Computationally Efficient Spatio-Temporal Dynamic Texture Recognition for Volatile Organic Compound (VOC) Leakage Detection in Industrial Plants

Diaa Badawi, Hongyi Pan, Sinan Cem Cetin, and A. Enis Çetin, *Fellow, IEEE*

***Abstract*—In this article, we present a computationally efficient algorithm to detect Volatile Organic Compounds (VOC) leaking out of components used in chemical processes in petrochemical refineries and chemical plants. A leaking VOC plume from a damaged component appears as a dynamic dark cloud in infrared videos. We describe a two-stage deep neural network structure, taking advantage of both spatial and temporal structure of the dynamic texture regions created by the leaking VOC plume. We first detect moving pixels which are darker then their neighboring pixels. We extract one-dimensional (1-D) signals representing the temporal history of such pixels from video and feed the 1-D signals to a 1-D convolutional neural network. If those pixels are near the edge of a VOC plume, their 1-D temporal signals exhibit high-frequency behavior. The neural network generates high probability estimates for such pixels. If 1-D neural network generates high confidence values, final decision is reached using a deep convolutional neural network (CNN) which processes image frames. The overall structure is computationally efficient because the spatio-temporal CNN does not process all of the image frames of the captured video. Experimental results are presented.**

***Index Terms*—VOC plume detection, IR video, CNN, time-series.**

I. Introduction

THE US Environmental Protection Agency (EPA) estimates that more than 70,000 tons of Volatile Organic Compounds (VOC) are emitted from leaking equipment, such as valves, pumps, and connectors, at petroleum refineries and and chemical manufacturing facilities annually [1]. Some types of VOCs such as acetaldehyde, benzene, formaldehyde, methylene chloride, naphthalene, toluene, and xylene are Volatile Hazardous Air Pollutants (VHAPs), which cause cancer, birth defects and reproductive effects. VOCs also contribute to the formation of ozone, which is a major source of smog, and one of the main causes of respiratory diseases in urban areas and areas close to refineries and chemical plants [2].

Manuscript received July 1, 2019; revised December 28, 2019 and February 14, 2020; accepted February 17, 2020. Date of publication February 27, 2020; date of current version August 10, 2020. This work was supported in part by NSF Grant 1739396 (UIC). This paper was presented in part at the IEEE International Conference on Acoustics, Signal, and Speech Processing, Brighton, U.K., May 12, 2019–May 17, 2019. The associate editor coordinating the review of this manuscript and approving it for publication was Mr. Werner Bailer. *(Corresponding author: Diaa Badawi.)*

Diaa Badawi, Hongyi Pan, and A. Enis Çetin are with the Department of Electrical and Computer Engineering, University of Illinois at Chicago, Chicago, IL 60607 USA (e-mail: dbadaw2@uic.edu; hpan21@uic.edu; aecyy@uic.edu).

Sinan Cem Cetin is with the University of California, Los Angeles, Los Angeles, CA 90024 USA (e-mail: akasinan@ucla.edu).

Digital Object Identifier 10.1109/JSTSP.2020.2976555

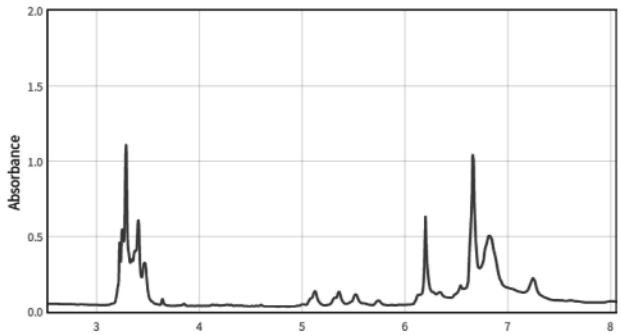


Fig. 1. Toluene absorbance as a function of the wavelength in infrared range. The scale of the wavelengths (x-axis) is in micrometers. The plot is downloaded from [3].

