J Inf Process Syst, Vol.18, No.5, pp.677~687, October 2022
https://doi.org/10.3745/JIPS.04.0253

ISSN 1976-913X (Print)
ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

TsCNNs-Based Inappropriate Image and Video Detection System for a Social Network

Youngsoo Kim[1], Taehong Kim[2], and Seong-eun Yoo[3,*]

Abstract

We propose a detection algorithm based on tree-structured convolutional neural networks (TsCNNs) that finds pornography, propaganda, or other inappropriate content on a social media network. The algorithm sequentially applies the typical convolutional neural network (CNN) algorithm in a tree-like structure to minimize classification errors in similar classes, and thus improves accuracy. We implemented the detection system and conducted experiments on a data set comprised of 6 ordinary classes and 11 inappropriate classes collected from the Korean military social network. Each model of the proposed algorithm was trained, and the performance was then evaluated according to the images and videos identified. Experimental results with 20,005 new images showed that the overall accuracy in image identification achieved a high-performance level of 99.51%, and the effectiveness of the algorithm reduced identification errors by the typical CNN algorithm by 64.87 %. By reducing false alarms in video identification from the domain, the TsCNNs achieved optimal performance of 98.11% when using 10 minutes frame-sampling intervals. This indicates that classification through proper sampling contributes to the reduction of computational burden and false alarms.

Keywords

CNN, Intelligent Image and Video Detection System, Tree-Structured Convolutional Neural Networks (TsCNNs)

1. Introduction

The rapid development of the Internet and network infrastructure technology in recent years has given rise to social computing, which has become a significant societal culture [1]. Social computing connects people's cultures, and aids social interaction through media such as blogs, wikis, social networking sites, and discussion forums on the Internet. As of 2020, average daily social media usage worldwide was 145 minutes per day. In particular, young people use social media extensively [2]. Social computing is increasingly affecting many organizations. The negative effects of social computing, such as malicious communications and sharing pornography, are social issues that require urgent intervention.

Image classification technology is often used to identify and block sexual content from spreading on the Internet. Initially, this technology determined obscenity (or the lack thereof) in images and videos by using skin color and shape recognition. However, performance improvement is limited by the technology's susceptibility to color modulation, lighting, etc. Various machine learning algorithms, such

Manuscript received November 6, 2021; first revision March 4, 2022; accepted March 13, 2022.
* **Corresponding Author:** Seong-eun Yoo (seyoo@daegu.ac.kr)
[1] Dept. of Artificial Intelligence, Jeonju University, Jeonju, Korea (pineland@jj.ac.kr)
[2] School of Information and Communication Engineering, Chungbuk National University, Cheongju, Korea (taehongkim@cbnu.ac.kr)
[3] School of Artificial Intelligence, Daegu University, Gyeongsan, Korea (seyoo@daegu.ac.kr)

as the support vector machine (SVM), have been applied to overcome the limitations in technology and improve accuracy to increase intelligent process automation [3]. Recently, the convolutional neural network (CNN), a type of deep learning algorithm, has drawn attention because of its near-human capacity in image classification, and it has been applied in many fields [4].

Depending on their nature, every organization may differ in the type and prioritization of the images they seek to detect. For instance, middle schools and high schools might prioritize blocking pornographic content, while religious institutions might prioritize the filtering of heretical information. In this paper, we consider an algorithm for detecting and filtering content considered inappropriate in South Korean military social networks. Sexual content may be prioritized as inappropriate because it overloads military networks and undermines a healthy culture. In addition, enemy propaganda is considered inappropriate because it can weaken a soldier's mental attitude toward the enemy. The word inappropriate in this paper includes all these adverse aspects.

This study proposes a tree-structured convolutional neural networks (TsCNNs)-based detection system that identifies inappropriate content in the military social network. The TsCNNs algorithm is made up of a very effective, though simple, structure comprising multiple CNNs to minimize confusion between similar classes. We collected social network images and videos and categorized them into 6 ordinary classes (labeled NORMAL) and 11 inappropriate classes (labeled ABNORMAL) that occur frequently. These manually categorized data sets were used as sample data to enhance the performance of the proposed algorithm, which was subsequently evaluated using newly collected data from the military social network. In addition, optimal sampling intervals were considered in terms of computing overload and accuracy during video classification. Consequently, we found the detection system to be effective in identifying inappropriate content in the military social network.

