PORNOGRAPHIC IMAGE RECOGNITION BY STRONGLY-SUPERVISED DEEP MULTIPLE INSTANCE LEARNING

Yuhui Wang[a,b], *Xin Jin*[a,b], *Xiaoyang Tan*[a,b]

[a]Department of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics

[b]Collaborative Innovation Center of Novel Software Technology and Industrialization, Nanjing University of Aeronautics and Astronautics

ABSTRACT

In this paper, we propose a principled framework for pornographic image recognition. Specifically, we present our definition of pornographic images, which characterizes the pornographic contents in images as the exposure of private body parts. As the private body parts often lie in local image regions, we model each image as a bag of local image patches (instances), and assume that for each pornographic image at least one instance accounts for the pornographic content within it. This treatment allows us to cast the model training as a Multiple Instance Learning (MIL) problem. Furthermore, we propose a strongly-supervised setting for MIL by identifying the most likely pornographic instances in positive bags, which effectively prevents the algorithm from getting trapped in a bad local optima. Last but not least, we formulate our strongly-supervised MIL under the deep CNN framework to learn deep representations; hence we call it Strongly-supervised Deep MIL (SD-MIL). We demonstrate that our SD-MIL based system produces remarkable accuracy with 97.01% TPR at 1% FPR, testing on 117K pornographic images and 117K normal images from our newly-collected large scale dataset.

Index Terms— Pornographic image recognition, Multiple Instance Learning, Deep learning

1. INTRODUCTION

It is of great significance to recognize and filter out pornographic images nowadays [1]. Although extensively studied, recognizing pornographic images remains a challenging problem in computer vision. Various factors may pose major challenges, such as the subjectivity involved in people's definition of pornography, the large variations exhibited by pornographic images in background, scenario, lighting, scale, pose of persons, and the high similarity between some special pornographic images and normal images.

To tackle these challenges, many authors advocate to detect Regions of Interest (ROI) first, typically through skin detection [2, 3]. Based on the detected ROI, some kind of shape or geometric analysis are then applied to recognize possible body postures related to nudity or pornography. However, accurate skin detection remains very challenging in itself, and ROI obtained by skin detection may have semantic gaps with the real pornographic contents. Some authors advocate to use a bag-of-features (BOF) approach, which represents an image as histograms built from a sparse set of visual features [4, 5]. As local features, image patches are typically extracted around difference-of-Gaussian interest points. When the interest points cannot capture the pornographic contents in images, the BOF approach, however, faces great risks.

Although above methods have achieved some success, essentially, they are lack of an appropriate definition of pornographic images. As a result, these methods can hardly distinguish pornography from some related concept such as body exposure, while in fact many activities that involve a high degree of body exposure (bathing, swimming, etc.) have nothing to do with pornography.

In this paper, we first attempt to provide a definition of pornographic images (in Section 2). Building upon this definition, we model each image as a bag of overlapped image patches (instances), and assume that for each pornographic image at least one instance accounts for the pornographic contents within it. This treatment allows us to cast the model training as a Multiple Instance Learning (MIL) problem. Furthermore, we propose a strongly-supervised setting for MIL by identifying the most likely pornographic instances in positive bags, which effectively prevents the algorithm from getting trapped in a bad local optima. Inspired by recent success achieved by convolutional neural network (CNNs) in image classification [6, 7], we formulate our strongly-supervised MIL under the deep CNN framework to learn deep representations; hence we call it Strongly-supervised Deep MIL (SD-MIL). We note that the MIL and deep learning techniques have already been used for pornographic image recognition by [8] and [9] respectively, but our SD-MIL method significantly differs from them in incorporating our definition of pornographic images into the system design.

To evaluate the performance of our SD-MIL based system, we have collected a large dataset and test our system on 117K pornographic images and 117K normal images. Our system produces remarkable accuracy with 97.01% TPR at 1% FPR, and achieves 55 FPS with GPU.

2. OUR DEFINITION OF PORNOGRAPHIC IMAGES

We argue that a good definition of pornographic images should on one hand be consistent with most people's understanding, while on the other hand make the concept of pornography operational in practice. Considering these, we attempt to set up two criteria for a pornographic image.

1. it appeals to prurient interest or intends to make people feel sexually excited, usually in a way that many other people find offensive.
2. it visually contains naked people or sexual acts. Particularly it should explicitly show at least one exposed private body part, including female breast, female and male sexual organ.

