# Smoking Image Detection Based on Convolutional Neural Networks

Dongyan Zhang
School of Computer & Communication Engineering
University of Science & Technology Beijing
Beijing, China
e-mail: zhangdy@ustb.edu.cn

Cheng Jiao
Cisco Systems, Inc.
Beijing, China
e-mail: chejiao@cisco.com

Shuo Wang
School of Computer & Communication Engineering
University of Science & Technology Beijing
Beijing, China
e-mail: m13120375758@163.com

***Abstract*—With the development of Internet technology and the improvement of network quality, online videos have become increasingly popular. In particular, online live broadcast has become a hotspot in recent years, and smoking behavior in these broadcasts is harmful to smokers and the surrounding environment. Therefore, it is necessary to detect and thereby effectively control smoking behaviors in video content. Traditionally, smoking images are detected based on the detection algorithms of cigarette smoke. Given the limited resolution of live broadcast videos, cigarette smoke is not visually apparent in the video content. This paper proposes a smoking image detection model based on a convolutional neural network, referred to as SmokingNet, which automatically detects smoking behaviors in video content through images. This method can detect smoking images by utilizing only the information of human smoking gestures and cigarette image characteristics without requiring the detection of cigarette smoke, showing high accuracy and superior performance for real-time monitoring.**

***Keywords- live streaming; streaming media; image detection; CNN***

## I. INTRODUCTION

With the development of Internet technology and the improvement of network quality, online videos have become increasingly popular. In particular, online live broadcast has become a hotspot in recent years. Social apps such as Twitter and Facebook and mobile personal livecast (MPL) services have emerged and received much attention. With such social apps as Periscope[1] and Facebook Live in the U.S. and Inke1 in China, numerous geo-distributed amateur broadcasters can broadcast their video content live to viewers around the world. It is well known that smoking behaviors are harmful to both smokers and the surrounding environment. Therefore, it is necessary to use images to automatically detect whether there are smoking behaviors in video content.

Convolutional neural networks (CNNs) are a deep learning model. Here, "deep" indicates that, compared with shallow learning models, deep learning models involve neural networks with more hidden layers, and thus, the neural networks used for deep learning are called deep neural networks (DNNs). With the deepening of the research and the progress of computer hardware conditions, the number of layers of the deep learning models has increased from the initial value of 6[2] to more than 100 nowadays[3]. In this study, we design a CNN-based model called SmokingNet, which can automatically detect smoking behaviors in video content through images. Based on GoogLeNet, the model optimizes the characteristics of smoking images. With non-square convolution kernels, the model enhances the ability of feature extraction of the target images. Before model training, a super-large data set similar to the target image is used for pre-training the model. When the trained model is used to detect smoking images in the system, the full connection layers in the model are converted into convolution layers, which improves the detection ability of the model for local small targets while maintaining considerable detection efficiency.

## II. RELATED WORK

In recent years, some researchers have proposed smoking image detection methods based on image recognition technology. Inoue et al. assigned eigenvectors to low-dimensional spaces using subspace theory, and thereafter used feature clustering to classify cigarette smoke[4]. Although this method can achieve smoking identification through smoke classification, the threshold in the algorithm is empirically set and its value will change with the background, leading to high false detection rates and poor applicability. Wu et al. proposed a method for detecting smoking images by first subtracting the current frame from the background obtained from a Gaussian mixture model to generate the foreground of the motion, subsequently using the shape and color features to identify human hands, faces, and cigarettes in the foreground, combining the color features to detect smoke, and finally conducting a comprehensive assessment of the relative positions of the

978-1-5386-8339-2/18/$31.00 ©2018 IEEE

human hands, faces, cigarettes, and smoke to identify whether smoking behaviors appear[5]. In the method proposed by Iwamoto et al., each frame of the video was first divided into several image blocks; subsequently, the blocks were subjected to comprehensive analysis based on image features such as gradient, variance, kurtosis, and skewness, and the blocks conforming to smoke characteristics were labeled; finally, the morphological changes and area changes of the labeled blocks of each frame in the entire video time domain were subjected to statistics to determine whether the marked areas contain smoke so as to detect smoking images[6]. Odetallah et al. focused on the analysis of color characteristics of cigarette smoke by performing background differentiation in the three channels of the RGB color space so as to extract the foreground images conforming to color characteristics of cigarette smoke, and subsequently conducted comprehensive analysis of the area changes and distance changes of human faces versus smoke-like images so as to detect smoking in the video[7]. Bien et al. analyzed indoor smoking images based on the recognition of human gestures, in which human hands and heads were first identified through skin color along with a detection of their movements, and subsequently, a support vector machine was employed as a classifier for the probability characteristics of smoking behaviors, with the accuracy rate of classification reaching 83.33%[8].

