Decoupled Deep Neural Network for Smoke Detection

Sheng Luo, Xiaoqin Zhang, Ling Zhen and Muchou Wang

January 30, 2021

Decoupled Deep Neural Network for Smoke Detection

Sheng Luo, Xiaoqin Zhang, Ling Zhen, Muchou Wang

January 2021

Abstract

Smoke detection is a practical technology to protect people's lives and property. Traditional methods to detect smoke are usually based on human-crafted features, such as color, texture and shape. Although these methods do work in some cases, they are not always effective because color, texture and shape of smoke are diverse. In recent years, deep learning have achieved the most accuracy with more complex structures and more numerous parameters. However, the existing methods based on human-craft features are not accurate enough, and the ones based on deep learning often take too much computing resources. To improve the detection accuracy and reduce the computational cost, inspired by the aforementioned works, we propose a decoupled sub-network to extract color and texture separately just following the procedure of the traditional human-crafted methods. The color sub-network, consisted of several 1×1 convolution layers, tries to find the most suitable color model by nonlinear functions. The next sub-network, based on a series of depth-wise separable convolution layers, extracts texture features and assembles them into shape features. After integrating these features, the proposed network can comprehensively determine whether there is smoke or fire. Experimental results demonstrate that our network is compact, efficient and effective, and the decoupling trick offers a critical capability needed to catalyze widespread implementation.

1 Introduction

Fire is a common and frequent natural disaster in human society. It makes human life and property suffer big losses. Compared with flame, smoke occurs much earlier, spreads faster, and always rises to sky [1–4]. More importantly, the volume of smoke is much larger than the one of fire. So, there is an urgent need for a efficient visual way to detect smoke. On the other hand, surveillance cameras provide a cost-effective manner to detect these visible accident. Therefore, the smoke detection algorithms should be lightweight enough to be deployed with high accuracy and low cost [5–9].

In the past decades, two categories algorithms have been proposed in the field of smoke detection: traditional methods and deep learning based methods [10–15]. The

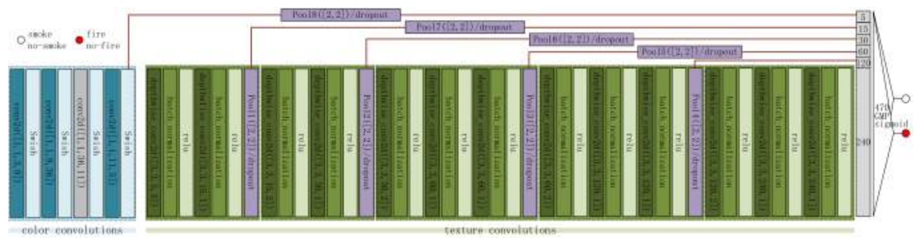

Figure 1: The architecture of our proposed decoupled neural network for smoke detection.

traditional methods are usually based on such human-crafted features, as color, texture, shape. Color features are usually extracted from different color spaces [16–23], such as RGB, HSV, YCbCr. Texture features are extracted by Fractal analysis [24–29], Wavelet decomposition [30–34], Gabor transform [35–39], and local gradient orientation histogram [40–43] methods. Smoke has very rich texture features, and most human-crafted methods only use a few of them. In fact, texture features are much more difficult to extract than color features. On the other hand, these human-crafted features vary greatly from different scenarios, which leads to unsatisfactory detection accuracy. To improve the accuracy, multiple human-crafted features are integrated. Even so, there is a bottleneck. In the past decade, deep learning methods have demonstrated the potential to achieve better accuracy with self-learning features. For example, Frizi *et al.* [44], Yin *et al.* [45], and Yin *et al.* [46] used multilayer convolutional neural networks (CNNs) to detect smoke and fire. However, those deep learning algorithms still have some limitations. First, classic CNNs trained on opaque samples can not be directly applied to smoke, because smoke has no definite shape. Second, deep neural networks have too many parameters, which leads to high computational cost and deploying difficulty. Last but not the least, smoke samples are insufficient to train these heavy networks, even complemented with such technology as data enhancement and transfer learning [47–49].

In fact, smoke does have distinctive colors, textures and shapes, which is the reason the traditional methods work. Since color, texture and shape are effective to distinguish smoke from background, there is no suitable human-crafted feature to cover all scenarios and convolutional networks can automatically extract features from a large number of samples, researchers try to take the both advantages of these methods. Wang *et al.* [50] input RGB and HSI images into 2 residual networks and made judgments by the integrated outputs. Maksymiv *et al.* [51] assumed that smoke textures are unique, so they located the candidate areas by AdaBoost and LBP and determined whether there is smoke by a classic convolutional network. Chen *et al.* [52] extracted static textures using a convolutional network and integrated them with dynamic textures to reduce false detection. Zhao *et al.* [53] found candidate areas by saliency technology and determine whether the saliency regions have smoke by AlexNet [54–56].

Inspired by these works, we propose a smoke recognition neural network which try to learn separate features by decoupled sub-networks and make judgements by comprehensive features [57–61]. Two separate features, color and texture, are extracted by 2 sub-networks. The color sub-network automatically extracts color features with maximum inter-class differences and minimum intra-class differences based on multi-layer 1×1 convolution. The texture sub-network automatically extracts spatial features such as textures and shapes on each color channel. The network is trained alternately by strong supervision with pixel-level labels and weak supervision with image-level labels [62–68]. Validated by experiments, the proposed network achieves a good performance than the previously mentioned smoke detection methods.

To summarize, we make the following contributions to the field:

- We achieve an outstanding accuracy for smoke detection by decoupling a end-to-end neural network, a black block, into several functional sub-networks, a serial of gray blocks. That means that the composite features are decoupled into almost independent ones, which leads to a more lightweight and accurate network. These features are still a bit more than the human-crafted features, but cover almost all cases and improve the detecting accuracy remarkably. On the other hand, though the detection accuracy of the proposed method is just a bit higher than the classic neural networks, the weight is much lighter than the one of the latter. Slight weight means less data dependency, less computation and easy deploying.

- The color sub-network generates a color mode with maximum interclass differences and minimum intraclass differences by a neural network, which cover wider range than such traditional color models as RGB, YCbCr, CIE Lab, HSI, YUV and dark channel. After all the neural network with nonlinear function could learn to map a complex space into a simple one.

- The texture sub-network disassembles smoke textures into multiply channel without mixture and finally assembles these texture features into shape features, which is a impossible task for human-crafted methods and difficult to interpret but good at discriminating smoke from background.

This paper is organized as follows. Section 2 presents related work. Our proposed approach for smoke detection is illustrated in Section 3 and Section 4. Section 5 introduces the details of the training process, and the the experiments of the proposed network and other compared methods are described in Section 6. Finally, we conclude the paper in Section 7.

2 Related Work

2.1 Traditional Human-Crafted Features

Color, texture, and shape are typical human-cratfed features used to detect smoke. In pioneering work [16–19, 69–72], authors built smoke recoging model with multiple color information, including RGB, YCbCr, CIE Lab, HSV, HSI, YUV, and dark channels. Chen *et al.* [73–76] built smoke recognition model on every channel in RGB color space. Appana *et al.* [35]. thought that HSV is more suitable than RGB for smoke detection and designed a color model in this color space. Zhao *et al.* [77] attempted to integrate the both advantages of RGB and HSV color spaces and made up a smoke color model in these two color spaces. Deng *et al.* [78] interpreted the empirical thresholds in a new model using k-means algorithm, and the proposed method clustered smoke pixels in an experimental color space. Zhang *et al.* [79] used gray and red values to find fire and smoke, and segmented the interesting regions using the Otsu method. Their experiments showed that color feature was too sensitive to thresholds. If the similarity among adjacent pixels is taken into consideration, the detection may be more robust. In fact, texture is such a feature dependent on adjacent context information. Fujiwara *et al.* [24] thought smoke was self-similar and discovered the distinguishing features by the fractal encoding method. Maruta *et al.* [30] thought that smoke was a self-affine fractal and differed smoke texture from non-smoke texture through the wavelet transform and the Hurst exponent. Toreyin *et al.* [80] thought that smoke regions were convex and their edges produced local extrema in the wavelet domain. Appana *et al.* [35] modeled smoke texture using the coefficients of Gabor. Yuan *et al.* [40] quantized the directional derivatives into ternary values to generate local ternary patterns (LTP), concatenated all joint histograms from different orders to propose high-order local ternary patterns (HLTP) and proposed HLTP based on magnitudes of noise-removed derivatives and values of center pixels (HLTPMC). Alamgir *et al.* [81] proposed a method that combined local binary patterns with the cooccurrence of texture features in the RGB color space to characterize the diverse manifestations of smoke. Piccinini *et al.* [82] proposed that the decreases of energy ratios in the wavelet domain between the background and current images represented the variations in the texture level and provided a clue for detecting smoke; they modeled this texture ratio for temporal evolution using a mixture of Gaussians. Tian *et al.* [83] separated a frame into quasi-smoke and quasi-background components, represented these components by dual dictionaries, and solved the detection as a convex optimization. In addition, they constructed texture features as a concatenation of the respective sparse coefficients. Wu *et al.* [84] represented smoke components in sparse coefficients on a learned smoke dictionary for block candidates, and selected the discriminative feature with respect to the sparse coefficients [85–89].

2.2 Convolutional Neural networks

Some researchers have adopted multiple convolutional neural networks (CNNs) to smoke detection in recent years [90–94]. In 2015, Hohberg [95] used LeNet, CaffeNet, and GoogleNet with diverse inception modules to detect smoke. In 2016, Frizzi *et al.* [44] built a network that was very similar to the well-known LeNet-5 with increased feature maps in the convolution layers. In the same year, Tao *et al.* proposed a network consisted of 5 convolutional layers and 3 fully connected layers, which in fact was the transferred AlexNet for binary classification task. [96]. In 2017, Filonenko *et al.* [97] evaluated such CNNs as AlexNet, Inception-V3, Inception-V4, ResNet, VGG, Xception, in a diverse range of possible scenarios, which was similar to the work of Hohberg [95]. The experments showed that inception-based networks achieved the highest performance. Yin *et al.* [45] proposed a 14 layers' normalized convolutional neural network (D-NCNN), which improved the convolutional layer in the traditional CNNs into a batch-normalized convolutional layer and alleviated the data imbalance problem with data augmentation skills. Based on saliency technology, Zhao *et al.* [53] first found candidate regions in the saliency image, then made the decision about whether there existed smoke in the regions using a CNN modified from AlexNet. Similarly, Maksymiv *et al.* [51] located candidate regions using traditional methods, such as AdaBoost and LBP, and determined whether there were smoke by a convolutional network. In 2018, Dung *et al.* [98] first located moving areas, then determined whether these moving areas were similar to smoke using a series of cascaded classifiers to integrate features such as color, region-growing, area size, and edge energy, and finally made decisions using a CNN [99–108]. In 2019, Yuan *et al.* [109] cascaded 11 basic blocks followed by a global average pooling and a 2D fully connected layer to detect smoke. The basic block was consisted of several parallel convolutional layers with the same number of filters but different kernel sizes for handling scale invariance. Then they added all normalized outputs from multiscale parallel layers and activated the result as the final output of the block. Gu *et al.* [110] established a CNN with two paths, one path was used to extract texture and the other one was used to extract contours. Then, the output of the two sub-networks were integrated to detect smoke. Wang *et al.* [50] thought that color information was important for the task of smoke detection, thus they built a parallel deep residual network based on the R, G and B components of RGB image and the H, S and I components of HSI transform image to adaptively extract color features. Based on this strategy, the discriminative ability for distiguishing smoke-like objects and background was enhanced. Ba *et al.* [**?**] noticed that different color channels had different discriminative abilities and they applied an attention model to these color channels. Besides, to the best of our knowledge, no poineering work has adopted CNNs to extract texture features separately. Usually, texture features are extracted simultaneously with color features using deep networks in an end-to-end fashion.

