Fire Detection in Infrared Video Surveillance Based on Convolutional Neural Network and SVM

Kewei Wang, Yongming Zhang, Jinjun Wang, Qixing Zhang*
State Key Laboratory of Fire Science
University of Science and Technology of China
Hefei, China
e-mail: qixing@ustc.edu.cn

Bing Chen
China Academy of Safety Science and Technology
Beijing, China

Dongcai Liu
School of Electronic and Information Engineering
Hefei Normal University
Hefei, China

***Abstract*—In this paper, a novel algorithm based on convolutional neural network (CNN) and support vector machine (SVM) for fire detection in infrared (IR) video surveillance is proposed. To improve the performance of IR fire detection, we develop a 9-layer convolutional neural network named IRCNN instead of traditional empirically handcrafted methods to extract IR image features. Then, a linear support vector machine is trained with extracted features to achieve fire detection. Our network adopts data augmentation technique and Adam optimization to deal with problems caused by the insufficient dataset, and accelerate the training process. Experimental results show that our method achieved both high precision (98.82%) and high recall (98.58%) on our IR flame dataset and real-time detection on the ordinary infrared surveillance cameras.**

Keywords-Fire detection; Infrared video surveillance; Feature extraction; Convolutional neural network; SVM

I. INTRODUCTION

Fire is one of the most frequent and serious threats to public safety and social development. Due to the time consumption and limited detection distance, traditional point smoke and temperature sensors can not satisfy our needs in large and open spaces such as forests, shopping centers and airports. With the development of computer vision and pattern recognition, video-based fire detection is increasingly applied in our life [1]. When there is no or little visible light, or the color of object to be detected is similar to the background, visual fire detection will not work [2]. In contrast, infrared thermal cameras could monitor around the clock, regardless of illumination and climate conditions. With the development of UAVs and continuous decline of the infrared cameras cost, IR fire detection technology will be increasingly applied in our daily life. Nevertheless, traditional infrared fire detection algorithms belong to empirically handcrafted methods that are time consuming and labor-intensive [3-6].

In recent years, deep convolutional neural networks (CNNs) have demonstrated state-of-the-art performance on image classification and object detection. With the breakthrough of CNNs in computer vision, some researchers applied this technology into fire smoke and flame detection in video surveillance. Inspired by R-CNN, Shi *et al.* [7] proposed a method of video-based flame detection by combining image saliency detection with convolutional neural network. Xu *et al.* [8] applied the domain adaptation method to build deep architecture, which acted as confusing the distributions of features extracted from synthetic and real smoke images, expanding the domain-invariant feature space of smoke images off non-smoke images. As far as we know, CNNs have achieved great performance in visible smoke and flame detection, but there are no researchers applying it to the fire detection based on infrared videos so far. Compared with traditional empirically handcrafted methods, CNNs could extract more essential features of fire images automatically, which contributes to the subsequent classification. The CNNs can be regarded as complicated filters, the parameters of which are learned from train set. They extract the shallow features firstly and combine them into high-level abstract features layer by layer. Yin *et al.* [9] developed a deep normalization and convolutional neural network (DNCNN) and achieved very low false alarm rates below 0.6% with detection rates above 96.73% on their smoke datasets. And in [10], the researcher replaced the last softmax layer of CNNs with support vector machine and achieved better performance in MINST and Cifar-10 datasets. In our work, we also found that the accuracy of CNN-SVM is better than single convolutional neural network. Therefore, we propose the fire detection algorithm in infrared video surveillance based on CNN-SVM in this paper, which combines a 9-layer convolutional neural network with SVM to complete thermal image features extraction and flame classification.

The rest of this paper is organized as follows. Section II reviews deep convolutional neural networks and representation learning. In Section III, the 9-layer convolutional neural network (IRCNN) is described in detail. Section IV presents experimental results and discussion. Finally, conclusions are given in Section V.