We note that it is very difficult to automatically evaluate whether a given image satisfies criteria 1 purely through its visual appearance. Ideally such judgement should be made by human beings, but doing this tends to be both laborious and erroneous. To make the definition operational in practice, in this work we bypass the above trouble by collecting a large scale dataset manually that each pornographic image in it satisfies above two criteria. This ensures that our model is learnt from the right data, but for testing we mainly focus on the verification of the second criteria due to the above reason. *That is, our current work is built on the assumption that once an image satisfies the second criteria, it is considered as pornographic. In this setting, our second criteria is sufficient to distinguish pornography from some related concept such as body exposure.*

Our definition characterizes the pornographic contents in images as the exposure of private body parts, which naturally transfers the pornographic image recognition problem into the detection of exposed private body parts. For this, the simplest way is to train independent detector for each exposed private part (e.g., a breast detector). However, due to the large inter-class variations and the lack of distinct patterns, these private part detectors often have poor performance in practice. As shown in [10], the accuracy of semantic classifiers of breast and sexual organ is only 86% and 76%. To overcome this, we choose to train a *generic* pornographic content detector that do not distinguish the types of private body parts, but can utilize more useful context information.

3. METHODOLOGY

3.1. Motivation and Overview of the Proposed System

Our overall goal is to train a generic pornographic content detector, which recognizes an image as pornographic if a sub-image of it is found to contain at least one exposed private body part. Here the sub-image, in appropriate size, may contain more useful context information that can aid detection, compared to individual private body part.

According to our definition of pornographic images, an image is considered as pornographic as long as it contains one exposed private part. This criteria can naturally shift to the image patches. But, the problem is that while original pornographic images contains at least one *complete* private part, it is often the case that some image patches (e.g., segmented by sliding window) contain only *part* of a private part, see Fig. 1 (e). These image patches inherently suffer from ambiguity, as it is hard to define the exact region of private parts (e.g., sex organ) and the exact degree of exposure of them towards pornography. Considering these, we formulate the training of generic pornographic content detector as a MIL problem, since MIL is a weakly supervised learning framework by nature, which can accommodate some instance-level ambiguity.

Most MIL algorithms start from a rough initialization and then perform some form of local optimization. A number of methods have been proposed for better initialization [11, 12, 13] and better heuristics for local optimization [14, 15, 16]. However, an implicit assumption behind these methods is that the positive instances across positive images are similar in appropriate feature spaces [13]. Unfortunately, this assumption does not hold for pornographic images, for example the exposed female breasts and sexual organs looks quite different. To solve this dilemma, we propose an efficient strategy to select the most likely positive instances in positive bags with the assistance of manual keypoint annotation, which effectively prevents training from prematurely locking onto erroneous instances. In addition, we formulate our strongly-supervised MIL under the deep CNN (DCNN) framework to learn deep representations.

In summary, there are three key components in constructing our Strongly-supervised Deep MIL (SD-MIL) based pornographic image recognition system. They are *instance generation*, *instance selection* and *DCNN-based learning*. We will detail them in the following.

3.2. Instance Generation and Selection

Instance generation is a key component in MIL and has significant impact of overall performance [17]. In this work we use the most simple way, i.e., sliding window, to generate multiple instances from an image. Specifically, we resize each image into 434×434, and then extract 16 image patches of 224×224 with the step of 70, see Fig. 1 (a) and (b).

With the multiple instances obtained by sliding windows, we want to select the *most likely* positive instances in positive bags to prevent training from prematurely locking onto erroneous instances. Considering this, we develop an efficient semi-automatic strategy for instance selection. In particular, we annotate the centers of private parts with keypoints, see Fig. 1 (a). Then, we empirically define that if an image patch contains a keypoint, or has a keypoint of female breast within 10 pixels from its boundary, or a keypoint of sex organ within 20 pixels from its boundary, we will treat it as the *most likely*

