Convolutional Neural Network for Video Fire and Smoke Detection

Sebastien Frizzi[1] Rabeb Kaabi[2,3,4] Moez Bouchouicha[2,3] Jean-Marc Ginoux[2,3] Eric Moreau[2,3] Farhat Fnaiech[4]
frizzi@univ-tln.fr rabebkaabi89@gmail.com moez@univ-tln.fr ginoux@univ-tln.fr moreau@univ-tln.fr fnaiech@ieee.org

[1]Université de Toulon, Département Génie Biologie- IUT, 83957 La Garde, France
[2]Aix Marseille Université, CNRS, ENSAM, LSIS, UMR 7296, 13397 Marseille, France
[3]Université de Toulon, CNRS, LSIS, UMR 7296, 83957 La Garde, France
[4]Université de Tunis, ENSIT, LR13ES03, SIME, 1008, Montfleury, Tunisia

***Abstract*—Research on video analysis for fire detection has become a hot topic in computer vision. However, the conventional algorithms use exclusively rule-based models and features vector to classify whether a frame is fire or not. These features are difficult to define and depend largely on the kind of fire observed. The outcome leads to low detection rate and high false-alarm rate. A different approach for this problem is to use a learning algorithm to extract the useful features instead of using an expert to build them. In this paper, we propose a convolutional neural network (CNN) for identifying fire in videos. Convolutional neural network are shown to perform very well in the area of object classification. This network has the ability to perform feature extraction and classification within the same architecture. Tested on real video sequences, the proposed approach achieves better classification performance as some of relevant conventional video fire detection methods and indicates that using CNN to detect fire in videos is very promising.**

Keywords—Fire and smoke detection, deep learning, convolutional neural network, feature maps, max pooling, dropout

I. INTRODUCTION

Fire detection task is crucial for people safety. Several fire detection systems were developed to prevent damages caused by fire. One can find different technical solutions. Most of them are sensors based and are also generally limited to indoors. They detect the presence of particles generated by smoke and fire by ionization, which requires a close proximity to the fire. Consequently, they cannot be used in large covered area. Moreover, they cannot provide information about initial fire location, direction of smoke propagation, size of the fire, growth rate of the fire, etc. To get over such limitations video fire detection systems are used.

Due to rapid developments in digital camera technology and video processing techniques, there is a significant tendency to replace standard fire detection methods with computer vision based systems. Video based fire detection techniques are well suited to detect fire in large and open spaces. Furthermore, thanks to these systems one can analyze the fire behavior and perform a three dimensional localization of the fire. In addition, closed circuit television surveillance systems are nowadays installed in various places monitoring indoors and outdoors. In this circumstance, it would be advantageous to develop a video fire detection system which could use this existing equipment without introducing any extra cost.

The research in this domain was started since the nineties. There are several video-based fire and flame detection algorithms in the literature. The majority of these algorithms focuses on the color and the shape characteristics together combined to the temporal behavior of smoke and flames [1],[2],[3],[4],[5], [6] and [7]. Afterward, the goal is to build a rule-based algorithm or a multi-dimensional feature vector which is used as an input to a conventional algorithm for classification: SVM, Neural Network, Adaboost, etc. Therefore, conventional video fire detection methods address the problem by relying on expert knowledge to build relevant features extractors. Experts are required to create the rule-based models and the discriminative features. A different approach for this problem is to use a learning algorithm to extract the useful features instead of using an expert to build them. Deep learning algorithms can learn such useful features to detect fire and smoke in video. Convolutional Neural Networks are a variant of deep learning that can extract topological properties from an image.

Thereby, our approach is conceptually very simple. We use a Convolutional Neural Network as a powerful fire/smoke frame detector. The CNN operates directly on raw RGB frame without the need of the feature extraction stage. The CNN automatically learns a set of visual features from the training data.

The rest of the paper is organized as follow. The next section reviews related work. Section III briefly introduces convolutional neural network. Section IV describes the proposed CNN architecture. The experimental results and performance analysis is given in section V. Finally, section VI discusses the limitations and concludes this paper.