978-1-5386-6396-7/18/$31.00 ©2018 IEEE

II. Related Work

Deep convolutional neural networks have a long history in computer vision. LeNet-5 [11], the pioneering 7-layer convolutional network was proposed by LeCun *et al.* in 1998 and was applied by several banks to recognize hand-written numbers on checks digitized in 32x32 pixel images. In [12], Krizhevsky *et al.* developed the 8-layer AlexNet and won the first place in ImageNet Large Scale Visual Recognition Challenge (ILSVRC 2012). With the development of computer resources and large scale datasets, the architectures of CNNs become more and more deeper (19-layer VGG, 22-layer GoogLeNet and 152-layer ResNet) and extract more essential high-level abstract features, contributing to the better performance on the baseline. Due to the excellent performance of CNNs, many researchers applied them to their own domains. In [13], researchers presented a classifier for marine animal classification using the combination of CNN with handcrafted features. Huynh *et al.* [14] constructed a convolutional neural network for the vehicle detection tasks and achieved positive and stable performance on the motorbike test dataset. Wu *et al.* [15] proposed a novel pipeline built upon deep CNN features to harvest discriminative visual objects and parts for scene classification. Simonyan *et al.* [16] developed a two-stream ConvNet architecture which incorporated spatial and temporal networks for action recognition in videos and achieved competitive performance with the state of the art on the standard video actions benchmark of UCF-101 and HMDB-51. It turns out that by means of the CNNs, excellent performance could be obtained in most specific fields.

The performance of machine learning methods critically depends on the data representation (or features), and for this reason, much of the actual efforts in deploying machine learning algorithms go into the design of preprocessing pipelines and data transformations [17]. Girshick *et al.* [18] used a large convolutional neural network to extract a fixed-length feature representation from candidate region proposals and fed them into SVM. In [19], researchers studied automatic extraction of feature representation through deep neural network (DNN) for medical image analysis and confirmed automatic feature extraction outperformed manual methods. Xu *et al.* [20] leveraged deep convolutional neural networks to advance event detection and introduced a discriminative video representation for event over a large scale video dataset when only limited hardware resources were available. Handcrafted features are labor-intensive and need prior knowledge based on expert system, which highlights the drawbacks of the traditional algorithms which are inefficient to extract the discriminative features of different fields. In contrast, the CNNs have a strong ability to extract image features and output the more essential representation.

III. Our Method

In this paper, we proposed a novel algorithm for fire detection in infrared video surveillance based on the deep convolutional neural network and support vector machine. Fig. 1 shows the overall flowchart of our method, which consists of three modules. Firstly, we build our IR flame dataset with images from infrared videos and expand dataset with data augmentation. Then, a 9-layer convolutional neural network named IRCNN is developed to extract the essential features from IR images. Finally, extracted features are sent into the classifier SVM to detect flames.