2.3 Networks with Decoupled Convolutions

When CNNs extract features, multiple kernels on multi-channel 3D feature maps always are applied to color and spatial simultaneously [**?**, 111–116]. In recent years, diverse convolution operations are proposed to extract useful features in a more flexible manner. For example, Chollet *et al.* [**?**] decomposed the standard convolution so that the feature extraction operations can be separately and independently performed on single channel feature maps and different channels. More specifically, multiple 1×1 convolution kernels were applied on feature maps, then the outputs were organized into 3 or 4 independent spaces to reduce the number of feature maps. Subsequently, these individual feature maps were handled with standard 3×3 or 5×5 convolution kernels. Besides, Howard *et al.* [**?**] also proved that depth-wise separable convolution was superior to decoupling the channel and spatial dimension.

Inspired by the aforementioned methods, We decouple the composite features into color features and texture features through two individual steps [117–121]. In order to find the best color mode, the single layer 1×1 convolution is cascaded into a multilayer 1×1 convolution, and a nonlinear operation is added after every 1×1 convolution. Therefore, this step does not reduce the number of channels but finds significant color channels. Then, the texture features are extracted on the channels with significant interclass difference. Finally, whether smoke or fire occurs in the scenario is determined according to the synthetic features [122–127].

3 Methods

The space of human-crafted features is too simple to cover all smoke samples, on the other hand the classic CNNs have too many parameters. To find suitable features to represent smoke, the proposed algorithm based on color features and shape features can be described as follows:

$$\begin{cases} J_s(x_s) \begin{cases} >= 0 & smoke \\ < 0 & non-smoke \end{cases} \\ x_s = x_c \cup x_a \end{cases} \tag{1}$$

where J_s is the classification function in our decoupled network, x_c and x_a are the color features and the texture features respectively, and x_s is the composite features. We define human-crafted features x_h as the union of the human-crafted color features x_{hc}, human-crafted shape features x_{ha} and other features x_{hx} as follows:

$$x_h = x_{hc} \cup x_{ha} \ldots \cup x_{hx} \tag{2}$$

Generally, the human-crafted features x_h is a small subset of the proposed features x_s, which is a small subset of the compound and huge features x_n. x_n are usually extracted by aforementioned classic deep neural networks. So,

$$\begin{cases} x_{hc} \subset x_c \\ x_{ha} \subset x_a \\ x_h \subset x_s \subset x_n \end{cases} \tag{3}$$

The features x_s are more accurate to describe smoke than the human-crafted features x_h, and much more compact than the network features x_n. So the proposed network try to achieve a better trade off between efficiency and effectiveness [128–133].

4 Decoupled Neural Network for Smoke Detection

The proposed network is composed of a color sub-network and a texture sub-network. The color sub-network is trained to select the color features x_c with maximum inter-class difference and minimum intra-class difference; the texture sub-network is used to extract the texture features x_a on every color channel. A concatenation layer is used to integrate color features from the color sub-network and texture features from the texture sub-network. Then, estimate every pixel of this composite feature map being smoke or not. Later these estimation is globally and maximally pooled into a judgement.

4.1 Color Sub-network

The color channels are transformed with 1×1 convolutions and then nonlinearly activated:

$$f_k^i = g^{i-1}(\sum_m w_{km}^{i-1} f_m^{i-1} + b_k^{i-1}) \tag{4}$$

where f_k^i is the k^{th} color channel of the i^{th} layer, f_m^{i-1} is the m^{th} color channel of the $(i-1)^{th}$ layer, w_{km}^{i-1} is the k^{th} convolution kernel from the $(i-1)^{th}$ layer to the i^{th} layer, w_{km}^{i-1} is the weight that operates on the m^{th} color channel and output the k^{th} layer, b_k^{i-1} is the k^{th} bias in the $(i-1)^{th}$ layer, and g^{i-1} is the nonlinear activation function of the $(i-1)^{th}$ layer.

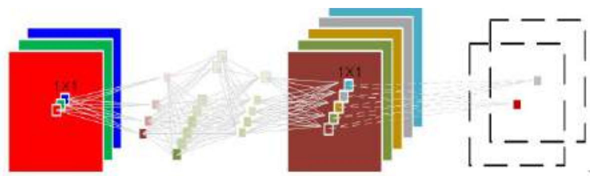

Figure 2: Architecture of the proposed color sub-network.

We adopt a neural network with 4 convolution layers to conduct color-conversion, and the details are shown in Figure (2). There are 2 special points: (1) No pooling operation is adopted in this sub-network and (2) Swish is adopted as activating operation to explore the complex color space. So the pixels between the input layer and the output layer of the color sub-network is one-to-one corresponding. Two dashed boxes are added at the end of Figure (2) to indicate the probabilities of smoke and fire at every pixel only based on the color features f^4, which are exploited in the pixel-level training process but not included in the image-level training process. Figure (3) shows the activation of this sub-network when a typical sample is fed into this network. It can be seen that each channel focus on different color information, and the 2^{nd} channel has no any response with all black.

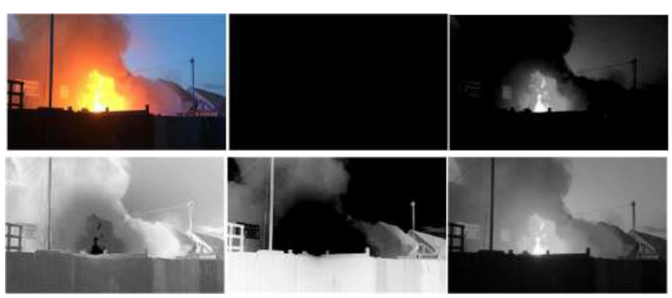

Figure 3: Five color channels generated by the color sub-network. The top left picture is the input image, and the others are the five color channels images of the output layer.

4.2 Texture Sub-network

The feature flow of the texture sub-network is shown in Figure (4), which extract texture features on spatial space without inter-channel mixing.

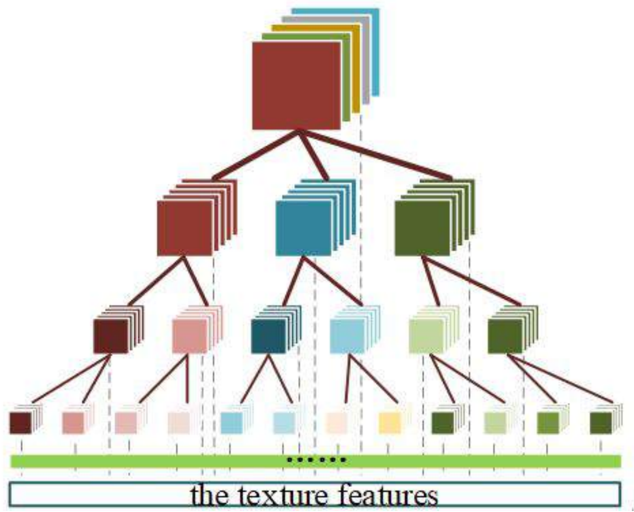

Figure 4: Architecture of the proposed texture sub-network.

In order to enrich the texture features, the texture sub-network expands each channel to 2 or 3 channels in each layer and the max pooling is performed to enlarge the receptive field. Regularization is conducted after each convolution operation, and the activating function Relu is adopted after the regularization operation. To avoid overfitting, some neural units are dropout after activation function. These operations make up a convolution block.

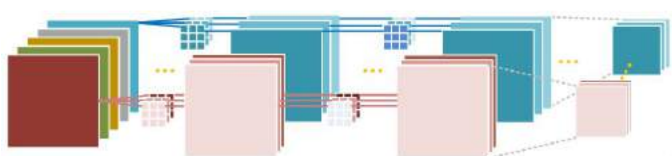

Figure 5: A typical layer of the texture sub-network which is denoted as 3_2. The texture features are extracted in each channel without any combination among channels.

The texture sub-network is stacked by 5 convolution blocks, and each block has two or three convolution kernels and a max pooling operation. The size of all convolution kernels is 3×3. And a typical layer is shown in Figure (5).

The convolution operation adopted in the i^{th} layer is formulated as follows:

$$f_{km}^{i} = g^{i-1}(w_{km}^{i-1} f_{m}^{i-1} + b_{km}^{i-1}) \tag{5}$$

where f_{km}^{i} is the $(k*m)^{th}$ texture channel in the i^{th} layer, f_m^{i-1} is the m^{th} texture channel in the $(i-1)^{th}$ layer, w_{km}^{i-1} is the k^{th} convolution kernel from the $(i-1)^{th}$ layer to the m^{th} texture channel of the i^{th} layer, b_{km}^{i-1} is the $(k*m)^{th}$ bias term in the $(i-1)^{th}$ layer, and g^{i-1} is the nonlinear activation function in the $(i-1)^{th}$ layer.

4.3 Batch Regularization

Currently, researchers generally utilize mini-batch stochastic gradient descent algorithms to learn network parameter when training deep neural networks. The training effect is correlated with the covariance in batches. So scaling and migration are adopted to reduces the variation in intra-covariance in batches before the nonlinear activation function. The operation is called batch regularization. The mean $\bar{f}_m^i$ and the variance $(\delta^2)^i$ in batches are computed as follows:

$$\begin{cases} \bar{f}_m^i = \frac{1}{N_b}\sum_{j=1}^{N_b} f_{jm}^i \\ (\delta^2)^i = \frac{1}{N_b}\sum_{j=1}^{N_b}(f_{jm}^i - \bar{f}_m^i) \end{cases} \tag{6}$$

where $\bar{f}_m^i$ is the mean of the m^{th} feature of the i^{th} layer, N_b is the number of samples in a batch , and $\bar{f}_m^i$ is the m^{th} feature of the j^{th} sample in the i^{th} layer. Thus, each feature is regularized as follows:

$$\hat{f}_m^i = \frac{(f_m^i - \bar{f}_m^i)}{\sqrt{(\delta^2)^i + \zeta}} \tag{7}$$

where ε is a small positive constant to improve stability.

Regularization reduces the difference between samples. Therefore, r_m and β_m are introduced to recover the original samples for each feature. Batch regularization consists of 2 steps, scaling and migrating:

$$BN(f_m^i) = \gamma_m \hat{f}_m^i + \beta_m \tag{8}$$

where $BN(f_m^i)$ is the output of batch regularization.

4.4 The Concat Layer

The feature maps of the color sub-network and the outputs of every pooling operation in the texture sub-network are pooled into same size and then integrated into 470 feature maps. That is,

$$f^{19} = [MP_{16}(f^4),\ MP_8(f^7),\ MP_4(f^{10}),\ MP_2(f^{14}), f^{18},\ f^{21}] \tag{9}$$

where f^4 is the output of the color sub-network, f^7, f^{10}, f^{14}, f^{18} and f^{21} are the features after the 1^{st}, 2^{nd}, 3^{rd} and 4^{th} pooling layer. In the network, f^{22} is the concat feature, and MP_16, MP_8, MP_4 and MP_2 are the max pooling with striding sizes of 16×16, 8×8, 4×4 and 2×2, respectively.

4.5 Global Max Pooling

The concatenated feature maps, f^{22}, are assembled into 4 feature maps, f^{23}, with a Depthwise Conv2d operation. In fact, this layer integrates the 470 features into a combo feature. As before, a dropout operation is adopted to reduce overfitting and improve the generalization ability. Inspired by [134], global max pooling is adopted to select the largest value from the 2D feature map as the output, and the 2D feature map is transformed into a one-dimensional vector. Thus, this process is formulated as:

$$f_m^{24} = global\_max\_pooling(f_{j,m}^{23}) \tag{10}$$

where f_m^{24} is a vector with 4 scalar, $f_{j,m}^{23}$ is the value of the j^{th} pixel in the m^{th} feature of the 23^{th} layer, whose size is 16×16, and m is the serial number of the channels.