Image recognition is the first breakthrough in deep learning. In 1995, LeCun et al., for the first time, used CNNs of deep learning models to successfully recognize handwritten digits[2]. As LeCun et al. incorporated image convolution operation into the learning models, the networks used by these models are called CNNs.

In 2012, Krizhevsky et al. used an improved CNN (named AlexNet)[9] to secure the first place in the ImageNet image recognition competition[10] with an error rate ten percent lower than that of the model that secured the second place. AlexNet not only inherits the advantages of CNNs, but also overcomes their side effects. At present, the ReLU activation function, LRN operation, and dropout technology together with the convolutional layers and the pooling layers constitute the basic structure of CNNs. In 2013, Zeiler et al. won the ImageNet competition by adjusting the network structure through a visualization technology of CNNs [11]. In 2014, Google introduced the GoogLeNet model [12] and won the first place in the ImageNet competition that year. Since GoogLeNet, it has been accepted in the academic community that a further increase in the number of CNN layers can improve the recognition accuracy, but the increase would make sample training more difficult [12-14]. In 2015, Srivastava et al. proposed highway networks[15], which allowed undecayed information flow across several layers through the gate mechanism so as to solve the problem of difficulty in training deep networks. He et al. proposed a residual neural network based on highway networks[3] and increased the number of network layers to 152, winning first place in the 2015 ImageNet competition..

## III. SMOKING IMAGE DETECTION BASED ON CNNS

The CNNs in deep learning have been widely used in image detection. The features to be extracted through CNNs for image recognition no longer need to be defined manually, and the feature extraction is achieved via automatic fitting through training. Each convolution operation can be regarded as a process of feature extraction, in which the weights of the convolution kernels are not preset but are continuously updated through training until the model converges, when the weights constitute the optimal feature extraction scheme. SmokingNet, a detection model based on CNNs, optimizes the characteristics of smoking images based on GoogLeNet and enhances the ability of feature extraction of the target images using non-square convolution kernels. This model is pre-trained with a super-large data set similar to target images prior to model training, and the trained model is used to detect smoking images.

### *A. Training Samples and Testing Samples*

In the detection of smoking images based on CNNs, the entire sample set is divided into three parts: training samples and validation samples in the training phase and testing samples in the testing phase. Moreover, the sample set is divided into positive and negative samples—a positive sample is an image showing smoking behaviors, whereas a negative sample is a background image.

#### *1) Positive samples*

Positive samples are collected from online smoking videos and the smoking videos made by our research group. Positive samples are acquired for training by playing these videos frame by frame. When a smoking frame appears, the image containing the complete cigarette is manually captured using a screenshot tool. The screenshot tool is specifically developed for this collection task based on the computer vision library OpenCV[16]. The screenshot tool contains a candidate box (as shown in the red box in Figure 1, which can be manually adjusted in size and moved to various positions with the mouse. When the left mouse button is pressed, the tool will save the local image in the candidate box to the local disk. To achieve sample expansion from a limited number of video images[17], the tool automatically performs three image transformations when capturing an image—horizontal mirror transformation, 45-degree clockwise rotation, and 45-degree counterclockwise rotation—generating local images at the same location of the original image and saving all of them to the disk, which indicates that a screenshot operation generates four images simultaneously (as shown in Figure 1). In addition, as all image samples are to be scaled to squares of the same size during the training phase, the candidate frame of the screenshot tool is always square, i.e., the saved training samples are all images with an aspect ratio of 1:1, so that image stretching can be avoided during the training phase to ensure the quality of the training samples.

