Using a CNN Ensemble for Detecting Pornographic and Upskirt Images

Yi Huang
Nanyang Technological University
Block N4, Nanyang Avenue, Singapore
yhuang@ntu.edu.sg

Adams Wai Kin Kong
Nanyang Technological University
Block N4, Nanyang Avenue, Singapore
adamskong@ntu.edu.sg

Abstract

The recent popularity of smartphones causes significant increase of upskirt filming cases in many countries and regions. To search upskirt images in suspects' IT devices and block them on social media, an effective detector is demanded, but it is neglected by both academic and commercial communities. Three commercial pornographic image detectors and one detector distributed by the U.S. National Institute of Justice are first examined on upskirt images. They classify 24.56%-41.90% of the testing upskirt images as pornographic images and the rest as normal images. These results imply that the detectors cannot be used to search upskirt images. To effectively use unbalanced training data, a convolution neural network (CNN) ensemble is proposed for classifying pornographic, upskirt and normal images. Training images for this classification problem are always unbalanced, because pornographic and normal images are numerous on the Internet, but upskirt images are significantly fewer. In this study, the ensemble is formed by multiple CNNs, which are trained on balanced datasets with different pornographic and normal images, but the same upskirt images. The probabilities from the CNNs are fused as the ensemble outputs. To train and test the CNN ensemble, 10,997 pornographic images, 1,673 upskirt images and 38,832 normal images are collected from the Internet. Experimental results show that the ensemble with six CNNs outperforms a single CNN by 2.93% and achieves detection accuracy of 90.23%. Comparing with the four pornographic image detectors, the ensemble also performs better in classifying pornographic images from normal images, with a significant margin of over 10%.

1. Introduction

Smartphones bring great joy to many people. Taking photos and sharing them on social media have become an important part of many people's lives. However, smartphones, like other technologies, can be misused. The popularity of smartphones has caused an increase of upskirt filming cases in many countries and regions such as Hong Kong, Japan and Singapore [1-3, 22]. In addition to taking upskirt images, in some countries such as New Zealand, it is illegal to possess and distribute these images [4]. Thus, upskirt images in IT devices of suspects have to be detected.

Many computational forensic technologies have been deployed to law enforcement agencies. Automatic fingerprint identification systems (AIFS) are the most well-known one. Law enforcement agencies use also many other technologies in their investigation such as face recognition software for identifying suspects, Efit for compositing face images based on eyewitness descriptions [5] and shoeprint matcher software for determining the size and brand of a latent shoe print. When searching images in suspects' IT devices, e.g., laptops and smartphones, face, pornographic image and child sexual abuse image detectors are utilized [6-7]. Though different image detection methods have been developed, upskirt images are neglected by both academic and commercial communities.

Pornographic image detection, also called adult image detection, has been studied by many researchers and commercial products are available. The detection methods can be categorized into color-based, shape-based and local feature approaches [6]. The color-based approach relies on the initiative assumption that pixels in pornographic images are mainly skin. Color models are usually used to define a robust skin color representation. Skin ratio, histogram, color probabilities and connected components were used to describe skin color [8-10]. The shape-based approach utilizes the shapes extracted from skin regions. Coefficients of Fourier Transform, Hu moments, Zernike moments and MPEG7 visual descriptors were employed to represent the shape information [11-13]. Skin color and shape are usually used together to enhance detection performance. The local feature approach applies local descriptors to classify pornographic images and normal images. Hue-SIFT, local PCA and local DCT were utilized with bag of words models [14-16]. Some of these research results were transferred to commercial products and are widely used by law enforcement agencies for searching pornographic images in suspects' computers. Note that in some countries, especially those in Middle East and some in South Asia, possessing pornographic images is illegal.

Upskirt images and pornographic images have different properties. One clear difference is that the percentages of skin pixels in upskirt images are likely fewer than those in pornographic images. However, the two legs of a victim are usually captured in upskirt images. Fig. 1 shows some typical upskirt images. How do the current pornographic image detectors respond to upskirt images? To answer this question, in the first part of this paper, three commercial pornographic image detectors and one pornographic image detector distributed by the U.S. National Institute of Justice are examined on upskirt images. There are numerous pornographic and normal images available on the Internet, but upskirt images are significantly fewer. To effectively use limited upskirt images in training sets for developing an upskirt and pornographic image detector, in the second part of this paper, a CNN ensemble is proposed.