The rest of this paper is structured as follows. Section 2 contains a discussion on related work, and Section 3 describes the TsCNNs algorithm and the TsCNNs-based detection system, including the performance evaluation methods. Section 4 describes the experimental environment setup (including data collection) and compares and evaluates the performance of the proposed algorithm using a typical CNN algorithm. Finally, Section 5 concludes our experiments and discusses future works.

2. Related Work

A lot of research on adult image detection systems has been conducted. Typically, there are three approaches: using color and shape, using bag-of-words (BoW), or using deep learning.

The approach based on color and shape is founded on the premise that pornographic images contain a lot of skin color and sexual outlines. Fleck et al. [5] proposed a method of determining whether an input image contains nudity by deriving a range of skin colors through skin filters, and by using geometric constraints on human structure to compare outlines with a resulting range of body models. Based on the method mentioned above, various ideas on improving performance have been suggested [6-8]. However, the performance of these approaches is limited by their susceptibility to color modulation, lighting, camera angle, the diversity in human skin color, confusion with materials of color similar to human skin, and so on.

Another widely used approach in the studies on obscenity detection is employment of the BoW technique because it is relatively more robust to rotation, shape scaling, or illumination as they relate to local features, such scale invariant feature transform (SIFT). Originally, BoW was one of the ways of

classifying types of documents that were mainly used in the field of computer vision to categorize or search for images. Deselaers et al. [9] first applied this technique in detecting obscene content, and used SIFT descriptors as local features. As a result, they proposed application of principal component analysis (PCA) and an SVM for reduction of computations and performance improvement. Avila et al. [10] described BossaNova, a representation of the content-based concept as a way of classifying pornographic videos. Because BossaNova relies on HueSIFT descriptors that represent both color and shape, it outperformed standard BoW models that primarily depend on outlines or edge cues. Additionally, Caetano et al. [11] upgraded BossaNova to make it more suitable for video classification. They applied binary descriptors in conjunction with BossaNova to improve accuracy on the same data set. However, the performance of this approach is significantly different from that of human-level identification.

Because AlexNet, a CNN-based deep learning model proposed by Krizhevsky et al. [4], won the ImageNet Large-Scale Visual Recording Challenge (ILSVRC) by a wide margin in 2012, the CNN has drawn considerable attention in image classification. VGGNet (2014) [12], GoogLeNet (2015) [13], and ResNet (2016) [14], which have improved further since AlexNet was released, showed human-level classification accuracy. In recent years, there have been cases in which the CNN was applied to adult image classification due to its remarkable performance in overcoming the limitations of the two approaches mentioned above [15]. However, previous studies only focused on identifying pornographic images, thus necessitating an additional algorithmic design to identify unauthorized content other than pornography.

In this study, we propose the TsCNNs algorithm and design a detection system based on it to identify in a military social network unauthorized content that could waste resources and undermine the culture of the military. We analyzed its performance using data from the Korean military social network. In addition, an efficient video identification mechanism was established through analysis of performance changes considering sampling intervals between sampling frames.

3. TsCNNs-based Inappropriate Image and Video Detection

3.1 The Convolutional Neural Network Algorithm

The typical CNN algorithm [4] basically consists of a convolutional layer, a pooling layer, and a fully connected layer. The first layer extracts a feature map from an input image through a convolution operation, and the same filter is applied to each pixel. Convolution refers to a series of image processes involving the filtering of an image or extraction of features by applying a filter (mask) to the input image, as expressed in Eq. (1):

$$S(t) = (f * g)(t) = \int f(\tau)w(t-\tau)d\tau \tag{1}$$

where f denotes an input image, g denotes a filter, and $S(t)$ denotes a feature map of the input. An image consists of pixels on a two-dimensional plane (height X, width Y) presented with $\sum$ as shown in Eq. (2), which calculates the sum of element-specific multiplications while one function overrides another function:

$$(f * g)(i,j) = \sum_{x=1}^{X}\sum_{y=1}^{Y} f(x,y)g(i-x, i-y) \tag{2}$$

The second layer involves an abstraction whereby a plurality of pixels is mapped to a single value through sub-sampling, called pooling, to reduce the data. Pooling combined with striding is a common way to reduce the spatial size of feature maps while preserving the degree of invariance. Among various pooling methods, maximum pooling and average pooling are mostly used, but the former is generally preferred in terms of performance. They are obtained with Eqs. (3) and (4), respectively, where h and w are the height and width of the pooling window:

$$y = max_{i,j=1}^{\mathrm{h},w} x_{i,j} \tag{3}$$

$$y = \frac{1}{\mathrm{h} \cdot w} \sum_{i,j=1}^{\mathrm{h},w} x_{i,j} \tag{4}$$

Now, let the size of the pooling window be (H_p, W_p) (typically, $H_p = W_p$) and let the stride of the window be S (in general, $S = H_p$). The output size (height, width, depth) of a pooling layer for a convolution layer's output volume (H_c, W_c, D_c) is (H_q, W_q, D_q), where H_c represents the height, W_c is the width, and D_c is the depth. The output size of a pooling layer is calculated with Eqs. (5), (6), and (7):

$$H_q = \frac{H_c - H_p}{S} + 1 \tag{5}$$

$$W_q = \frac{W_c - W_p}{S} + 1 \tag{6}$$

$$D_q = D_c \tag{7}$$

These convolutional and pooling layers alternate continuously, with the aim being to remarkably reduce the data and focus on the significant features, thus significantly improving classification performance. This is the most distinguishing part of a CNN.

Finally, the fully connected layer, which is similar to multi-layer perceptron in the existing neural network, and learning that focuses on minimizing the error between input and output through the back-propagation algorithm are executed. When an input is received, the probability for each class is calculated using Eq. (8), and the input is automatically classified as having the highest probability, where K represents the number of classes. This method is called a softmax algorithm, which is widely used to convert raw values into probabilities of classes:

$$P(y = i | x, w_1, \cdots, w_K, b_1, \cdots, b_K) = \frac{e^{w_i x + b_i}}{\sum_{j=1}^{K} e^{w_j x + b_j}} \tag{8}$$

3.2 The Tree-structured Convolutional Neural Networks Algorithm

In a CNN model, the more classes it has, the greater the confusion between similar classes, because the classes influence each other in the learning process. Therefore, one advantage is that a model having a small number of classes is able to generate better performance than a large number of classes, because it models more-detailed features of the classes. The proposed TsCNNs algorithm gives each CNN a small number of classes through a tree-structured model in order to minimize confusion between similar classes.

3.2.1 Network architecture

The architecture of the TsCNNs algorithm is organized in tree form, which consists of nodes connected

by a directed acyclic graph, as shown in Fig. 1. The tree is a sort of unidirectional and acyclic graph, consisting of a single root node as the starting point, with parent and child nodes connected by edges, and leaf nodes with no children. A parent node has one or more child nodes, whereas a child node is connected to only one parent node. Each node stands for a classifier, and predicts the class at that level for the input image. The first classification occurs at the root node (in orange) located at left. The root node's output branches by connecting to M nodes (in turquoise) in the next stage, which means the former node becomes a parent node, and the latter nodes become its child nodes. Also, the child nodes become sub-classes of the parent node. Such a relationship and process is repeated along the edge until reaching a leaf node (in gray) that matches the final classification result for the input image. Each leaf node is associated with each class, and no two leaf nodes share the same class simultaneously. The 2,*N* index of the green node indicates the *N*-th child node of the CNN2 node in Fig. 1. In other words, the index of each node shows the path taken from the root node along the edge, as shown in the indexes of the gray nodes.

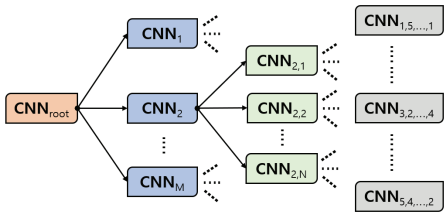

Fig. 1. Network architecture of TsCNNs.

3.2.2 Algorithm

The proposed TsCNNs algorithm aims to increase the accuracy, rather than the efficiency, of the computation under the premise that it is supported by an environment with sufficient computing resources. We start with the assumption that we have different CNN models trained in advance on different data sets to recognize as many classes as the number of child nodes. Each model associated with a specific node is trained with a segmented data set according to the class of the child nodes. The child nodes branching further out in the tree are trained to classify images into finer classes.