4.6 Output Layer

A softmax function is used to determine whether smoke or fire happens in the input image. The probability of smoke p_s is

$$p_s = \frac{1}{f_{j,m}^{24}} \sum [e^{f_{j,m}^{24}}] \tag{11}$$

where $f_{j,m}^{24}$, a scalar, is the output of the 24^{st} layer.

4.7 Loss Function

The loss of the network $J_A(W)$ is the sum of the smoke loss $J_s(W_s)$, the fire loss $J_f(W_f)$, and the L_2 norm of all trainable parameters in the network

$$\begin{cases} J_s(W_s) = -\frac{1}{N}\sum[L_s \log(p_s) + (1-L_s)\log(1-p_s)] \\ J_f(W_f) = -\frac{1}{N}\sum[L_s \log(p_f) + (1-L_f)\log(1-p_f)] \\ J(W_A) = \lambda_s J_s(W_s) + \lambda_f J_f(W_f) + \lambda_w ||W_A||_2 \end{cases} \tag{12}$$

where p_s and p_f are the smoke probability and the fire probability detected by the network; L_s and L_f are the image labels of smoke and fire; W_s, W_f and W_A are the trainable parameters of the smoke path, fire path and the union, respectively, and λ_s, λ_f and λ_A are the weight coefficients. The L_2 norm attempts to smooth the training process and constrain the parameter space.

$$\begin{cases} W_{t+1} = W_t + V_t \\ V_t = M_u W_t - L_r \nabla d(W_t) \end{cases} \tag{13}$$

where t is the number of iterations, W_{t+1} and W_t are the network parameters at the $t+1$ and t turns, respectively, V_t is the parameter adjustment, M_u is the momentum, which is generally 0.9, L_r is the learning rate, and $\nabla d(W_t)$ is the parameter gradient.

4.8 Validation accuracy

The validation accuracy A_s and A_f are defined as:

$$\begin{cases} A_s = \frac{1}{N_b}\sum_{q=1}^{N_b}|L_q^s - P_q^s| \\ A_f = \frac{1}{N_b}\sum_{q=1}^{N_b}|L_q^f - P_q^f| \end{cases} \tag{14}$$

where N_b is the number of samples in a batch, L_q^s and L_q^f are the labels of smoke and fire, and P_q^s and P_q^f are the predicted probabilities of smoke and fire.

5 Network Training

This network, especially the color sub-network, should be sufficiently trained by abundant pixel-level samples. Labelling pixel-level samples is time consuming, difficult and expensive. Thus, image-level annotated samples, which are cheap and plentiful, are adopted to make up the pixel-level samples. So, a complex training process is adopted to train the color and texture sub-networks with both pixel-level annotated images and image-level annotated samples.

5.1 Training Dataset

Our training dataset contains two subsets. The pixel-level labeled subset includes 241 images containing smoke and fire, 1,283 images with smoke but no fire, 359 images with fire only, and 1,042 images without smoke and fire, totally 2,925 samples. The image-level annotated subset contains 1,085 images of ”smoke-fire”, 29,476 images of ”smoke-no-fire”, 653 images of ”no-smoke-fire”, and 41,537 images of ”no-smoke-no-fire”, which are selected from ImageNet2012, COCO2014 and ILSVRC2012, totally 83,751 samples.

To reduce the influence of illumination, a min-max regularization is adopted as following:

$$f_i^1 = (f_i^0 - \min(f^0))/\max((f^0) - \min(f^0)) \tag{15}$$

where f_i^0 is the value of the p^{th} pixel of layer 0, f_i^1 is the normalized value, and $\max(f^0)$ and $\min(f^0)$ are the maximum and the minimum pixel values in each sample.

To increase the number of categories with less samples, 210-degree random rotations and random brightness shifts within the range of +10 and -10 are performed. So every categories have similar image samples.

5.2 Training Process

There are 4 steps in the network training process. At the 1^{st} step, the color sub-network is trained N_1 times with the pixel-level labeled subset; at the 2^{nd} step, the color sub-network and texture sub-network are trained N_2 times with the pixel-level labeled subset; at the 3^{rd} step, the whole network is trained N_3 times with the image-level labeled subset; Finally, the entire training is conducted N_4 times.

The pixel-level labeled subset is used at the 1^{st} and 2^{nd} steps to classify every pixel according to the pixel features. Therefore, 2 classification layers are inserted after the color sub-network and the texture sub-network. The output of the color sub-network is convoluted into a 2-channel feature map with the same size by $[1,1,5,2]$ convolutional kernels. One channel is the smoke probability of every pixel, and the other channel is the fire probability. The loss functions J_1^s of smoke and J_1^f of fire are

$$\begin{aligned} J_1^{s/f}(W^1) = \sum_{s/f}\sum_{ij}[L_{ij}^{s/f}\log(p_{ij}^{s/f})\ + \\ (1-L_{ij}^{s/f})\log(1-p_{ij}^{s/f})] + \lambda_1||W^1||_2 \end{aligned} \tag{16}$$

where s/f denotes smoke or fire in the 2 output channels and in the labels, (i,j) is the pixel coordinate, $p_{ij}^{s/f}$ is the probability of pixel p_{ij} in the s or f channel, $L_{ij}^{s/f}$ is the label of p_{ij} for smoke or fire, λ_1 is a weight factor, and $||W^1||_2$ is the L_2 norm of the trainable parameters of the color sub-network.

The output of the texture sub-network, f^{18}, is also convoluted into a 2-channel map by $[1,1,120,2]$ convolutional kernels. The classification layers similar to the ones of color sub-network are added after the texture sub-network. However, the sizes of the feature maps are different. The loss functions J_2^s and J_2^f of smoke and fire are

$$\begin{aligned} J_2^{s/f}(W^2) = \sum_{(s/f)_2}\sum_{ij}[L_{ij}^{(s/f)_2}\log(p_{ij}^{(s/f)_2})\ + \\ (1-L_{ij}^{(s/f)_2})\log(1-p_{ij}^{(s/f)_2})] + \lambda_2||W^2||_2 \end{aligned} \tag{17}$$

whose parameters are similar to the ones of Equation (16).

The validation accuracy of smoke (A_1^s) and fire (A_1^f) are

$$\begin{cases} R_{ij}^{s/f} = \begin{cases} 0, & p_{ij}^{s/f} < 0.5 \\ 1, & p_{ij}^{s/f} > 0.5 \end{cases} \\ A_{1/2}^{s/f}(W^1) = (2\times\frac{1}{N}\sum_{ij}L\times R+\delta) \\ \qquad /(\frac{1}{N}\sum_{ij}L\times L+\frac{1}{N}\sum_{ij}R\times R+\delta) \end{cases} \tag{18}$$

where N is the number of pixels in the image sample, and δ is a very small constant to avoid dividing by zero.

The image-level labeled subset is used at the 3^{rd} stage to train the whole network through weak supervision.

The detail training procedures is shown in Figure (6).

5.3 Hyperparameters

The input image size is [256,256,3], the batch size is 32, and the learning rate decreases exponentially with an initial learning rate of 0.01 and a decaying coefficient of 0.9. The dropout rate is set as 0.6 in the training stage and 1 in the predicting stage. The Adam optimizer is adopted to optimize the network parameters.

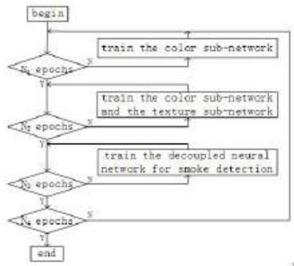

Figure 6: Training procedures of the proposed network.

5.4 Learning Curve

Taking N_1, N_2, N_3 and N_4 as 1, 1, 10 and 16,000, respectively, the training and validating accuracy during the training process is shown in Figure (7). Figure (7a) illustrates the training process of smoke with 16,000 iterations, and the highest validating accuracy of smoke is 0.96; Figure (7b) shows the same training process of fire, and the highest validating accuracy is 0.99.

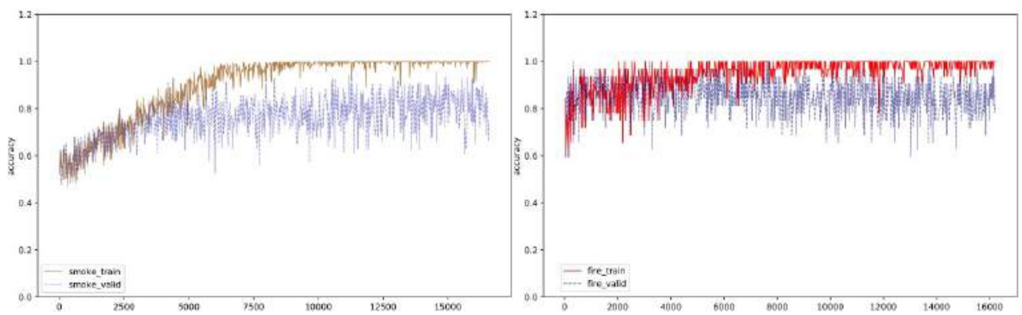

Figure 7: Learning process of the decoupled neural network with the color sub-network $3 \times 9 \times 36 \times 11 \times 5$. (a) The highest validating accuracy of smoke is 0.96. (b) The highest validating accuracy of fire is 0.99.

6 Experiment

To test the proposed network, which is abbreviated as DNNSD, three experiments are conducted: the 1^{st} compares the detecting accuracy between DNNSD and some classic traditional algorithms; the 2^{nd} compares the detecting accuracy between DNNSD and some classic neural networks; and the 3^{rd} analyzes the impact of the network architecture on detecting accuracy. Before these comparison, the experimental settings are introduced. At the end of this section, what the network focuses on are displayed. Last several successful and failure cases of DNNSD are both demonstrated.

6.1 Experimental settings

6.1.1 Validating Dataset

The validating dataset also consists of images with four different classes: “smoke-with-fire”, “smoke-without-fire”, “non-smoke-with-fire”, and “non-smoke-without-fire”. The images with smoke or fire are randomly selected from the publicly available smoke database [135–138]. The non-smoke-without-fire images are randomly selected from ImageNet2012, COCO2014 and ILSVRC2012. The dataset comprises of 7231 images with 1917 smoke-with-fire images, 1949 smoke-without-fire images, 1541 non-smoke-with-fire images and 1824 non-smoke-without-fire images. These samples are not exerted any data augmentation.

6.1.2 Evaluation Metrics

To evaluate the proposed method, three metrics including hit rate HR, false-alarm rate FAR and detection accuracy DR are defined as follows:

$$\begin{cases} HR = \frac{P_p}{Q_p} * 100\% \\ FAR = \frac{N_p}{P_p} * 100\% \\ DR = \frac{P_p + N_n}{Q_n + Q_p} * 100\% \end{cases} \tag{19}$$

where Q_p and Q_n are the number of positive and negative samples; P_p is the number of correctly detected positive samples, N_p is the number of negative samples classified as positive samples, and N_n is the number of correctly classified negative samples. The higher the hit accuracy HR and detection rate DR are, or the lower the false-alarm rate FAR is, the better detection results the network can achieve.

6.1.3 Computing Platform

All training and validating phases are performed on a computer with an Intel(R) Core i7-6700 CPU at 3.40 GHz, an NVIDIA GeForce GTX 1080Ti, the operation system of ubuntu 16.04 and the framework pytorch 1.4.0.

6.2 Comparison with Classic Human-Crafted Algorithms

In order to evaluate DNNSD, some traditional methods are adopted to train and validate on the same datasets. Inspired by [51] and [84], the HSV color model is utilized to locate the candidate regions, and LBP and AdaBoost are adopted to describe the smoke texture. HSV is transformed from the RGB color space, and the pixels whose saturation component is between 0 and 0.28 and whose value component is between 0.38 and 0.985 are thought to be smoke. This experience is referenced from [35]. Inspired by [139] and [140], MSER and SLIC are utilized to locate the candidate regions, and the texture features are extracted using wavelet transform. Later, SVM is conducted on the histograms of all texture features in the candidate regions to determine whether there is smoke. So, there are 4 methods, color+LBP+SVM, color+AdaBoost+SVM, MSER+Wavelet+SVM and SLIC+Wavelet+SVM, are conducted for the comparison.