II. RELEVANT WORK

The number of papers dealing with video fire detection in literature is growing exponentially. Several researchers have played a significant role in the development of useful video fire detection algorithms [6]. Verstockt [1] proposed a multi-sensor fire detector which fuses visual and non-visual flame

features from moving objects. He used ordinary video and long wave infrared (LWIR) thermal images. First, he operates a dynamic background subtraction to extract moving objects. Also, LWIR moving objects are filtered by histogram-based hot object segmentation. A set of flame features analyze these moving objects with focus on distinctive geometric, temporal and spatial disorder characteristics of flame regions. Then, a LWIR flame probability is calculated by combining the probability of the bounding box disorder, the principal orientation disorder and the histogram roughness of the hot moving objects in LWIR. Similarly, the same calculus is operated on ordinary video to get a video flame probability. Finally, he combines both the LWIR and the video flame probability to give an indication about the presence of flames. Toreyin [2] uses a four steps video-based detection algorithm. First he estimated moving pixels and regions by using a hybrid background method: a three-frame differencing operation is performed to determine regions of legitimate motion, followed by adaptive background subtraction to extract the entire moving region. Second, he used a Gaussian mixture model in the RGB colour space to detect fire-coloured pixels. The fire-colour distribution is obtained from sample images containing fire regions. In the third step, a temporal wavelet transform is performed to analyse the flame flicker. Finally, a spatial wavelet analysis of moving regions containing fire mask pixels to evaluate colour variations in pixel values is performed. Significant spatial variations presuppose fire region. Celik [3]. developed two models: one for fire detection and the other for smoke detection. A rule-based fuzzy logic model was used instead of existing heuristic rules. This choice made the classification more robust in effectively discriminating fire and fire like coloured objects. For smoke detection, a statistical analysis was carried out based on the idea that smoke shows grayish colour with different illumination. Borges [4] used a multidimensional features vector as input to a Bayes classifier. Features are: the boundary roughness of the potential fire regions, the third order statistical moment of the potential fire regions which defines the skewness, the variance, and finally, the amount of fire from frame to frame (varies because of flame flickering). There are more recent works inspired by previous related research [5], [6]. As said before, all these methods are based-rule or require the build of discriminative features to detect fire. A further solution would be the use of deep learning algorithm. In the next section we introduce a variant of deep learning: Convolutional Neural Network.

III. CONVOLUTIONAL NEURAL NETWORK

CNN were first introduced by Fukushima [8], he derived a hierarchical neural network architecture inspired by Hubel's research work [9]. Lecun [10] generalized them to classify digits successfully and to recognizing hand-written check numbers by LeNet-5 which is shown in Fig. 1. Ciresan [11], used CNN and realized the best performance in the literature for multiple object recognition for multiple image databases: MNIST, NORB, HWDB1.0, CIFAR10 and the ImageNet dataset.

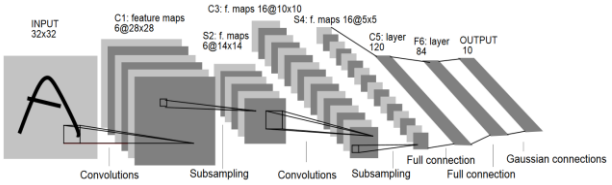

Fig. 1. LeNet-5, a Convolutional Neural Network for digits recognition

A convolutional Neural Network consists of several layers. Fig. 2, shows these different layers.

A. *Convolutional layers:*

It's the core building block of the CNN. These layers consist of a rectangular grid of neurons which have a small receptive field extended through the full depth of the input volume. Thus, the convolutional layer is just an image convolution of the previous layer, where the weights specify the convolution filter.