Algorithm 1. Class prediction

```
INPUT rawImage, SET node_idx to root
img ← RESIZE(rawImage)
BEGIN procedure PREDICT_CLASS(img, node_idx)
    class ← CNN_node_idx(img)
    IF num(node_idx.children) == 0 THEN
        RETURN class
    ELSE
        new_node_idx ← GET_CHILDNODE_IDX(node_idx, class)
        RETURN PREDICT_CLASS(img, new_node_idx)
    END IF
END procedure
```

The algorithm for predicting the class of a new input image is in Algorithm 1. First, when a new image is input, the starting point is initialized with the root node, and then, the image needs to match the input size of the CNN model, so it is adjusted through the resize function. For instance, AlexNet [4], a type of CNN model, has an input image size of 227×227×3, and VGGNet [12] and GoogLeNet [13] have a size of 224×224×3. Next, a recursive algorithm, denoted by PREDICT_CLASS, is applied to a tree-based class prediction algorithm according to the network model until a leaf node is reached. More specifically, a classified result for the image is first received from the CNN model corresponding to the current node. If the current node does not have child nodes (that is, if it is a leaf node), the result becomes the final class of the input image. Otherwise, it gets the child node index of the current node according to the resulting class and calls PREDICT_CLASS recursively. As a result, the CNN models corresponding to the nodes according to the edges of the tree-structured network model are called, and the resulting value for the input image is returned when a leaf node is reached.

3.3 Implementing Inappropriate Image and Video Detection

We implemented an "unauthorized image" file-detection framework using Google's open source-based machine learning library TensorFlow [16]. Each neural network node conducted an analysis through simultaneous multiple graphical computations, and generated boundary values. Finally, a neural network consisting of such nodes yielded the results.

Fig. 2. A shape of the TsCNNs algorithm.

The TsCNNs algorithm has a tree-based model to minimize confusion among similar classes. If there are similar classes, the input image is first classified by grouping it into a higher node class. Then, when the input image is classified into the upper class, it is repeatedly classified into lower-node, similar classes as candidates. For instance, since two classes, the face of North Korean dictator Kim Jong-un and the face of a person in a crowd, can be confused with each other, as illustrated in Fig. 2, the input is first put into the higher node class called FACE. Then, it is classified repeatedly as FACE (PUBLIC) and FACE (KIM JONG-UN). In order to identify inappropriate images and videos, they are broadly categorized as NORMAL and ABNORMAL, which are respectively classified into 6 and 11 classes at the leaf node level, as shown in Table 1. The ABNORMAL category is more granular to better identify inappropriate images and videos. The construction of the ABNORMAL category is based on unauthorized image files that appear within the Korean military, whereas classes in the NORMAL category are based on images that were openly shared among the soldiers.

Table 1. All classes in the NORMAL category and in the ABNORMAL category

Category	Class
NORMAL (6)	MILITARY ACTIVITY, SPORTS, EQUIPMENT, CLOTHES (ORDINARY), FACE (PUBLIC), ETC.
ABNORMAL (11)	SEXUAL ACTIVITY, QUISI-SEXUAL ACTIVITY, ADULT CARTOON, GENITAL EXPOSURE, FEMALE UPPER BODY, FEMALE LOWER BODY, MALE WHOLE BODY, HOMOSEXUALITY, CLOTHES (EROTIC), NORTH KOREAN FLAG, FACE (KIM JONG-UN)

Transfer learning was conducted based on Google's Inception model (version 3) to reduce the time and computing resources needed for training and achieving good performance. We constructed 17 classes, and trained a model using about 5,000 pieces of data in the ABNORMAL category. The discriminant formulas of NORMAL and ABNORMAL are shown in Eq. (9). When an *i*-th image file is input, each probability is calculated from 1 to J based on 15 classes, where FACE (PUBLIC) and FACE (KIM JONG-UN) are integrated into FACE, and CLOTHES (ORDINARY) and CLOTHES (EROTIC) are integrated into CLOTHES, as mentioned above. CNN-based classification at the child node level is then performed within each category, making it the final class and giving it a parent-node classification probability.

