Temporal Deep Learning for Video Smoke Detection

Andreas, Wellhausen

Bosch GmbH, Hildesheim, Germany
University of Duisburg Essen, Duisburg, Germany

Abstract

In the proposed paper, temporal approaches based on Deep Learning are applied to Video Smoke Detection. Two Deep Learning methods are investigated: Firstly, a combination of convolutional neural networks (CNN) and long-short-term-memory networks (LSTM), secondly the inflated 3D architecture (i3D), which consists of 3D convolutions. These are two state of the art approaches to extract spatial and temporal information out of video sequences. These temporal approaches are compared to a single frame CNN method. Based on these information, a new way to detect and localize smoke within such sequences is presented, called cell-wise classification. It is shown that the temporal approaches work better for smoke than single frame approaches and that the cell-wise classification approach is suitable for Video Smoke Detection. A dataset consisting of about 350 hours of sequences is utilized. This is the first large scale dataset for the purpose of Video Smoke Detection, on which temporal Deep Learning methods are applied.

Keywords: Video Smoke Detection, Deep Learning, CNN, LSTM, i3D

Introduction

Video Smoke Detection (VSD) is a promising solution to detect fires in buildings with high ceilings (e.g., factories, warehouses, train stations, tunnels, etc.) or outdoor areas (e.g., landing stripes, harbors, pedestrian areas, etc.). VSD algorithms can detect smoke very fast and prevent higher human or property damage. These benefits are the reason why an increasing amount of research groups and companies are aiming to develop reliable algorithms for VSD.

In classic approaches, physical or visual characteristics of smoke are identified and extracted by ordinary Computer Vision algorithms to distinct smoke from non smoke events. A comprehensive overview is given in [1]. Such characteristics are called features, which transform the problem of VSD in an easier to handle lower-dimensional domain. In the
best case, smoke is separable from non smoke events by a simple decision function in this domain, which is usually yielded by Machine Learning approaches. Typical smoke characteristics for feature development are color [2], static texture [3], dynamic textures [4], moving and growing behavior [5].

Due to the success Deep Learning in object detection, neural networks are more and more applied to VSD in research. Such Deep Learning methods require databases of smoke and non smoke samples for development and comparison.

Temporal information is very characteristic for visual smoke appearance. Therefore Deep Learning-based VSD algorithms could benefit from the information of multiple frames. This requires that the underlying database consists of sequences of consecutive frames. In this paper a large scale dataset containing smoke and non smoke sequences is introduced, which is provided by the Bosch Building Technology GmbH.

For sequence classification, several Deep Learning architectures are available. The most common are a combination of CNN+LSTM and 3D convolutional architectures [6], which are investigated. This paper takes advantage of transfer learning and uses pretrained weights for initialization. As CNN Inception [8] is applied, for which ImageNet [7] pretrained weights are available. For 3D convolutions, only one pretrained public accessible architecture is known to the author, the i3D [6]. The i3D achieves state of the art results on Kinetics [9], which is the largest open human action recognition dataset consisting of sequences.

Deep Learning research for VSD started by observing single frame methods. It is successfully shown that smoke can be detected by using CNNs [10]. Different state of the art CNN architectures are compared like VGG, ResNet, Inception, and Xception. The experimental result is that Xception works best [11].

Temporal approaches are also investigated. [12] analyzes a two-stream approach using two networks one for RGB and one for optical flow. The resulting features are merged to classify smoke. A combination of CNN and a recurrent network is observed by [13] with promising results, but on small image crops only.

To overcome the issue of small datasets [14] enlarge their dataset by artificial smoke data and showed that this could increase the performance of single frame based smoke detection methods using Deep Learning.

All these approaches only predict, if there is smoke in the whole frame/sequence or within a crop. In this paper a new cell-wise classification approach is introduced, which is a straightforward extension of ordinary classification approaches.
The paper is structured as follows. Firstly, the large scale dataset of Bosch is proposed in Section 2. In Section 3 the cell-wise classification approach is introduced. Section 4 presents the investigated Deep Learning architectures. Afterward, the experiments how the architectures are trained is explained in Section 5. The evaluation concept and results are described in Section 6. In Section 7 the conclusion is finally given.

The Bosch Dataset

Bosch provides a large scale dataset of smoke, and non smoke sequences, in total 357 hours of video material. The recordings were taken in 82 different locations. At 56 of them, smoke experiments are conducted.

