The 4th International Conference on Digital Arts, Media and Technology
and 2nd ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering

Recognizing Pornographic Images using Deep Convolutional Neural Networks

Olarik Surinta
Multi-agent Intelligent Simulation Laboratory (MISL)
Faculty of Informatics, Mahasarakham University
Maha Sarakham, Thailand
olarik.s@msu.ac.th

Thananchai Khamket
Applied Informatics Group
Department of Information Technology
Faculty of Informatics, Mahasarakham University
Maha Sarakham, Thailand
thananchai.k@msu.ac.th

***Abstract*—In this paper, we propose to use deep convolutional neural network (CNN) architectures, namely the deep residual networks (ResNet), the GoogLeNet, the AlexNet, and the AlexNet architectures, for pornographic image dataset. Also, the local descriptors, namely the local binary patterns (LBP), the histogram of oriented gradients, and the scale invariant feature transform (SIFT) combined with a support vector machine (SVM), a multi-layer perceptron (MLP), or a K-nearest neighbor (KNN) techniques are proposed. Additionally, a bag of visual words (BOW) and the BOW using extracted HOG features (HOG-BOW) are compared. To classify the pornographic images, we compare the CNN architectures to well-known local descriptor techniques combined with the SVM, the MLP, and the MLP methods. Experimental results indicate that the ResNet architecture yields higher accuracies than all other approaches.**

***Index Terms*—pornographic classification, d eep convolutional neural network, deep residual networks, local descriptor technique, bag of visual words**

I. Introduction

Most research in image classification has focused on applications such as a face, object, scene, plant, animal, character, and even pornographic recognition [1]–[3]. In pornographic image recognition, image processing and machine learning techniques are proposed to use. Due to the image processing techniques, most of the pornographic image recognition researches, the human skin, called a region of interest (ROI), is first extracted from the whole image. In this process, the RGB color space is converted into HSV and YCbCr color spaces to extract the skin color [4]–[6]. In [4], the ratio value between the ROI and the whole image region calculated and decided as the pornographic image when the ratio is more than the threshold value. The threshold set to 50% and the accuracy result of the YCbCr and the HSV color space achieved a 76.25% and 77.5%, respectively. Wijaya et al. [7] presented a fusion of scale variant descriptor to extract the ROI, and a principal component analysis is used to reduce the feature vector. Then, matching the ROI with the training set using the nearest neighbor approach. This technique obtained an accuracy of 80%.

As for the machine learning technique, in [5], the color image is converted into YCbCr color space. Two feature extraction techniques including color-based and texture-based features are extracted a feature vector from the ROI. Hence, the results show that the combination of color- and texture-based features combined with a support vector machine (SVM) technique obtain more than 96%.

Contribution: In this paper, we evaluate the performance of 16 different techniques on a TI-UNRAM pornographic image dataset [7], [8]. The use of existing deep convolutional neural network (CNN) architectures (ResNet, GoogLeNet, and AlexNet) and a bag of visual words method are presented. In addition, this paper is combining three well-known local descriptor methods, called the local binary pattern (LBP) [9], the histogram of oriented gradients (HOG) [10], and the scale invariant feature transform (SIFT) [11] with three machine learning techniques include the SVM [12] using RBF kernel, multi-layer perceptron (MLP), and k-nearest neighbor (KNN). In deep learning, the results have shown that the ResNet architecture [13] are competitive when compared to the other deep CNN architectures. Furthermore, the results also show that the local binary pattern (LBP) when combined with the SVM using the RBF kernel, has obtained an accuracy of 87.80%, which is slightly lower than the ResNet architecture.

Paper outline: The remaining parts of the paper is organized as follows: In Section II, the pornographic image recognition approaches are described. In Section III, experimental settings and the results are presented. Finally, a conclusion and future work is given in Section IV.

II. Pornographic Image Recognition Methods

We will explain in this section the deep residual networks (ResNet) [13] and the local binary patterns (LBP) [9] with support vector machine (SVM) [12], which are used as image recognition methods.

A. Deep Residual Networks (ResNet)

ResNet architecture has very deep network [13]. It has shown good performance in many image recognition such as Hanja Chinese character [14], gastric pathology image [15].