Most VOCs absorb infrared light in Medium Wave Infrared (MWIR) or Long Wave Infrared (LWIR) bands. For example, toluene absorbs infrared (IR) light in both MWIR and LWIR bands as shown in Fig. 1.

As a result a MWIR or LWIR thermal camera can image leaking VOC plumes from a faulty component [4]–[6]. In Fig. 2(a), an image frame from a thermal camera is shown. The leaking VOC region is darker than neighboring regions because the VOC absorbs IR light. As a result, MWIR and LWIR thermal cameras make VOC leaks visible which are normally invisible with a regular camera as shown in Fig. 2(b). In Fig. 3(a) and (b) two image frames containing VOC leaks are shown. In these image frames it is not possible to identify the leak by examining the location of the leak. However, leak locations can be easily spotted when we watch the video clips. This is because VOC leak regions are not stationary in the thermal IR video. They move in an erratic manner due to wind and/or other factors. This is demonstrated by the frame sequence shown in Fig. 4. Therefore, a computer vision algorithm should use both the spatial and temporal information to detect VOC leak regions. The VOC gas leak detection problem in infrared video is similar to related to wildfire smoke detection problem [5], [7]–[13]. Smoke and flames are also dynamic textures in the video [14]–[19]. Ideally, a neural network taking a cube of video data should be trained for VOC leakage detection. However, such an approach would not be computationally efficient and it would not be possible to implement such a system in real-time.

In this paper, we design two types of neural networks for VOC detection. We implement the neural networks in two stages one

(a)

(b)

Fig. 2. Example IR thermal image (a) and a corresponding ordinary camera image (b) for the same scene. As we can see, the VOC leak is not visible in the case of visible light image. Images (a) and (b) are taken from [20].

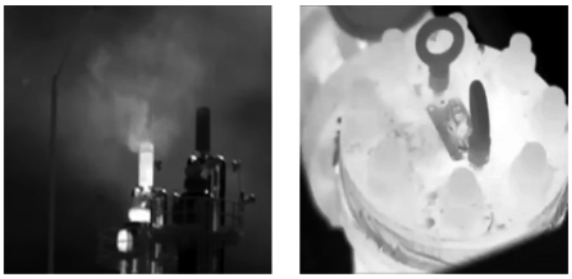


Fig. 3. Example IR-thermal frames of VOC leaks. As we can see, it is not easy to figure out the VOC leak from a single image frame.

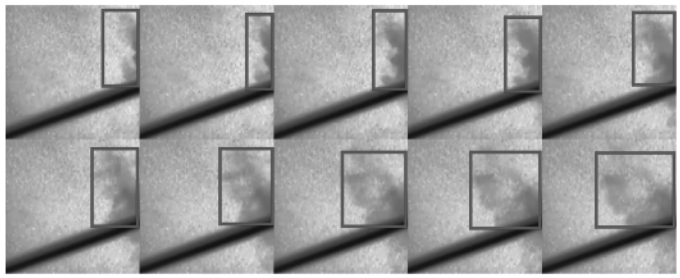


Fig. 4. Example IR-thermal frame sequence of VOC leaks.

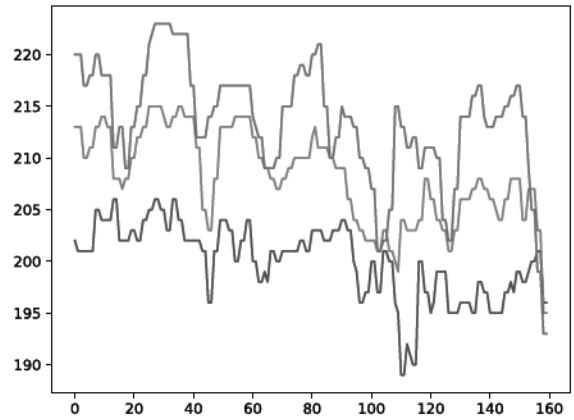


Fig. 5. Time-series data of three pixels in VOC leakage regions in IR video.