The rest of this paper is organized as follows. Section 2 evaluates the four detectors. Section 3 presents the architecture of the proposed CNN ensemble. Section 4 reports the experimental results. Section 5 gives conclusive remarks.

Fig 1: Samples of upskirt images.

2. Detector Evaluation

To evaluate the existing pornographic image detectors, 5,019 pornographic, upskirt and normal images [1] were collected from the Internet. Each category of the images has 1,673 images. The definition of pornographic images in this paper is that at least one of the subjects in the images exposes her nipples or his/her genitalia. It may be a strict definition for countries in North America and Europe, but it is a suitable definition for conservative countries, in particular those in Middle East and Asia. The definition of upskirt images in this paper is images capturing a woman's underwear, crotch area, or genitalia without her authorization. Fig. 2 shows some samples of upskirt and normal images employed in this study and Fig. 3 shows the H-color distributions in the HSV color space estimated from the 5,019 testing images. The S and V-color distributions are not given because they are not very informative. The H-color distribution of the pornographic images has a high peak around 0.1 because of their high percentages of skin pixels. This peak can also be found in the distributions of upskirt and normal images. However, their peak values are lower. The peak value of the upskirt images is higher than that of the normal images, implying that the upskirt images have more skin pixels than the normal images, but fewer skin pixels than the pornographic images, in general. Fig. 3 indicates that the H-color distribution of the upskirt images is in between the H-color distributions of the pornographic and normal images. To understand how the current pornographic image detectors respond to upskirt images, four detectors, SnitchPlus, RedLight, Media Detective and Smutsnifer, are employed in this evaluation [17-20]. SnitchPlus, Media Detective and Smutsnifer are commercial detectors. RedLight was developed by the Digital Forensics and Cyber Security Center, The University of Rhode Island and is distributed by the U.S. National Institute of Justice. The four detectors do not perform hard classification between pornographic images and normal images. They use the terms, possible pornographic images and suspicious images to describe their classification results. Table 1 shows the percentages of the testing images being classified as "pornographic images". The second column lists the percentages of the pornographic images being correctly classified and the fourth column lists the percentages of the normal images being incorrectly classified as "pornographic images". The overall accuracy for pornographic and normal image classification is given in the last column. SnitchPlus, RedLight and Media Detective perform similarly in terms of overall accuracy and achieve accuracy of 79%. Media Detective and SnitchPlus perform similarly not only in terms of the overall accuracy, but also in terms of the percentages of the pornographic images and the normal images being classified as pornographic images. They correctly classify 88-90% pornographic images but incorrectly classify 28-31% normal images. RedLight and Smutsnifer perform similarly in terms of the percentages of the pornographic images being correctly classified, only 0.78% (72.92% – 72.14%) difference, but they perform differently in terms of the percentages of the normal images being incorrectly classified, 5.1% (18.41% – 13.32%) difference. Table 1 pinpoints that there is room for improving their pornographic image detection performance and the four detectors classify upskirt images neither as pornographic images nor as normal images. It is hard to use them to search evidence images in upskirt filming cases, because on average, they classify 66% upskirt images as normal images and 34% upskirt images as pornographic

[1] In Singapore, it is illegal to distribute pornographic images and it may be also illegal to distribute upskirt images. We are prohibited to distribute the database. However, the trained networks will be shared with other researchers.

images. Consequently, forensic investigators have to check all images in suspects' IT devices. Thus, an effective pornographic and upskirt image detector is demanded.

Table 1. Percentages of images classified as *pornographic images.

Software	Porn	Upskirt	Normal	$^{\&}$Accuracy
SnitchPlus	88.34	38.37	28.63	**79.86**
RedLight	72.14	30.9	**13.32**	79.41
Media Detective	**89.96**	41.90	30.66	79.65
Smutsnifer	72.92	24.56	18.41	77.26

* SnitchPlus, RedLight, Media Detective and Smutsnifer use the terms suspicious images and possible pornographic images to describe their results.

& The accuracy is calculated from porn and normal images only.The best performance in each column, except for upskirt images is highlighted.