Methods	HR(%)	DR(%)	FAR(%)
Color+LBP+SVM	79.50	77.69	9.21
Color+AdaBoost+SVM	74.31	75.76	8.38
MSER+Wavelet+SVM	81.06	80.18	6.13
SLIC+Wavelet+SVM	83.85	82.21	7.92
DSNN(Ours)	**98.13**	**97.47**	**1.76**

Table 1: Detection accuracy of the decoupled neural network and the 4 traditional methods

Deep CNNs	ResNet50 [141]	Xception [142]	InceptionV3 [143]	MobileNet [144]	**DNNSD**
Weight	99M	88M	92M	17M	**0.56M**
Parameters	25,636,712	22,910,480	24,851,784	4,253,864	**136,873**
A_s	0.84	0.86	0.90	0.80	**0.96**
A_f	0.87	0.88	0.89	0.78	**0.99**

Table 2: Performance of the decoupled neural network and classic deep neural networks

The results are shown in Table 1. The *AR*, *HR* and *FAR* of Color+LBP+SVM are 79.50%, 77.69% and 9.21%, respectively; the corresponding *AR*, *HR* and *FAR* of Color+AdaBoost+SVM are 74.31%, 75.76% and 8.38%; the *AR*, *HR* and *FAR* of MSER+Wavelet+SVM are 81.06%, 80.18% and 6.13%; and the *AR*, *HR* and *FAR* of SLIC+Wavelet+SVM are 83.85%, 82.21%, and 9.92%. The best result is achieved by DNNSD, and the corresponding measures, *AR*, *HR* and *FAR*, are 98.13%, 97.47% and 1.76%. It can be seen that the proposed method is much more accurate than these human-crafted methods.

6.3 Comparison with Classic Deep Neural Networks

To evaluate DNNSD, we fine-tune the pretrained classic deep neural networks, such as ResNet50 [141], Xception [142], Inception V3 [143], and MobileNet [144], with 2 outputs. One output is the smoke probability, and the other is the fire probability. Just like DNNSD, one fully connected layer is connect on the bottleneck features of each model. Each model is trained for 50 epochs and 1,000 batches in a epoch. The results are shown in Figure (8) and Table 2. The accuracy of the all these fine-tuned models, ResNet50, Inception V3, Xception, and MobileNet, are inferior to the one of DNNSD. DNNSD achieves the highest accuracy, $A_s = 0.96/A_f = 0.99$. In these fine-tuned models, Inception V3 achieves the highest accuracy, $A_s = 0.90/A_f = 0.89$ and MobileNet achieves the lowest A_s, 0.80, and the lowest A_f, 0.79. These results are different from those reported in [97], in which Xception achieved the highest accuracy of 1.0 and Inception V3 achieved the lowest accuracy of 0.76. Another very important benefit of DNNSD is that its weight is only 0.56M, but the corresponding one of the second best model, Inception V3, is 92M. It can be seen that the proposed network is much more compact with higher accuracy.

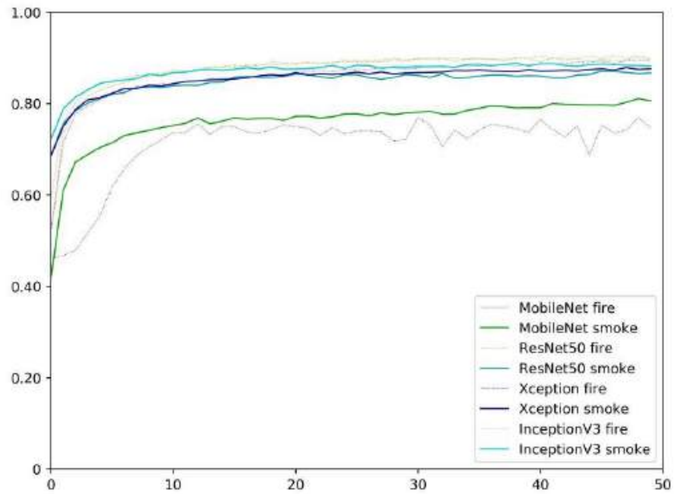

Figure 8: Learning curves of the fine-tuned ResNet50, Xception, Inception V3, and MobileNet. The datasets for training and validating are same as the one of the decoupled neural network. Among these models, Inception V3 achieves the highest accuracy, $A_s = 0.90/A_f = 0.89$, MobileNet achieves the lowest A_s, 0.80 and the lowest A_f 0.78.

6.4 Impact of Network Architecture on Detection Accuracy

To validate the decoupled thought, different network architectures of the color sub-network and the texture sub-network are compared. When the color sub-network is varied, the texture sub-network is fixed with the structure $3_2 \times 2_2 \times 2_3 \times 2_3$. When the texture sub-network is varied, the color sub-network is fixed with the structure $3 \times 9 \times 36 \times 11 \times 5$.

6.4.1 Different Architectures of Color Sub-network

	$3\times36\times5$	$3\times9\times36\times5$	$3\times9\times64\times5$	$3\times9\times36\times11\times5$	$3\times9\times36\times11\times3$	$3\times9\times53\times17\times5$	$3\times9\times64\times27\times13\times5$
A_s	0.75	0.82	0.89	0.96	0.94	0.96	0.96
A_f	0.92	0.95	0.99	0.99	0.96	0.99	0.99

Table 3: Performance of the decoupled neural network with different color sub-networks.

	$3_2\times2_2\times2_3$	$3_3\times2_3\times2_3$	$2_2\times2_2\times2_3\times2_3$	$3_2\times2_2\times2_3\times2_3$	$3_3\times2_3\times2_3\times2_3$	$3_2\times2_2\times2_3\times2_3\times2_3$
A_s	0.81	0.83	0.85	0.96	0.96	0.96
A_f	0.91	0.96	0.96	0.99	0.99	0.99

Table 4: Performance of the decoupled neural network with different texture sub-networks.

In order to find a suitable network structure, the architecture of the color sub-network is adjusted to change network width and depth, which mean the number of layers of the sub-network and the number of channels per layer. We use $\prod_{i=1}^{R} n_i$ to describe the structure of the sub-network, which has R layers and n_i channels at the i^{th} layer. For example, a $3\times9\times36\times11\times5$ sub-network has 5 layers, the number of channels of each layer are 3, 9, 36, 11 and 5, and the convolution kernels are $1\times1\times3\times9$, $1\times1\times9\times36$, $1\times1\times36\times11$ and $1\times1\times11\times5$, respectively. The experimental results are shown in Table 4. The detection accuracy gradually increases from the $3\times36\times5$ network with the accuracy $A_s = 0.75$ and $A_f = 0.92$; when the structure is $3\times9\times36\times11\times5$, the accuracy is stable with an approximate $A_s = 0.96$ and $A_f = 0.99$. If the sub-network is changed wider and deeper, the accuracy is almost unchanged. However, the fire accuracy A_f is always higher than the smoke accuracy A_s and never drop below 0.92. The smoke accuracy A_s is always lower than the one of fire, especially when the sub-network is shallow, which shows that the fire color space is relatively convergent and the one of smoke is more emanative. Therefore, shallow sub-networks, which mean simple transformations in color space, may be insufficient to reach the smoke space. This is also the reason why these traditional methods which make use of color models in RGB [16], YCbCr [17], CIE Lab [18], HSV [19], HSI [145], YUV [146], dark channels [69] could not achieved satisfied accuracy. When the sub-network is deep enough to contain the whole smoke color space, wider or deeper structure would have little effects on these 2 accuracy. The structure with sufficient width and depth is $3\times9\times36\times11\times5$. Two examples of the training process whose structures of the color sub-networks are $3\times9\times64\times5$ and $3\times9\times53\times17\times5$, as shown in Figure (9a) and (9b). The detection accuracy is $A_s = 0.87/A_f = 0.99$ and $A_s = 0.96/A_f = 0.99$, respectively. In the figure, the red lines indicate the learning curve of fire, the blue and green lines indicate the ones of smoke, the solid lines denote the training accuracy, and the dotted lines denote the validating accuracy.

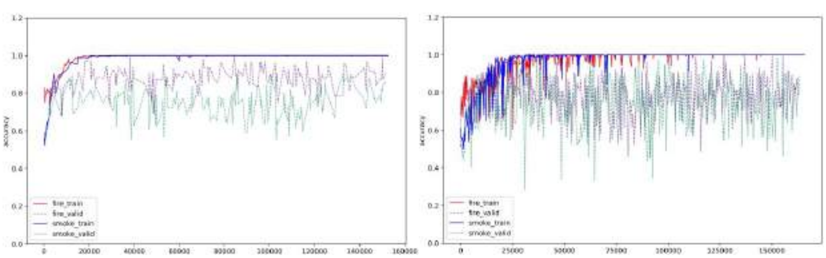

Figure 9: Learning curves of the decoupled neural network whose color sub-networks are $3\times9\times64\times5$ and $3\times9\times53\times17\times5$. (a) $3\times9\times64\times5$. (b) $3\times9\times64\times5$. The red lines indicate fire, the blue and green lines indicate smoke, the solid lines denote the training precision, and the dotted lines denote the validating precision.

6.4.2 Different Architecture of Texture Sub-network

All convolutions in the texture sub-network are depth-wise convolutions in which the convolution kernels operate on only one channel without mixing among other channels. The structure of the texture sub-network is recorded as $\prod_1^R C_K$, indicating that the sub-network has R layers, one feature map are transformed into C feature maps in a layer and convolutions are conducted K times successively in this layer. In each layer, the convolution expands the channel numbers by 2 or 3 times, and the convolution kernel is recorded as $3 \times 3 \times C_{i-1} \times C$, where C_{i-1} is the number of channels of the previous layer and C is the expanding time. The following $K-1$ convolutions do not expand the channels, and the convolution kernels are $3 \times 3 \times C_{i-1} \times 1$.

We vary the sub-network depth R, the channel number per layer C, and the convolution number per layer K to find the best structure of the texture sub-network, and 6 typical results are shown in Table **??**. If the network depth R is increased, the highest detection accuracy of this network improves. But after the sub-network depth is increased to 5 layers, the detection accuracy increases slowly. On the other hand, if the channel number per layer C increases, the accuracy also improves. Analogously, after the channel number increases to a certain level, the accuracy increases slowly. If the convolution number per layer increases, the accuracy improves slightly. Among these sub-networks, the $3_2 \times 2_2 \times 2_3 \times 2_3$ structure achieve the best accuracy and the lightest weight.

The training process of 2 examples whose texture sub-networks are $3_2 \times 2_2 \times 2_3$ and $2_2 \times 2_2 \times 2_3 \times 2_3$, just as shown in Figure (10), and the detection accuracy is $A_s = 0.81/A_f = 0.91$ and $A_s = 0.85/A_f = 0.96$, respectively. In this figure, the colors and line types have the same meaning as in Figure (9).

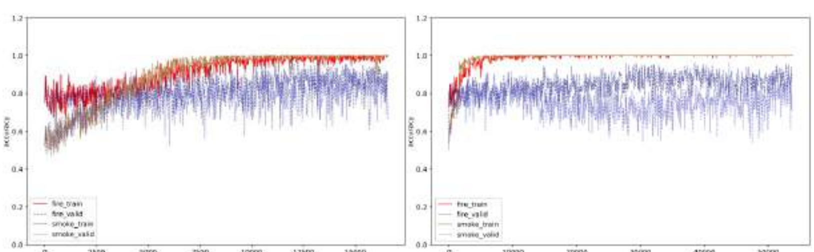

Figure 10: Learning curves of the decoupled neural network whose texture sub-networks are $3_2 \times 2_2 \times 2_3$ and $2_2 \times 2_2 \times 2_3 \times 2_3$. (a) $3_2 \times 2_2 \times 2_3$. (b) $2_2 \times 2_2 \times 2_3 \times 2_3$. The red lines indicate fire, the blue and green lines indicate smoke, the solid lines denote the training precision, and the dotted lines denote the validating precision.