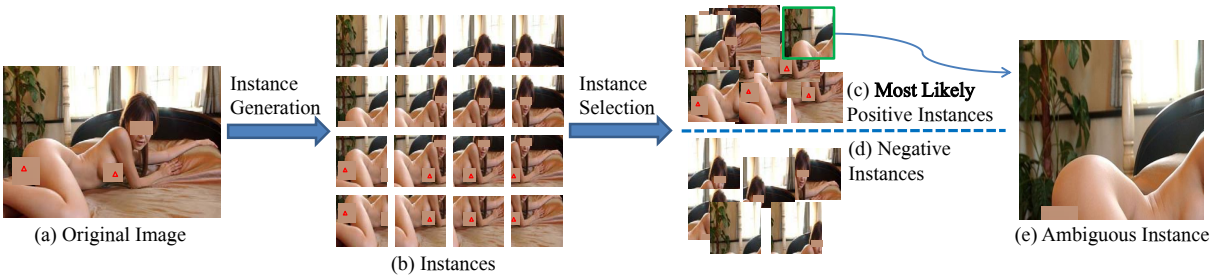

Fig. 1. Illustration of instance generation and selection. From left to right, (a) shows the original image with the keypoint annotation of exposed private parts denoted by red triangles; (b) shows the instances generated by sliding window; (c) and (d) show the most likely positive instances and negative instances. Note that due to our greedy segmentation manner (sliding window), some instance suffer from inevitable ambiguity, which contains only part of some exposed private part, see (e).

positive instance. The remaining instances are treated as negative. Fig. 1 (c) and (d) show the instance selection results.

3.3. DCNN-based Learning

We begin the learning process by briefly introducing some notations. We model each image as a bag consisting of n image patches (instance). Let $\mathbf{X} = \{\mathbf{x}_i | i = 1, ..., n\}$ be a bag of n instances, and $Y \in \{1, 2\}$ be the label of this bag, denoting whether the image is pornographic ($Y = 1$) or normal ($Y = 2$). Given one instance $\mathbf{x}_i$, the deep CNN extracts layer-wise representations of it. We denote the output of the last fully connected layer as $\{h_c | c = 1, 2\}$, where c is the index of class. We let $c = 1$ denote pornographic and $c = 2$ denote normal. Then, given a bag $\mathbf{X}$, a multiple deep CNN extracts representations of $\mathbf{X}$: $h = \{h_{c,i}\} \in R^{2\times n}$, in which each column is the representation of an instance. Specifically, we use the open-source package Caffe [18] to extract deep features, redefine the objective and fine-tune our CNNs based on the GoogLeNet model [7].

In the MIL setting, the aggregated representation of the bag for positive class ($c = 1$) is:

$$\hat{h_1} = \max_i(\{h_{1,i} | i = 1, \ldots, n\}), \tag{1}$$

Let $i^* = \text{argmax}_i(\{h_{1,i} | i = 1, \ldots, n\})$, and let $\hat{h_2} = h_{2,i^*}$, then we use softmax function to transform aggregated representation of the bag for positive class into a probability:

$$p_1 = \frac{\exp(\hat{h_1})}{\sum_{c=1}^{2} \exp(\hat{h_c})}. \tag{2}$$

The probability of each instance i for class c can be computed as follows:

$$p_{c,i} = \frac{\exp(h_{c,i})}{\sum_{c=1}^{2} \exp(h_{c,i})}. \tag{3}$$

Since we should ensure that at least one instance in the positive bag is positive, while all instances in the negative bag are negative. We can reach the loss function as follows:

$$L = -1\{Y = 1\} \log(p_1) - 1\{Y = 2\} \frac{1}{n} \sum_{i=1}^{n} \log(p_{2,i}), \tag{4}$$

The gradient is calculated via back propagation,

$$\begin{aligned} \frac{\partial L}{\partial h_{c,i}} = & -1\{Y = 1\} (p_c - 1\{c = 1\}) \frac{\partial \hat{h_c}}{\partial h_{c,i}} \\ & - 1\{Y = 2\} \frac{1}{n} (p_{c,i} - 1\{c = 2\}), \end{aligned} \tag{5}$$

where

$$\frac{\partial \hat{h_c}}{\partial h_{c,i}} = \begin{cases} 1, & h_{c,i} = \hat{h_c}, \\ 0, & else. \end{cases} \tag{6}$$

Eq. 1 to Eq. 6 are the derivation of Deep Multiple Instance Learning for two-class problem. We now move to our Strong-supervised Deep MIL algorithm. We let y_i denote the label of instance $\mathbf{x}_i$, Y^+ be the index set of our most likely positive instances, and Y^- be the index set of the negative instances. We further let n^+ and n^- be the number of elements in Y^+ and Y^-, respectively.