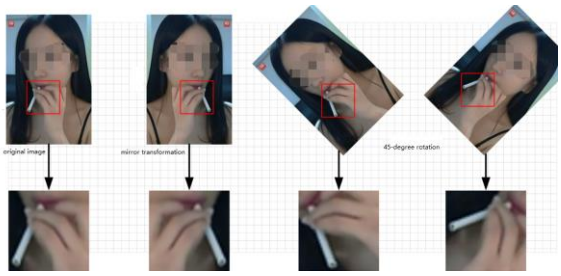


Figure 1. Example of acquisition of positive samples

*2) Negative samples*

As shown in Figure 2, there are five sources of negative samples: smoking-free images of the above smoking videos (hereafter referred to as smoking-free background images), LFW face data set[18], Oxford hand data sets[19], VOC2012 data set[20], and RWTH-PHOENIX-Weather sign language data set[21]. The VOC2012 data set is composed of natural images, including indoor and outdoor images of humans, animals, vehicles, and daily necessities. When processing samples of the VOC2012 data set, one randomly collects 50*50 to 150*150 pixels from each source image to generate its four local images as negative samples, and the smoking-free background images are processed similarly. For the images of face, hand, and sign language, local images are extracted according to the position information given in the data set. Given that the positive samples contain local images of the hand and the face in addition to cigarettes, in order to improve the detection accuracy of the model, the negative samples are not simply collected from smoking-irrelevant background images and are instead specifically enriched with data sets of the face and hand.


Figure 2. Example of acquisition of negative samples

*3) Sample sets*

In the testing phase, complete image frames are subjected to image detection using a trained CNN, and therefore, the testing samples here only refer to complete images of the above positive and negative samples. Accordingly, the entire sample set is divided into three parts—the training samples and validation samples in the training phase, and the testing samples in the testing phase—with their distribution shown in Table I. The number of images in the training phase is nine times that in the testing phase, and the samples are randomly assigned; the testing samples and the validation samples are also randomly assigned with a ratio of 9:1. As complete images are used in the testing phase, they are also referred to as testing samples.

TABLE I. DISTRIBUTION OF SAMPLE SETS

| | Source | Training phase | | | Testing phase | |
|---|---|---|---|---|---|---|
| | | *Image* | *Training sample* | *Validation sample* | *Image* | *Testing sample* |
| **Positive sample** | Internet video | 9303 | 33490 | 3722 | 1034 | 1034 |
| | On-site video | 2658 | 9568 | 1064 | 295 | 295 |
| **Sub-total** | | **11961** | **43058** | **4786** | **1329** | **1329** |
| **Negative sample** | VOC2012 | 10377 | 37357 | 4151 | 1153 | 1153 |
| | Oxford hand | 10074 | 9066 | 2008 | 1120 | 1120 |
| | RWTH-PHOENIX-Weather | 11030 | 9927 | 1103 | 1226 | 1226 |
| | LFW | 11910 | 10719 | 1191 | 1323 | 1323 |
| | Smoking-free background image | 17418 | 62704 | 6968 | 1935 | 1935 |
| **Sub-total** | | **60809** | **129773** | **15421** | **6757** | **6757** |
| **Total** | | **72770** | **172831** | **20207** | **8086** | **8086** |

## B. SmokingNet Structures

The convolution kernels of the CNN convolutional layers have been used to extract local features of a given image, and the features extracted by the first convolutional layer directly affect the feature fusion of the deep network. In most cases, the convolution kernels of a classical CNN model are squares[9,11,12], but, as cigarettes are strip-shaped, the square convolution kernels are not suitable for extracting the shape characteristics of cigarettes. Based on the shape characteristics of cigarettes, convolution kernels of four sizes are included in the first convolutional layer of SmokingNet, as shown in Figure 3. affect the feature fusion of the deep network. In most cases, the convolution kernels of a classical CNN model are squares[9,11,12], but, as cigarettes are strip-shaped, the square convolution kernels are not suitable for extracting the shape characteristics of cigarettes. Based on the shape characteristics of cigarettes, convolution kernels of four sizes are included in the first convolutional layer of SmokingNet, as shown in Figure 3.