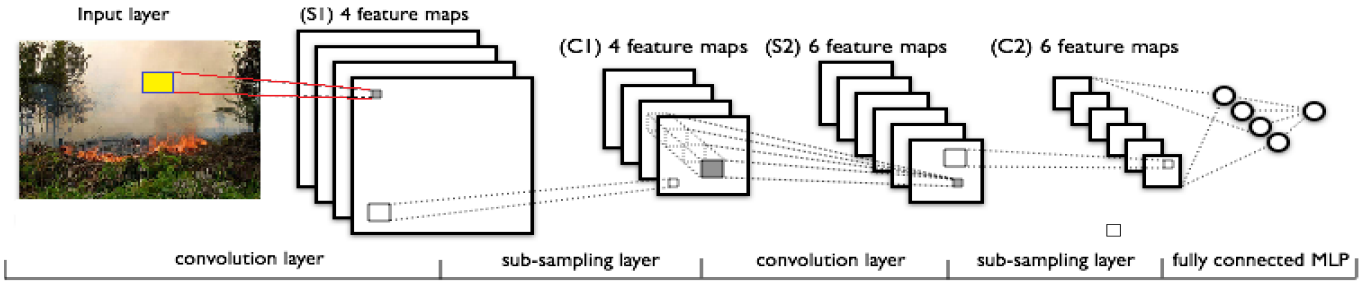

Fig. 2. CNN layers

B. *Pooling layers:*

After each convolutional layer, there may be a pooling layer. Pooling layers subsample their input. There are several ways to do this pooling, such as taking the average or the maximum, or a learned linear combination of the neurons in the block. For example, the Fig. 3. shows max pooling for a 2×2 window.

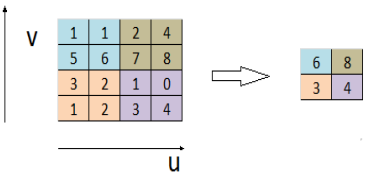

Fig. 3. Max pooling in CNN

C. *Fully connected layers:*

Finally, after several convolutional and max pooling layers, the high-level reasoning in the neural network is done via fully connected layers.

In convolutional neural networks every layer acts as a detection filter for the presence of specific features or patterns present in the original data. The first layers in a CNN detect features that can be recognized and interpreted relatively easy. Later layers detect increasingly features that are more abstract. The last layer of the CNN is able to make an ultra-specific classification by combining all the specific features detected by the previous layers in the input data. In the next section, the proposed CNN architecture for video fire and smoke detection is presented.

IV. CNN FOR VIDEO FIRE AND SMOKE DETECTION

A. *Structure*

Our classification architecture is classical for Convolutional Neural Network, combining convolution and Max pooling. However, to get a fast classification we choose a small network. Fig. 4 shows the nine layers CNN. An RGB color image goes through successively two convolutional operations

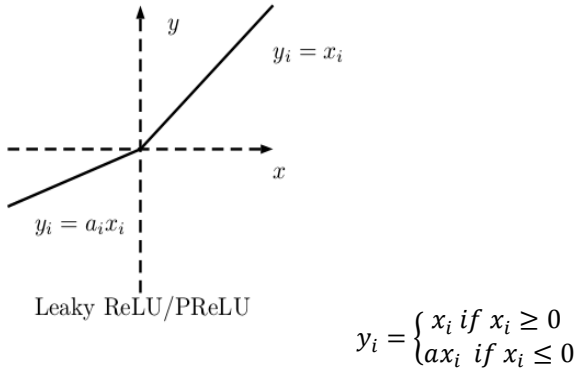

Fig. 5. Leaky ReLu

B. *Training*

The goal of our classification is to decide whether an image contains fire or/and smoke. To solve this problem, the classifier is learned from a collection of images labeled. Furthermore, we want to locate the position of the fire and smoke in an image or in a video. The training set is composed of 27919 RGB labeled images of size 64x64 pixels. 8915 for smoke, 7257 for fire and 11752 negative (no fire or smoke). We create 3 subsets: training 60% of images, Validation 20% and test 20%. The training data has been realized with a computer composed of a microprocessor Intel Xeon (frequency CPU 3,1Ghz, RAM 16Go) and a graphic card GTX 980 Ti (2816 cores, 6 GB memories). We used a stochastic gradient descent (SGD) with mini-batches of size 100. The weight in the network is initialized randomly. The initial learning rate is 0.01 and momentum 0.9. The learning rate decreases by a factor 0.95