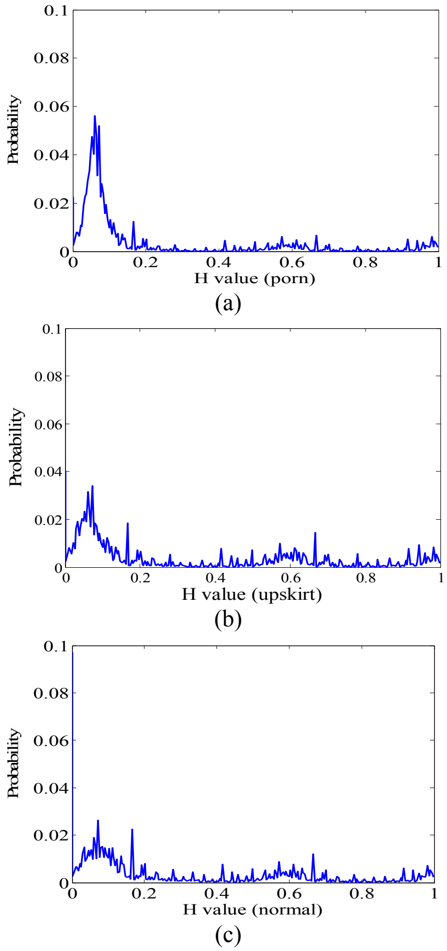
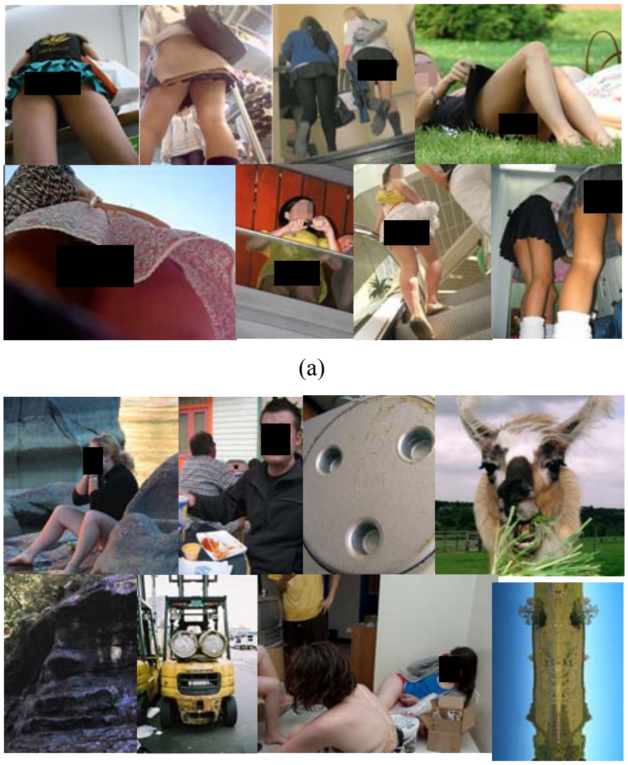


Fig. 3 H-color distributions of (a) pornographic images, (b) upskirt images and (c) normal images in the HVS space.



(a)

(b)

Fig. 2 Samples of the testing images. (a)-(b) are respectively upskirt and normal images

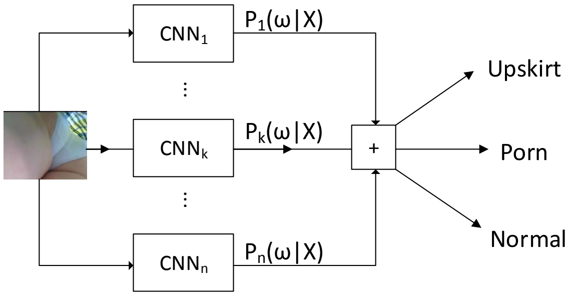


Fig . 4 Illustration of the CNN ensemble. The label X represents the input image and the label ω represents a class.

3. A CNN Ensemble

To develop an effective detector for classifying pornographic, upskirt and normal images, a large training database is essential. Numerous pornographic and normal images are available on the Internet, but upskirt images are fewer. How to effectively use limited upskirt training images but almost unlimited pornographic and normal images is an important question. In this paper, for using training data effectively, a convolution neural network (CNN) ensemble is proposed for classifying pornographic, upskirt and normal images. The ensemble is constituted by