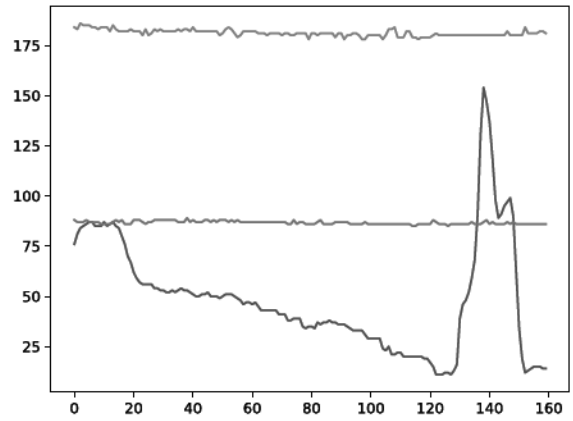


Fig. 6. Time-series data of three pixels in thermal IR video. The time-series shown in blue corresponds to a moving object.

after another by taking advantage of both spatial and temporal structure of the dynamic texture created by the leaking VOC plume. We first detect moving pixels which are darker then some of their neighboring pixels. These pixels may be at the boundary of a VOC plume. We extract one-dimensional (1-D) signals representing the temporal (history) signals of such pixels from video and feed these 1-D signals to the neural network. If those pixels are near the edge of a VOC plume their 1-D temporal signals exhibit high-frequency behavior because VOC clouds exhibit erratic movements similar to ordinary smoke. A typical 1-D signal corresponding to a pixel at the edges (or near the edges) of a VOC plume is shown in Fig. 5. On the other hand regular background pixels have pretty stationary behaviors as shown in Fig. 6 and motion patterns of ordinary moving objects are different from the VOC gas leak pixels shown in Fig. 5. The neural network is trained in such a way that it generates high probability estimates for such pixels. If the 1-D neural network generates high confidence values, we feed the corresponding video frame to a deep convolutional neural network (CNN). The final decision is reached using the CNN which processes image frames. The overall structure is computationally efficient because the CNN does not process all of the image frames of the captured video. It only processes data after a suspicious activity is detected by the 1-D neural network which processes temporal history of dark pixels.

The organization of the paper is as follows. In Section II-A we present the 1-D neural network analyzing the temporal history of a pixel. In Section II-B we present the spatio-temporal 2-D

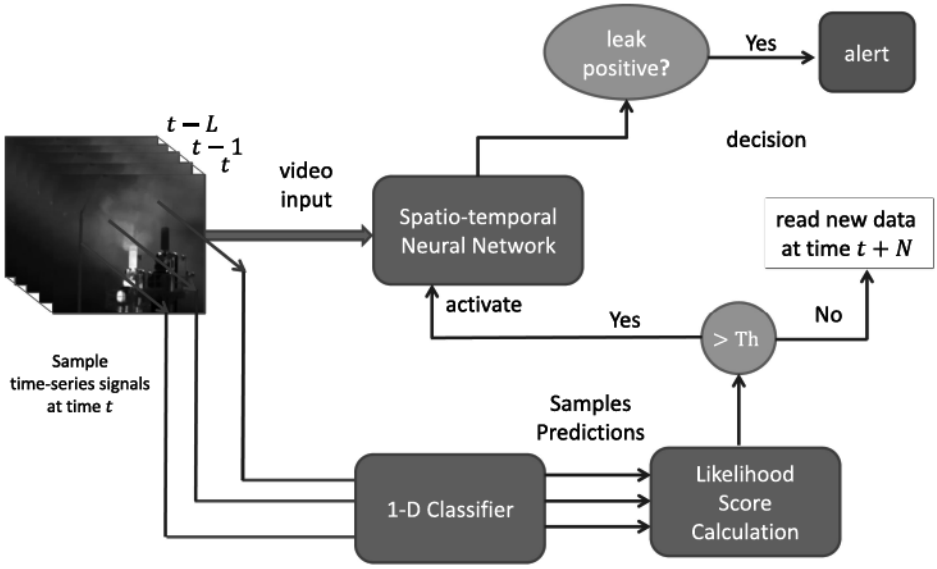


Fig. 7. Block diagram of our proposed system. "Th" stands for threshold.