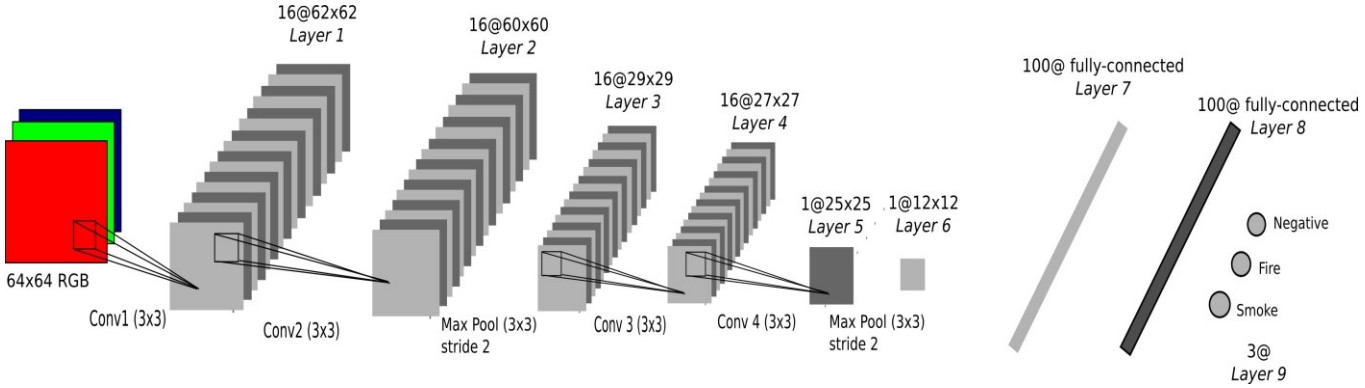

Fig. 4. CNN architecture

with kernel of size 3x3. The same structure is applied after layer three. A Max pooling 3x3 with stride 2 follows the convolutional layer two and five. The layers one to four have 16 feature maps. The layers five and six have only one feature map. The layer seven and eight are fully connected. The output of the last fully connected layer is fed to a 3 way Softmax which produces a distribution over 3 class labels.

Similarly to [12] and [13] we choose for convolutional and fully connected layers a Leaky ReLu activation function with coefficient a=1/3 (see Fig. 5).

each 5 epochs. By contrast, the momentum increases to reach 0.9999. To obtain the best accuracy for these parameters, several trials were done. We implemented CNN with Theano [14], [15] and Lasagne [16].

We use dropout of 0.5 in the two fully connected layers to avoid overfitting. We trained the network for roughly 100 cycles.

V. RESULTS

The classification accuracy on the test set is 97.9%. The test set is composed of 1427 fire images, 1758 smoke images and

2399 negative images, so 5584 images. Tables 1 to 3 give the confusion matrix for each class. On the fire confusion matrix, the false negatives and false positives do not contain smoke images. In the same way, the smoke confusion matrix do not contain fire image for false negatives and false positives. We can conclude that the parameters of our CNN model allow good classification distinction between fire and smoke. Furthermore, the surfaces under the ROC (Receiver Operating Characteristic) curve of Fig. 6, for the three classes are close to unity, pointing to a good classification on the test set. The fire ROC curve has a greater area than the others therefore indicating a better performance classification for the fire.

TABLE I. CONFUSION MATRIX FOR FIRE

Fire	**True class**		
		True	***False***
Hypothesis class	**True**	1400	3[a]
	False	27[a]	4154

[a]not smoke images

TABLE II. CONFUSION MATRIX FOR SMOKE

Smoke	**True class**		
		True	***False***
Hypothesis class	**True**	1698	26[a]
	False	60[a]	3800

[a]not fire images

TABLE III. CONFUSION MATRIX FOR NEGATIVE (NO FIRE/SMOKE)

No Fire/Smoke	**True class**		
		True	***False***
Hypothesis class	**True**	2370	87[a]
	False	29[b]	3098

[a]Image Fire 27 – image Smoke 60
[b]Image Fire 3 – image Smoke 26

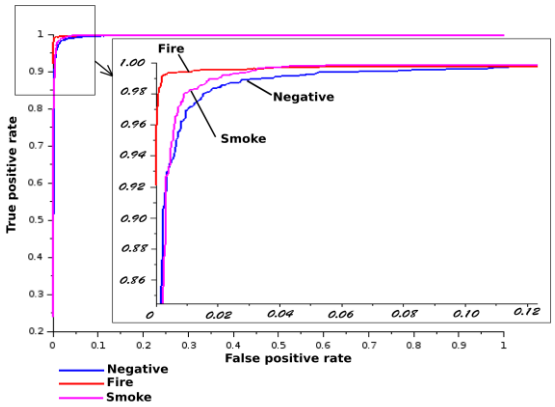

Fig. 6. ROC curves for the 3 classes: smoke, fire and negative

Our aim is to detect starting fire or characterize a fire on a video. The processing time for detection is a key factor with the accuracy. Therefore, we decide to use the "light structure" described in Fig. 4. The actual methods use sliding windows to detect and classify object on original or reshaped images. These windows go through the convolutional neural network and the fully connected layers to be classified. To analyze the entire image of a video frame, the window position must change and go again through convolutional neural network. Our approach is quite different, instead of sliding a window of 64x64 pixels in the image to locate the fire and the smoke; we decide to work on the last feature map. We divide the network in two parts. The first part is composed of 6 layers: 1 to 6 (convolutional and Max pooling layers), the second part is the fully connected layers Fig. 7 shows this CNN architecture.