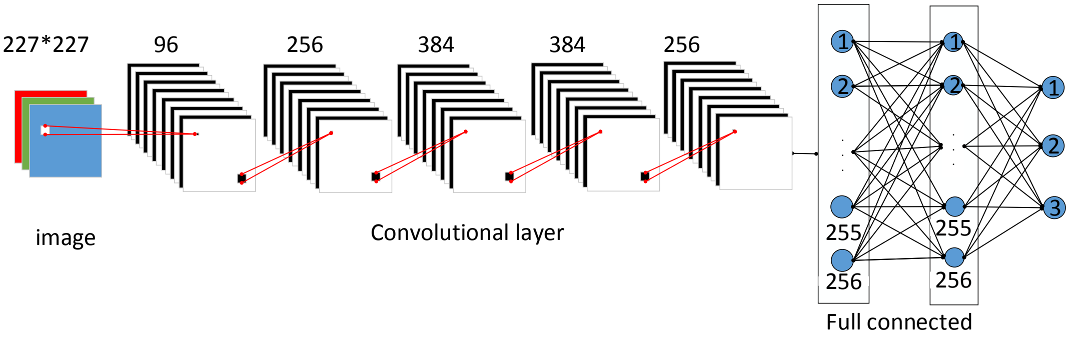


Figure 5: A basic CNN unit employed in the CNN ensemble

a set of CNNs, each of which is trained independently. Given a testing image, each of the CNNs in the ensemble determinates its class and gives the corresponding probability. The probabilities from different CNNs associated with the same class are summed. Finally, the testing image is assigned to the class with the highest sum probability. Fig. 4 illustrates the CNN ensemble. All the CNNs in the ensemble have same architecture, which is modified from the CNN used by Krizhevsky et al [21] in the ImageNet Challenge 2012. Fig. 5 shows the architecture of CNNs. The network includes five convolutional layers and three fully connected layers. The first five convolutional layers have 96, 256, 384, 384 and 256 kernels and their sizes are respectively 11×11×3, 5×5×48, 3×3×256, 3×3×192 and 3×3×192. The first fully connected layer has 256 neurons and the second fully connected layer has 256 neurons. In the last fully connected layer, there are three neurons corresponding to pornographic, upskirt and normal images. The response normalization processes follow the first and second convolutional layers. The max-pooling with a size of 3 and a stride of 2 follows the response normalization processes and the fifth convolutional layer. Dropouts are done in the first and second fully connected layers and the dropout rate is 0.5. As with Krizhevsky et al.'s [21] network, our network maximizes the multinomial logistic regression objective. The shorter side of the training and testing images is rescaled to 256 pixels and then, the central patch with a size of 227×227 pixels is cropped. The mean of training images is subtracted as a preprocessing step. 256 training images form a batch to optimize the CNN through stochastic gradient descent (SDG). The initialized learning rate is 0.01; the momentum is 0.9 and the weight decay is 0.0005. The number of iteration in each experiment is 10,000. The weights in the CNN are initialized with a zero-mean Gaussian distribution with a standard deviation of 0.01. The bias is set to zero in the first and third layers, but it is set to one in the other layers.

4. Experimental Results

For this study, 10,997 pornographic images, 1,673 upskirt images and 38,832 normal images were collected from the Internet to form a database. In total, the database contains 51,503 images. Upskirt and normal image samples in this database have been shown in Figs. 1 and 2. These images are re-organized to form several datasets for training and testing the CNN ensemble. In all the experiments, the testing set is the same, containing 5,019 images and each class has 1,673 images. This testing set has been used to evaluate the four detectors. A single CNN trained on balanced and unbalanced datasets are employed for comparison. The proposed CNN ensemble is also compared with the four detectors for pornographic image and normal image classification. We do not compare other pornographic image detection methods from the academic community because different research groups use different databases and their databases are not publicly available for the nature of the subject [6].

Before presenting the performance of the CNN ensemble, the accuracy of the CNN illustrated in Fig. 5 trained on balanced and unbalanced datasets is first reported, because it is the basic unit of the CNN ensemble. The training parameters listed in Section 3 were also used. Seven balanced datasets were constructed and each of them contains 1,673 images per class, 5,019 images in total. In all the datasets, the upskirt images are same as those used in Section 2, but their pornographic images and normal images are different. The seven datasets were used to train seven CNNs named CNN 1, CNN 2,…, and CNN 7. In each training, 90% pornographic, upskirt and normal images in one of the datasets were used. The rest of the 10% upskirt images were used for testing. 10% of the pornographic images and 10% of the normal images in the testing set were selected to test the trained CNN. For each CNN, this training and testing process were repeated 10 times. Table 2 lists the performance of the CNNs trained on the balanced datasets. The average classification accuracy for pornographic, upskirt and normal images are respectively 91.96%, 85.05% and 84.88%. The pornographic image classification accuracy is defined as the number of pornographic images being correctly classified over the total number of testing pornographic images. Other classification accuracy is defined in the same way. The average overall accuracy is 87.30%. The classification accuracy of normal images is the lowest, which is 0.17% lower than the classification accuracy of upskirt images. The CNNs achieve the highest accuracy on pornographic images. The classification accuracy of normal images is lower than pornographic and upskirt images, likely because they are more diverse and the CNNs are harder to find common features to represent them.