neural network. In Section II-C, we present the energy-efficient additive-correlation based spatio-temporal neural network. In Section III we present our experimental results for the different algorithms. Finally, we present our conclusion and discussion.

II. Computationally Efficient Spatio-Temporal Video Analysis Framework

As pointed out above, plumes of VOC leaks appear as dark regions in a white-hot mode thermal IR video (VOC leaks appear as bright regions in black-hot mode). Such regions do not have a stationary shape over time. They expand and move in an erratic manner. Similar to flame and smoke detection in regular video [10], [12], [21]–[27] VOC leakage in IR video can be determined using spatio-temporal analysis. Deep neural networks have demonstrated their abilities in complex pattern recognition tasks. However, they are computationally demanding and may not be a practical solution to be used on ordinary low-cost computers for real-time applications. For this purpose, we develop a computationally efficient framework consisting of two components in cascade. The first component is computationally efficient and can be used for real-time monitoring, while the latter is more computationally demanding and is invoked by the first sub-system only when needed.

In details, the system first samples IR video and feeds those samples of 1-D time-series signals of pixels of possible VOC regions to a 1-D neural network. Afterwards, the system quantifies the likelihood of VOC leakage from the results obtained by the 1-D neural network. If the confidence of having VOC leak exceeds a certain threshold, it feeds corresponding image frames to the second component, which is a spatio-temporal convolutional neural network as shown in Fig. 7.

In these settings, the task of continuous monitoring is assigned to the relatively efficient 1-D neural network instead of an ordinary 2-D CNN. In order to see the computational saving, one can compare the computational complexity of convolution carried out in 1-D vs 2-D CNN. For an input of size $N \times N \times D$, and a filter of size $k \times k \times D$, the realization of a single 2-D feature map has a computational complexity $O(N \times N \times k \times k \times D)$. On the other hand, the realization of a single 1-D feature map with an input size $M \times D$ and a filter size of $l \times D$ has a computational complexity of $O(M \times l \times D)$. Therefore, as long as $M < N^2$ and $l < k^2$, the 1-D based convolution is more efficient. As a matter of fact, we have $M << N^2$ in our system, thus the computational saving of the 1-D neural network is enormous.

A. One-Dimensional Temporal Analysis of Dark Moving Pixels in IR Video

The first step of our VOC leak detection method is to identify dark moving regions in IR video. We do not use a threshold for dark region detection because the background level affects the average value of the VOC leak region. We process a given image frame in small windows and identify the minimum or local minima. After this step we construct 1-D temporal signals corresponding to such pixels in IR video.

In other words, we extract 1-D temporal records from the original spatio-temporal (video) data. We process these 1-D history signals separately using a neural network in order to identify whether these history signals are part of a VOC leakage scene or not. As shown in Fig. 5, if the pixel is at the boundary or near the boundary of a VOC leak region, it will exhibit an erratic (high-frequency) behavior over time. On the other hand, if the pixel is from an ordinary object absorbing IR light or a cold place, it will be stationary and exhibit low-frequency behavior most of the time as shown in Fig. 6.