As shown in Figure 3(a), if only a small convolution kernel of 3*3 pixels is used, the cigarette-like part of the positive sample background (as indicated by the black box) will interfere with feature extraction; if a large convolution kernel of 5*5 pixels as shown in Figure 3(b) is used, it is impossible to extract fine edge features; similarly, although the convolution kernels in Figure 3(c)(d) can well extract cigarette features in the horizontal and vertical directions, they cannot deal with other angles at which the cigarette may be positioned. Therefore, SmokingNet simultaneously uses these four convolution kernels in the first layer to enhance

the capability of CNNs for feature extraction, and in the second layer, fuses the feature graphs generated by different convolution kernels.

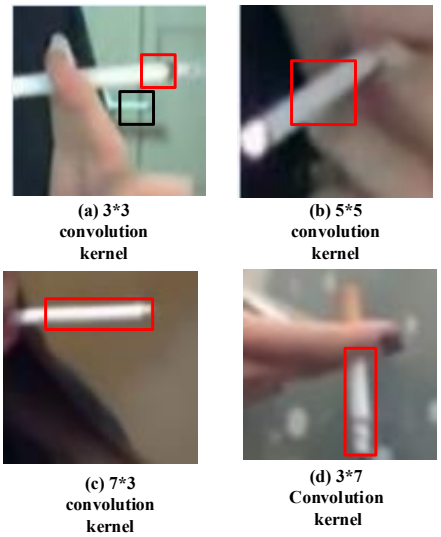


Figure 3. Example of four convolution kernels.

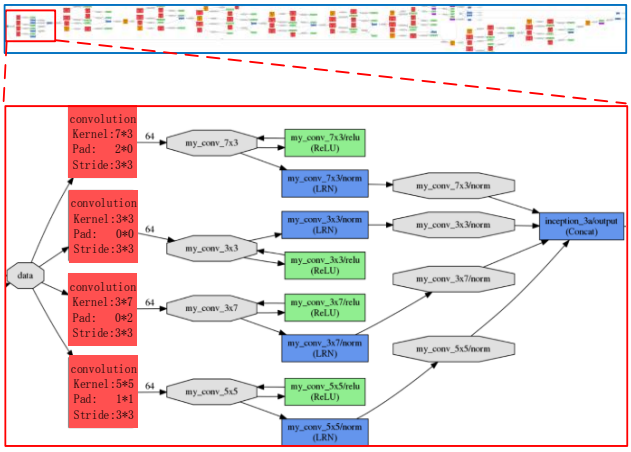


Figure 4. SmokingNet structure

As the average resolution of the original images of the training samples is 76*76 pixels, the input image of the SmokingNet training phase is set to 84*84 pixels in order to avoid image distortion owing to large-scale zoom. As shown by the top blue boxes in Figure 4, the structure of SmokingNet is the same as that of GoogLeNet except for the first two layers. The red boxes in Figure 4 refer to the first two layers of SmokingNet. The first layer is a convolution layer, whose convolution kernels of four sizes convolve the input images simultaneously, with each convolution kernel outputting 64 feature graphs, after which the feature graphs are normalized by the ReLU activation function and LRN and subsequently aggregated into the inception_3a/output layer. This layer does not perform any operation on the feature graphs but concatenates those feature graphs derived from four image sources (i.e., from one original image + three transformed images) into a single output, resulting in a total of 256 feature-graph outputs. The size of the feature-graph output from the convolutional layers can be calculated according to equation (4-1)[22]:

$$output_i = \frac{input_i + 2 \times pad_i - kernel_i}{stride_i} + 1, \quad i \in \{width, height\} \tag{4–1}$$

where kernel represents the size of a convolution kernel, pad represents the filled pixels at the edge of an image (usually filled with the mean of training sample pixels), and stride represents the step size of the convolution operation. The parameters of the first convolutional layer are shown in the red box in Figure 4, where the width values in the horizontal direction and the height values in the vertical direction are presented on the left and right sides of "*," respectively. Accordingly, it can be calculated that the size of each feature-graph output from the first layer is 28*28 pixels, which is consistent with the input size of the inception_3a/output layer of GoogLeNe, indicating that SmokingNet has the same network structure as GoogLeNet after the inception_3a/output layer. However, GoogLeNet is aimed at the classification of 1,000 types of ImageNet datasets, whereas SmokingNet is aimed at binary classification of smoking images, and hence, it is necessary to change the output of the last full connection layer of GoogLeNet from 1,000 to 2.