He et al. [13] proposed a novel architecture called *shortcut connections*, the shortcut directly uses the input of the previous layer to the next output and sum the output in ReLu activation function. In [13], the deep ResNet architecture is proposed with a depth of 18, 34, 50, 101, and 152 layers on the ImageNet dataset. The ResNet-152 architecture is deeper 22 and 7 times than AlexNet [16] and GoogLeNet [17] architectures, respectively. The results on the ImageNet dataset show that the top-5 error rate decrease from 15.3% [16] to 3.6%.

Actually, the plain network (see Fig. 1(a)) uses in may deep learning such as VGGNet [18]. This network, however, the error increases when adding more layers. In the ResNet architecture, a building block as shown in Fig. 1(b) is proposed to due with the deeper layers.

B. Local Binary Pattern with Support Vector Machine (LBP-SVM)

Local Binary Patterns (LBP) is a non-parametric feature method and first proposed by Wang and He [19] for texture classification. In [19], a *texture specturm method* computed 8 neighbor pixels in a 3×3 neighborhood. Consequently, in [9] improved the performance of the LBP by use a two-level of the LBP. This method decreased the possible number of the texture units (TU) into $2^8 = 256$ TU. In the first level, the 8 neighbor pixels around the center pixel value are convert into 0 and 1 according to the threshold T. In this case, the center pixel value is defined as the T. The pixel value is assianed to 1 if the pixel more than T, otherwise, it assigns to 0. In the second level, thus, the 8 neighbor pixels are multiplied by the weights $(2^n, n = 0, 1, ..., 7)$ that given to the corresponding

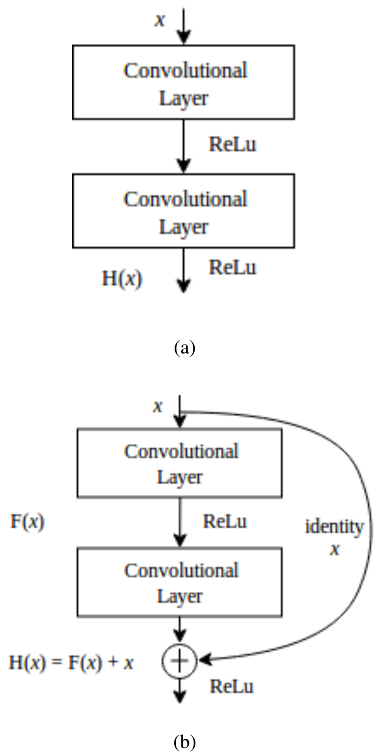


Fig. 1: Illustration of the residual learning between (a) plain network and (b) a residual network.

pixels. As a result, the values of the 8 neighbor pixels are summed and use as the texture unit. Nowaday, the LBP is applied to many applications such as face recognition [20] and character recognition [3].

Support Vector Machine (SVM) algorithm [12] has been applied to many image recognition problems. The SVM algorithm finds the maximizing margin between the separating hyperplane and the training data, called *the optimal separating hyperplane*. The training set is $(\mathbf{x}_i, y_i), i = 1, ..., l$, where $\mathbf{x}_i \in \mathbf{R}^n$ with corresponding labels $y_i \in \{1, -1\}$. It can be split by the hyperplane $\mathbf{w}^T\mathbf{x} + b = 0$, where $\mathbf{w}$ is the weight vector and b is the bias. The optimal separating hyperplane obtains the largest distance to the closest positives $\mathbf{w}^T\mathbf{x} + b = +1$ and negatives $\mathbf{w}^T\mathbf{x} + b = -1$.

In this paper, we chose the radial basis function (RBF) kernel as a non-linear similarity function. The RBF kernel is defined by:

$$K(\mathbf{x}_i, \mathbf{x}_j) = exp(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|)^2 \quad (1)$$

The hyperparameters that need to be tuned in the SVM with the RBF kernel include the cost parameter C and the gamma γ. The C parameter has a significant effect on the decision boundary. It controls the with of the margin. The γ parameter directly affects overfitting. This cause large γ values to increase the number of support vectors.