The motivation behind using simple temporal 1-D signals instead of the entire video data is to greatly reduce the computational complexity of the detection algorithm. This approach also works if the IR camera is slowly moving, zooming or panning. Since we only process a relatively few number of 1-D temporal

TABLE I
ARCHITECTURE OF THE 1-D CONVOLUTIONAL NEURAL NETWORK USED IN TEMPORAL SIGNAL CLASSIFICATION

Layer	Specification
Input Layer	input size: 160×1
Conv Layer	32 5×1 filters, strides=2
Batch-norm Layer	applied
Conv Layer	64 3×32 filters, no strides
Max-pooling Layer	down-sampling by 2
Batch-norm Layer	applied
Conv Layer	128 3×64 filters, no strides
Max-pooling Layer	down-sampling by 2
Batch-norm Layer	applied
Global Average-pooling	output size: 128
Dense Layer	output size: 128
Batch-normalization Layer	applied
Output Layer	output size: 1 (soft prediction)

pixel signals per image frame, processing 1-D time-series data is much more efficient than processing a single image or image frames.

We tried two classifiers of the 1-D temporal data. The first is a regular 1-D convolutional neural network. The second is an LSTM-based classifier. Our input is a single temporal time-series history signal generated by a moving dark pixel of the IR camera and the output is a binary prediction value predicting whether this specific time-series signal comes from a VOC leakage region or not.

In our settings, we read temporal signals of size 160 at a frame rate of 25 fps, which roughly corresponds to 6.4 seconds from IR videos with spatial dimensions of 224×224. The time span of 6.4 seconds is sufficient for any gas leakage to have a noticeable spread-out across the scene as shown in Fig. 5. Therefore, we expect that a sufficient number of time history signals to have time-varying intensity values informative of VOC leakage. Therefore, if our 1-D neural network is trained to perform a binary classification task to distinguish VOC leakages from other events based on time-series signals, we expect it to be able to recognize any potential leakage from other moving objects in the IR video. The architecture of the CNN used is given in Table I.

Since our decision will be based on the collective prediction results of the 1-D time-series signals, we devise a confidence score that quantifies our confidence as to whether the scene contains VOC leaks or not. In this regard, let $\mathbf{x} \in \mathbb{R}^{160}$ represent the input signal, which is a vector of length 160, and let $D(x) \in \{0, 1\}$ be the hard decision made by the 1-D convolutional classifier, where "0" corresponds to predicting ordinary temporal signals, and "1" corresponds to predicting that the time-series signal is from a VOC leak region in IR video. Let N be the number of sample trajectories extracted from the entire spatio-temporal video data. The VOC-leak confidence score is defined as follows:

$$L := \frac{\sum_{i=1}^{N} I\left(D(x_i) = 1\right)}{N} \tag{1}$$

where $I(.)$ is the indicator function and x_i is the i-th signal. In other words, if there are enough time-series signals identified as positive class (VOC leak), the VOC-leak confidence score L will be high. On the other hand, if the sampled time-series signals do not contain any leakage, the score will be low. In case the confidence score L exceeds a certain threshold, the system recognizes a suspicious event, and then invokes the deep neural network that analyzes spatial data in order to verify the entire scene.

The architecture of the 1-D convolutional neural network is summarized in Table I.

B. Two-Dimensional (2-D) Spatio-Temporal Analysis Network

As mentioned earlier, we also utilize the image-based deep convolutional network whenever the VOC confidence score L exceeds a certain threshold.

Thanks to the 1-D network, we need to feed only a relatively few consecutive frames to the image-based deep CNN. In other words, we do not need to process the entire stream captured by the surveillance camera. The input fed to the 2D CNN is 3-D spatio-temporal images, where the first two dimensions correspond to the height and width, and the last dimension corresponds to the number of successive temporal frames. The reason for feeding spatio-temporal images rather that signle frames is two-fold: First, in some VOC leak scenes, the leakage is very weak and barely discernible by the human eye from a single frame. Therefore, we may run the risk of having the network recognizing the background scene (e.g. pipelines) as a scene belonging to the VOC leak class. Secondly, we incorporate temporal frames in order to ensure that the network does not over fit the non-VOC scenes, but learn to classify the image frames using both spatial and temporal information. The architecture of our network is shown in Table II.

