijing 100190, P. R. China
{kai.li, jlxing, bli, wmhu}@nlpr.ia.ac.cn

ABSTRACT

Automatically recognizing pornographic images from the Web is a vital step to purify Internet environment. Inspired by the rapid developments of deep learning models, we present a deep architecture of convolutional neural network (CNN) for high accuracy pornographic image recognition. The proposed architecture is built upon existing CNNs which accepts input images of different sizes and incorporates features from different hierarchy to perform prediction. To effectively train the model, we propose a two-stage training strategy to learn the model parameters from scratch and end-to-end. During the training procedure, we also employ a hard negative sampling strategy to further reduce the false positive rate of the model. Experimental results on a large dataset demonstrate good performance of the proposed model and the effectiveness of our training strategies, with a considerable improvement over some traditional methods using hand-crafted features and deep learning method using mainstream CNN architecture.

Index Terms— Pornographic image recognition, deep learning, bootstrap

1. INTRODUCTION

With the explosive growth of images and videos on the Internet, pornographic image recognition has becoming an increasingly important task. One of its major applications is to automatically recognize pornographs and prevent them from being exposed to unsuitable crowd like children and teenagers. It can be deployed into many popular Internet products like image search engines, photo sharing social networks, file hosting service providers, live video broadcasting websites, and so on.

Most of the existing pornographic image recognition methods consist of two stages: feature extraction and classifier learning. Existing methods can be divided into three main categories based on the underlying features they use: skin color based methods, shape based methods, and local feature based methods. Skin color based methods are based on the assumption that the majority of pornographic images have a large fraction of skin color pixels [1, 2]. However, skin color information alone is not reliable since skin color pixels appear on many normal images, such as face closeup images. So, based on the observation that most pornographic images share some characteristic shapes, researchers augmented the color features with shape features [3, 4] and show some performance gains. With the popularity of *bag-of-visual-words* (BOVW) methods [5] in object classification, many local feature based methods have been proposed for pornographic image recognition [6, 7, 8] and yield promising results. Specifically, local feature based methods represent an image by a set of discrete *visual words*, which are obtained by quantization of local descriptors. After image feature extraction, Support Vector Machine (SVM) is the most widely used classifier in recent methods. For a systematically survey of existing methods for pornographic image recognition, please refer to [9].

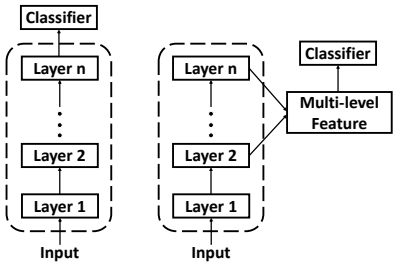


Fig. 1. Schematic diagrams of mainstream CNNs for image classification (**left**), and our proposed MLFF-CNN for pornographic image recognition (**right**). Our architecture incorporates features from multiple levels to perform recognition.

Despite the success of existing pornographic image recognition methods, they are all based on hand-crafted features. In these methods, the hand-crafted feature extraction procedure is independent of the classifier learning procedure, and then the resulted features might not be optimal. The recent developments in CNNs have demonstrated the effectiveness of feature learning and classifier learning in an end-to-end manner. Most notably, Krizhevsky *et al*. propose a CNN architecture,

i.e., AlexNet [10] which shows significant improvements upon traditional methods on image classification tasks. Since then, CNNs have been applied to achieve state-of-the-art performance on many computer vision tasks.

Motivated by the great potential of CNNs, in this paper we explore CNN based approach for pornographic image recognition. Unlike other image classification tasks (e.g., ImageNet [11]) which require invariant high level features to distinguish between different categories (e.g., cat versus dog), for pornographic image recognition task, it's beneficial to incorporate low level cues (e.g., color and shape) which are discriminative features of pornographic images. Previous works show that CNN can learn hierarchical features directly from raw input data [12]. For example, the lower layers respond to edge/color conjunctions, the middle layers capture textures and shape patterns, and the top layers show class-specific high level features. As low level features are effective for pornographic image recognition as discussed above, meanwhile, we can get this low level features easily from the lower layers in CNN. So, based on these observations, we propose a multi-level feature fusion CNN (MLFF-CNN) which enjoys the best of both worlds. More specifically, unlike mainstream CNNs which only use top high level features for classification, our MLFF-CNN exploits multi-level features which are more effective for pornographic image recognition, meanwhile, unlike traditional methods, our features come from CNN feature maps which can be learned from raw input data without any handcrafting. The schematic diagram of our MLFF-CNN is shown in Figure 1.