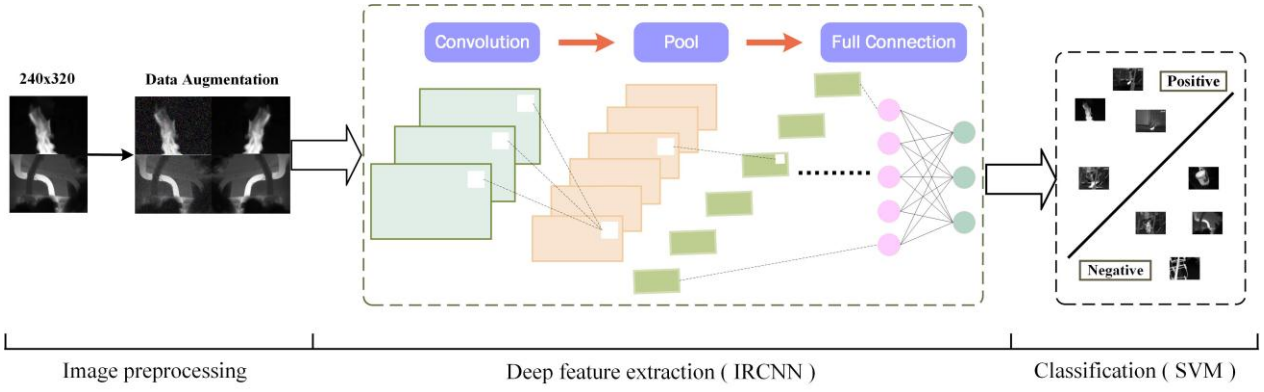

Figure 1. Framework of the infrared flame detection algorithm

A. IRCNN Architecture

Inspired by [12, 21, 22], we proposed the 9-layer IRCNN to extract features from infrared images. The details of the architecture are listed in Table I. IRCNN contains nine layers with weights, the first six are convolutional layers and the remaining three are fully-connected layers. Image features are computed by forward propagation and the output of last fully-connected layer is fed to a 2-way softmax layer which produces a distribution over the 2 way labels. The kernels of the convolutional layers are connected to all kernel maps in previous layer and the neurons in the fully-connected layers are connected to all neurons in the previous layer. Each layer contains learnable weights and consists of a linear transformation followed by a nonlinear mapping, which is implemented by Rectified Linear Unit (ReLU) to reduce the training time with gradient descent. Max pooling is applied to the first, second, fourth and sixth layers for translational invariance, and we add dropout layer in the first and second fully-connected layers to avoid overfitting.

In [21], the researchers noted that Local Response Normalization (LRN) did not improve the performance on the dataset, but led to increased memory consumption and computation time. Moreover, in [22], by visualizing the first and second layers of AlexNet, Zeiler *et al.* found that the first layer filters were a mixture of extremely high and low frequency information and the second layer showed aliasing artifacts because of the large kernel size 11×11 and large stride 4. Therefore, we removed the LRN and changed the kernel size to 5×5 and stride to 2 in our convolutional neural network. The first convolutional layer filters the 240×320×3 input raw image with 128 kernels of size 5×5 with a stride of 2 pixels. The second convolutional layer takes as input the outputs of the first layer and filters them with 384 kernels of size 3×3. At last, the fully-connected layers have 2048 neurons each.

TABLE I. " ARCHITECTURE DETAILS OF IRCNN (C: CONVOLUTIONAL LAYER; R: RELU; P: POOLING LAYER; F: FULLY-CONNECTED LAYER; D: DROPOUT; S: SOFTMAX LAYER)

IRCNN Architecture									
Layer	1	2	3	4	5	6	7	8	9
Input size	240×320×3	59×79×128	15×20×384	15×20×512	4×5×1024	4×5×512	4×5×256	2048×1	2048×1
Type	C + R + P	C + R + P	C + R	C + R + P	C + R	C + R + P	F + R + D	F + R + D	S
Kernel size	5×5	3×3	3×3	3×3	3×3	3×3	-	-	-
Max pooling	2×2	2×2	-	2×2	-	2×2	-	-	-
Output size	59×79×128	15×20×384	15×20×512	4×5×1024	4×5×512	4×5×256	2048×1	2048×1	2×1

Gradient descent is significantly important for optimization of deep neural networks. In the AlexNet architecture, the stochastic gradient descent (SGD) algorithm was adopted to optimize the convolutional neural network. However, the SGD algorithm is difficult to select the proper learning rate and easy to be trapped in their numerous suboptimal local minima, which leads to bad convergence. A learning rate that is too small leads to painfully slow convergence, while a learning rate that is too large can hinder convergence and cause the loss function to fluctuate around the minimum or even to diverge. In our architecture, we adopted the Adaptive Moment Estimation (Adam) [23] algorithm to avoid the aforementioned problems. In addition to storing an exponentially decaying average of past squared gradients v_t like RMSprop, Adam also keeps an exponentially decaying average of past gradients m_t, similar to Momentum:

$$m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t \tag{1}$$

$$v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2 \tag{2}$$

where m_t and v_t are estimates of the first moment (the mean) and the second moment (the uncentered variance) of the gradients g_t respectively, and $\beta_1, \beta_2 \in [0,1)$ are the hyper-parameters that control the exponential decay rates of these moving averages.

