EasyChair Preprint

№ 4306

Video Based Fire Detection Using Xception and ConvLSTM

Tanmay Verlekar and Alexandre Bernardino

October 1, 2020

Video based fire detection using Xception and Conv-LSTM

Tanmay T. Verlekar[1] and Alexandre Bernardino[1]

[1] ISR - Instituto de Sistemas e Robótica, Av. Rovisco Pais 1, 1049-001 Lisboa, Portugal
tverlekar@isr.tecnico.ulisboa.pt

Abstract. Immediate detection of wildfires can aid firefighters in saving lives. The research community has invested a lot of their efforts in detecting fires using vision-based systems, due to their ability to monitor vast open spaces. Most of the state-of-the-art vision-based fire detection systems operate on individual images, limiting them to only spatial features. This paper presents a novel system that explores the spatio-temporal information available within a video sequence to perform classification of a scene into fire or non-fire category. The system, in its initial step, selects 15 key frames from an input video sequence. The frame selection step allows the system to capture the entire movement available in a video sequence regardless of the duration. The spatio-temporal information among those frames can then be captured using a deep convolutional neural network (CNN) called Xception, which is pre-trained on the ImageNet, and a convolutional long short term memory network (ConvLSTM). The system is evaluated on a challenging new dataset, presented in this paper, containing 70 fire and 70 non-fire sequences. The dataset contains aerial shots of fire and fire-like sequences, such as fog, sunrise and bright flashing objects, captured using a dynamic/moving camera for an average duration of 13 sec. The classification accuracy of 95.83% highlights the effectiveness of the proposed system in tackling such challenging scenarios.

Keywords: Fire detection, Video processing, Deep learning.

1 Introduction

A series of wildfires erupting across a country can result in a large number of deaths, injuries and destruction of properties. In recent years, an increase in heat waves, droughts, climate variabilities and changes in regional weather patterns has dramatically increased the risk of wildfires. Human activities, demographics, territorial and forest management changes have also contributed to this increase. A large number of firefighters risk their lives to mitigate the destruction caused by such fires. Thus, immediate detection of wildfires can play a significant role in the response of the firefighters in combating and controlling its spread.

Conventional systems for fire detection rely on sensors that detect an increase in temperature or smoke to trigger an alarm [1]. Such systems are designed to operate in closed environments, where sufficient heat or particles can reach their sensors. Since

closed proximity to fire and smoke is not possible in open spaces, the conventional sensor-based systems are ineffective in tackling wildfires. Conversely, the ability of vision-based systems to monitor vast open spaces has allowed the research community to develop several fire detection systems using a simple 2D camera [2].

Traditional vision-based fire detection systems rely on colour cues [3], [4] and image contours [5] to classify an image into a fire or non-fire category. The use of static features allows such systems to perform successful classification, only when the flames are prominently displayed in the image. The classification accuracy of the vision-based fire detection systems can be improved by using dynamic features such as motion [6] or dynamic textures [7] obtained from a video sequence. The best results among such systems are obtained by combining all available features into a multi-feature fusion system [8].

The field of computer vision has seen a significant improvement in its classification ability with the introduction of deep convolutional neural networks (CNN) [9]. Some of the state-of-the-art vision-based fire detection systems fine-tune popular CNNs, such as VGG16 and Resnet50 [10] to classify an image into fire or non-fire category. Such systems perform significantly better than the traditional vision-based fire detection systems, detecting fire even in small areas of an image while also being robust against objects having color or intensity similar to a fire in the scene. Some novel CNN based architectures, such as the densely dilated convolutional network, designed specifically to perform fire detection, perform even better than the fine-tuned CNNs, while also being lighter than them [11]. Apart from detecting, some CNNs allow segmenting the fire in an image using a bounding box [12] or pixel precision [13]. Other state-of-the-art systems rely on spatio-temporal information available from the entire video sequence to perform classification. The system presented in [14] stacks 64 frames to generate a tensor and uses it as an input to a deep convolutional generative adversarial neural network. The system presented in [15] uses a CNN to detect spatial features within a frame and then accumulates them across a video sequence using a Long Short Term Memory (LSTM) network. Although effective, these systems rely on a fixed duration video sequence for their input. They also process every available frame, which may contain a lot of redundant information, making them computationally expensive. And finally, they rely on a static camera setup to produce reliable results.