The main contributions of this work can be summarized in three-fold: (1) We propose a novel MLFF-CNN which exploits multi-level features for pornographic image recognition; (2) We design an effective two-stage training strategy for our MLFF-CNN; (3) We present a novel hard negative sampling strategy to further reduce false positive rate of our model.

2. THE PROPOSED APPROACH

Our MLFF-CNN is built on AlexNet [10], which is the winner of ILSVRC-2012 ImageNet challenge. AlexNet contains five convolutional layers, two fully connected layers (FC) and one SoftMax layer. Each convolutional layer is followed by rectified linear units (ReLu), contrast normalization (Norm) and max pooling layers (with the Norm and pooling layers being optional). It's worth noting that our MLFF-CNN is a general framework, which can easily use more recent CNNs (e.g., GoogLeNet [13] or VGGNet [14]) as the underlying network. There are two critical questions in our MLFF-CNN: one is how to get the multi-level features, and the other is how to effectively train it since it's more complex than mainstream CNNs (see Figure 1). The detailed answers to these questions are given as follows.

2.1. Detailed Architecture

Our multi-level features come from the convolutional layers of AlexNet. Since the feature maps of different convolutional layers have different sizes which are also high-dimensional, in order to fuse them into multi-level features we must encode them into fixed-length vectors first. The feature map of each convolutional layer can be formulated as a three-dimensional tensor of size $h \times w \times c$, which contains $h \times w$ cells and each cell represents one c-dimensional deep descriptor. For example, in AlexNet, we will get a $13 \times 13 \times 384$ feature map of the third convolutional layer for an input image of size 227×227. We can treat each of the c-dimensional deep descriptors $\boldsymbol{x}_i$ ($i \in \{1, ..., h \times w\}$) as local descriptor. In analogy to the BOVW methods [5], we can get a c-dimensional feature vector $\boldsymbol{f}_l^{global}$ for convolutional layer l by average pooling these $h \times w$ c-dimensional deep descriptors, i.e.,

$$\boldsymbol{f}_l^{global} = \frac{1}{h \times w} \sum_{i=1}^{h \times w} \boldsymbol{x}_i \tag{1}$$

Note that Equation 1 loses spatial information. Previous works show that adding spatial information through spatial pyramid pooling (SPP) [15] improves BOVW by pooling in local spatial bins which has also been used in CNN for ImageNet classification [16]. More specifically, we partition the feature map into four equal sub-regions A, B, C and D, and pool the deep descriptors inside each sub-region according to Equation 1 to get $\boldsymbol{f}_l^A, \boldsymbol{f}_l^B, \boldsymbol{f}_l^C, \boldsymbol{f}_l^D$. The final feature vector for convolutional layer l can be formulated as

$$\boldsymbol{f}_l = [\boldsymbol{f}_l^{global}, \boldsymbol{f}_l^A, \boldsymbol{f}_l^B, \boldsymbol{f}_l^C, \boldsymbol{f}_l^D] \tag{2}$$

After getting the feature vector of each convolutional layer, we concatenate them to form our multi-level feature which is followed by two fully-connected (FC) layers and one SoftMax layer for classification. The detailed architecture is shown in Figure 2(a).

2.2. Training Strategy

To effectively train the proposed model, we design a two-stage training procedure to pre-train and fine-tune the model. We also perform hard negative mining during the training process.

Two-stage training: Unlike AlexNet, our MLFF-CNN is not a chain-like net, but it's still a directed-acyclic graph (DAG) (Figure 2(a)) which can be trained end-to-end from scratch using back-propagation and stochastic gradient descent algorithm. Even so, we find that pre-training it properly and then fine-tuning it end-to-end can boost the performance. The network which is used for pre-training is shown in Figure 2(b) and we denote it as PT-CNN. One can see that we remove the feature fusion and classification part of MLFF-CNN (green-dotted box in Figure 2(a)), and insert supervision signal (red-dotted boxes in Figure 2(b)) upon every SPP layer.