### C. *SmokingNet Training Methods*

For CNNs, the training data used in this test task are still not sufficient in quantity. If the model is trained by directly using randomized network weights, over-fitting is likely to occur, and therefore, it is necessary to conduct fine-tuning of the model after it has been subjected to pre-training based on a super-large data set so as to improve network performance[17,23] Generally, the closer the image content of the pre-training data set is to the detection target, the better the training effect[23], and hence, SmokingNet does not use the commonly-used ImageNet data sets for pre-training but selects 1mHand—a trained model developed by Koller et al. to identify sign language[24]—as the initial weight of the network, owing to the following three major reasons: first, the training samples of 1mHand contain and only contain human hands, trunks, and faces, and are thus similar to the cigarette-free background images of the positive smoking samples in which the parts in contact with cigarettes are all hands or mouths; second, 1mHand uses a super-large data set of more than 1 million data, which can improve the generalization ability of the model[23]; third, 1mHand directly uses GoogLeNet for training, and moreover, given that the overall structure of SmokingNet is similar to that of GoogLeNet, significant pre-training time can be saved, and fine-tuning can be initiated by directly using the network weight data file under the open Caffe framework of GoogLeNet.

TABLE II. TRAINING PARAMETERS

| batch | test_iter | test_interval | base_lr | lr_policy | gamma | stepsize | max_iter |
|---|---|---|---|---|---|---|---|
| 64 | 316 | 2700 | 0.01 | step | 0.1 | 230000 | 270000 |

The main parameters in the process of using the Caffe framework for training are shown in Table II, where batch indicates 64 training samples per iteration; test_interval indicates that the validation samples are subjected to accuracy testing every 2,700 iterations—given that the cycle in which all the training samples are trained once is called one epoch and that the 2,700 iterations result in the training of a total of 172,800 samples, which is a number similar to the total number of training samples, and therefore, the test cycle is considered one epoch; test_iter represents the number of iterations required for testing, and 316 iterations result in the testing of a total of 20,224 validation samples, i.e., all the validation samples are tested each time; max_iter represents the maximum number of iterations for training, which is set here to 270,000, a number approximately equivalent to 100 epochs; lr_policy represents the learning strategy, and when set to step, it will adjust the learning rate according to the formula (4-2)[22]:

$$learn\_rate = base\_lr \times gamma^{\lfloor iter/stepsize \rfloor} \quad (4\text{–}2)$$

Accordingly, it can be calculated that, when the number of iterations is less than 230,000 (approximately 85 epochs), the learning rate is base_lr, i.e., 0.01, whereas the learning rate becomes 0.001 when the number of iterations is greater than or equal to 230,000.

The performance indicators of a binary classification model usually include accuracy, precision, recall, and F-measure [25].

In each epoch of training, the validation samples are used to test the accuracy of SmokingNet, and the average (Loss) of the loss function values is calculated when all the validation samples are forward-propagated to the Softmax classifier, and hence, by the end of the training, a total of 101 tests will have been performed (the first test is performed before the training begins). As shown in Fig. 5, the accuracy and loss are inversely dependent on each other, with the model gradually converging after 85 epochs of training and the accuracy stabilizing at approximately 0.95 after 95 epochs.

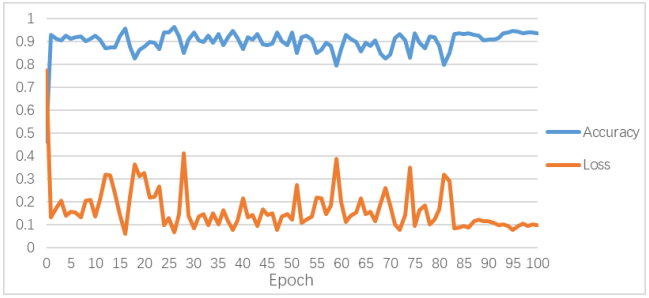


Figure 5. Variations of the accuracy and loss with respect to the epochs of training processes

## IV. EXPERIMENT

This section will further analyze the detection performance of this method through experimental comparison between SmokingNet and other models. The experimental software and hardware conditions are listed in Table III.