III. Experimental Settings and Results

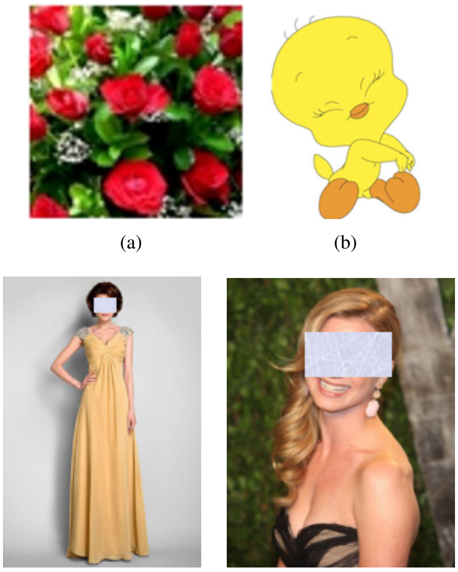
In this section, we first briefly explain the pornographic image dataset and the experimental setups used for the dataset. After that the results are presented and discussed.

A. The TI-UNRAM Pornographic Image Dataset

In this research, we evaluate the deep CNN methods and compare with the other methods including local feature descriptors combined with machine learning technique and bag of visual words on pornographic image dataset called *TI-UNRAM* [7], [8].

The TI-UNRAM dataset was introduced in [7], [8] to evaluate pornographic recognition algorithm. This dataset includes two classes (pornographic and non-pornographic), the dataset is randomly divided 50% of the whole dataset into training and test set. In the non-pornographic images include random objects such as flower, cartoon, bridge, planet, animal, plant, natural scene, and even food. Some examples of non-pornographic images are shown in Fig 2. The dataset contains 685 pornographic and 715 non-pornographic images, in total 1,400 images. These images are also collected from the Internet using a download tool. Then, images collected in numerous pixel resolutions. All images in the dataset are in RGB color space. Some of the pornographic images that show in the challenge dataset comprise single, couple, and even triple persons. An overview of the TI-UNRAM pornographic image dataset and the number of image training data are described in Table I.

In this dataset, mostly female appears in pornographic images. However, there is error occurred with this dataset.



(a) (b) (c) (d)

Fig. 2: Some examples of non-pornographic images, (a) flower image, (b) cartoon image, (c) and (d) woman images of the TI-UNRAM dataset.

TABLE I: Overview of the TI-UNRAM pornographic image dataset.

Dataset	Color format	Number of images			
		Total	Porn	Non-porn	Train/Test
TI-UNRAM [7], [8]	Color	1,400	685	715	700

Here, some of non-pornographic men images (see Fig. 3(a)) labeled as the pornographic class. Although, some of the pornographic women images are not the pornographic image, however, it labeled as the pornographic class, see Fig. 3(d)

B. Experimental Setup

In the experiments, we used 2-fold cross validation according to [7], [8] to evaluate the performance of the different methods. We compute the average and standard deviation for evaluating the test performance of the deep convolutional neural network (CNN) architectures and the variants of local descriptors combined with the machine learning methods, and a Bag of visual words (BOW) technique on the TI-UNRAM pornographic image dataset. For the local descriptors and the BOW methods, the images are resized to 90 $\times$ 90 pixel resolution, and 256 $\times$ 256 pixels for the deep CNN architectures.

C. Experimental Results

In this section, we compare the deep CNN architectures, local descriptors combined with machine learning methods, and the BOW technique for the pornographic image dataset.

As mentioned in many research studies [1], [2], [21], the deep CNN approaches are always obtained a high accuracy performance when a transfer learning and data augmentation are applied. To compare the deep CNN with the other methods, in our experiments, we decided not to use the transfer learning and data augmentation.

TABLE II: Recognition results (accuracy and standard deviation) using deep CNN methods for the TI-UNRAM pornographic image dataset.

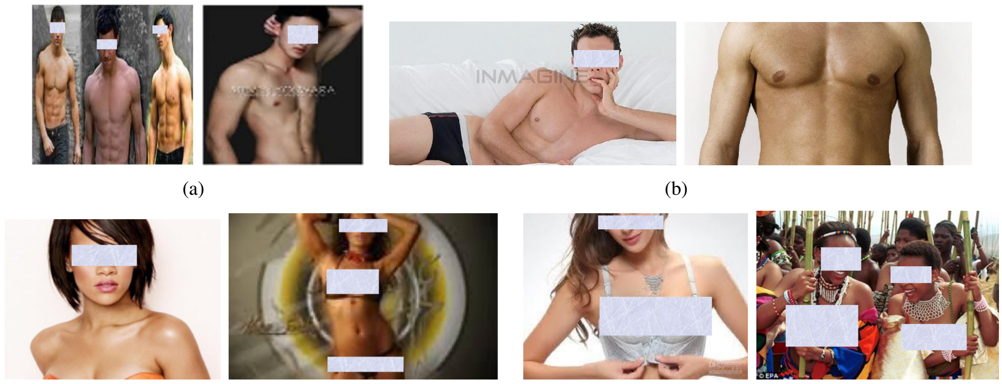
Deep CNN Methods	Layer	Test Accuracy (%)
ResNet	50	**88.00$\pm$ 0.37**
GoogLeNet	22	87.20$\pm$ 0.18
AlexNet	8	86.10$\pm$ 0.35
LeNet	5	85.90$\pm$ 0.04

TABLE III: Recognition results using different local descriptors and machine learning techniques for the TI-UNRAM pornographic image dataset.