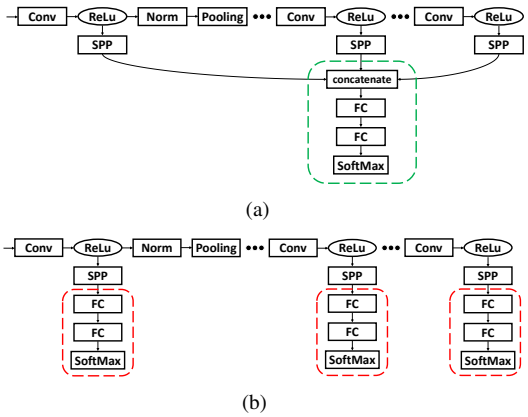


Fig. 2. The detailed architectures of our MLFF-CNN (**a**), and PT-CNN which is used for pre-training (**b**).

Note that our PT-CNN is similar to the *deeply supervised network* [17] which is used to improve the convergence rate of deep CNN. However, we use it with a different purpose here. We argue that these supervision signals (red-dotted boxes in PT-CNN) encourage the feature map of respective convolutional layer to be directly predictive of the final labels. After pre-training, the features at every SPP layers are discriminative to some extent. We use the parameters of PT-CNN after pre-training to initialize MLFF-CNN, then fine-tune the whole net end-to-end to further adjust the parameters to fully exploit the multi-level features for pornographic image recognition.

Hard negative sampling: For a pornographic image classification system, it's important to reduce the false positive rate. To this end, we propose a hard negative sampling strategy to train the network. In particular, at each epoch we update the set S_{hard_neg} which accumulates the misclassified negative samples by the historical models. We denote the set contains the remaining negative samples as S_{neg}, the set contains all of the positive samples as S_{pos}. Then at next epoch, we sample images from the above three sets to form a minibach for stochastic gradient descent training. We shuffle the set when we reach the last sample in that set. We repeat this process multiple epoches until the network converges. Since we feed the network with hard negatives at each iteration, the network is encouraged to classify them correctly and hence reduce false positive rate.

3. EXPERIMENTS AND ANALYSIS

3.1. Dataset

Since there is no released standard benchmark dataset for pornographic image recognition. In this paper, we build an image dataset from the Web for experiments. Our definition of pornographic image is as follows: image that contains human sensitive parts or shows direct sexual contact. We download 500,000 images from several adult sites and take them as pornographic images. Similarly, we download 600,000 images from three major image search engines (i.e., Google, Bing, Baidu) with diverse types of query words, such as objects, scenes, people etc, and take them as normal images. We invited 10 students in our laboratory to manually filter out the images which are wrong labeled or of low quality. At last, we compile a dataset consists of nearly 150,000 pornographic images and about 500,000 normal images. To the best of our knowledge, this is the largest dataset for pornographic image recognition in literatures.

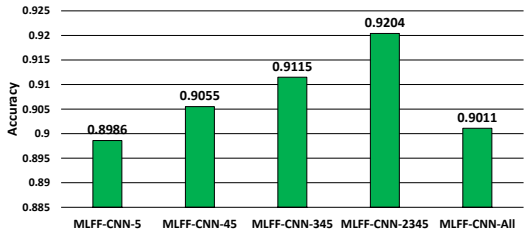


Fig. 3. Comparison between five networks which increasingly incorporate more layers.