6.5 Visualization of the Color Sub-network and the Texture Sub-network

6.5.1 Activation of a Typical Sample

To observe what DNNSD cares, feed the network with a sample shown in Figure (11a), and exhibit the feature maps of the color sub-network, f^4, in Figure (11b) and the ones of the texture sub-network, f^{20}, in Figure (11c), respectively. The actual size of f^4 is (256, 256), and the one of f^{20} is (16, 16). The feature maps of the color sub-network are reduced and the feature maps of the texture sub-network are enlarged by proper times to illustrate in a same figure. It is confirmed that the feature maps are seriously activated at the pixels where smoke or fire happen, and suppressed at other pixels. So the network is effective.

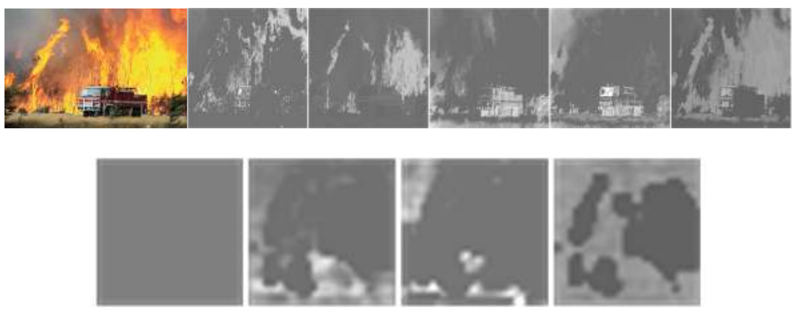

Figure 11: The feature maps from the color sub-network and the texture sub-network. (a) The original image with the size (256, 256), (b) the feature maps of the color sub-network with the size (256, 256), and (c) the feature maps of the whole network with the size (16, 16). (a) and (b) are zoomed out and (c) is zoom in by proper times for this display.

6.5.2 Maximize the Activation of the sub-networks

To further observe what DNNSD is interested in, a random image is fed into this network and later change its value to maximize the activation of the color sub-network and the whole network. The activation maximization is defined as

$$\hat{x} = \arg\max_{x} \quad \|\mathrm{h_{ij}}(x, \theta)\|. \tag{20}$$

where x is the random image, h_{ij} is a part of network from the input layer to the j^{th} channel of the i^{th} layer, and θ is the corresponding sub-network parameters. The 5 input images which maximize the activation of the color sub-network are shown in Figure (12a) and the ones corresponding to the whole network are shown in Figure (12b). It can be seen that Figure (12a) is not a pure color image with one single value, but a dominated color image with a small number of mottling. The typical color values, which maximize the activation of the color sub-network. So the color features cover a little texture information in the color sub-network. Compared with Figure (12a), Figure (12b) has rich texture. So the texture sub-network pay more attention on texture. On the other hand, though these 2 sub-network focus on different features, the network is incompletely decoupled. The first input image at Figure (12b) is corresponding to the label ”no-smoke-no-fire”, the second one corresponding to the label ”smoke-no-fire”, the third one corresponding to the label ”no-smoke-fire”, and the last one corresponding to the label ”smoke-fire”.

	1	2	3	4	5
R	111	153	107	89	111
G	111	188	85	164	133
B	111	71	152	140	174

Table 5: The typical values, the color of the most pixels, which could maximize the activation of the color sub-network.

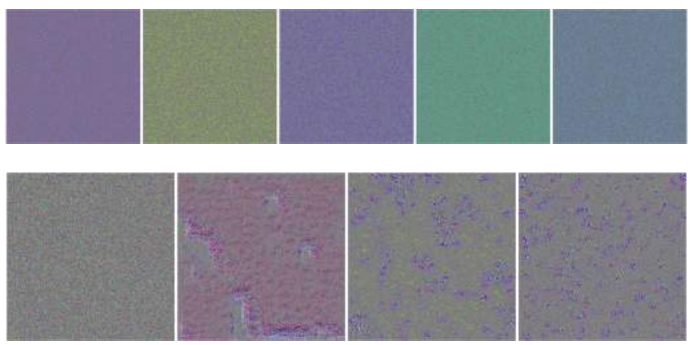

Figure 12: The input images which maximize the activation of the color sub-network and the whole network. (a) The 5 input images which maximize the activation of the color sub-network. (a) The 4 input images which maximize the activation of the whole network.

Since only 1×1 filters are adopted at the color sub-network and its pixel of the feature maps is one-by-one corresponding to the pixel of the input image, the color sub-network pays its main attention on color and some attention on fine texture. On the other hand, since the filters of the texture sub-network operate only the separate maps without any combination among these feature maps, the texture sub-network only pays its attention at texture. So the color features and the shape features are incompletely decoupled.

7 Conclusion

In some color modes, smoke can be observed obviously in some special channels. This is a common knowledge, which can be concluded that every object has its special color patterns. So it is helpful for object detection. But in general, color is not sufficient to distinguish these objects from background. So, color, texture, shape, and other features have to be integrated. On the other side, the fashionable neural networks extract complex and chaotic features to achieve the state-of-the-art accuracy. This method try to decouple these complex features into several kinds of features, which is effective to detect not only smoke but also other objects with special color patterns. The novelty of this work lies in decoupling of an end-to-end complex network to several sub-networks, each of which extracts only one type of feature. Because these sub-networks are serially stacked, and the network is trained by some pixel-level samples, the features are not completely decoupled. The texture features unavoidably make use of the color features. In spite of the incomplete decoupling, the proposed network is more effective and efficient than the other classic networks.

8 Acknowledgement

This research was sponsored by National Science Foundation of China (No. 51375012; 51521064), Zhejiang Provincial Public Technology Research Project of China (Project No. 2016C31117), Project of science and technology plans of Wenzhou City (Grants Nos. 2018ZG021, G20150017, ZG2017016), and the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry (Project No. R20150404), which are greatly appreciated by the authors.

References

[1] D. Yuan, X. Chang, P. Huang, Q. Liu, and Z. He, "Self-supervised deep correlation tracking," *IEEE Trans. Image Process.*, vol. 30, pp. 976–985, 2021. [Online]. Available: https://doi.org/10.1109/TIP.2020.3037518

[2] S. Hu, F. Zhu, X. Chang, and X. Liang, "Updet: Universal multi-agent reinforcement learning via policy decoupling with transformers," *CoRR*, vol. abs/2101.08001, 2021. [Online]. Available: https://arxiv.org/abs/2101.08001

[3] F. Wang, L. Zhu, C. Liang, J. Li, X. Chang, and K. Lu, "Robust optimal graph clustering," *Neurocomputing*, vol. 378, pp. 153–165, 2020. [Online]. Available: https://doi.org/10.1016/j.neucom.2019.07.102

[4] X. Bai, L. Zhu, C. Liang, J. Li, X. Nie, and X. Chang, "Multi-view feature selection via nonnegative structured graph learning," *Neurocomputing*, vol. 387, pp. 110–122, 2020. [Online]. Available: https://doi.org/10.1016/j.neucom.2020.01.044

[5] A. K. Singh, Z. Lv, H. Lu, and X. Chang, "Guest editorial: Recent trends in multimedia data-hiding: a reliable mean for secure communications," *J. Ambient Intell. Humaniz. Comput.*, vol. 11, no. 5, pp. 1795–1797, 2020. [Online]. Available: https://doi.org/10.1007/s12652-019-01499-5

[6] H. Liu, Q. Zheng, M. Luo, X. Chang, C. Yan, and L. Yao, "Memory transformation networks for weakly supervised visual classification," *Knowl. Based Syst.*, vol. 210, p. 106432, 2020. [Online]. Available: https://doi.org/10.1016/j.knosys.2020.106432

[7] Z. Ge, D. Mahapatra, X. Chang, Z. Chen, L. Chi, and H. Lu, "Improving multi-label chest x-ray disease diagnosis by exploiting disease and health labels dependencies,"

Multim. Tools Appl., vol. 79, no. 21-22, pp. 14 889–14 902, 2020. [Online]. Available: https://doi.org/10.1007/s11042-019-08260-2

[8] L. Zhang, X. Chang, J. Liu, M. Luo, M. Prakash, and A. G. Hauptmann, "Few-shot activity recognition with cross-modal memory network," *Pattern Recognit.*, vol. 108, p. 107348, 2020. [Online]. Available: https://doi.org/10.1016/j.patcog.2020.107348

[9] X. Chang, X. Liang, Y. Yan, and L. Nie, "Guest editorial: Image/video understanding and analysis," *Pattern Recognit. Lett.*, vol. 130, pp. 1–3, 2020. [Online]. Available: https://doi.org/10.1016/j.patrec.2019.07.003

[10] D. Zhang, L. Yao, K. Chen, S. Wang, X. Chang, and Y. Liu, "Making sense of spatio-temporal preserving representations for eeg-based human intention recognition," *IEEE Trans. Cybern.*, vol. 50, no. 7, pp. 3033–3044, 2020. [Online]. Available: https://doi.org/10.1109/TCYB.2019.2905157

[11] C. Yan, Q. Zheng, X. Chang, M. Luo, C. Yeh, and A. G. Hauptmann, "Semantics-preserving graph propagation for zero-shot object detection," *IEEE Trans. Image Process.*, vol. 29, pp. 8163–8176, 2020. [Online]. Available: https://doi.org/10.1109/TIP.2020.3011807

[12] W. Liu, X. Chang, L. Chen, D. Phung, X. Zhang, Y. Yang, and A. G. Hauptmann, "Pair-based uncertainty and diversity promoting early active learning for person re-identification," *ACM Trans. Intell. Syst. Technol.*, vol. 11, no. 2, pp. 21:1–21:15, 2020. [Online]. Available: https://doi.org/10.1145/3372121

[13] L. Zhang, M. Luo, J. Liu, X. Chang, Y. Yang, and A. G. Hauptmann, "Deep top-k ranking for image-sentence matching," *IEEE Trans. Multim.*, vol. 22, no. 3, pp. 775–785, 2020. [Online]. Available: https://doi.org/10.1109/TMM.2019.2931352

[14] R. Zhou, X. Chang, L. Shi, Y. Shen, Y. Yang, and F. Nie, "Person reidentification via multi-feature fusion with adaptive graph learning," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 31, no. 5, pp. 1592–1601, 2020. [Online]. Available: https://doi.org/10.1109/TNNLS.2019.2920905

[15] K. Chen, L. Yao, D. Zhang, X. Wang, X. Chang, and F. Nie, "A semisupervised recurrent convolutional attention model for human activity recognition," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 31, no. 5, pp. 1747–1756, 2020. [Online]. Available: https://doi.org/10.1109/TNNLS.2019.2927224

[16] J. Chen, Y. Wang, Y. Tian, and T. Huang, "Wavelet based smoke detection method with rgb contrast-image and shape constrain," in *Proceedings of IEEE International Conference on Visual Communications and Image Processing*, 2013.

[17] X. Qi and J. Ebert, "A computer vision-based method for fire detection in color videos," *International Journal of Imaging*, vol. 2, no. 9 S, pp. 22–34, Jan. 2009.

[18] L. Millan-Garcia, G. Sanchez-Perez, M. Nakano, K. Toscano-Medina, H. Perez-Meana, and L. Rojas-Cardenas, "An early fire detection algorithm using ip cameras," *Sensors*, vol. 12, no. 5, pp. 5670–5686, 2012.

[19] D. Krstinić, D. Stipaničev, and T. Jakovčević, "Histogram-based smoke segmentation in forest fire detection system," *Information Technology and Control*, vol. 38, no. 3, 2009.