This paper presents a novel system to classify a video sequence into a fire or non-fire category using a CNN called Xception and a convolutional long short term memory network (ConvLSTM). It also presents a frame selection step which allows the system to process video sequences of varying duration. The system is evaluated on a challenging new dataset containing 70 fire and 70 non-fire sequences. The dataset contains video sequences of fog, intense sunshine, very small areas under fire and others, captured using a dynamic/moving camera. The proposed architecture allows the system to correctly classify sequences captured under such challenging scenarios.

2 Proposed System

The proposed system performs binary classification of an input video sequence into fire or non-fire category. The system operates in 4 steps, as illustrated by the system architecture presented in Fig. 1.

- During the first step the proposed system selects 15 key frames from an input video sequence;
- Each frame is processed using a CNN called Xception, pre-trained on the ImageNet dataset [9] to obtain static features;
- Spatio-temporal features across frames can then be obtained using a ConvLSTM;
- Final classification is performed using a fully connected (FC) network.

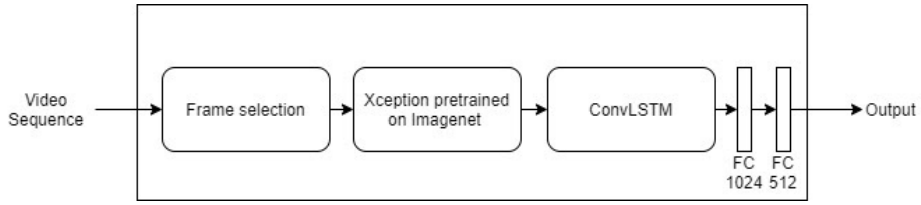


Fig. 1. Proposed System architecture.

2.1 Frame Selection

Due to advancements in camera technology, even a simple inexpensive camera can capture 30 frames per second. A high frame rate can capture a lot of redundant information, especially when observing fire sequences. To avoid processing similar looking frames, the proposed system automatically selects 15 frames uniformly distributed over the entire video sequence. The frame selection step allows the system to:

- Avoid all the starting/ending frames where the video might not contain any significant information;
- Capture the entire movement available in a video sequence;
- Operate on video sequences with varying duration.

However, in its current implementation the proposed system operates under the following assumptions for an input sequence:

- At least one frame is selected for each second;
- Positive sequences contain fire in majority of their frames, while negatives sequences are without fires;
- In dynamic sequences the camera movement isn't abrupt.

The selected frames are then resized to 299×299 pixels, so that they can be used as input to Xception.

2.2 Xception

Xception is a CNN architecture based entirely on depthwise separable convolution layers as illustrated in Fig. 2. The network consists of repeated pointwise convolution followed by a depthwise convolution [9]. A pointwise convolution is a 1×1 convolution used to change input dimensions, and a depthwise convolution is a channel-wise n×n spatial convolution. The two types of convolutions reduce the number of connections in Xception, making it lighter than most other CNNs. Xception is also one of the best CNNs in classifying the ImageNet [9]. ImageNet is a dataset of over 15 million labeled high-resolution images with around 22,000 categories. Xception uses a subset of ImageNet of 1000 categories with roughly 1.3 million training images, 50,000 validation images and 100,000 testing images to provide a classification accuracy of 79%. Thus, the proposed system uses Xception, pre-trained on ImageNet, to obtain a 2048-dimensional feature vector from its final convolutional layer. The feature vector is used as an input to the ConvLSTM in the following step.