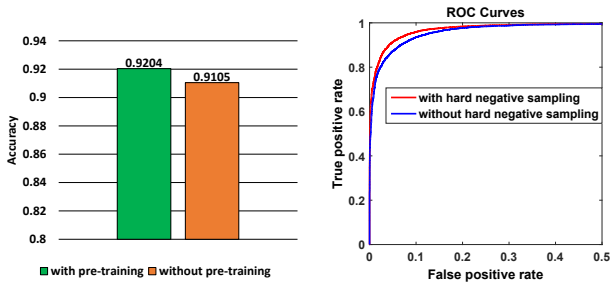
3.2. Experiment Setting

From the whole dataset, we randomly select (from each class) 25,000 images for validation, 10,000 images for testing. We repeat this procedure ten times and report the average performance. We use accuracy and receiver operating characteristic (ROC) curve to evaluate the performance of different methods. We use the Caffe toolbox [18] for all the following experiments. We train all networks using mini-batch (256) stochastic gradient descent with momentum (0.9) and weight decay (5×10^{-4}). For fully-connected layers we use a dropout ratio of 0.5. We use data augmentation similar to [10], i.e., randomly cropping of 227×227 pixels from the 256×256 input image, then randomly mirroring it before feeding it to the network. The learning rate starts from 10^{-2} and is divided by 10 when the training curve reaches a plateau.

3.3. Results and Analysis

We conduct three different sets of experiments to respectively evaluate the effectiveness of the multi-level features, the bootstrapping two-stage training strategies, as well as the overall performance with comparison to other methods.

The effectiveness of multi-level features: In order to show the effectiveness of multi-level features for pornographic image recognition, we iteratively train five networks by adding more layers. We denote them as MLFF-CNN-5, MLFF-CNN-45, MLFF-CNN-345, MLFF-CNN-2345 and MLFF-CNN-All. The numbers represent the layers being used for multi-level features construction. For example, 45 represent the fourth and fifth convolutional layers. The comparison between these variants is shown in Figure 3. From



(a) (b)

Fig. 4. The effectiveness of our proposed two-stage training strategy **(a)**, and hard negative sampling strategy **(b)**.

the chat, we can see that the accuracy increases as we incorporate more layers. This clearly validate our hypothesis that multi-level features are effective for pornographic image recognition. Interestingly, we notice that MLFF-CNN-All network performs worse. This is not surprising, as the features of the first convolutional layer are too primitive which can hurt the performance. So it's better to use mid- and high-level layers in practice.

The effectiveness of our training strategies: Since the MLFF-CNN-2345 network works best, we use it as default and denote it as MLFF-CNN for short in the following. In order to verify the effectiveness of our two-stage training strategy, we train another MLFF-CNN network from scratch without our deeply supervised style pre-training procedure. The results are shown in Figure 4(a). We can see that comparing with training from scratch, our two-stage training strategy can boost the performance significantly. Our deeply supervised style pre-training procedure initializes the model to a better point in the parameter space than random initialization, the following end-to-end fine-tune procedure further adjust the parameters for better classification.

In order to verify whether our hard negative sampling strategy can reduce false positive rate. We conduct another experiment using MLFF-CNN with/without hard negative sampling. We plot the ROC curves in Figure 4(b). We can see that the network training with hard negative sampling has lower false positive rate than its counterpart when the true positive rates are the same. This clearly validate our hard negative sampling strategy can reduce the false positive rate of the model which is very important for pornographic image recognition system in practice.

Compare with traditional methods and the underlying AlexNet:

To show the superiority of our MLFF-CNN. We compare it with other pornographic image recognition methods in literatures and the underlying AlexNet. We choose two representative methods [19, 8] based on hand-crafted features. [19] is a skin color and shape based algorithm which use patch-based skin color detector (we denote it as PBSC) for pornographic image recognition. [8] is a local feature based method

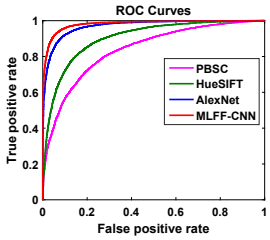


Fig. 5. The ROC curves of different methods.

Table 1. Compare to traditional methods and the underlying AlexNet.