TABLE III. EXPERIMENTAL HARDWARE AND SOFTWARE CONDITIONS

| | | | |
|---|---|---|---|
| **Hardware** | **CPU** | **Memory** | **GPU** |
| | Intel Xeon E3-1230 v5 | 32GB | NVIDIA GTX1070 |
| **Software** | **Operating System** | **GPU Driver** | **Others** |
| | Windows7 sp1 64bit | GeForce 372.90 CUDA 8.0 + cuDNN v5 | Caffe(20160922) OpenCV 2.4.13 |

First, the true positive rate (TPR) and false positive rate (FPR) of the six models i.e., SmokingNet, GoogLeNet, AlexNet, Haar feature + Adaboost classifier [26], LBP feature + Adaboost classifier [27], and HOG feature + SVM classifier [28] calculated according to the methods shown in Figure 6 are tested under different thresholds, and the results are used to generate receiver operating characteristics (ROC) curves of each model [25] as shown in Figure 6. Both SmokingNet and GoogLeNet are tested with two pre-trained data sets—the 1mHand data set [24] and the ImageNet data set [10]—whereas AlexNet is tested with only the ImageNet data set. Given that uneven numbers of positive testing samples versus negative testing samples may cause distortion of other performance indicators, the positive and negative samples are maintained at a ratio of 1:1 in this experiment; with 1,329 positive testing samples as the reference, 1,329 images are randomly selected from the negative testing samples. Each model is tested five times, and subsequently, each indicator is reported as a mean over the five tests.

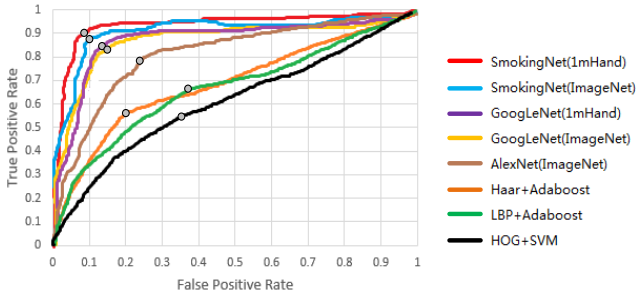


Figure 6. ROC curves of various models

The closer the ROC curve is to the upper left corner in the above figure, the better the model classification performance[25]. As shown in Figure 6, SmokingNet pre-trained with 1mHand is superior to other models. The calculated area under an ROC curve (AUC) of a model is an important quantitative indicator for comparing the detection performances of the models [25]. In addition, to further comprehensively compare the performances of the models, the probability threshold at the upper left turning point (shown by the circles in Figure 6 on each ROC curve is used to calculate the precision, accuracy, recall, and F-measure of each model. In addition, the average number of images detected each second, expressed in frames per second (FPS),

is tested for each model to compare the detection efficiencies of the models.

TABLE IV. PERFORMANCE DATA OF EACH MODEL

| Model | Pre-training | Detection mode | AUC | Threshold | Precision | Recall | Accuracy | F-measure | FPS |
|---|---|---|---|---|---|---|---|---|---|
| **Smoking Net** | 1mHand | FCN | 0.95 | 0.74 | 0.90 | 0.90 | 0.90 | 0.90 | 88 |
| | ImageNet | FCN | 0.92 | 0.71 | 0.89 | 0.88 | 0.89 | 0.88 | 89 |
| **GoogLe Net** | 1mHand | FCN | 0.88 | 0.69 | 0.87 | 0.84 | 0.85 | 0.85 | 73 |
| | ImageNet | FCN | 0.87 | 0.75 | 0.85 | 0.83 | 0.84 | 0.84 | 71 |
| **Alex Net** | ImageNet | FCN | 0.79 | 0.69 | 0.77 | 0.79 | 0.78 | 0.78 | 164 |
| **Haar+ Adaboost** | | Sliding window | 0.68 | 0.81 | 0.74 | 0.56 | 0.68 | 0.64 | 31 |
| **LBP+ Adaboost** | | Sliding window | 0.66 | 0.57 | 0.65 | 0.67 | 0.66 | 0.66 | 46 |
| **HOG+ SVM** | | Sliding window | 0.61 | 0.62 | 0.61 | 0.55 | 0.60 | 0.58 | 35 |