The smoke experiments are inspired by the tests, which are conducted to certify ordinary smoke detectors. This mainly concerns the combustible material, e.g., wood (open flame, smoldering), liquids (ethanol, gasoline, decalin), cotton and paper. These combustible materials are ignited by a flame, such that an open fire occurs, or by a hot plate so that only smoldering smoke is visible. Fig. 1 shows some examples.

![Fig. 1. Examples of smoke sequences within the Bosch dataset.](image1)

Experiments, according to standards, are costly. Therefore the most smoke experiments are done with different colored and sized smoke cartridges, to simulate the visual behavior of real smoke events.

Beyond smoke experiments many non smoke scenarios are recorded. Fig. 2 shows some examples for non smoke recordings.

![Fig. 2. Some non smoke sequences within the Bosch dataset.](image2)

These non smoke events can cause the VSD system to throw a false alarm. The whole dataset is split into events, which last from 30 s to 500 s: 1016 smoke and 1241 non smoke events. Each event is one sample to train or test the Deep Learning algorithms. Therefore the
events are divided into a train and a test set. To evaluate the generalization of the algorithms the locations of the train and test set are disjoint, i.e. locations of the training set must not occur in the test set. Table 1 gives an overview of the resulting sets.

Table 1: Distribution of events in dataset into train and test samples.

<table>
<thead>
<tr>
<th>Set</th>
<th>Smoke</th>
<th>Non Smoke</th>
<th>Locations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train</td>
<td>704</td>
<td>879</td>
<td>48</td>
</tr>
<tr>
<td>Test</td>
<td>312</td>
<td>362</td>
<td>37</td>
</tr>
</tbody>
</table>

In the training set all smoke plumes are labeled by bounding boxes. For the test events it is only known, if there is smoke in a sequence or not.

**Cell-wise Classification Approach**

The VSD problem differs from other object detection problems, because it is not necessary to fit an accurate bounding box to the smoke region. It is sufficient to classify at least one part of the smoke region for sure. The idea of the proposed concept is to map a binary label mask to each frame. To do so, each frame is covered by a rough grid of cells. If a cell intersects with the bounding box the binary label mask is set to 1 and 0 otherwise (see Fig. 3). The binary mask can be interpreted as the probability of smoke existence.

![Cell-wise classification labels](image)

Fig. 3. Cell-wise classification labels: A frame is separated into a grid. Each cell yields the label 1, if it intersects with a bounding box and 0 otherwise.

An algorithm has to predict the probability for smoke of each cell. All proposed Deep Learning architectures are designed, such that they extract spatial feature maps, where each feature vector in this map corresponds to a cell. These feature vectors are used to predict the smoke probability for each cell. To predict the smoke probability these feature vectors are fed into a classifier. Fig. 4 gives an overview of this concept.
Investigated Deep Learning Methods

Three different state of the art Deep Learning concepts are investigated as feature extractors within the cell-wise classification approach, a CNN, a CNN+LSTM and the i3D. Fig. 5 illustrates these approaches.

The CNN does not use any temporal information, each frame is considered independently without information about the past.

The CNN+LSTM approach first extracts a spatial feature map for each frame independently. Then, the feature vectors for each cell are fed into an LSTM, which combines spatial with temporal information from the last frames of this cell. In theory a LSTM can carry temporal information from an arbitrary past.

The i3D uses 3D convolutions and 3D pooling to extract spatial and temporal information of fixed time windows simultaneously.

Fig. 5. The arrows illustrate the information flow of the proposed models. The input is a sequence of RGB images. (a) The CNN model processes and predicts a feature map of smoke for each frame independently. (b) The CNN+LSTM model first extracts spatial information by CNN for each frame independently. This information is propagated to an LSTM, which adds previous and extracts new temporal information. Here each frame also yields a feature map. (c) The i3D approach combines spatial and temporal information directly. Due to temporal pooling, the first frames are ignored.
As CNN an Inception is chosen, which is initialized with weights pretrained on ImageNet. The i3D architecture is similar to the Inception, but the 2D convolutions and pooling layers are replaced by 3D counterparts. For the i3D, weights pretrained on Kinetics are utilized. The architectures are presented in Fig. 6.