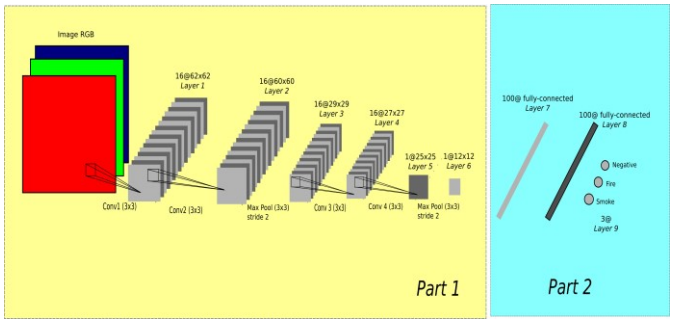

Fig. 7. Two parts CNN. Part 1: 6 layers (convolutional, maxpooling). Part 2: Two fully connected layers and the output layer.

Using the first part of the network, we evaluate the last feature map (layer 6) of the entire image. We know that based on the network structure of the CNN, a sliding window with size 64x64 pixels in the RGB image corresponds to a window size of 12x12 pixels in the last feature map Fig. 4.

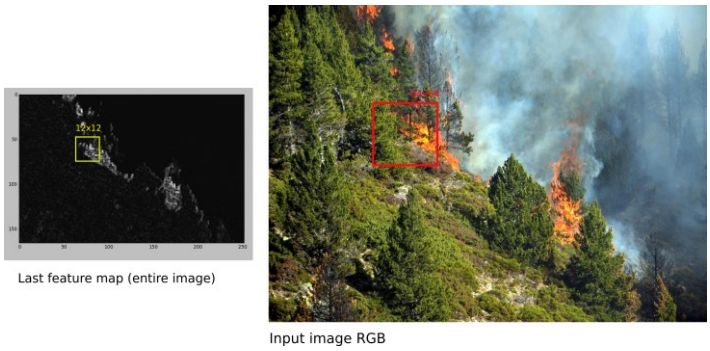

Fig. 8. Sliding window: Original image vs Feature map

To detect the fire and smoke in a frame video, we apply a sliding window of size 12x12 pixels on the last feature map (see Fig. 8). To speed up the prediction for each window 12x12, we realize a tensor 12x12x1xN (N : number of

windows) from the last feature map and we use the GPU of the graphic card. With this method, the accuracy seems unchanged and the speed of detection and prediction increase according of the original size image and the number of windows predicted as shown in Fig. 9. and Fig. 10.

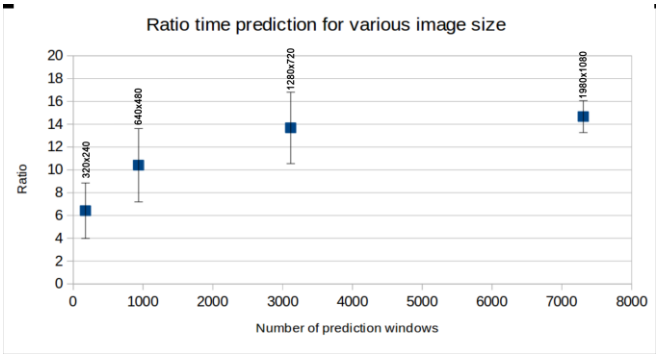

Fig. 9. Ratio time prediction for various image size.
Ratio= (time prediction for the entire original image)/(time prediction on the last feature map+time to make the last feature map). Sliding windows with step 16px on the original image and 4px on the last feature map. Processed over 200 frames on a video.