In this study, we used a regular CNN, a recurrent LSTM and a novel energy efficient network which performs only one multiplication per convolution operation.

C. Additive-Correlation Based Spatio-Temporal Neural Network

In this subsection we review an energy efficient neural network which can be used in mobile systems or cameras. We implemented a neural network, which we call the Additive Neural Network (AddNet). The AddNet was first introduced in [28], [29] and it performs what we call Additive-Correlations (AC) in its neurons. The additive-correlation is based on the following arithmetic operation:

$$x \oplus w := \text{sgn}(xw)(|x| + |w|) \tag{2}$$

where x and w are two real-valued scalars, $sgn(.)$ is the signum function. We extend the scalar definition to the case of real vectors in order to construct a dot product like operation. Let $\mathbf{x}$ and $\mathbf{w} \in \mathbb{R}^d$. We define the additive-correlation of two vectors

TABLE II
ARCHITECTURE OF THE 2-DIMENSIONAL SPATIO-TEMPORAL NEURAL NETWORK AND THE ADDNET. "N" REFERS TO THE NUMBER OF SUCCESSIVE FRAMES FED TO THE CNN (TEMPORAL DEPTH DOMAIN). THROUGHOUT OUR EXPERIMENTS, WE SET N TO 3,4 AND 5

Layer	Specification
Input Layer Conv Layer Batch-norm Layer	input size: $112 \times 112 \times N$ 64 5×5 filters, strides=2 applied
Conv Layer Max-pooling Layer Batch-norm Layer	128 3×3 filters, no strides pooling size: 2 applied
Conv Layer Max-pooling Layer	256 3×3 filters, no strides pooling size: 2
Global Average-pooling Batch-norm Layer	output size: 256 applied
Dense Layer Batch-norm Layer	output size: 256 applied
Output Layer	output size: 1 (soft prediction)

as follows:

$$\mathbf{x} \oplus \mathbf{w} := \sum_{i=1}^{d} \text{sgn}(x_i w_i)(|x_i| + |w_i|) \tag{3}$$

where each entry of the above equation has the same sign of regular multiplication. As a result, whenever x_i and w_i have the same sign they positively contribute to the AC. On the other hand, if they have different signs they negatively contribute to the AC as in regular correlation operation between the two input vectors. The above operation avoids the use of multiplication operation which consumes significant amount of energy in many mobile systems. It is straightforward to show that Eq. (3) can be also expressed as follows:

$$\mathbf{x} \oplus \mathbf{w} = \sum_{i=1}^{d} \text{sgn}(x_i) w_i + x_i \text{sgn}(w_i) \tag{4}$$

As the regular dot product induces the ℓ_2 norm, the AC operation induces a scaled version of the ℓ_1 norm as follows:

$$\mathbf{x} \oplus \mathbf{x} = \sum_{i=1}^{d} \text{sgn}(x_i x_i)(|x_i| + |x_i|) = 2||\mathbf{x}||_1 \tag{5}$$

For convenience, we define the corresponding matrix-vector operation as follows: let the vector $\mathbf{x} \in \mathbb{R}^d$ and the matrix $\mathbf{W} \in \mathbb{R}^{d \times W}$. We then define the matrix-vector AC operation as follows:

$$\mathbf{y} := \mathbf{W} \oplus \mathbf{x} = [\mathbf{x} \oplus \mathbf{w_1}\ \mathbf{x} \oplus \mathbf{w_2}\ \ldots\ \mathbf{x} \oplus \mathbf{w_W}]^T \tag{6}$$

where $\mathbf{w_i}$ is the i^{th} column of $\mathbf{W}$ for $i = 1, 2, \ldots, W$ and $\mathbf{y} \in \mathbb{R}^W$ is the resulting vector.