The unbalanced dataset contains 10,997 pornographic images, 1,673 upskirt images and 38,832 normal images. In total, the unbalanced dataset has 51,503 images. The ratio of pornographic, upskirt and normal images is 657:100:2321. The same training and testing protocol was used to estimate the accuracy. In each training, 10% of the upskirt images were kept for testing and the same numbers of pornographic and normal images were also kept for testing. This training and testing scheme was repeated ten times. The pornographic, upskirt and normal image classification accuracy is 88.11%, 68.93% and 97.67%, respectively and the overall accuracy is 84.90%. These results show that the classification accuracy is highly influenced by the distribution of the training data. The classes with more training images perform better. Thus, the CNN achieves very high accuracy for normal images, but cannot achieve a satisfactory result for upskirt images. The upskirt image detection accuracy is too low for operational use and therefore, all the CNNs in the ensemble were trained on balanced datasets.

CNN 1, CNN 2, … and CNN 7 trained on the balanced datasets formed ensembles. The same testing set was employed to evaluate their performance. ECNN-n represents the ensemble constituted by n CNNs. Since ECNN-n can be formed by many different combinations of the CNNs, all the possible combinations were used to form ECNN-n and the average accuracy is reported. For example, the accuracy of ECNN-2 is the average of 21 (C_2^7) ensembles, each of which was formed by two CNNs. Table 3 lists the accuracy of the CNN ensembles and Fig. 6 gives a graphical representation. They show that the CNN ensembles outperform the CNN trained on the balanced and unbalanced datasets. From ECNN-1 to ECNN-6, the overall accuracy and accuracy of upskirt and normal images keep improving when the number of CNN increases. However, the accuracy of pornographic images slightly fluctuates around 93.6% after ECNN 2. This slight fluctuation is also found in the accuracy of normal images after ECNN-5, but the accuracy of upskirt images keep increasing until ECNN-7. Comparing with the CNN trained on the balanced datasets, ECNN-7 offers improvement of 1.71% for pornographic images, 3.35% for upskirt images, 3.64% for normal images and 2.89% for overall accuracy. These results show that the CNN ensembles can increase the accuracy, especially for upskirt and normal images. However, the performance of the CNN ensembles would be saturated and further improvement is limited. Though CNN is used to compare with the CNN ensembles, it should be emphasized that according to the survey done by Ries and Lienhart in 2014, no one used CNN for classifying pornographic images and normal images. Moreover, we have not noted any scientific papers about classifying upskirt images from other images.

Table 4 compares the performance of ECNN-6, ECNN-7 and the four pornographic image detectors. Both ECNN-6 and ECNN-7 outperform all the four detectors in classifying pornographic and normal images with a significant margin of over 10%. In terms of the detection accuracy of pornographic and normal images, ECNN-6 performs the best. Table 5 shows the confusion matrix of ECNN-7. These experimental results show that the CNN ensembles not only offer a new function, classifying upskirt images, but also outperform the four detectors in classifying pornographic and normal images.

To further improve the accuracy, a two-class CNN classifier whose output labels are normal and porn images is added to the output of the CNN ensemble. It re-classifiers the porn and normal images outputted from the CNN ensemble. Except for the 1673 testing images for each class, all the rest 9324 porn images and the same number of normal images randomly selected were used to train the two-class CNN. Table 6 shows the confusion matrix of this hierarchical CNN using ECNN-7 as the first level classifier. The average accuracy is 90.56% which is slightly higher than ECNN-6 and ECNN-7. The accuracy can be improved more by using more porn and normal images.