[20] P. Ren, Y. Xiao, X. Chang, M. Prakash, F. Nie, X. Wang, and X. Chen, "Structured optimal graph-based clustering with flexible embedding," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 31, no. 10, pp. 3801–3813, 2020. [Online]. Available: https://doi.org/10.1109/TNNLS.2019.2946329

[21] D. R. Nayak, R. Dash, X. Chang, B. Majhi, and S. Bakshi, "Automated diagnosis of pathological brain using fast curvelet entropy features," *IEEE Trans. Sustain. Comput.*, vol. 5, no. 3, pp. 416–427, 2020. [Online]. Available: https://doi.org/10.1109/TSUSC.2018.2883822

[22] P. Huang, J. Hu, X. Chang, and A. G. Hauptmann, "Unsupervised multimodal neural machine translation with pseudo visual pivoting," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, 2020, pp. 8226–8237. [Online]. Available: https://doi.org/10.18653/v1/2020.acl-main.731

[23] L. Zhang, X. Chang, J. Liu, M. Luo, S. Wang, Z. Ge, and A. G. Hauptmann, "ZSTAD: zero-shot temporal activity detection," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, 2020, pp. 876–885. [Online]. Available: https://doi.org/10.1109/CVPR42600.2020.00096

[24] N. Fujiwara and K. Terada, "Extraction of a smoke region using fractal coding," in *Proceedings of IEEE International Symposium on Communications and Information Technology*, vol. 2, Jan. 2004, pp. 659–662 vol.2.

[25] C. Li, J. Peng, L. Yuan, G. Wang, X. Liang, L. Lin, and X. Chang, "Blockwisely supervised neural architecture search with knowledge distillation," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, 2020, pp. 1986–1995. [Online]. Available: https://doi.org/10.1109/CVPR42600.2020.00206

[26] C. Liu, X. Chang, and Y. Shen, "Unity style transfer for person re-identification," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, 2020, pp. 6886–6895. [Online]. Available: https://doi.org/10.1109/CVPR42600.2020.00692

[27] M. Zhang, H. Li, S. Pan, X. Chang, and S. W. Su, "Overcoming multi-model forgetting in one-shot NAS with diversity maximization," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, 2020, pp. 7806–7815. [Online]. Available: https://doi.org/10.1109/CVPR42600.2020.00783

[28] F. Zhu, Y. Zhu, X. Chang, and X. Liang, "Vision-language navigation with self-supervised auxiliary reasoning tasks," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, 2020, pp. 10 009–10 019. [Online]. Available: https://doi.org/10.1109/CVPR42600.2020.01003

[29] Y. Zhu, F. Zhu, Z. Zhan, B. Lin, J. Jiao, X. Chang, and X. Liang, "Vision-dialog navigation by exploring cross-modal memory," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, 2020, pp. 10 727–10 736. [Online]. Available: https://doi.org/10.1109/CVPR42600.2020.01074

[30] H. Maruta, A. Nakamura, T. Yamamichi, and F. Kurokawa, "Image based smoke detection with local Hurst exponent," *Proceedings of International Conference on Image Processing*, pp. 4653–4656, Jan. 2010.

[31] M. Han, Y. Wang, X. Chang, and Y. Qiao, "Mining inter-video proposal relations for video object detection," in *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXI*, 2020, pp. 431–446. [Online]. Available: https://doi.org/10.1007/978-3-030-58589-1_26

[32] J. Zhang, M. Wang, Q. Li, S. Wang, X. Chang, and B. Wang, "Quadratic sparse gaussian graphical model estimation method for massive variables," in *Proceedings of*

the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020, 2020, pp. 2964–2972. [Online]. Available: https://doi.org/10.24963/ijcai.2020/410

[33] Z. Li, X. Chang, L. Yao, S. Pan, Z. Ge, and H. Zhang, "Grounding visual concepts for zero-shot event detection and event captioning," in *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, 2020, pp. 297–305. [Online]. Available: https://dl.acm.org/doi/10.1145/3394486.3403072

[34] Z. Wu, S. Pan, G. Long, J. Jiang, X. Chang, and C. Zhang, "Connecting the dots: Multivariate time series forecasting with graph neural networks," in *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, 2020, pp. 753–763. [Online]. Available: https://dl.acm.org/doi/10.1145/3394486.3403118

[35] D. K. Appana, R. Islam, S. A. Khan, and J.-M. Kim, "A video-based smoke detection using smoke flow pattern and spatial-temporal energy analyses for alarm systems," *Information Sciences*, vol. 418-419, pp. 91–101, Dec. 2017.

[36] P. Huang, X. Chang, A. G. Hauptmann, and E. H. Hovy, "Forward and backward multimodal NMT for improved monolingual and multilingual cross-modal retrieval," in *Proceedings of the 2020 on International Conference on Multimedia Retrieval, ICMR 2020, Dublin, Ireland, June 8-11, 2020*, 2020, pp. 53–62. [Online]. Available: https://doi.org/10.1145/3372278.3390674

[37] W. Wang, R. Liu, M. Wang, S. Wang, X. Chang, and Y. Chen, "Memory-based network for scene graph with unbalanced relations," in *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*, 2020, pp. 2400–2408. [Online]. Available: https://doi.org/10.1145/3394171.3413507

[38] X. Cheng, Y. Zhong, M. Harandi, Y. Dai, X. Chang, H. Li, T. Drummond, and Z. Ge, "Hierarchical neural architecture search for deep stereo matching," in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/fc146be0b230d7e0a92e66a6114b840d-Abstract.html

[39] M. Zhang, H. Li, S. Pan, X. Chang, Z. Ge, and S. W. Su, "Differentiable neural architecture search in equivalent space with exploration enhancement," in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/9a96a2c73c0d477ff2a6da3bf538f4f4-Abstract.html

[40] F. Yuan, J. Shi, X. Xia, and Y. Fang, "High-order local ternary patterns with locality preserving projection for smoke detection and image classification," *Information Sciences*, vol. 372, pp. 225–240, Dec. 2016.

[41] W. Liu, G. Kang, P. Huang, X. Chang, L. Yu, Y. Qian, J. Liang, L. Gui, J. Wen, P. Chen, and A. G. Hauptmann, "Argus: Efficient activity detection system for extended video analysis," in *IEEE Winter Applications of Computer Vision Workshops, WACV Workshops 2020, Snowmass Village, CO, USA, March 1-5, 2020*, 2020, pp. 126–133. [Online]. Available: https://doi.org/10.1109/WACVW50321.2020.9096929

[42] M. Wu, S. Pan, C. Zhou, X. Chang, and X. Zhu, "Unsupervised domain adaptive graph convolutional networks," in *WWW '20: The Web Conference 2020,*

Taipei, Taiwan, April 20-24, 2020, 2020, pp. 1457–1467. [Online]. Available: https://doi.org/10.1145/3366423.3380219

[43] P. Ren, Y. Xiao, X. Chang, P. Huang, Z. Li, X. Chen, and X. Wang, "A comprehensive survey of neural architecture search: Challenges and solutions," *CoRR*, vol. abs/2006.02903, 2020. [Online]. Available: https://arxiv.org/abs/2006.02903

[44] S. Frizzi, R. Kaabi, M. Bouchouicha, J.-M. Ginoux, E. Moreau, and F. Fnaiech, "Convolutional neural network for video fire and smoke detectoin," *IEEE Industrial Electronics Society*, pp. 877–882, 2016.

[45] Z. Yin, B. Wan, F. Yuan, and X. Xia, "A deep normalization and convolutional neural network for image smoke detection," *IEEE Access*, vol. 5, pp. 18 429–18 438, 2017.

[46] M. Yin, C. Lang, Z. Li, and S. Feng, "Recurrent convolutional network for video-based smoke detection," *Multimedia Tools and Applications*, vol. 78, no. 1, pp. 237–256, Jan. 2019.

[47] M. Li, F. Wang, X. Chang, and X. Liang, "Auxiliary signal-guided knowledge encoder-decoder for medical report generation," *CoRR*, vol. abs/2006.03744, 2020. [Online]. Available: https://arxiv.org/abs/2006.03744

[48] Z. Yu, J. Nguyen, X. Chang, J. Kelly, C. McLean, L. Zhang, V. Mar, and Z. Ge, "Melanoma diagnosis with spatio-temporal feature learning on sequential dermoscopic images," *CoRR*, vol. abs/2006.10950, 2020. [Online]. Available: https://arxiv.org/abs/2006.10950

[49] S. Hu and X. Chang, "Multi-view drone-based geo-localization via style and spatial alignment," *CoRR*, vol. abs/2006.13681, 2020. [Online]. Available: https://arxiv.org/abs/2006.13681

[50] Z. Wang, M. Huang, Q. Zhu, and S. Jiang, "Smoke detection in storage yard using parallel deep residual network," *Laser and Optoelectronics Progress*, May 2018.

[51] O. Maksymiv, T. Rak, and D. Peleshko, "Real-time fire detection method combining adaboost, lbp and convolutional neural network in video sequence," in *Proceedings of International Conference The Experience of Designing and Application of CAD Systems in Microelectronics*. IEEE, 2017, pp. 351–353.

[52] J. Chen, Z. Wang, H. Chen, and L. Zuo, "Dynamic smoke detection using cascaded convolutional neural network for surveillance videos," *University of Electronic Science and Technology of China*, vol. 46, no. 6, pp. 992–096, 2016.

[53] Y. Zhao, J. Ma, X. Li, and J. Zhang, "Saliency detection and deep learning-based wildfire identification in UAV imagery," *Sensors*, vol. 18, no. 3, p. 712, Feb. 2018.

[54] D. Yuan, X. Chang, and Z. He, "Accurate bounding-box regression with distance-iou loss for visual tracking," *CoRR*, vol. abs/2007.01864, 2020. [Online]. Available: https://arxiv.org/abs/2007.01864

[55] P. Ren, Y. Xiao, X. Chang, P. Huang, Z. Li, X. Chen, and X. Wang, "A survey of deep active learning," *CoRR*, vol. abs/2009.00236, 2020. [Online]. Available: https://arxiv.org/abs/2009.00236

[56] C. Yan, X. Chang, M. Luo, Q. Zheng, X. Zhang, Z. Li, and F. Nie, "Self-weighted robust LDA for multiclass classification with edge classes," *CoRR*, vol. abs/2009.12362, 2020. [Online]. Available: https://arxiv.org/abs/2009.12362

[57] Z. Li, L. Yao, X. Chang, K. Zhan, J. Sun, and H. Zhang, "Zero-shot event detection via event-adaptive concept relevance mining," *Pattern Recognit.*, vol. 88, pp. 595–603, 2019. [Online]. Available: https://doi.org/10.1016/j.patcog.2018.12.010

[58] L. Zhang, J. Liu, M. Luo, X. Chang, Q. Zheng, and A. G. Hauptmann, "Scheduled sampling for one-shot learning via matching network," *Pattern Recognit.*, vol. 96, 2019. [Online]. Available: https://doi.org/10.1016/j.patcog.2019.07.007

[59] R. P. Padhy, X. Chang, S. K. Choudhury, P. K. Sa, and S. Bakshi, "Multi-stage cascaded deconvolution for depth map and surface normal prediction from single image," *Pattern Recognit. Lett.*, vol. 127, pp. 165–173, 2019. [Online]. Available: https://doi.org/10.1016/j.patrec.2018.07.012

[60] C. Gong, D. Tao, X. Chang, and J. Yang, "Ensemble teaching for hybrid label propagation," *IEEE Trans. Cybern.*, vol. 49, no. 2, pp. 388–402, 2019. [Online]. Available: https://doi.org/10.1109/TCYB.2017.2773562

[61] K. Zhan, X. Chang, J. Guan, L. Chen, Z. Ma, and Y. Yang, "Adaptive structure discovery for multimedia analysis using multiple features," *IEEE Trans. Cybern.*, vol. 49, no. 5, pp. 1826–1834, 2019. [Online]. Available: https://doi.org/10.1109/TCYB.2018.2815012