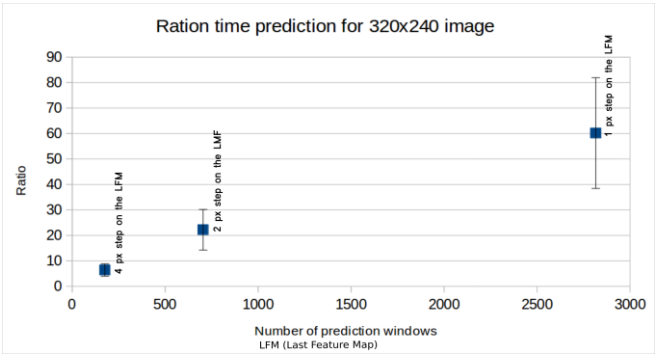

Fig. 10. Ratio time prediction for 320x240 image size and various sliding steps. Processed over 200 frames on a video.

The classification and localization result in Fig. 11.1 and 11.2, was done by sliding the feature map with windows 12x12 pixels and a step of 2 pixels. Fig. 9.1 (c) and Fig. 9.2.(c) describe the detection and localisation mask. Red colour represent the fire detected and green colour represent the smoke detected. The intensity of red and green colours varies with the probability of fire and/or smoke detection.Knowing the localization of fire and smoke in the feature map makes it possible to project these positions on the RGB original image (Rectangles red and green).

VI. CONCLUSION

In this paper, a vision-based method for fire and smoke detection was presented. The proposed algorithm uses a deep learning approach based on convolutional neural network (CNN). The confusion matrix and ROC curves indicates a very good overall accuracy for the detection stage. We showed that during the detection test, scaning directly the feature map instead of scanning the full original frame, could decrease the time cost to a ratio from 6 to 60.

In future work, we expect to improve the method by using a 3D convolutional neural network. Indeed, CNN are currently limited to handle 2D inputs which leads us to process the video input only frame by frame. Contrariwise, 3D CNN extracts features from both spatial and temporal dimensions by performing 3D convolutions. Thus the motion information of fire and smoke could be encoded, which makes it possible to decrease considerably the time cost. Moreover, to optimize the detection and localization of smoke and fire on a video, we must improve our training set. Smoke is more difficult to detect and localize due to the nature of his shape and texture. Our model detect only the red fire, to detect other color of fire, we have to increase our training set with others fire colors such blue one, etc ... In addition, we plan to compare our algorithm to conventional methods over a wider variety of video fire images: different material, sources and ventilations.

REFERENCES

[1] S. Verstockt, A. Vanoosthuyse, S. Van Hoecke, P. Lambert, and R. Van de Walle, Multi-sensor fire detection by fusing visual and non-visual flame features, In Proceedings of International Conference on Image and Signal Processing, June 2010, pp. 333 –341.

[2] B. U. Toreyin, Y. Dedeoglu,U. Gudukbay, A. E. Cetin, ¸Computer vision based method for real-time fire and flame detection,Pattern recognition letters, 2006, 27,1,pp. 49-58.

[3] T. Çelik, H. Özkaramanlı and H. Demirel, Fire and smoke detection without sensors: Image processing based approach, Signal Processing Conference, 2007 15th European, Poznan, 2007, pp. 1794-1798.

[4] K. Borges, P. Vinicius, J. Mayer and E. Izquierdo, Efficient visual fire detection applied for video retrieval, Signal Processing Conference, 2008 16th European, IEEE, 2008.

[5] K. Poobalan and S. Liew, Fire detection algorithm using image processing techniques, Proceedings of the 3rd International Conference on Artificial Intelligence and Computer Science (AICS2015), October 2015, pp. 160-168

[6] P. Gomes, P. Santana and J. Barata, A vision-based approach to fire detection, International Journal of Advanced Robotic Systems, 09-2014.

[7] E. Çetin et al, Video fire detection – Review, Digital Signal Processing, Volume 23, Issue 6, December 2013, pp. 1827-1843

[8] K. Fukushima, Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biological Cybernetics, 1980, 36(4),pp.193–202.

[9] D. H. Hubel and T. N. Wiesel, Ferrier lecture: Functional architecture of macaque monkey visual cortex, Proceedings of the Royal Society of London, Series B, Biological Sciences, 1977, 198(1130):pp. 1–59.

[10] Y. Lecun, L. Bottou, Y. Bengio and P. Haffner, Gradient-based learning applied to document recognition, in Proceedings of the IEEE, vol. 86, no. 11, pp. 2278-2324, Nov 1998.