The calculated probability is scaled to a value between 0 and 1 and is used as the weight, *w*. If the corresponding class belongs to the NORMAL category, *c* is -1; otherwise, *c* is 1. Finally, if the summation of the results is greater than 0, it is considered NORMAL, but ABNORMAL otherwise. This is because different classes may appear at different rates in the same picture. For instance, the classification results would vary between cases where the face in a woman's photo is small and her erotic body outline large, and vice versa, depending on the ratio of these differences.

$$d_i = \begin{cases} NORMAL, & if \sum_{j=1}^{J} c_{ij} w_{ij} > 0 \\ ABNORMAL, & otherwise \end{cases} \tag{9}$$

In videos, image frames are extracted at each time interval and are classified according to the detection mechanism described above. Many frames are extracted and identified according to the extraction interval, necessitating both detection accuracy and identification of the computing burden when determining the extraction interval. To evaluate performance, certain variables are measured, as follows:

- TN (true negative): NORMAL cases identified as NORMAL.
- FN (false negative): ABNORMAL cases identified as NORMAL.
- TP (true positive): ABNORMAL cases identified as ABNORMAL.
- FP (false positive): NORMAL cases identified as ABNORMAL.
- TPR (true positive rate): the ratio of actual ABNORMAL cases correctly identified (also referred to as sensitivity or recall):

$$TPR = \frac{TP}{TP + FN} \tag{10}$$

- FPR (false positive rate): the ratio of NORMAL cases identified incorrectly:

$$FPR = \frac{FP}{FP + TN} \tag{11}$$

- Precision: the ratio of actual ABNORMAL cases to cases classified as ABNORMAL.

$$Precision = \frac{TP}{TP + FP} \tag{12}$$

- Accuracy: the ratio of cases correctly identified overall:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{13}$$

4. Experiments and Evaluations

4.1 Detection of Abnormal Images

For a training data set, 45,062 image files related to obscenity and rebellion were collected from the military social network and classified manually. We collected 35,786 image files and tested 20,005 of them after eliminating duplicates. After a typical CNN algorithm was first experimented with, the proposed TsCNNs was applied, and the results yielded by the two algorithms were compared. The overall classification results based on experiments with the sample image files are shown in Table 2. There was no significant difference between the accuracy of the CNN and that of the TsCNNs algorithm because the ABNORMAL files were too few, owing to the nature of the military. When the typical CNN algorithm was applied, the total accuracy level (including the ABNORMAL classification rate) was 98.61%. The accuracy of the TsCNNs algorithm was 99.51%, exhibiting an improvement of 0.90%. Even though the improvement level was low, the proposed algorithm significantly reduced false alarms, compared to those generated by the typical CNN. In particular, of the 279 false alarms generated by the CNN, the proposed algorithm eliminated 181 of them (approximately 65%). All ABNORMAL images were correctly classified regardless of the algorithm used, even if the number (only 153) was relatively small.

Table 2. Confusion matrix for image classification

Actual		Predicted		
		NORMAL	**ABNORMAL**	**Accuracy (%)**
NORMAL	CNN	19,573	279	98.59
	TsCNNs	19,754	98	99.51
ABNORMAL	CNN	-	153	100
	TsCNNs	-	153	100

A thorough examination of false alarms showed that the typical CNN produced 279 false positives in four classes: CLOTHES (ORDINARY), CLOTHES (EROTIC), FACE (PUBLIC) and FACE (KIM JONG-UN), as shown in Table 3. The most frequent errors occurred between CLOTHES (ORDINARY) and CLOTHES (EROTIC) and stemmed from the basic ambiguity between CLOTHES (ORDINARY) and CLOTHES (EROTIC). When the TsCNNs was used to address the misclassifications, errors were reduced by 87.7% to only 15. In addition, 81 false positives from the typical CNN resulted from confusion between FACE (KIM JONG-UN) and FACE (PUBLIC). We achieved an improvement of 91.36% by using the TsCNNs algorithm, reducing false positives to only seven images. The other confusion errors were caused by differences in people's perception of what constitutes pornography (e.g., celebrity photos and artwork). With the exception of such errors, false alarms were reduced by 89.16%, from 203 to 22, with the proposed algorithm. Therefore, we found the TsCNNs algorithm to be very effective in reducing false alarms.