[62] M. Luo, C. Yan, Q. Zheng, X. Chang, L. Chen, and F. Nie, "Discrete multi-graph clustering," *IEEE Trans. Image Process.*, vol. 28, no. 9, pp. 4701–4712, 2019. [Online]. Available: https://doi.org/10.1109/TIP.2019.2913081

[63] E. Yu, J. Sun, J. Li, X. Chang, X. Han, and A. G. Hauptmann, "Adaptive semi-supervised feature selection for cross-modal retrieval," *IEEE Trans. Multim.*, vol. 21, no. 5, pp. 1276–1288, 2019. [Online]. Available: https://doi.org/10.1109/TMM.2018.2877127

[64] Z. Cheng, X. Chang, L. Zhu, R. Catherine Kanjirathinkal, and M. S. Kankanhalli, "MMALFM: explainable recommendation by leveraging reviews and images," *ACM Trans. Inf. Syst.*, vol. 37, no. 2, pp. 16:1–16:28, 2019. [Online]. Available: https://doi.org/10.1145/3291060

[65] K. Chen, L. Yao, D. Zhang, X. Chang, G. Long, and S. Wang, "Distributionally robust semi-supervised learning for people-centric sensing," in *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, 2019, pp. 3321–3328. [Online]. Available: https://doi.org/10.1609/aaai.v33i01.33013321

[66] Z. Li, W. Liu, X. Chang, L. Yao, M. Prakash, and H. Zhang, "Domain-aware unsupervised cross-dataset person re-identification," in *Advanced Data Mining and Applications - 15th International Conference, ADMA 2019, Dalian, China, November 21-23, 2019, Proceedings*, 2019, pp. 406–420. [Online]. Available: https://doi.org/10.1007/978-3-030-35231-8_29

[67] P. Huang, X. Chang, and A. G. Hauptmann, "Multi-head attention with diversity for learning grounded multilingual multimodal representations," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, 2019, pp. 1461–1467. [Online]. Available: https://doi.org/10.18653/v1/D19-1154

[68] E. Yu, J. Sun, L. Wang, X. Chang, H. Zhang, and A. G. Hauptmann, "Cross-modal transfer hashing based on coherent projection," in *IEEE International Conference on Multimedia & Expo Workshops, ICME Workshops 2019, Shanghai, China, July 8-12, 2019*, 2019, pp. 477–482. [Online]. Available: https://doi.org/10.1109/ICMEW.2019.00088

[69] K. He, J. Sun, and X. Tang, "Single image haze removal using dark channel prior," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 12, pp. 2341–2353, 2011.

[70] P. Huang, Vaibhav, X. Chang, and A. G. Hauptmann, "Improving what cross-modal retrieval models learn through object-oriented inter- and intra-modal attention networks," in *Proceedings of the 2019 on International Conference on Multimedia Retrieval, ICMR 2019, Ottawa, ON, Canada, June 10-13, 2019*, 2019, pp. 244–252. [Online]. Available: https://doi.org/10.1145/3323873.3325043

[71] P. Huang, G. Kang, W. Liu, X. Chang, and A. G. Hauptmann, "Annotation efficient cross-modal retrieval with adversarial attentive alignment," in *Proceedings of the 27th ACM International Conference on Multimedia, MM 2019, Nice, France, October 21-25, 2019*, 2019, pp. 1758–1767. [Online]. Available: https://doi.org/10.1145/3343031.3350894

[72] X. Chang, W. Liu, P. Huang, C. Li, F. Zhu, M. Han, M. Li, M. Ma, S. Hu, G. Kang, J. Liang, L. Gui, L. Yu, Y. Qian, J. Wen, and A. G. Hauptmann, "Mmvg-inf-etrol@trecvid 2019: Activities in extended video," in *2019 TREC Video Retrieval Evaluation, TRECVID 2019, Gaithersburg, MD, USA, November 12-13, 2019*, 2019. [Online]. Available: https://www-nlpir.nist.gov/projects/tvpubs/tv19.papers/mmvg-Informedia.pdf

[73] W. Chen, S. Yuan, G. Tsai, and H. Wang, "Color channel-based smoke removal algorithm using machine learning for static images," in *Proceedings of IEEE International Conference on Image Processing*. IEEE, Oct. 2018, pp. 2855–2859.

[74] E. Yu, W. Liu, G. Kang, X. Chang, J. Sun, and A. G. Hauptmann, "Inf@trecvid 2019: Instance search task," in *2019 TREC Video Retrieval Evaluation, TRECVID 2019, Gaithersburg, MD, USA, November 12-13, 2019*, 2019. [Online]. Available: https://www-nlpir.nist.gov/projects/tvpubs/tv19.papers/inf_ins.pdf

[75] I. Rida, S. Bakshi, X. Chang, and H. Proença, "Forensic shoe-print identification: A brief survey," *CoRR*, vol. abs/1901.01431, 2019. [Online]. Available: http://arxiv.org/abs/1901.01431

[76] F. Zhu, X. Chang, R. Zeng, and M. Tan, "Continual reinforcement learning with diversity exploration and adversarial self-correction," *CoRR*, vol. abs/1906.09205, 2019. [Online]. Available: http://arxiv.org/abs/1906.09205

[77] Y. Zhao, "Candidate smoke region segmentation of fire video based on rough set theory," *Journal of Electrical and Computer Engineering*, vol. 2015, pp. 1–8, 2015.

[78] D. Xing, Y. Zhongming, W. Lin, and L. Jinlan, "Smoke image segmentation based on color model," *Innovation and Sustainability*, vol. 6, no. 2, p. 130, Aug. 2015.

[79] Z. Dengyi, H. Aike, R. Yujie, and Z. Jinming, "Forest fire and smoke detection based on video image segmentation," in *Proceedings of International Society for Optical Engineering*, vol. 6788, 2007.

[80] B. U. Töreyin, Y. Dedeoğlu, and A. E. Cetin, "Wavelet based real-time smoke detection in video," in *Proceedings of European Signal Processing Conference*, Jan. 2005, pp. 293–296.

[81] N. Alamgir, K. Nguyen, V. Chandran, and W. Boles, "Combining multi-channel color space with local binary co-occurrence feature descriptors for accurate smoke detection from surveillance videos," *Fire Safety*, vol. 102, pp. 1–10, Dec. 2018.

[82] P. Piccinini, S. Calderara, and R. Cucchiara, "Reliable smoke detection system in the domains of image energy and color," in *Proceedings of International Conference on Image Processing*, Jan. 2008, pp. 1376–1379.

[83] H. Tian, W. Li, P. O. Ogunbona, and L. Wang, "Detection and separation of smoke from single image frames," *IEEE Transactions on Image Processing*, vol. 27, pp. 1164–1177, Mar. 2018.

[84] X. Wu, X. Lu, and H. Leung, "Video smoke separation and detection via sparse representation," *Neurocomputing*, vol. 360, pp. 61–74, Sep. 2019.

[85] P. Huang, X. Chang, and A. G. Hauptmann, "Multi-head attention with diversity for learning grounded multilingual multimodal representations," *CoRR*, vol. abs/1910.00058, 2019. [Online]. Available: http://arxiv.org/abs/1910.00058

[86] A. K. Singh, Z. Lv, S. Rho, S. K. Singh, X. Chang, and W. Puech, "IEEE access special section editorial: Information security solutions for telemedicine applications," *IEEE Access*, vol. 6, pp. 79 005–79 009, 2018. [Online]. Available: https://doi.org/10.1109/ACCESS.2018.2885256

[87] C. Gong, Z. Li, X. Chang, and Y. Luo, "Learning-based multimedia analyses and applications," *Adv. Multim.*, vol. 2018, pp. 2 705 839:1–2 705 839:2, 2018. [Online]. Available: https://doi.org/10.1155/2018/2705839

[88] Z. Zhao, X. Li, X. Du, Q. Chen, Y. Zhao, F. Su, X. Chang, and A. G. Hauptmann, "A unified framework with a benchmark dataset for surveillance event detection," *Neurocomputing*, vol. 278, pp. 62–74, 2018. [Online]. Available: https://doi.org/10.1016/j.neucom.2017.04.079

[89] X. Chang, Y. Yan, and L. Nie, "Guest editorial: Semantic concept discovery in MM data," *Multim. Tools Appl.*, vol. 77, no. 3, pp. 2945–2946, 2018. [Online]. Available: https://doi.org/10.1007/s11042-018-5646-9

[90] L. Zhang, J. Liu, M. Luo, X. Chang, and Q. Zheng, "Deep semisupervised zero-shot learning with maximum mean discrepancy," *Neural Comput.*, vol. 30, no. 5, 2018. [Online]. Available: https://doi.org/10.1162/neco_a_01071

[91] D. Cheng, Y. Gong, X. Chang, W. Shi, A. G. Hauptmann, and N. Zheng, "Deep feature learning via structured graph laplacian embedding for person re-identification," *Pattern Recognit.*, vol. 82, pp. 94–104, 2018. [Online]. Available: https://doi.org/10.1016/j.patcog.2018.05.007

[92] M. Luo, X. Chang, L. Nie, Y. Yang, A. G. Hauptmann, and Q. Zheng, "An adaptive semisupervised feature analysis for video semantic recognition," *IEEE Trans. Cybern.*, vol. 48, no. 2, pp. 648–660, 2018. [Online]. Available: https://doi.org/10.1109/TCYB.2017.2647904

[93] D. P. Agrawal, B. B. Gupta, H. Wang, X. Chang, S. Yamaguchi, and G. M. Pérez, "Guest editorial deep learning models for industry informatics," *IEEE Trans. Ind. Informatics*, vol. 14, no. 7, pp. 3166–3169, 2018. [Online]. Available: https://doi.org/10.1109/TII.2018.2834547

[94] W. Liu, X. Chang, Y. Yan, Y. Yang, and A. G. Hauptmann, "Few-shot text and image classification via analogical transfer learning," *ACM Trans. Intell. Syst. Technol.*, vol. 9, no. 6, pp. 71:1–71:20, 2018. [Online]. Available: https://doi.org/10.1145/3230709

[95] S. P. Hohberg, "Wildfire smoke detection using convolutional neural networks," 2015.

[96] C. Tao, J. Zhang, and P. Wang, "Smoke detection based on deep convolutional neural networks," in *Proceedings of International Conference on Industrial Informatics - Computing Technology, Intelligent Technology, Industrial Information Integration*, Jan. 2017, pp. 150–153.

[97] A. Filonenko, L. Kurnianggoro, and K.-H. Jo, "Comparative study of modern convolutional neural networks for smoke detection on image data," in *Proceedings of International Conference on Industrial Informatics - Computing Technology, Intelligent Technology, Industrial Human system Integration*, Jul. 2017, pp. 64–68.

[98] N. M. Dung, D. Kim, and S. Ro, "A video smoke detection algorithm based on cascade classification and deep learning," *KSII Transactions on Internet and Information Systems*, vol. 12, pp. 6018–6033, Dec. 2018.