[11] D. Ciresan, U. Meier; J. Masci; L.M. Gambardella and J. Schmidhuber, Flexible, High Performance Convolutional Neural Networks for Image Classification, Proceedings of the Twenty-Second international joint conference on Artificial Intelligence-Volume Volume Two **2**: pp. 1237–1242, *November 2013*.

[12] X. Bing, N. Wang, T. Chen and M. Li, Empirical Evaluation of Rectified Activations in Convolutional Network, CoRR abs/1505.00853 (2015): n. pag.

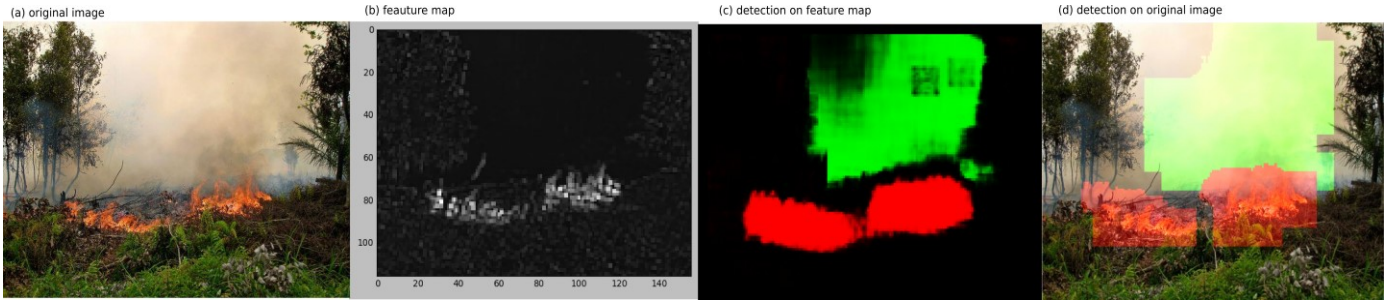

Fig. 11.1. Fire/smoke detection on forest image

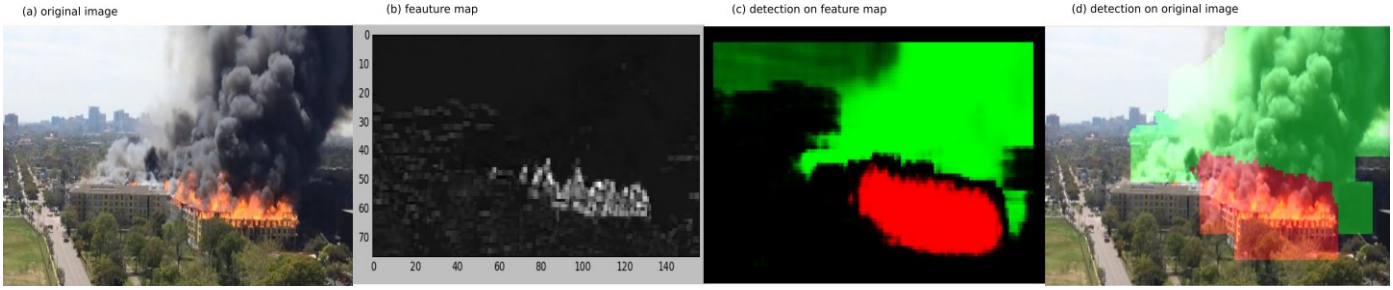

Fig. 11.2. Fire/smoke detection on building image

[13] A. L. Maas, A. Y. Hannun and A. Y. Ng, Rectify nonlinearities improve neural network acoustic model, ICML 2013 Workshop on Deep Learning for Audio, Speech, and Language Processing, June 2013, Atlanta

[14] F. Bastien, P. Lamblin, R. Pascanu, J. Bergstra, I. Goodfellow, A. Bergeron, N. Bouchard, D. Warde-Farley and Y. Bengio. "Theano: new features and speed improvements". NIPS 2012 deep learning workshop.

[15] J. Bergstra, O. Breuleux, F. Bastien, P. Lamblin, R. Pascanu, G. Desjardins, J. Turian, D. Warde-Farley and Y. Bengio. "Theano: A CPU and GPU Math Expression Compiler". Proceedings of the Python for Scientific Computing Conference (SciPy) 2010. June 30 - July 3, Austin

[16] LASAGNE, Lightweight library to build and train neural networks in Theano, https://github.com/Lasagne/Lasagne, 13 August 2015.