5. Conclusion

Because of the popularity of smartphones, cases of upskirt filming increase significantly in some countries and regions, especially those in East and South Asia. To search evidence images in IT devices of suspects and separate them from normal and pornographic images, an upskirt image detector is demanded. However, this demand is neglected by both academic and commercial communities. In this study, three commercial pornographic image detectors and the detector developed by the Digital Forensics and Cyber Security Center, The University of Rhode Island and distributed by the U.S. National Institute of Justice are evaluated on 5,019 pornographic, upskirt and normal images. The results show that there is room for improving their pornographic image detection performance and they all do not offer functions to detect upskirt images. These detectors classify 24.56%-41.90% upskirt images as pornographic images and the rest as normal images. This classification performance implies that very likely investigators have to check every images in IT devices of suspects for searching upskirt images. To address this need and use the training data more effectively, a CNN ensemble is proposed. 10,997 pornographic images, 1,673 upskirt images and 38,832 normal images are collected for this study. Comparing with a single CNN, the CNN ensemble provides improvement of 1.71% for pornographic images, 3.29% for upskirt images, 3.64% for normal images and 2.88% for overall accuracy. It also outperforms the four detectors in classifying pornographic and normal images with a significant margin of over 10%. By using a two-class CNN to re-classify porn and normal images outputted from the CNN ensemble, the accuracy can be further improved to 90.56%.

Table 2 Classification accuracy (%) of a single CNN trained on balanced datasets.

	Porn	Upskirt	Normal	Overall
CNN 1	92.53	85.24	86.85	88.21
CNN 2	91.93	86.91	87.57	88.80
CNN 3	90.86	85.89	84.40	87.05
CNN 4	92.23	84.04	83.98	86.75
CNN 5	91.45	84.88	85.83	87.39
CNN 6	91.81	85.77	85.00	87.53
CNN 7	92.89	82.61	80.51	85.34
Average	91.96	85.05	84.88	87.30

Table 3 Classification accuracy (%) of CNN ensembles.

	Porn	Upskirt	Normal	Overall
ECNN-1	91.96	85.05	84.88	87.30
ECNN-2	93.48	85.89	87.01	88.79
ECNN-3	93.25	87.36	87.77	89.46
ECNN-4	93.62	87.80	88.39	89.94
ECNN-5	93.61	88.10	88.53	90.08
ECNN-6	93.84	88.12	88.72	90.23
ECNN-7	93.66	88.40	88.52	90.19

Table 4 A summary of the classification accuracy (%) of different methods.

	Porn	Upskirt	Normal	*2-Class Accuracy	&3-Class Accuracy
SnitchPlus	88.34	^N.A.	71.37	79.86	N.A.
RedLight	72.14	N.A.	86.68	79.41	N.A.
Media Detective	89.96	N.A.	69.34	79.65	N.A.
Smutsnifer	72.92	N.A.	81.59	77.26	N.A.
ECNN-6	**93.84**	88.12	**88.72**	**91.28**	**90.23**
ECNN-7	93.67	**88.40**	88.52	91.10	90.18

*classification accuracy of pornographic and normal images

&classification accuracy of pornographic, upskirt and normal images.

^ N.A. represents that the detectors do not offer functions to classify upskirt images.

The highest accuracy in the columns of pornographic, normal and 2-class accuracy is highlighted.

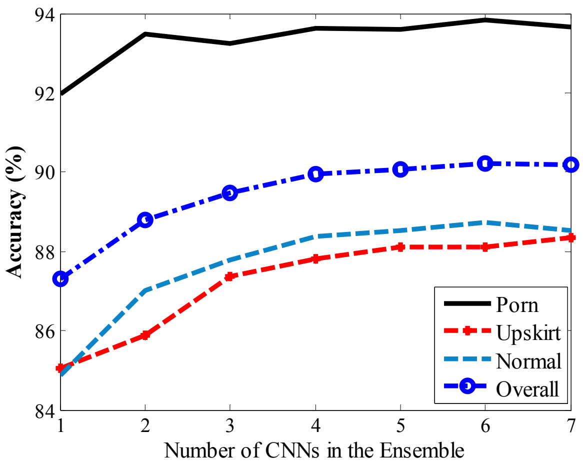


Fig. 6 The performance of the CNN ensembles.