m_t and v_t are biased towards zero, especially during the initial time steps and when the decay rates are small. So Adam counteracts these biases by computing bias-corrected first and second moment estimates:

$$\hat{m}_t = \frac{m_t}{1-\beta_1^t} \tag{3}$$

$$\hat{v}_t = \frac{v_t}{1-\beta_2^t} \tag{4}$$

Then, these bias-corrected first and second moment estimates are used to update the parameters, which yields the Adam update rule:

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{\hat{v}_t} + \varepsilon} \hat{m}_t \tag{5}$$

where θ_t, η are the model's parameters and learning rate and ε is a smoothing term that avoids division by zero. (usually on the order of $1e-8$.)

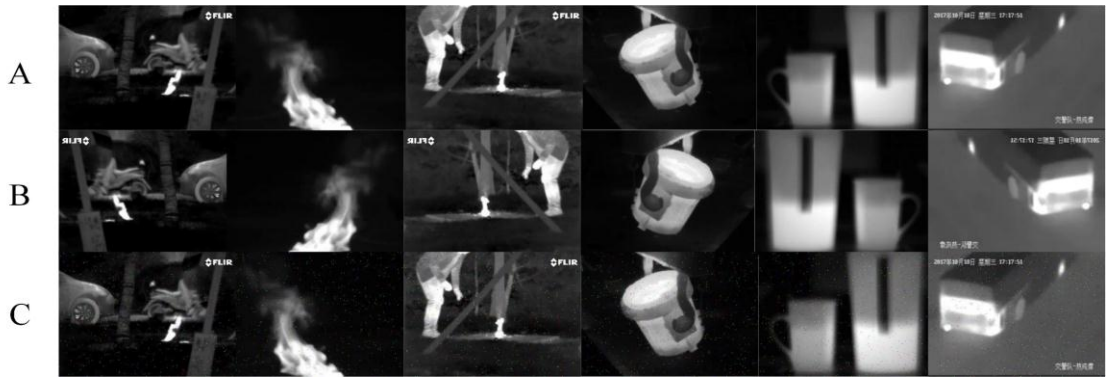

Figure 2. IR images in the dataset. A: original images; B: images generated by horizontal flip; C: images generated by salt and pepper noise.

IV. EXPERIMENTS

We used the open source TensorFlow platform to construct and train the proposed IRCNN and scikit-learn to implement the linear support vector machine. Features extraction (IRCNN) experiments were carried out in the Ubuntu 16.04 operative system running on a PC with Intel(R) Core(TM) i7-6850 CPU @ 3.60GHz and 4 NVIDIA GTX 1080 Ti GPUs and image classification (SVM) experiments were carried out in the Ubuntu 16.04 operative system running on a PC with Intel(R) Core(TM) i5-6500 CPU @ 3.20GHz.

A. IR Image Dataset

Because there were no public IR datasets for fire detection in previous literatures, we built an infrared flame dataset to train the IRCNN. 135 infrared flame video sequences (about 1-2 minutes) were captured with the FLIR-A310 camera (7.5-13μm), which provided most of the positive flame samples. For the negative samples without flames, we captured 42 infrared video sequences (about 1 minutes) and collected 118 video sequences from the internet. Moreover, 1044 negative infrared images were obtained from FLIR corporation and the public infrared video surveillance, which contributed to the diversity of our dataset. Then, we extracted one frame every two seconds from the IR sequences and obtained an infrared flame dataset composed of 5300 positive images and 6100 negative images at last. Considering the low resolution of the infrared camera, we directly adopted the resolution 240x320 as the input image size of IRCNN and the rest of images in our dataset were resized to 240x320.

1) Data Augmentation: The scale of train set directly determines the performance of deep convolutional neural networks since they often contain millions of parameters to be learned. And it is labor-intensive and exhausting to generate plenty of samples just by capturing and collecting infrared videos. Therefore, we produced more IR images from the original dataset by using data augmentation technique. Considering the surveillance camera orientation and complex conversion of infrared videos, two strategies of data augmentation were adopted to generate more data, including horizontal flip and salt and pepper noise, as shown in Fig. 2. Then, we mixed newly-generated samples and original samples to obtain the final and larger train set. Table II shows three sets for training and testing. The number of Set2 generated by data augmentation technique is much larger than the original Set1, which contributes to enhancing performance of IRCNN. As shown in Table III, experiment 1 and experiment 4 demonstrated that the IRCNN precision performance was improved by 3.86% using data augmentation technique in this paper.