Then, in training phase, we only aggregate instance representations of the bag for positive class ($c = 1$) from Y^+:

$$\hat{h_1} = \max_i(\{h_{1,i} | i \in Y^+\}). \tag{7}$$

Besides the aggregated bag representation for positive class, the loss function of our strongly-supervised deep MIL is also different from that of deep MIL, by considering the negative instances in positive bags. Considering that these negative instances are naturally error-prone instances, we impose the constraint that all negative instances in positive bags should be correctly classified. Then, we reach the following loss function for our strongly-supervised deep MIL:

$$\begin{aligned} L = & -1\{n^+ > 0\} \log(p_1) \\ & - 1\{n^- > 0\} \frac{1}{n^-} \sum_{i=1}^{n} 1\{i \in Y^-\} \log(p_{2,i}), \end{aligned} \tag{8}$$

The gradient is calculated via back propagation,

$$\begin{aligned} \frac{\partial L}{\partial h_{c,i}} = & -1\{i \in Y^+\} (p_c - 1\{c = 1\}) \frac{\partial \hat{h_c}}{\partial h_{c,i}} \\ & - 1\{i \in Y^-\} \frac{1}{n^-} (p_{c,i} - 1\{c = 2\}). \end{aligned} \tag{9}$$

Fig. 2. Some example images from our database. The first row shows pornographic images, while the second row are normal images.

4. EXPERIMENTS

4.1. Datasets and Experimental Setting

We have collected a large scale dataset consisting of 155,000 pornographic images and 222,000 normal images from Internet, see Fig. 2. For these pornographic images, we randomly select 33,000 images to annotate exposed private parts with keypoints, which serves as strong supervision in the training phase. To conduct our experiments, we use the annotated 33,000 pornographic images and 100,000 randomly selected normal images as the training set, randomly select 5,000 pornographic images and 5,000 normal images from the remaining images as the validation set, while the remaining 117,000 pornographic images and 117,000 normal images are used as the test set.

The NPDI Pornography database [19] contains 16,727 key frames selected from the videos, 10,340 normal images and 6,387 pornographic images. However, according to our definition, 1,198 of the 6,387 pornographic images are incorrectly labeled; hence we remove them from our experiment. Since the NPDI database has no pornographic keypoint annotation, we do not use it for model training, but just use it for cross-database testing.

4.2. Main Results

Table 1. Comparison of Detection Rate (%)

Methods	Porn	Normal	All
Retrieval-based Method (Retrieval) [3]	67.74	59.62	63.78
Bag-of-Feature based Method (BoF) [5]	79.78	71.88	75.92
Deep Holistic Image (D-Holistic)	89.91	94.14	91.98
Deep Part Detector (D-Part Detector)	92.49	50.65	72.08
Deep MIL (D-MIL)	94.15	95.83	94.97
Strongly-supervised Deep MIL (SD-MIL)	**98.28**	**98.55**	**98.41**

On our newly-collected large scale dataset, we compare our method with two traditional methods using shallow low-level features, i.e., the retrieval-based method [3] and the Bag-of-Feature based method [5], then we compare it with some in-house baselines using deep learning to verify the effectiveness of the proposed algorithm components. In particular, we implement three in-house baselines. 1) Deep holistic (D-Holistic) image method by train CNNs with the holistic images rather than the multiple instance based representation; 2) Deep part detector (D-Part Detector) method by training independent part detector for female breast, female sexual organ and male sexual organ with 70×70 patches centered at keypoints. The trained part detectors are then used to scan the image when testing; 3) Deep MIL (D-MIL) method without using additional supervision on instances.

Tab. 1 shows the comparison results of detection rate. We can observe that the proposed method significantly outperform the retrieval-based and Bag-of-Feature (BoF) methods, with an improvement of detection rate on all test set by 34.63% and 22.49% respectively. We also see that all these deep learning based baselines, except for the deep part detector method, outperform the traditional methods by a large margin. Even so, the proposed strongly-supervised deep MIL achieves the best performance. The deep part detector method has very high false positive rate, which is consistent with our intuition since the small local patch based representation lacks useful context information. The deep holistic image method also achieves good performance compared to the traditional methods, which witnesses the power of deep learning, but still has a large gap with our method. The deep MIL method without using additional supervision of instances are prone to get in bad optima, since positive instances of pornographic images vary largely in appearance.

On the NPDI Pornography database, we test the SD-MIL model trained on our database. The performance of our model is 97.5% measured by the MAP, while the best performance in literature on this database is $96.4\pm1\%$ by [20].