Methods	Test Accuracy (%)
LBP+SVM	**87.80$\pm$ 0.13**
HOG+SVM	78.00$\pm$ 0.02
SIFT+SVM	78.00$\pm$ 0.01
LBP+MLP	85.80$\pm$ 0.30
HOG+MLP	75.87$\pm$ 0.01
SIFT+MLP	74.28$\pm$ 0.02
HOG+BOW	80.71$\pm$ 0.34
BOW	79.00$\pm$ 0.21
LBP+KNN	73.50$\pm$ 0.12
HOG+KNN	70.00$\pm$ 0.01
SIFT+KNN	66.43$\pm$ 0.02
FD+YCbCr [7]	83.97

The recognition results using deep CNN methods include accuracy and standard deviation are shown in Table II. According to the experimental results, if compare only the deep CNNs, the ResNet architecture outperforms the other deep CNN architectures. This approach has obtained an accuracy of 88% which is the best performance on the TI-UNRAM pornographic image dataset. Furthermore, the ResNet slightly better than the GoogLeNet. The ResNet architecture and other deep CNNs, however, takes a massive amount of time for computing for which the use of GPUs is necessary. We note that all deep CNNs experiments were carried out using the Keras with TensorFlow backend platform on a GeForce GTX 1050 Ti 4G GPU model.

Table III the experiments show when the local descriptors combined with the machine learning techniques; support vector machine (SVM), multi-layer perceptron (MLP), and K-nearest neighbor (KNN). The local binary pattern (LBP) outperforms the scale invariant feature transform (SIFT) and the histogram of oriented gradients (HOG) methods. Consequently, the LBP combined with the SVM (LBP+SVM) obtain an accuracy of 87.80% which is the highest result when the local descriptor combined with the machine learning technique. Also, in [7] proposed a fusion descriptor on YCbCr color space (FD+YCbCr) method and obtained a recognition rate of 83.97% on the TI-UNRAM dataset. It is clear, that when compared the FC+YCbCr and the LBP+SVM methods,



(a) (b) (c) (d)

Fig. 3: Illustration of the complex images between (a), (c) non-pornographic and (b), (d) pornographic images of the TI-UNRAM dataset.

the LBP+SVM method has obtained more accurately.

IV. Conclusion

In this paper, we have presented a comparative study of local feature descriptors combined with machine learning techniques and a bag of visual words (BOW) to deep convolutional neural networks (CNNs) on the TI-UNRAM pornographic image dataset.

First, we proposed to use the local binary pattern (LBP), the histogram of oriented gradients, and the scale invariant feature transform as for the local descriptor methods. We combined these methods with three machine learning techniques; support vector machine (SVM), multi-layer perceptron, and K-nearest neighbor. The results show that the combination between the LBP and the SVM, called *LBP+SVM* outperforms the other combinations. The LBP+SVM method also gives a better result than the BOW method.

Second, three deep CNN architectures include the ResNet, GoogLeNet, and AlexNet architectures are compared. However, In many researches show that a transfer learning and a data augmentation obtained the highest results on the image classification. For this reason, to make a fair comparison, the transfer learning and the data augmentation are not performed on the deep CNN experiments. The deep CNN results have also shown that the best recognition accuracy is the ResNet, GoogLeNet, and AlexNet, respectively. The results show that the ResNet architecture obtains the best results. Finally, the ResNet architecture which is the best result in our experiment, also slightly higher than the LBP+SVM.

In future work we want to improve the result of the deep convolutional neural network (CNN) by using transfer learning [21] and data augmentation [1], [22]. We also consider the deep learning approach that requires less memory usage and a decrease in training computing time.

Acknowledgment

This research was supported by the Faculty of Informatics, Mahasarakham University, Thailand.