TABLE II. DATASET FOR TRAINING AND TESTING

Dataset	Number of positive samples	Number of negative samples	Sum of samples	Purpose
Set1	4000	4600	8600	Train
Set2	12000	13800	25800	Train
Set3	1300	1500	2800	Test

TABLE III. EXPERIMENTAL RESULTS OF DIFFERENT METHODS

Experiment	Method	Input size	Optimization	P	R	F1	Dataset
1	IRCNN (A)	240x320	Adam	95.01%	97.45%	96.51%	Set1 and Set3
2	AlexNet (B)	224x224	SGD	96.99%	97.41%	97.24%	Set2 and Set3
3	IRCNN (C)	240x320	SGD	97.70%	97.08%	97.39%	Set2 and Set3
4	IRCNN (D)	240x320	Adam	98.87%	97.68%	98.27%	Set2 and Set3
5	IRCNN + SVM (E)	240x320	Adam	98.82%	98.58%	98.70%	Set2 and Set3

B. Results and Discussion

In this paper, we carried out plenty of experiments to select the best method for infrared flame detection. Precision (P), Recall (R) and F1 score were measured to quantitatively compare the performance of different methods on the test set. We focus on the infrared flame images and our goal is to achieve high precision and recall at the same time.

To test the performance of IRCNN, Adam optimization, data augmentation and combination with SVM, we conducted a large number of experiments under five different conditions. All experiments were tested on the Set3 and the results are listed in the Table III. At least five repeated tests were conducted under same conditions to reduce error. On the Set3, method E proposed in this work, achieved the highest F1 score and recall among the five conditions and obtained higher precision than other methods, except method D. Although our method E obtained slightly lower precision than method D, the recall of our method E was much higher than the method D. Moreover, Table III showed that the method D achieved both higher precision and recall than method B, which demonstrated that IRCNN proposed in this paper was superior to the classic AlexNet network for IR flame detection. We guess that the large kernel size 11x11 and stride 4 in the first convolutional layer lead to the loss of key features. The network IRCNN obviously became worse when the training set was replaced with Set1. Hence, data augmentation technique plays an important role in our network and contribute to extracting more essential features of infrared images.

To sum up, in combination with SVM, IRCNN displays the best performance than other methods, and data augmentation helps to extract salient features of IR images. In addition, the IRCNN architecture proposed in this paper is

better than AlexNet for IR flame detection on this limited dataset.

The training loss behaviors of different methods in 10000 steps are illustrated in Fig. 3. Compared to method B and C with SGD optimization, the training loss of our IRCNN network with Adam decreases smoothly, and the training loss of IRCNN with Adam optimization reaches 0.085 after just 1000 steps. This indicates that the Adam optimization algorithm could adjust the learning rate more wisely during the training process, leading to the fast convergence. In addition, it can be observed that compared to the IRCNN architecture, the AlexNet network fluctuates strongly except slow convergence. This demonstrates that the IRCNN architecture is better than AlexNet network for IR flame detection again.

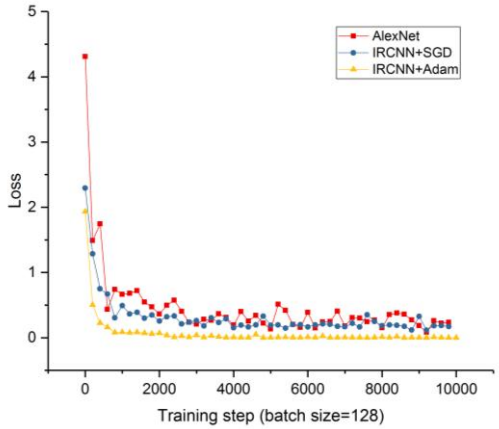

Figure 3. Training loss behaviors of method B, C and D.