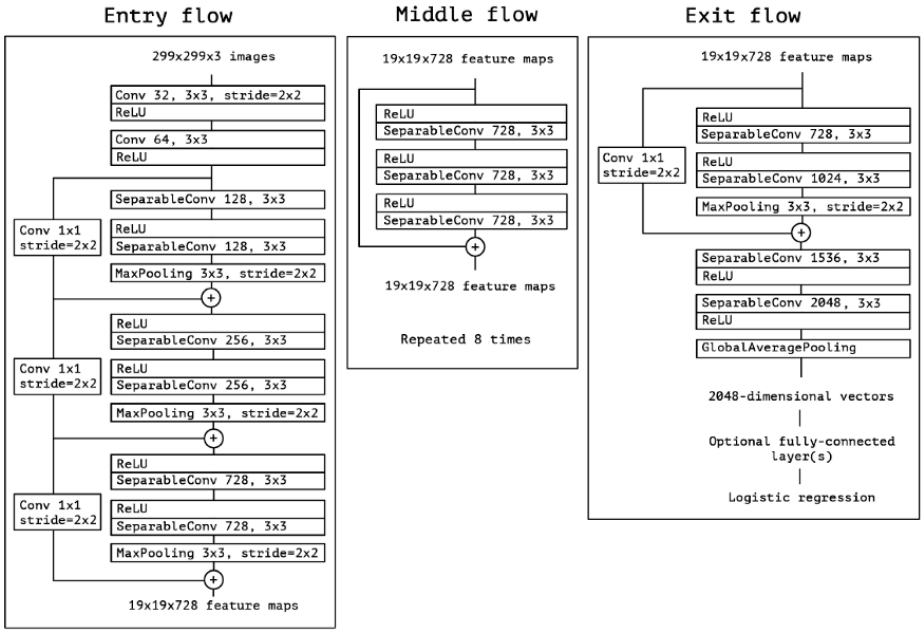


Fig. 2. The Xception architecture [9].

2.3 ConvLSTM

LSTM is a deep recurrent neural network (RNN) explicitly designed to remember information for long periods of time [16]. It is configured as a chain of repeating cells connected to each other using a cell state (C), as illustrated in Fig. 3. Each cell has four neural network layers interacting with each other to decide on what information must be discarded from the cell state, what new information must be added to the cell state,

and the output of each cell. A ConvLSTM is a LSTM architecture specifically designed for sequence prediction problems with spatial inputs, like images or video sequences. It operates similar to a LSTM, but the internal matrix multiplications are replaced with convolution operations, illustrated with red colour in Fig. 3. As a result, the information flowing through the ConvLSTM cells maintains the input dimension, allowing the network to obtain better spatio-temporal correlations [16].
The proposed system uses a single layer of ConvLSTM, with 15 ConvLSTM cells. The number of output filters in the convolution is set to 64, with a kernel size of 7×7 and strides of 2×2.

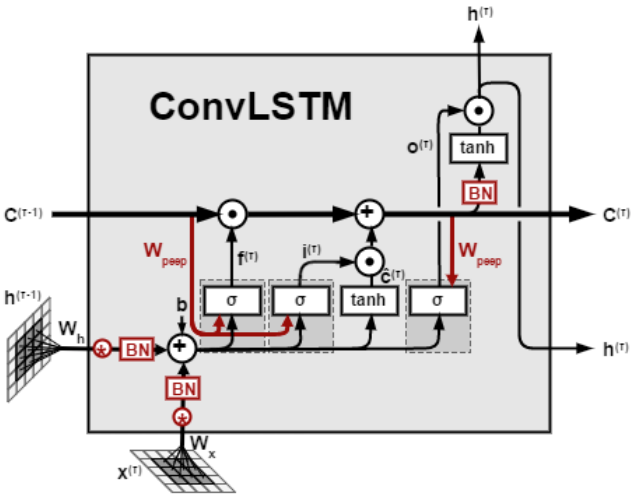


Fig. 3. A ConvLSTM cell [17].

2.4 FC network

The output obtained from the final ConvLSTM cell is used as an input to the FC network. The FC network of the proposed system consists of two FC layers of dimensions 1024 and 512 respectively, with a dropout of 0.2 between them to prevent overfitting. It also contains a SoftMax layer to perform classification.

3 Experimental Results

The scarceness of publicly available video-based fire detection datasets prevents a thorough evaluation of most state-of-the-art vision-based fire detection systems. The largest video dataset currently available (to our best knowledge) is MIVIA [18] with only 14 fire and 17 non-fire video sequences.

3.1 ISR fire video dataset

To evaluate the proposed system a novel dataset is collected containing 70 fire and 70 non-fire video sequences, called the ISR fire video dataset. The video sequences for the dataset are acquired by segmenting videos from You-tube. To make the classification process challenging, and effective in tackling wildfires, the dataset is populated with sequences such as - see Fig. 4:

- Aerial shots of trees, houses and fields on fire;
- Flames occluded by heavy smoke;
- Small section of an area under fire.
- Clouded sky, fog covering an area or smoke in the absence of a fire;
- Sunrise and sunsets;
- Cars flashing their headlight and other bright red fire-like objects.