The results of each testing are shown in Table IV. Among them, the precision, accuracy, recall, and F-measure of each model are generated at a certain threshold so that they only reflect the local performance of the model, but because the threshold is optimal, the results are still of some reference significance. However, AUC is not affected by the ratio of positive samples to negative samples[25], and its value is based on the entire ROC curve, which allows AUC to reflect the overall performance of a model more accurately. As shown in Table IV, SmokingNet is superior to other models in terms of classification performance, and the model pre-trained with 1mHand is slightly better than that pre-trained with ImageNet, indicating that the 1mHand data set is more similar to the detection target than the ImageNet data set and thereby provides initial weights closer to the optimal weight. Moreover, the size of the input images (224*244 pixels) for GoogLeNet during the training phase is too different from the original size of the training samples and GoogLeNet uses a single-size convolution kernel, which accounts for the experimental observation that the detection performance of GoogLeNet is slightly worse than that of SmokingNet. AlexNet is smaller than half of GoogLeNet in terms of the number of layers, and each of its performance indicators is lower, indicating that the number of layers in a neural network model is a key factor affecting the detection performance of the model. Compared with the first three deep learning models in Table IV, the latter three shallow learning models only extract shallow features of images and use specific classifiers to detect images, thereby making their detection performances significantly worse than those of the deep learning models. Furthermore, the HOG feature is sensitive to the direction of a target object[28], and hence, the model HOG feature + SVM classifier exhibits worse performance than the other two shallow learning models. Notably, fully convolutional network (FCN)-based deep learning models can use GPU for parallel computing whereas shallow learning models have no GPU-based solution yet and can only use CPU for computing, and there is a difference between FCN algorithms and sliding-window algorithms. Therefore, with respect to detection efficiency, FCN-based deep learning models are evidently superior to shallow learning models, which adopt a sliding-window detection mode. The factor affecting the FCN detection efficiency is usually the model depth. As shown in Table IV the FPS of AlexNet is more than twice that of GoogLeNet and is nearly twice that of SmokingNet. SmokingNet is a streamlined form of GoogLeNet, and despite its far lower detection efficiency compared with that of AlexNet, it has still achieved a detection efficiency more than three times the target FPS (25), indicating that it is fully capable of achieving real-time detection of smoking images

## V. CONCLUSION

In this study, we design and implement a deep learning model, i.e., SmokingNet, which is specially optimized for smoking images. In addition, we conduct large-scale model training and testing sample collection with respect to smoking as a specific detection target. Based on GoogLeNet, the detection accuracy of smoking images is enhanced by improving the network structure and using special convolution layers to better extract the characteristics of smoking images. In addition, specific pre-training models are selected for SmokingNet based on detailed analysis of the characteristics of smoking images, and subsequently, the training parameters and training process are elaborated. Finally, the detection performance of SmokingNet is tested by comparison with those of other deep and shallow learning models. The experimental results show that, compared with the classical deep learning models, SmokingNet shows significantly improved detection performance with precision and recall over one percent higher than those of the second-best model and with a detection efficiency as high as 80 FPS, indicating that SmokingNet is fully capable of achieving real-time detection of smoking images during live webcast.

## REFERENCES


[1] B. Wang, X. Zhang, G. Wang, H. Zheng, and B. Y. Zhao, "Anatomy of a personalized livestreaming system," in *IMC*, pp. 485-498, ACM, 2016.

[2] YL. Cun, and Y. Bengio, "Convolutional networks for images, speech, and time series," The handbook of brain theory and neural networks, vol 3361(10), 1995.

[3] K. He , X. Zhang , S. Ren, and J Sun, "Deep Residual Learning for Image Recognition," Computer Science, pp. 770-778, 2015.