In regular neural networks, a dense feed-forwarding pass can be expressed as follows:

$$\mathbf{y} = \phi(\mathbf{W}^T \mathbf{x} + \mathbf{b}) \tag{7}$$

where $\mathbf{x} \in \mathbb{R}^d$ is the input vector, $\mathbf{W} \in \mathbb{R}^{d \times K}$ is the weights matrix, $\mathbf{b} \in \mathbb{R}^K$ is the bias vector and $\phi(.)$ is the element-wise nonlinear activation. In AddNet, replace the matrix-vector multiplication in feed-forwarding by the operation defined in Eq. (6). Furthermore, we introduce a normalization vector $\mathbf{a} \in \mathbb{R}^K$ so as to control the range of the responses of the term $\mathbf{W} \oplus x$. We express the feed-forwarding pass of a dense layer as follows:

$$\mathbf{y} = \phi(\mathbf{a} \odot (\mathbf{W} \oplus \mathbf{x} + \mathbf{b})) \tag{8}$$

where $\odot$ represents the element-wise product between the vector $\mathbf{W}$ and the vector $\mathbf{a} \in \mathbb{R}^K$. Realizing the element-wise product between $\mathbf{a}$ and $\mathbf{W} \oplus \mathbf{x} + \mathbf{b}$ is inexpensive because we only carry out d multiplications compared to $K \times d$ multiplication operations in the case of $\mathbf{W}^T\mathbf{x}$. We can construct AddNet convolutional layers in a straightforward manner by substituting each convolution (dot product) operation with the equivalent AC operation. AddNet is more energy efficient than regular neural networks because it performs only one multiplication per "convolution" operation.

In our experiments, we used the architecture described in Table II, i.e., the same as in the case of our regular convolutional network used in analyzing IR video frames.

III. DATA SET AND EXPERIMENTAL RESULTS

We compiled a data set of hundreds of frames and a data set of thousands of time-series pixel history signals from 29 publicly available IR thermal videos and 12 videos that we recorded using a low resolution bolometer type IR camera. We used these data sets to train the two components of the systems and evaluate their recognition capabilities separately and jointly. It is worth mentioning that some videos contain more than one scene of interest. For information about the data sources, the reader may refer to Appendix A.

A. One-Dimensional (1-D) Data Set

We gathered a time-series data set of 15,000 pixel history signals from 6 normal IR videos, and 7 IR videos that contain VOC leaks in order to create a binary-class data set for training the 1-D classifier. In order to obtain an accurate data set, we carefully extracted the temporal data from VOC leakage regions from the 7 video clips containing VOC leaks. On the other hand, we randomly extracted 1-D signals from different locations in 6 normal video clips. This is to ensure our normal data cover different motion and intensity patterns. One can also select one pixel out of each 8 by 8 block or 16 by 16 block of the relatively dark regions of the video clip. IR cameras produce Discrete Cosine Transform (DCT) compressed data. Therefore it is also possible to use the DC value of each 8 by 8 or 16 by 16 image blocks. Example time-series signals corresponding to pixels in VOC leaks and ordinary pixels are shown in Fig. 5 and 6, respectively.

TABLE III
RESULTS OF THE CONFIDENCE SCORE OVER DIFFERENT SCENES. THE MEAN SCORES AND THE STANDARD DEVIATIONS ARE ESTIMATED FROM 10 DIFFERENT TRIALS