A small portion of the dataset (37%) contains dynamic shots of the scene distributed evenly across the two categories. Such sequences are significantly more challenging to classify as they lack a static background across frames. The dynamic shots in the dataset include a gradual zoom in, zoom out, panning and forward movement of the camera, as illustrated in Fig. 4. (c). All the sequences are captured at 30 fps with a mean duration of 13 sec. Since the proposed system resizes the frames to 299×299 pixels, the dataset includes sequences ranging from 400×256 to 1920×1080 pixels.

(a)

(b)

(c)

Fig. 4. Dataset samples, (a) fire, (b) non-fire, (c) dynamic shot.

3.2 Evaluation

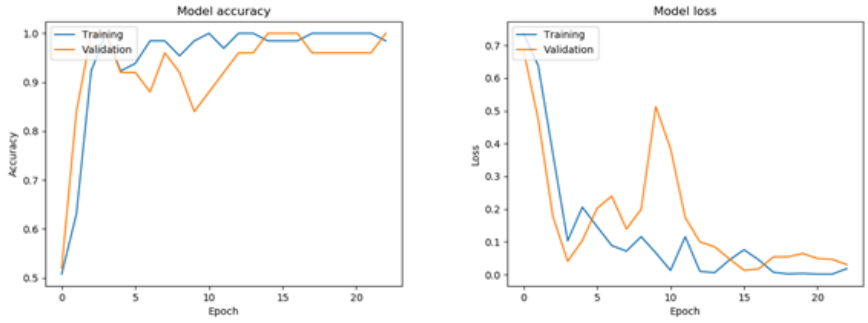
To evaluate the proposed system the ISR fire video dataset is randomly split into 3 sets, such that each set contains equal number of fire and non-fire sequences. Out of the available 140 sequences, 44 sequences are selected for testing, 68 sequences are selected for training and remaining 28 sequences are selected for validation. The training set is augmented using shift, zoom and rotate to further increase the size of the training set to 204 sequence. The proposed system is trained using an Intel® Core™ i7-9700 CPU with GeForce RTX 2080 Ti. Training is performed using the cross-entropy loss function. The batch size is set to 5 and the number of epochs is set to 50, with the early stopping criteria set to monitor accuracy. To assess variance, the evaluation is repeated 3 times by randomly selecting new training, validation and test sets for each iteration. The results of the evaluation are reported in Table 1.

Table 1. Classification accuracy of the proposed system (%).

Systems	**Train**	**Validation**	**Test**
Proposed (Xception+ConvLSTM)	100.0±0.0	98.7±1.9	95.8±1.2

Xception+LSTM	100.0±0.0	94.7±4.9	90.0±0.0
VGG16+ConvLSTM	100.0±0.0	90.7±4.9	85.0±0.0
VGG16+LSTM [15]	98.3±0.2	89.3±1.9	78.3±3.1

The classification accuracy and the corresponding loss of the proposed system is illustrated in Fig. 5. From the figure it can be seen that the classification accuracy of the system over the training set is at 100%, suggesting that the proposed system is effective in learning to classify fire and non-fire sequences. The validation accuracy of the proposed system reduces marginally across the three iterations with a mean score of 98.66%. The ability of the system to classify new sequences can be inferred from its performance over the test set. As reported in Table 1, the proposed system performs remarkably with a mean classification accuracy of 95.83%. The classification accuracy is significantly better than the VGG16+LSTM architecture, employed by the state-of-the-art systems, such as [15]. The classification accuracy improves with the use of Xception in place of VGG16 and ConvLSTM in place of LSTM, with the proposed system providing the best results. Thus, the quality of features obtained using Xception can be considered better than VGG16. While the improvement in performance using ConvLSTM suggests that there exists a strong spatio-temporal correlation between features obtained from sequential frames.
It should also be noted that the state-of-the-art system presented in [14] uses all available frames from a video sequence. Thus, for a 15 sec video sequence captured at 30 fps, it will process 450 frames. The proposed system provides equivalent results using just 15 frames, making it computationally inexpensive.



(a) (b)

Fig. 5. Plots representing (a) accuracy and (b) loss.