[99] M. Luo, F. Nie, X. Chang, Y. Yang, A. G. Hauptmann, and Q. Zheng, "Adaptive unsupervised feature selection with structure regularization," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 29, no. 4, pp. 944–956, 2018. [Online]. Available: https://doi.org/10.1109/TNNLS.2017.2650978

[100] Z. Ma, X. Chang, Z. Xu, N. Sebe, and A. G. Hauptmann, "Joint attributes and event analysis for multimedia event detection," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 29, no. 7, pp. 2921–2930, 2018. [Online]. Available: https://doi.org/10.1109/TNNLS.2017.2709308

[101] Z. Li, F. Nie, X. Chang, L. Nie, H. Zhang, and Y. Yang, "Rank-constrained spectral clustering with flexible embedding," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 29, no. 12, pp. 6073–6082, 2018. [Online]. Available: https://doi.org/10.1109/TNNLS.2018.2817538

[102] Z. Li, F. Nie, X. Chang, Y. Yang, C. Zhang, and N. Sebe, "Dynamic affinity graph construction for spectral clustering using multiple features," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 29, no. 12, pp. 6323–6332, 2018. [Online]. Available: https://doi.org/10.1109/TNNLS.2018.2829867

[103] X. Chen, G. Yuan, W. Wang, F. Nie, X. Chang, and J. Z. Huang, "Local adaptive projection framework for feature selection of labeled and unlabeled data," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 29, no. 12, pp. 6362–6373, 2018. [Online]. Available: https://doi.org/10.1109/TNNLS.2018.2830186

[104] Z. Li, F. Nie, X. Chang, Z. Ma, and Y. Yang, "Balanced clustering via exclusive lasso: A pragmatic approach," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, 2018, pp. 3596–3603. [Online]. Available: https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16711

[105] W. Liu, X. Chang, L. Chen, and Y. Yang, "Semi-supervised bayesian attribute learning for person re-identification," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, 2018, pp. 7162–7169. [Online]. Available: https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17325

[106] L. Gui, X. Liang, X. Chang, and A. G. Hauptmann, "Adaptive context-aware reinforced agent for handwritten text recognition," in *British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018*, 2018, p. 207. [Online]. Available: http://bmvc2018.org/contents/papers/0628.pdf

[107] J. Han, L. Yang, D. Zhang, X. Chang, and X. Liang, "Reinforcement cutting-agent learning for video object segmentation," in *2018 IEEE Conference on Computer Vision*

and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, 2018, pp. 9080–9089. [Online]. Available: http://openaccess.thecvf.com/content_cvpr_2018/html/Han_Reinforcement_Cutting-Agent_Learning_CVPR_2018_paper.html

[108] X. Chang, P. Huang, Y. Shen, X. Liang, Y. Yang, and A. G. Hauptmann, "RCAA: relational context-aware agents for person search," in *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part IX*, 2018, pp. 86–102. [Online]. Available: https://doi.org/10.1007/978-3-030-01240-3_6

[109] F. Yuan, L. Zhang, B. Wan, and X. Xia, "Convolutional neural networks based on multi-scale additive merging layers for visual smoke recognition," *Machine Vision and Applications*, vol. 30, pp. 345–358, Mar. 2019.

[110] K. Gu, Z. Xia, J. Qiao, and W. Lin, "Deep dual-channel neural network for image-based smoke detection," *IEEE Transactions on Multimedia*, p. 1, 2019.

[111] L. Liu, D. Jing, and X. Chang, "Particle filtering based visual attention model for moving target detection," in *14th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery, ICNC-FSKD 2018, Huangshan, China, July 28-30, 2018*, 2018, pp. 1387–1391. [Online]. Available: https://doi.org/10.1109/FSKD.2018.8686873

[112] H. Wang, X. Chang, L. Shi, Y. Yang, and Y. Shen, "Uncertainty sampling for action recognition via maximizing expected average precision," in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, 2018, pp. 964–970. [Online]. Available: https://doi.org/10.24963/ijcai.2018/134

[113] C. Gong, X. Chang, M. Fang, and J. Yang, "Teaching semi-supervised classifier via generalized distillation," in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, 2018, pp. 2156–2162. [Online]. Available: https://doi.org/10.24963/ijcai.2018/298

[114] J. Chen, S. Chen, Q. Jin, A. G. Hauptmann, P. Huang, J. Liang, Vaibhav, X. Chang, J. Liu, T. Hu, W. Liu, W. Ke, W. Barrios, H. Idrees, D. Yoo, Y. Sheikh, R. Salakhutdinov, K. Kitani, and D. Huang, "Informedia @ TRECVID 2018: Ad-hoc video search, video to text description, activities in extended video," in *2018 TREC Video Retrieval Evaluation, TRECVID 2018, Gaithersburg, MD, USA, November 13-15, 2018*, 2018. [Online]. Available: https://www-nlpir.nist.gov/projects/tvpubs/tv18.papers/inf.pdf

[115] Z. Cheng, X. Chang, L. Zhu, R. Catherine Kanjirathinkal, and M. S. Kankanhalli, "MMALFM: explainable recommendation by leveraging reviews and images," *CoRR*, vol. abs/1811.05318, 2018. [Online]. Available: http://arxiv.org/abs/1811.05318

[116] M. Luo, X. Chang, Z. Li, L. Nie, A. G. Hauptmann, and Q. Zheng, "Simple to complex cross-modal learning to rank," *Comput. Vis. Image Underst.*, vol. 163, pp. 67–77, 2017. [Online]. Available: https://doi.org/10.1016/j.cviu.2017.07.001

[117] M. Luo, F. Nie, X. Chang, Y. Yang, A. G. Hauptmann, and Q. Zheng, "Avoiding optimal mean $l_{2,1}$-norm maximization-based robust PCA for reconstruction," *Neural Comput.*, vol. 29, no. 4, pp. 1124–1150, 2017. [Online]. Available: https://doi.org/10.1162/NECO_a_00937

[118] Y. Chang, F. Nie, Z. Li, X. Chang, and H. Huang, "Refined spectral clustering via embedded label propagation," *Neural Comput.*, vol. 29, no. 12, 2017. [Online]. Available: https://doi.org/10.1162/neco_a_01022

[119] X. Chang, Y. Yu, Y. Yang, and E. P. Xing, "Semantic pooling for complex event analysis in untrimmed videos," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 8, pp. 1617–1632, 2017. [Online]. Available: https://doi.org/10.1109/TPAMI.2016.2608901

[120] X. Chang, Z. Ma, Y. Yang, Z. Zeng, and A. G. Hauptmann, "Bi-level semantic representation analysis for multimedia event detection," *IEEE Trans. Cybern.*, vol. 47, no. 5, pp. 1180–1197, 2017. [Online]. Available: https://doi.org/10.1109/TCYB.2016.2539546

[121] L. Nie, L. Zhang, Y. Yan, X. Chang, M. Liu, and L. Shaoling, "Multiview physician-specific attributes fusion for health seeking," *IEEE Trans. Cybern.*, vol. 47, no. 11, pp. 3680–3691, 2017. [Online]. Available: https://doi.org/10.1109/TCYB.2016.2577590

[122] D. Zhang, J. Han, L. Jiang, S. Ye, and X. Chang, "Revealing event saliency in unconstrained video collection," *IEEE Trans. Image Process.*, vol. 26, no. 4, pp. 1746–1758, 2017. [Online]. Available: https://doi.org/10.1109/TIP.2017.2658957

[123] X. Chang, Z. Ma, M. Lin, Y. Yang, and A. G. Hauptmann, "Feature interaction augmented sparse learning for fast kinect motion detection," *IEEE Trans. Image Process.*, vol. 26, no. 8, pp. 3911–3920, 2017. [Online]. Available: https://doi.org/10.1109/TIP.2017.2708506

[124] R. Wang, F. Nie, R. Hong, X. Chang, X. Yang, and W. Yu, "Fast and orthogonal locality preserving projections for dimensionality reduction," *IEEE Trans. Image Process.*, vol. 26, no. 10, pp. 5019–5030, 2017. [Online]. Available: https://doi.org/10.1109/TIP.2017.2726188

[125] S. Wang, X. Li, X. Chang, L. Yao, Q. Z. Sheng, and G. Long, "Learning multiple diagnosis codes for ICU patients with local disease correlation mining," *ACM Trans. Knowl. Discov. Data*, vol. 11, no. 3, pp. 31:1–31:21, 2017. [Online]. Available: https://doi.org/10.1145/3003729

[126] Z. Li, F. Nie, X. Chang, and Y. Yang, "Beyond trace ratio: Weighted harmonic mean of trace ratios for multiclass discriminant analysis," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 10, pp. 2100–2110, 2017. [Online]. Available: https://doi.org/10.1109/TKDE.2017.2728531

[127] Z. Ma, X. Chang, Y. Yang, N. Sebe, and A. G. Hauptmann, "The many shades of negativity," *IEEE Trans. Multim.*, vol. 19, no. 7, pp. 1558–1568, 2017. [Online]. Available: https://doi.org/10.1109/TMM.2017.2659221

[128] L. Nie, L. Zhang, L. Meng, X. Song, X. Chang, and X. Li, "Modeling disease progression via multisource multitask learners: A case study with alzheimer's disease," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 28, no. 7, pp. 1508–1519, 2017. [Online]. Available: https://doi.org/10.1109/TNNLS.2016.2520964

[129] X. Chang and Y. Yang, "Semisupervised feature analysis by mining correlations among multiple tasks," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 28, no. 10, pp. 2294–2305, 2017. [Online]. Available: https://doi.org/10.1109/TNNLS.2016.2582746

[130] M. Fan, X. Chang, and D. Tao, "Structure regularized unsupervised discriminant feature analysis," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, 2017, pp. 1870–1876. [Online]. Available: http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14288

[131] M. Luo, F. Nie, X. Chang, Y. Yang, A. G. Hauptmann, and Q. Zheng, "Probabilistic non-negative matrix factorization and its robust extensions for topic modeling," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February*

4-9, 2017, San Francisco, California, USA, 2017, pp. 2308–2314. [Online]. Available: http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14469

[132] X. Xue, F. Nie, S. Wang, X. Chang, B. Stantic, and M. Yao, "Multi-view correlated feature learning by uncovering shared component," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, 2017, pp. 2810–2816. [Online]. Available: http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14439

[133] H. Fan, X. Chang, D. Cheng, Y. Yang, D. Xu, and A. G. Hauptmann, "Complex event detection by identifying reliable shots from untrimmed videos," in *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, 2017, pp. 736–744. [Online]. Available: https://doi.org/10.1109/ICCV.2017.86

[134] B. Zhou, A. Khosla, A. Lapedriza, and A. Oliva, "Learning deep features for discriminative localization," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Jun. 2016, pp. 2921–2929.

[135] "visor," http://imagelab.ing.unimore.it/visor/video_categories.asp.

[136] "fesb," http://wildfire.fesb.hr/index.php?option=com_content&view=article&id=65%3Aadaptive-estimation-of-detection-parameters-video-database&Itemid=72.

[137] "yfn," http://staff.ustc.edu.cn/~yfn/vsd.html.

[138] "Visifire," http://signal.ee.bilkent.edu.tr/VisiFire/Demo.

[139] F. Yuan, L. Zhang, X. Xia, B. Wan, Q. Huang, and X. Li, "Deep smoke segmentation," *Neurocomputing*, vol. 357, pp. 248–260, Sep. 2019.

[140] D. Xiong and L. Yan, "Early smoke detection of forest fires based on SVM image segmentation," *Forest Science*, vol. 65, pp. 150–159, Apr. 2019.

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

[142] F. Chollet, "Xception: deep dearning with depthwise separable convolutions," Oct. 2016.

[143] C. Szegedy, V. Vanhoucke, S. Ioffe, and J. Shlens, "Rethinking the inception architecture for computer vision," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Jun. 2016, pp. 2818–2826.

[144] A. G. Howard, M. Zhu, B. Chen, and D. Kalenichenko, "Mobilenets: efficient convolutional neural networks for mobile vision applications," Apr. 2017.

[145] S. Rinsurongkawong, M. Ekpanyapong, and M. N. Dailey, "Fire detection for early fire alarm based on optical flow video processing," in *Proceedings of International Conference on Electrical Engineering/Electronics Computer Telecommunications and Information Technology*. IEEE, May 2012, pp. 1–4.

[146] R. D. Labati, A. Genovese, V. Piuri, and F. Scotti, "Wildfire smoke detection using computational intelligence techniques enhanced with synthetic smoke plume generation," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 43, no. 4, pp. 1003–1012, 2013.