C. Time Analysis

It is generally believed that convolutional neural networks take a significant amount of time to execute. In this section, we evaluated the computation time over multiple layers and SVM classifier in detail to demonstrate the practicality of our algorithm. The cumulative computation time of IRCNN layers and SVM are illustrated in Fig. 4 when classifying ten input IR images each time. The total time of ten IR images classification is 0.487s, which means that our algorithm could classify at least twenty IR images per second. Therefore, the proposed algorithm completely satisfies the requirements of ordinary infrared surveillance cameras. In addition, it can be observed that the layers C1-C4 consume most of the computation time. The reason is that the first four layers transform the input size 240x320 to 4x5 while the last layers just alter the number of feature maps or dimensions. Therefore, we can conclude that the size of input image determines the classification time and that's why classic deep convolutional networks adopt the little patches of images to train and test models.

The computation time distribution over different layer types and SVM classifier is illustrated in Fig. 5. As shown in Fig. 5, convolutional layers consume most of the computation time, and pooling layers occupy more time than fully-connected layers. As we all know, the fully connected layers usually need more time to learn their parameters that is far more than the parameters of convolutional and pooling layers when training the convolutional neural networks. However, when testing the images with the trained model, the results turned to be reversed. Moreover, the last classifier layer replaced by SVM consume the least time, which demonstrates that our method proposed in this paper is effective and feasible.

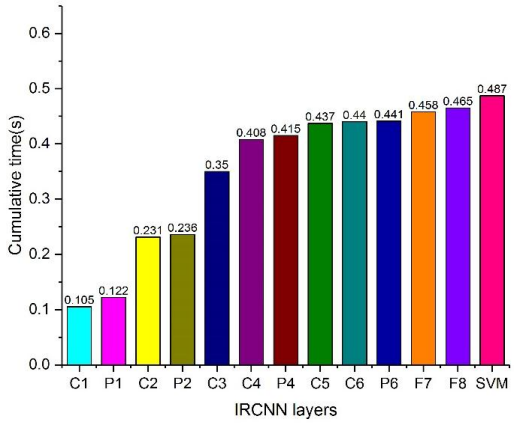

Figure 4. Cumulative time of IRCNN layers and SVM classifier. (C: convolutional layer; P: pooling layer; F: fully connected layer.)

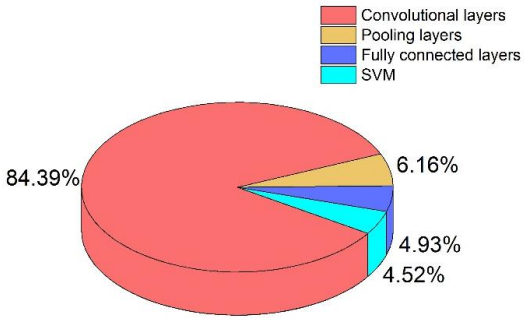

Figure 5. The distribution of computation time over different layer types and SVM classifier.

V. CONCLUSION

It is a challenging task to recognize fire from infrared video surveillance. In this paper, a 9-layer convolutional neural network named IRCNN was proposed to extract IR image features, automatically. We adopted kernel size 5x5 and stride 2 in the first convolutional layer to avoid the loss of key features, and Adam optimization to accelerate the training process. Compared to AlexNet network, our IRCNN architecture is better and achieves the state-of-the-art performance on the IR flame dataset with precision 98.82%, recall 98.58% and F1 score 98.70%. In addition, we evaluated the computation time over multiple layers and SVM classifier and found that the size of input image determined total computation time, and the convolutional layers consumed most of the time. Moreover, our algorithm could classify at least twenty IR images per second, indicating that we can achieve real-time fire detection on the ordinary infrared surveillance cameras.

As it is labor-intensive and exhausting to build IR flame dataset just by capturing and collecting infrared videos, our future work will focus on the methods to generate more data, such as GANs [24] or domain adaptation [25].