4 Conclusion

This paper presents a novel system that explores the spatio-temporal information available in a video sequence to perform fire detection. The proposed system can operate on video sequences of varying duration using the frame selection step. The spatial features

are obtained using a popular CNN called Xception pre-trained on ImageNet. It then uses a ConvLSTM which performs significantly better than most RNNs in obtaining spatio-temporal correlations between frames. The system operates on a challenging dataset presented in this paper containing 70 fire and 70 non-fire video sequences captured using a dynamic camera. The results suggest that the proposed system is effective in classifying fire and non-fire sequences in challenging scenarios. Thus, the proposed system can be mounted on drones to aid firefighters in detecting wildfires.
One limiting factor of this paper that can be improved further, is the size of the dataset used. As a future work, the dataset can be populated with more dynamic shots of fire and non-fire scenes, along with more difficult to classify sequences. The system can be improved to provide frame by frame decisions, while exploring the spatio-temporal information available in a video sequence.

5 Acknowledgement

This work was supported by FCT with the LARSyS - FCT Project UIDB/50009/202 and project FIREFRONT (PCIF/SSI/0096/2017).

References

1. Fonollosa, J., Solórzano, A., Santiago, M.: Chemical sensor systems and associated algorithms for fire detection: A review. Sensors, 18(2), 553, (2018).
2. Bu, F., Gharajeh, M.: Intelligent and vision-based fire detection systems: A survey. Image and Vision Computing, 91, (2019).
3. Chen, T., Wu, P., Chiou, Y.: An early fire-detection method based on image processing. In: International Conference on Image Processing, (2004).
4. Seebamrungsat, J., Suphachai, P., Riyamongkol, P.: Fire detection in the buildings using image processing. In: Third ICT International Student Project Conference, (2014).
5. Poobalan, K., Liew, S.: Fire detection algorithm using image processing techniques. In: International Conference on Artificial Intelligence and Computer Science, (2015).
6. Chunyu, Y., Jun, F., Jinjun, W., Yongming, Z.: Video fire smoke detection using motion and color features. Fire technology, 46(3), 651-663, (2010).
7. Ye, W., Zhao, J., Wang, S., Wang, Y., Zhang, D., Yuan, Z.: Dynamic texture-based smoke detection using Surfacelet transform and HMT model. Fire Safety Journal, 73, 91-101, (2015).
8. Gong, F., Li, C., Gong, W., Li, X., Yuan, X., Ma, Y., Song, T.: A real-time fire detection method from video with multifeature fusion. Computational intelligence and neuroscience, (2019).
9. Chollet, F.: Xception: Deep learning with depthwise separable convolutions. In: IEEE conference on computer vision and pattern recognition, (2017).
10. Sharma, J., Granmo, O., Goodwin, M., Fidje, J.: Deep convolutional neural networks for fire detection in images. In: International Conference on Engineering Applications of Neural Networks, (2017).
11. Li, T., Zhao, E., Zhang, J., Hu, C.: Detection of Wildfire Smoke Images Based on a Densely Dilated Convolutional Network. Electronics, 8(10), (2019).

12. Kang, L., Wang, I., Chou, K., Chen, S., Chang, C.: Image-Based Real-Time Fire Detection using Deep Learning with Data Augmentation for Vision-Based Surveillance Applications. In: IEEE International Conference on Advanced Video and Signal Based Surveillance, (2019).
13. Xu, Z., Wanguo, W., Xinrui, L., Bin, L., Yuan T.: Flame and Smoke Detection in Substation Based on Wavelet Analysis and Convolution Neural Network. In: 3rd International Conference on Innovation in Artificial Intelligence, (2019).
14. Aslan, S., Güdükbay, U., Töreyin, B., Çetin A.: Deep convolutional generative adversarial networks-based flame detection in video. arXiv preprint arXiv:1902.01824, (2019).
15. Kim, B., Lee J.: A Video-Based Fire Detection Using Deep Learning Models. Applied Sciences 9(14), (2019).
16. Xingjian, S., Chen, Z., Wang, H., Yeung, D., Wong, W., Woo, W.: Convolutional LSTM network: A machine learning approach for precipitation nowcasting. Advances in neural information processing systems, 802-810, (2015).
17. https://medium.com/neuronio/an-introduction-to-convlstm-55c9025563a7
18. https://mivia.unisa.it/datasets/video-analysis-datasets/fire-detection-dataset/