References

[1] P. Pawara, E. Okafor, L. Schomaker, and M. Wiering, "Data augmentation for plant classification," in *Advanced Concepts for Intelligent Vision Systems*, J. Blanc-Talon, R. Penne, W. Philips, D. Popescu, and P. Scheunders, Eds. Cham: Springer International Publishing, 2017, pp. 615–626.

[2] E. Okafor, L. Schomaker, and M. A. Wiering, "An analysis of rotation matrix and colour constancy data augmentation in classifying images of animals," *Journal of Information and Telecommunication*, vol. 2, no. 4, pp. 465–491, 2018.

[3] L. Liu, H. Zhang, A. Feng, X. Wan, and J. Guo, "Simplified local binary pattern descriptor for character recognition of vehicle license plate," in *Computer Graphics, Imaging and Visualization, Seventh International Conference on*, Aug 2010, pp. 157–161.

[4] J. A. Marcial-Basilio, G. Aguilar-Torres, G. Sánchez-Pérez, L. K. Toscano-Medina, and H. M. Pérez-Meana, "Detection of pornographic digital images," *International Journal of Computers*, vol. 5, no. 2, pp. 298–305, Sep 2011.

[5] C. Santos, E. M. dos Santos, and E. Souto, "Nudity detection based on image zoning," in *Information Science, Signal Processing and their Applications (ISSPA), 11th International Conference on*, Jul 2012, pp. 1098–1103.

[6] S. Karavarsamis, N. Ntarmos, K. Blekas, and I. Pitas, "Detection of pornographic digital images," *International Journal of Digital Crime and Forensics*, vol. 5, no. 1, pp. 39–51, Mar 2013.

[7] I. G. P. S. Wijaya, I. B. K. Widiartha, K. Uchimura, and G. Koutaki, "Phonographic image recognition using fusion of scale invariant descriptor," in *Frontiers of Computer Vision (FCV), 21st Korea-Japan Joint Workshop on*, Jan 2015, pp. 1–5.

[8] I. G. P. S. Wijaya, I. B. K. Widiartha, and S. E. Arjarwani, "Pornographic image recognition based on skin probability and Eigenporn of skin ROIs images," *TELEKOMNIKA, Telecommunication, Computing, Electronics and Control*, vol. 13, no. 3, pp. 985–995, Sep 2015.

[9] T. Ojala and M. Pietik "A comparative study of texture measures with classification based on featured distributions," *Pattern Recognition*, vol. 29, no. 1, pp. 51 – 59, 1996.

[10] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society Conference on*, vol. 1, Jun 2005, pp. 886–893.

[11] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, Nov 2004.

[12] V. N. Vapnik, *Statistical Learning Theory*. Wiley, 1998.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*, Jun 2016, pp. 770–778.

[14] S. Lee and G. Jang, "Recognition model based on residual networks for cursive hanja recognition," in *2017 International Conference on*

Information and Communication Technology Convergence (ICTC), Oct 2017, pp. 579–583.

[15] B. Liu, K. Yao, M. Huang, J. Zhang, Y. Li, and R. Li, "Gastric pathology image recognition based on deep residual networks," in *IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC)*, vol. 02, Jul 2018, pp. 408–412.

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Neural Information Processing Systems, the 25th International Conference on*, ser. NIPS'12. USA: Curran Associates Inc., 2012, pp. 1097–1105.

[17] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*, Jun 2015, pp. 1–9.

[18] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Learning Representations (ICLR), International Conference on*, 2015, pp. 1–14.

[19] L. Wang and D.-C. He, "Texture classification using texture spectrum," *Pattern Recognition*, vol. 23, no. 8, pp. 905 – 910, 1990.

[20] I. L. Kambi Beli and C. Guo, "Enhancing face identification using local binary patterns and k-nearest neighbors," *Journal of Imaging*, vol. 3, no. 3, 2017. [Online]. Available: http://www.mdpi.com/2313-433X/3/3/37

[21] M. M. Ghazi, B. Yanikoglu, and E. Aptoula, "Plant identification using deep neural networks via optimization of transfer learning parameters," *Neurocomputing*, vol. 235, pp. 228 – 235, 2017.

[22] E. Okafor, R. Smit, L. Schomaker, and M. Wiering, "Operational data augmentation in classifying single aerial images of animals," in *IEEE International Conference on INnovations in Intelligent SysTems and Applications (INISTA)*, July 2017, pp. 354–360.