[4] H. Inoue, "Image-based smoke detection with k-Subspaces clustering," International Workshop on Nonlinear Circuits and Signal Processing, pp. 321-324, 2009.

[5] P. Wu , JW. Hsieh , JC. Cheng , SC. Cheng , and SY. Tseng, "Human smoking event detection using visual interaction clues," Pattern Recognition (ICPR), 2010 20th International Conference, pp. 4344-4347, IEEE, 2010.

[6] K. Iwamoto, H. Inoue, and T. Matsubara, "Cigarette smoke detection from captured image sequences," International Society for Optics and Photonics,pp. 753813-753813-10, 2010.

[7] A. D. Odetallah, and S. S. Agaian, "Human visual system-based smoking event detection," SPIE Defense, Security, and Sensing. International Society for Optics and Photonics, pp. 840607-840607-12, 2012.

[8] T. L. Bien, and C. H. Lin, "Detection and recognition of indoor smoking events," Proc. of SPIE, vol. 8784, pp. 878424-1, 2013.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," Advances in neural information processing systems. 2012: 1097-1105.

[10] J. Deng, W. Dong, R. Socher, L. J. Li, and K. Li, "Imagenet: A large-scale hierarchical image database," Computer Vision and Pattern Recognition, pp. 248-255, IEEE, 2009:.

[11] M. D. Zeiler, and R. Fergus, "Visualizing and understanding convolutional networks," European Conference on Computer Vision. Springer International Publishing,vol.8689, pp. 818-833, 2014.

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

[13] K. Simonyan, and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," Computer Science, 2014.

[14] C. Y. Lee, S. Xie, P. Gallagher, et al, "Deeply-Supervised Nets," AISTATS, vol. 2(3) ,pp.6, 2015.

[15] R. K. Srivastava, and K. Greff, "Schmidhuber J. Highway Networks," Computer Science, 2015.

[16] OpenCV[EB/OL]. http://opencv.org.

[17] Y. Bengio, "Practical recommendations for gradient-based training of deep architectures," Neural Networks: Tricks of the Trade. Springer Berlin Heidelberg, vol. 7700 (1-3), pp. 437-478, Springer Berlin Heidelberg 2012.

[18] G. B. Huang, M. Ramesh, and T. Berg, et al, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," University of Massachusetts, Amherst, pp. 07-49, Technical Report, 2007.

[19] A. Mittal, A.Zisserman, and P. H. S. Torr, "Hand detection using multiple proposals," BMVC, vol. 40 (12), pp. 75.1-75.11, 2011.

[20] M. Everingham, L. Van Gool, and C .Williams, et al, "The PASCAL Visual Object Classes Challenge 2012 Results," Available from www.pascal-network.org, 2012.

[21] J. Forster, C. Schmidt, and T. Hoyoux, et al, "RWTH-PHOENIX-Weather: A Large Vocabulary Sign Language Recognition and Translation Corpus," LREC, vol. 9, pp 3785-3789 2012.

[22] Caffe | Deep Learning Framework[EB/OL]. http://caffe.berkeleyvision.org.

[23] F. F. Li, Andrej Karpathy, Justin Johnson. CNNs in Practice[EB/OL].http://cs231n.stanford.edu/slides/winter1516_lecture 11.pdf, 2016.

[24] O. Koller, H. Ney, and R. Bowden, "Deep Hand: How to Train a CNN on 1 Million Hand Images When Your Data Is Continuous and Weakly Labelled," The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3793-3802, IEEE, 2016.

[25] T .Fawcett, "An introduction to ROC analysis," Pattern recognition letters, vol.27(8), pp. 861-874, 2006.

[26] P. Viola, and M.Jones, "Rapid object detection using a boosted cascade of simple features," Computer Vision and Pattern Recognition, vol. 1(2), pp.511, IEEE, 2001.

[27] L. Zhang, R. Chu, and Xiang S, et al, "Face detection based on multi-block lbp representation," International Conference on Biometrics. Springer Berlin Heidelberg, vol. 4642, pp. 11-18, Proceedings. DBLP, 2007.

[28] N. Dalal, and B.Triggs, "Histograms of oriented gradients for human detection," Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). vol. 1 (12),pp. 886-893, IEEE, 2005.