ACKNOWLEDGMENT

This work is supported by the National Key Research and Development Plan (Grant No. 2017YFC0805100 and 2016YFC0800100), and the Anhui Provincial Key Research and Development Program (Grant No. 1704a0902030).

REFERENCES

[1] A. E. Çetin, K. Dimitropoulos, B. Gouverneur, N. Grammalidis, O. Günay, Y. H. Habiboğlu, et al., "Video fire detection – Review," Digital Signal Processing, vol. 23, pp. 1827-1843, 2013.

[2] J. Han and B. Bhanu, "Fusion of color and infrared video for moving human detection," Pattern Recognition, vol. 40, pp. 1771-1784, Jun 2007.

[3] B. U. Töreyin, R. G. Cinbiş, Y. Dedeoğlu, and A. E. Çetin, "Fire detection in infrared video using wavelet analysis," Optical Engineering, vol. 46, pp. 067204-067204-9, 2007.

[4] O. Günay and A. E. Çetin, "Compressive Sensing based flame detection in infrared videos," in Signal Processing and Communications Applications Conference (SIU), 2013 21st, 2013, pp. 1-4.

[5] W.-H. Kim, "DSP Embedded Early Fire Detection Method Using IR Thermal Video," KSII Transactions on Internet and Information Systems (TIIS), vol. 8, pp. 3475-3489, 2014.

[6] O. Günay and A. E. Çetin, "Real-time dynamic texture recognition using random sampling and dimension reduction," in Image Processing (ICIP), 2015 IEEE International Conference on, 2015, pp. 3087-3091.

[7] L. Shi, F. Long, C. Lin, and Y. Zhao, "Video-Based Fire Detection with Saliency Detection and Convolutional Neural Networks," in International Symposium on Neural Networks, 2017, pp. 299-309.

[8] G. Xu, Y. Zhang, Q. Zhang, G. Lin, and J. Wang, "Deep domain adaptation based video smoke detection using synthetic smoke images," Fire Safety Journal, vol. 93, pp. 53-59, 2017.

[9] Z. Yin, B. Wan, F. Yuan, X. Xia, and J. Shi, "A deep normalization and convolutional neural network for image smoke detection," IEEE Access, 2017.

[10] Y. Tang, "Deep learning using linear support vector machines," arXiv preprint arXiv:1306.0239, 2013.

[11] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," Proceedings of the IEEE, vol. 86, pp. 2278-2324, 1998.

[12] 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.

[13] Z. Cao, J. C. Principe, B. Ouyang, F. Dalgleish, and A. Vuorenkoski, "Marine animal classification using combined CNN and hand-designed image features," in OCEANS'15 MTS/IEEE Washington, 2015, pp. 1-6.

[14] C.-K. Huynh, T.-S. Le, and K. Hamamoto, "Convolutional neural network for motorbike detection in dense traffic," in Communications and Electronics (ICCE), 2016 IEEE Sixth International Conference on, 2016, pp. 369-374.

[15] R. Wu, B. Wang, W. Wang, and Y. Yu, "Harvesting discriminative meta objects with deep CNN features for scene classification," in Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1287-1295.

[16] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in Advances in neural information processing systems, 2014, pp. 568-576.

[17] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," IEEE transactions on pattern analysis and machine intelligence, vol. 35, pp. 1798-1828, 2013.

[18] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 580-587.

[19] Y. Xu, T. Mo, Q. Feng, P. Zhong, M. Lai, I. Eric, et al., "Deep learning of feature representation with multiple instance learning for medical image analysis," in Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on, 2014, pp. 1626-1630.

[20] Z. Xu, Y. Yang, and A. G. Hauptmann, "A discriminative CNN video representation for event detection," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1798-1807.

[21] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.

[22] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in European conference on computer vision, 2014, pp. 818-833.

[23] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.

[24] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, et al., "Generative adversarial nets," in Advances in neural information processing systems, 2014, pp. 2672-2680.

[25] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," arXiv preprint arXiv:1409.7495, 2014.