Fig. 6. The Inception and the i3D architectures. (a) and (c) The repeatedly used 2D Inception block (2D Inc.) and 3D Inception (3D Inc.) block, resp. (b) The complete Inception architecture. (d) The complete i3D architecture.
For both architectures the activation function always is a rectified linear unit (ReLU). Between a convolutional layer and an activation function batch normalization is placed. The stride is of the form time, x, y. Only one value means that it is the same. If not mentioned, the stride is 1. To yield a feature map the average pooling of the original architectures is removed.

In the CNN+LSTM approach the appended LSTM has 512 hidden units. For all approaches CNN and CNN+LSTM a sigmoid layer receives the feature vector of each cell and is used as classifier to predict the smoke probability.

The exact number of channels for each layer can be found in the original references. The number of parameters are 5.59 Mio for CNN, 8.76 Mio for CNN+LSTM and 12.26 Mio for i3D.

**Experiments**

During training the loss function penalizes wrong predicted cells and the weights are determined, such that the loss is minimized. As loss function the mean binary cross-entropy over all cells within a batch is used. Let $y \in \{0,1\}$ be the label and $p \in [0,1]$ the prediction of the smoke probability for a cell, than the loss $l$ for a cell is defined as

$$l = y \log(p) + (1 - y) \log(1 - p).$$

(\text{Eq. 1})

Stochastic gradient descent with a momentum of 0.9 is used for weight update. For training and testing all frames are resized to a resolution of $256 \times 256$ and down sampled to 3 fps. Note the choice of such small resolution and framerate is rather unusual, but suitable compromise between computational effort and performance.

Each batch consists of 6 sequences with a length of 30 frames, which are randomly cropped from the train sequences each epoch. For the i3D approach only the last frame is predicted. The training runs for 150 epochs (one epoch is finished, when all training sequences are processed once to update the weights). The learning rate is decayed by a factor of 0.95 every epoch starting with 0.01.

Several augmentation methods are used to increase the data artificially: Of the $256 \times 256$ images a $224 \times 224$ subpart is randomly cropped. The frames are randomly rotated or flipped and finally a gaussian noise with $\mu = 0$ and $\sigma = 0.5$ is added to each channel of the input, which is normalized between -127.5 and 127.5. The implementation is done in Keras [15] and all experiments are conducted on a Nvidia TitanX.

**Evaluation and Results**

For testing the full length of each sequence is processed. The maximal smoke probability of each sequence is chosen to draw a receiver operating characteristic (ROC). The ROC monitors the true positive rate (TPR) against the false positive rate (FPR). Each point on a ROC can be
interpreted as the detection rate of smoke events, while accepting a certain false alarm rate within the non smoke sequences.

A suitable measure for comparison of the different approaches is the area under curve (AUC). The higher the AUC, the better the result. Fig. 7 shows the resulting ROCs for the investigated approaches. It is not surprising that the temporal approaches significantly outperform the single frame approach. The i3D is by far the best. One can conclude that Deep Learning architectures based on 3D convolutions are most suitable for VSD.

The i3D detects 76 % of the smoke sequences without any false alarm. The 24 % not detected smoke sequences are usually sequences with less motion within the smoke region, low smoke density or small smoke plumes. The most challenging non smoke sequences are very slow moving shadows caused by cranes or garage doors. The algorithm is surprisingly robust against snow, rain or moving clouds. The cell-wise classification approach is suitable for different sized smoke events.

Fig. 7. ROC curves of all investigated approaches including the resulting AUC.

To get a qualitative impression of the i3D performance Fig. 8 shows some examples, where the alarm threshold is adjusted, such that there is no false alarm within the test set. These are typical application fields for VSD: Stadium, Tunnel, Airplane-Hangar, Outdoor and Wildfire. The i3D generalizes to these locations, even it has not seen them during training.
Fig. 8. Examples of i3D performance on test sequences. Yellow boxes are cells with predicted smoke probability > 0.8 and red cells with predicted smoke probability > 0.95. This is the threshold, where the i3D does not have any false alarm on the test set.

Conclusion

In the proposed paper a complete VSD system based on temporal Deep Learning approaches is trained and evaluated on a large scale dataset consisting of more than 2000 smoke and non smoke sequences between 30 s and 500 s.

It is shown that temporal information increases the performance of smoke detection compared to single frame approaches, whereas the i3D performs best. Furthermore, it is shown that a custom cell-wise classification approach is suitable to detect smoke events of different size.

References