5. CONCLUSION

In this paper, we propose a novel approach for pornographic image recognition. We model each image as a bag of image patches (instances), and assume that at least one instance accounts for the pornographic content in a pornographic image. This treatment allows us to cast our task as a Multiple Instance Learning problem. Our another primary innovation is a robust training strategy for MIL by narrowing down the range of positive instances in a positive bag, which effectively prevents training from prematurely locking onto erroneous instances. Furthermore, we implement our strongly-supervised MIL method under the deep learning framework to learn deep representations. Our system demonstrates strong performance on our newly-collected large dataset.

6. ACKNOWLEDGEMENTS

This work is partially supported by National Science Foundation of China (61373060), Qing Lan Project, and the Funding of Jiangsu Innovation Program for Graduate Education (KYLX_0289).

7. REFERENCES

[1] 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.

[2] Qing-Fang Zheng, Wei Zeng, Wei-Qiang Wang, and Wen Gao, "Shape-based adult image detection," *International Journal of Image and Graphics*, vol. 6, no. 01, pp. 115–124, 2006.

[3] Jau-Ling Shih, Chang-Hsing Lee, and Chang-Shen Yang, "An adult image identification system employing image retrieval technique," *Pattern Recognition Letters*, vol. 28, no. 16, pp. 2367–2374, 2007.

[4] Thomas Deselaers, Lexi Pimenidis, and Hermann Ney, "Bag-of-visual-words models for adult image classification and filtering," in *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*. IEEE, 2008, pp. 1–4.

[5] Ana PB Lopes, Sandra EF de Avila, Anderson NA Peixoto, Rodrigo S Oliveira, and Arnaldo de A Araujo, "A bag-of-features approach based on hue-sift descriptor for nude detection," in *Signal Processing Conference, 2009 17th European*. IEEE, 2009, pp. 1552–1556.

[6] 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.

[7] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich, "Going deeper with convolutions," *arXiv preprint arXiv:1409.4842*, 2014.

[8] Daxiang Li, Na Li, Jing Wang, and Tingge Zhu, "Pornographic images recognition based on spatial pyramid partition and multi-instance ensemble learning," *Knowledge-Based Systems*, vol. 84, pp. 214–223, 2015.

[9] Mohamed Moustafa, "Applying deep learning to classify pornographic images and videos," *arXiv preprint arXiv:1511.08899*, 2015.

[10] Semin Kim, Hyunseok Min, Jaehyun Jeon, Yong Man Ro, and Seungwan Han, "Malicious content filtering based on semantic features," in *Proceedings of the 2nd International Conference on Interaction Sciences: Information Technology, Culture and Human*. ACM, 2009, pp. 802–806.

[11] Saurabh Singh, Abhinav Gupta, and Alexei Efros, "Unsupervised discovery of mid-level discriminative patches," *Computer Vision–ECCV 2012*, pp. 73–86, 2012.

[12] Parthipan Siva, Chris Russell, and Tao Xiang, "In defence of negative mining for annotating weakly labelled data," in *Computer Vision–ECCV 2012*, pp. 594–608. Springer, 2012.

[13] Hyun Oh Song, Ross Girshick, Stefanie Jegelka, Julien Mairal, Zaid Harchaoui, and Trevor Darrell, "On learning to localize objects with minimal supervision," *arXiv preprint arXiv:1403.1024*, 2014.

[14] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston, "Curriculum learning," in *Proceedings of the 26th annual international conference on machine learning*. ACM, 2009, pp. 41–48.

[15] M Pawan Kumar, Benjamin Packer, and Daphne Koller, "Self-paced learning for latent variable models," in *Advances in Neural Information Processing Systems*, 2010, pp. 1189–1197.

[16] Ramazan Gokberk Cinbis, Jakob Verbeek, and Cordelia Schmid, "Multi-fold mil training for weakly supervised object localization," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. IEEE, 2014, pp. 2409–2416.

[17] Zhi-Hua Zhou, "Multi-instance learning: A survey," *Department of Computer Science & Technology, Nanjing University, Tech. Rep*, 2004.

[18] Yangqing Jia, "Caffe: An open source convolutional architecture for fast feature embedding," 2013.

[19] "Npdi pornography database https://sites.google.com/site/pornographydatabase/," 2013.

[20] 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.