Table 3. Improvements between similar classes (typical CNN → TsCNNs)

Actual	Predicted		
	CLOTHES (EROTIC)	**FACE (KIM JONG-UN)**	**ETC**
CLOTHES (ORDINARY)	122 → 15	-	-
FACE (PUBLIC)	-	81 → 7	-
ETC.	-	-	76

4.2 Detection of Abnormal Videos

A total of 265 videos (minimum length 3 minutes to as long as 96 minutes) were collected and classified. the results of video classification based on the interval between frame extractions are shown in Fig. 3. the sampling interval initially increased, gradually increasing the accuracy, and then decreased slightly for extraction intervals longer than 10 minutes. the optimum accuracy level was 98.11% with a sampling interval of 10 minutes. even if the sampling interval was less than 10 minutes, accuracy was reduced due to a large number of false alarms. FPR represents the rate at which false alarms occur and is inversely proportional to accuracy. that is, false positives appeared frequently for intervals of up to 10 minutes, significantly affecting the level of accuracy. this indicates that many normal videos were identified as abnormal when the sampling interval was smaller. as the sampling interval increased, the TPR dramatically decreased, indicating that abnormal videos were not correctly detected. however, with a sampling interval longer than 10 minutes, accuracy was not significantly reduced because the number of abnormal videos (11) was significantly smaller than the number of normal videos (254). thus, the general level of accuracy was also affected more by the FPR than by the TPR. as a result, because sampling frame rates that are too short not only increased false alarms but also linearly increased the overhead of the classification processing system, an appropriate sampling rate should be identified.

The confusion matrix from classification of videos is shown in Table 4 when the sampling interval was 10 minutes. the overall accuracy was 98.11%. three false negatives occurred owing to failure to sample obscene frames, indicating the need for denser sampling. however, further narrowing the sampling rate because of the occurrence of two false positives is unhelpful. that is why the optimum sampling interval was approximately 10 minutes. long videos generated false positives mainly because they were identified as abnormal whenever even a single abnormal frame of the continuously extracted frames occurred. both precision and TPR (recall) should be considered in object classification. we determined the sampling interval should be less than 10 minutes because TPR referring to detection of abnormal videos is more important than precision during actual operation. moreover, the difference between the classification accuracy of images and that of videos is attributable to false alarms and the inherent nature of images that contain many aspects on a single screen compared to videos.

Table 4. Confusion matrix for video classification

Actual	Predicted				
	NORMAL	ABNORMAL	Precision (%)	TPR (%)	Accuracy (%)
NORMAL	252	2	80.00	72.73	98.11
ABNORMAL	3	8			

5. Conclusion and Future Work

In this study, we proposed the TsCNNs-based detection system to identify inappropriate content in a military social network, experimenting with actual data, and evaluating the performance of the system. The proposed system exhibited a high detection rate (99.51%) for inappropriate content. In particular, identification errors among similar classes, caused by confusion in a typical CNN, decreased significantly. In addition, we obtained a high accuracy level of 98.11% by applying the TsCNNs algorithm to

video identification at a frame extraction sampling rate of 10 minutes. We found that sampling rates that are higher than 10 minutes not only burdened the system but also reduced accuracy due to false alarms during video classification. In terms of overhead and accuracy in our system, optimal results were obtained with a sampling interval of approximately 10 minutes. In summary, we found the proposed algorithm effective in identifying inappropriate content, even though there were slight variations depending on the nature of the domain.

Future studies will focus on improving the TsCNNs algorithm based on a better image identification model (such as ResNet [14]), from collection of a larger amount of data, and from more precise design and understanding of the structure to detect inappropriate content. To facilitate generalization of the experiment results, application to various domains should be evaluated as a supplement to the existing data.

Acknowledgement

This research was supported by the Daegu University Research Grant, 2018.

References

[1] S. Sharma, S. Bawa, and H. Lomash, "Proliferation of social computing: cultural computing paradigm," *International Journal of Computer Applications*, vol. 137, no. 9, pp. 27-30, 2016.

[2] Statista, "Daily time spent on social networking by internet users worldwide from 2012 to 2022," 2022 [Online]. Available: https://www.statista.com/statistics/433871/daily-social-media-usage-worldwide.

[3] Y. C. Lin, H. W. Tseng, and C. S. Fuh, "Pornography detection using support vector machine," in *Proceedings of the 16th IPPR Conference on Computer Vision, Graphics and Image Processing (CVGIP)*, Kinmen, China, 2003, pp. 123-130.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 25, pp. 1106-1114, 2012.