Table 5 Confusion matrix (%) of ECNN-7

		Predicted class		
		Porn	Upskirt	Normal
Actual Class	Porn	93.66	3.05	3.29
	Upskirt	5.74	88.40	5.86
	Normal	7.23	4.25	88.52

Table 6 Confusion matrix (%) of the hierarchical CNN based on ECNN-7

		Predicted class		
		Porn	Upskirt	Normal
Actual Class	Porn	92.47	3.06	4.47
	Upskirt	5.08	88.40	6.52
	Normal	4.96	4.24	90.80

Acknowledgment

This work is partially supported by the Ministry of Education, Singapore through Academic Research Fund Tier 1, RG14/14.

References

[1] S.M. Thompson and L Moore, "Hong Kong needs to outlaw growing practice of upskirting", *South China Morning Post*, 31 July, 2014 (access on 29 March 2015) http://www.scmp.com/comment/article/1563340/hong-kong-needs-outlawgrowing-practice-upskirting

[2] M. Yeo and T. Zhuo, "Alarming rise of 'upskirt perverts' in Singapore", *Stomp*, 14 June 2015 (access on 29 March 2015) http://singaporeseen.stomp.com.sg/this-urban-jungle/alarming-rise-of-upskirt-perverts-in-singapore

[3] "More than 100 cases of upskirt photo taking reported on MRT", *Ejinsight*, 9 Feb 2015 (access on 29 March 2015).

[4] http://www.ejinsight.com/20150209-more-than-100-cases-upskirt-photo-taking-reported-on-mtr/

[5] Upskirt, Wikepedia https://en.wikipedia.org/wiki/Upskirt

[6] Vision metric, http://www.visionmetric.com/products/about-efit-v/how-efit-v-works/ (access on 29 March 2015).

[7] C.X. Ries and R. Lienhart, "A survey on visual adult image recognition", *Multimedia Tools and Applications*, pp. 661-688, 2014.

[8] N. Sae-Bae, X. Sun, H.T. Sencar and N.D. Memon, "Towards automatic detection of child pornography", *IEEE International Conference on Image Processing*, pp. 5332-5336, 2014.

[9] L. Duan, G. Cui, W. Gao and H. Zhang, "Adult image detection method base-on skin color model and support vector machine." *In proceedings of the 5th Asian conference on computer vision*, pp. 797-800, 2002.

[10] M.L. Jones, J.M. Rehg, "Statistical color models with application to skin detection", *IJCV*, vol. 46, no. 1, pp. 81-96, 2002

[11] R. Lienhart and R. Hauke, "Filtering adult image content with topic models" *in proceedings of the IEEE International Conference on Multimedia and Expo*, pp. 1472-1475, 2009.

[12] W.A. Arentz, and B. Olstad "Classifying offensive sites based on image content", *Computer Vision and Image Understanding*, pp. 295–310, 2004.

[13] Q.F. Zheng, W. Zeng, G. Wen, and W.Q. Wang, "Shape-based adult images detection", *in Proceedings of the 3rd international conference on image and graphics*, pp 150–153, 2004

[14] W. Kim, H.K. Lee, S. Yoo, and S. Baik, "Neural network based adult image classification", in: Artificial neural networks: biological inspirations ICANN vol 3696. Springer, Berlin/Heidelberg, pp 481–486, 2005

[15] A.P.B. Lopes, S.E.F. de Avila, and A.N.A. Peixoto, R.S. Oliveira, A. de, A. Araújo, "A bag-of-features approach based on hue-sift descriptor for nude detection", *the 17th European Signal Processing Conference*, pp 1552–1556, 2009

[16] T. Deselaers, L. Pimenidis, and H. Ney, "Bag-of-visual-words models for adult image classification and filtering", *ICPR*, pp. 1-4, 2008

[17] A. Ulges, and A. Stahl, "Automatic detection of child pornography using color visual words", *ICME*, pp. 1-6, 2011.

[18] Snitch, Hyperdyne Software, https://hyperdynesoftware.com/products.html (access on 30 April 2016)

[19] Media Detective http://www.mediadetective.com/ (access on 30 April 2016)

[20] Smut Sniffer http://www.smutsniffer.com/ (access on 30 April 2016)

[21] Redlight – Pornography scanner http://dfcsc.uri.edu/research/redLight (access on 30 April 2016)

[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105.

[23] Riki Ozawa, "In Japan, silent camera apps open door for unsavory photos", *The Mercury News*, 30, Dec, 2011 http://www.mercurynews.com/business/ci_19647736. (access on 10 April 2016)