Approach	Accuracy (%)
PBSC [19]	77.35
HueSIFT [8]	80.07
AlexNet [10]	90.34
MLFF-CNN	**92.04**

which use HueSIFT [20] as local descriptors and an enhanced BOVW model for recognition. The results are show in Table 1. We can see that local feature based methods (i.e., [8]) are better than skin color based methods (i.e., [19]). CNN based feature learning methods (i.e., AlexNet) are better than traditional hand-crafted feature based methods. The main reason is the hand-crafted low level features are not optimal for pornographic image recognition. Our MLFF-CNN combines the merits of both traditional methods and CNN based methods, specifically, our MLFF-CNN integrates multi-level automatically learned features for pornographic image recognition which achieves the best results. It's worth noting that, although our MLFF-CNN use multi-level features, its size is smaller than the underlying AlexNet. What's more, thanks to the SPP pooling, our MLFF-CNN can handle images of different sizes. Thus, our model is very suitable for practical applications which is small and accurate.

4. CONCLUSION

In this paper, we address the pornographic image recognition task using CNN based approach. We propose a MLFF-CNN, and design a two-stage training strategy for our MLFF-CNN. Furthermore, we propose a hard negative sampling training strategy in order to reduce the false positive rate. Experimental results show the effectiveness of our MLFF-CNN model and the training strategies.

5. ACKNOWLEDGEMENTS

This work is partly supported by the 973 basic research program of China (Grant No. 2014CB349303), the Natural Science Foundation of China (Grant No. 61472421), and the Strategic Priority Research Program of the CAS (Grant No. XDB02070003).

6. REFERENCES

[1] Michael J Jones and James M Rehg, "Statistical color models with application to skin detection," *International Journal of Computer Vision*, vol. 46, no. 1, pp. 81–96, 2002.

[2] Lijuan Duan, Guoqin Cui, Wen Gao, and Hongming Zhang, "Adult image detection method base-on skin color model and support vector machine," in *Asian Conference on Computer Vision*, 2002, pp. 797–800.

[3] David A Forsyth, Margaret Fleck, and Chris Bregler, "Finding naked people," *International Journal of Computer Vision*, vol. 1065, no. 1, pp. 593–602, 1996.

[4] Weiming Hu, Ou Wu, Zhouyao Chen, Zhouyu Fu, and Steve Maybank, "Recognition of pornographic web pages by classifying texts and images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 6, pp. 1019–1034, 2007.

[5] Gabriella Csurka, Christopher R Dance, Lixin Fan, Jutta Willamowski, and Cédric Bray, "Visual categorization with bags of keypoints," in *European Conference on Computer Vision Workshop*, 2004, pp. 1–22.

[6] T Deselaers, L Pimenidis, and H Ney, "Bag-of-visual-words models for adult image classification and filtering," in *International Conference on Pattern Recognition*, 2008, pp. 1–4.

[7] Rainer Lienhart and Rudolf Hauke, "Filtering adult image content with topic models," in *IEEE International Conference on Multimedia and Expo*, 2009, pp. 1472–1475.

[8] Sandra Avila, Nicolas Thome, Matthieu Cord, Eduardo Valle, and Arnaldo de A Araujo, "Pooling in image representation: The visual codeword point of view," *Computer Vision and Image Understanding*, vol. 117, no. 5, pp. 453–465, 2013.

[9] Christian X Ries and Rainer Lienhart, "A survey on visual adult image recognition," *Multimedia Tools and Applications*, vol. 69, no. 3, pp. 661–688, 2014.

[10] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105.

[11] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al., "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[12] Matthew D Zeiler and Rob Fergus, "Visualizing and understanding convolutional networks," in *European Conference on Computer Vision*, 2014, pp. 818–833.

[13] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.

[14] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[15] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2006, pp. 2169–2178.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," in *European Conference on Computer Vision*, 2014, pp. 346–361.

[17] Chen-Yu Lee, Saining Xie, Patrick Gallagher, Zhengyou Zhang, and Zhuowen Tu, "Deeply-supervised nets," in *International Conference on Artificial Intelligence and Statistics*, 2015, pp. 562–570.

[18] 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 ACM International Conference on Multimedia*, 2014, pp. 675–678.

[19] Haiqiang Zuo, Weiming Hu, and Ou Wu, "Patch-based skin color detection and its application to pornography image filtering," in *International World Wide Web Conference*, 2010, pp. 1227–1228.

[20] Koen EA Van De Sande, Theo Gevers, and Cees GM Snoek, "Evaluating color descriptors for object and scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1582–1596, 2010.