[5] M. M. Fleck, D. A. Forsyth, and C. Bregler, "Finding naked people," in *Computer Vision – ECCV '96*. Heidelberg, Germany: Springer, 1996, pp. 593-602.

[6] C. Y. Jeong, J. S. Kim, and K. S. Hong, "Appearance-based nude image detection," in *Proceedings of the 17th International Conference on Pattern Recognition (ICPR)*, Cambridge, UK, 2004, pp. 467-470.

[7] Q. F. Zheng, W. Zeng, W. Q. Wang, and W. Gao, "Shape-based adult image detection," *International Journal of Image and Graphics*, vol. 6, no. 1, pp. 115-124, 2006.

[8] H. A. Rowley, Y. Jing, and S. Baluja, "Large scale image-based adult-content filtering," in *Proceedings of the First International Conference on Computer Vision Theory and Applications (VISAPP)*, 2006, pp. 290-296.

[9] T. Deselaers, L. Pimenidis, and H. Ney, "Bag-of-visual-words models for adult image classification and filtering," in *Proceedings of 2008 19th International Conference on Pattern Recognition*, Tampa, FL, 2008, pp. 1-4.

[10] S. Avila, N. Thome, M. Cord, E. Valle, and A. D. 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.

[11] C. Caetano, S. Avila, S. Guimaraes, and A. D. A. Araujo, "Pornography detection using BossaNova video descriptor," in *Proceedings of 2014 22nd European Signal Processing Conference (EUSIPCO)*, Lisbon, Portugal, 2014, pp. 1681-1685.

[12] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014 [Online]. Available: https://arxiv.org/abs/1409.1556.

[13] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, et al., "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Boston, MA, 2015, pp. 1-9.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, 2016, pp. 770-778.

[15] F. Nian, T. Li, Y. Wang, M. Xu, and J. Wu, "Pornographic image detection utilizing deep convolutional neural networks," *Neurocomputing*, vol. 210, pp. 283-293, 2016.

[16] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, et al., "TensorFlow: a system for large-scale machine learning," in *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI)*, Savannah, GA, 2016, pp. 265-283.

Youngsoo Kim **https://orcid.org/0000-0002-6214-7222**

He received a B.S. degree in computer science in 1994 from the Republic of Korea Air Force Academy (KAFA), an M.S. degree in computer science engineering from Sogang University in 2001, and a Ph.D. degree in computer science engineering from the Korea Advanced Institute of Science and Technology (KAIST) in 2009. He worked as a researcher in the Battlefield Informatization Lab of the Military Development Research Center of the Korea National Defense Research Institute (KIDA) until 2021. He has been an assistant professor in the Department of Artificial Intelligence, Jeonju University, Korea. His research interests include artificial intelligence, IoT, cloud computing, and edge/fog computing.

Taehong Kim **https://orcid.org/0000-0001-6246-6218**

He received the B.S. degree in computer science from Ajou University, Suwon, South Korea, in 2005, and the M.S. degree in information and communication engineering and the Ph.D. degree in computer science from the Korea Advanced Institute of Science and Technology, Daejeon, South Korea, in 2007 and 2012, respectively. He worked as a Research Staff Member with the Samsung Advanced Institute of Technology and Samsung DMC R&D Center from 2012 to 2014. He also worked as a Senior Researcher with the Electronics and Telecommunications Research Institute, Daejeon, from 2014 to 2016. Since 2016, he has been an associate professor with the School of Information and Communication Engineering, Chungbuk National University, Cheongju, South Korea. His research interests include edge computing, federated learning, Internet of Things, and wireless sensor networks. Dr. Kim has been an Associate Editor of IEEE ACCESS since 2020.

Seongeun Yoo **https://orcid.org/0000-0003-3626-342X**

He received a B.S. degree in electronics and computer engineering from Hanyang University, Seoul, Korea, in 2003, and his M.S. and Ph.D. degrees in information and communications engineering from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea, in 2005 and 2010, respectively. Since September 2010, he has been a faculty member with the School of Computer and Communication Engineering, Daegu University, Gyeongsan, Korea. His research interests include real-time communication in wireless sensor networks and Internet of Things, as well as real-time embedded systems.