Scene ID	Scene Description	Contains leak? (Y/N)	CNN Confidence Score		LSTM Confidence Score	
			mean	std.	mean	std.
Scene 1	wildlife	No	0.01	0.01	0.22	0.04
Scene 2	wildlife	No	0.02	0.02	0.23	0.04
Scene 3	wildlife	No	0.11	0.05	0.21	0.06
Scene 4	wildlife	No	0.16	0.08	0.20	0.06
Scene 5	road	No	0.05	0.04	0.23	0.05
Scene 6	road	No	0.03	0.01	0.31	0.08
Scene 7	road	No	0.02	0.01	0.18	0.05
Worst-Case Score (for VOC negative videos)			0.16		0.31	
Scene 8	pipe leak	Yes	0.42	0.1	0.71	0.09
Scene 9	pipe leak	Yes	0.41	0.08	0.40	0.08
Scene 10	jet engine	Yes	0.54	0.09	0.57	0.08
Scene 11	chimneys	Yes	0.36	0.13	0.50	0.07
Scene 12	pipe gas leak	Yes	0.37	0.1	0.35	0.08
Scene 13	natural gas leak	Yes	0.85	0.06	0.47	0.09
Scene 14	natural gas leak	Yes	0.26	0.11	0.47	0.07
Scene 15	chimneys	Yes	0.49	0.11	0.38	0.07
Scene 16	VOC leak	Yes	0.34	0.08	0.72	0.08
Scene 17	VOC leak	Yes	0.71	0.07	0.76	0.09
Scene 18	chimneys	Yes	0.37	0.07	0.68	0.07
Scene 19	natural gas leak	Yes	0.62	0.12	0.64	0.08
Worst-Case Score (for VOC positive videos)			0.26		0.35	

In a bid to enforce classification invariance to background intensity levels, we augmented our training data set by adding a constant value to the recorded signals while training. Since the 1-D temporal signals are pixel intensities, their values are bounded between 0–255. In this case, we used DC levels from 0–255 and made sure that the signal values are bounded below by 0 and bounded above by 255, respectively.

We also implemented an LSTM based classifier. Similar to the 1-D CNN case, we used time-series signals of length 160. The architecture is an LSTM layer, which has 20 cells that read the input signals. The output size of the LSTM layer is 20. We then feed the output vector to a dense linear layer, which serves as our output layer. The input to each LSTM cell is a feature vector extracted from a segment of the original temporal signal of length 16. The segments corresponding to adjacent LSTM cells overlap by 50%. The features fed to each cell are the magnitudes of the DFT of 16-sample long time-series segments. The reason for using DFT instead of time-series data segments is to achieve translation invariance.

We reserved 20% of the aforementioned 1-D data set to use it for early-stopping validation purposes. We trained our 1-D neural network for 5 epochs using Adam Optimizer. We were able to achieve 98% accuracy over the validation data set.

In order to establish a VOC-leak confidence score threshold as defined in Eq. (1), we carried out stochastic inference over another validation data set of video scenes. This data set is obtained from 19 video scenes (Scene 1–20 as in Table III and Table V),[1] each of which contains successive frames of a spatial size of 224. The videos have significantly different resolutions and scales because they were obtained using different IR cameras. We have 7 videos that do not contain VOC leaks. The remaining 12 videos contain scenes in which there are VOC leaks. The VOC-leak videos have different scenarios and vary greatly in the VOC gas eruption location and the scale. These videos are different from those used initially to train the 1-D classifier. In each trial, we sampled 40 1-D temporal signals randomly from each video and calculated the VOC confidence score defined in Eq. (1). We repeated the process 10 times for each video. We report the average score and the standard deviation of the confidence score for the different scenes in Table III. Example IR image frames of the two classes are shown in Fig. 8 and Fig. 9, respectively.

As we can see from Table III the VOC confidence scores of temporal signals obtained from ordinary scenes are significantly lower than those of VOC-leak scenes, with the highest confidence score being 0.16 for the 1-D CNN as shown in the 4-th column of Table III. On the other hand, the lowest confidence score for VOC leaks is 0.26. When the input 1-D

[1]For information about the data sources, see Table IX in the following link [Online]. Available: https://github.com/Diaa0/Volatile-Organic-Compound-VOC-Leakage-Detection/blob/master/VOC_leak_appendix.pdf

Fig. 8. Example IR image frames containing VOC leaks.

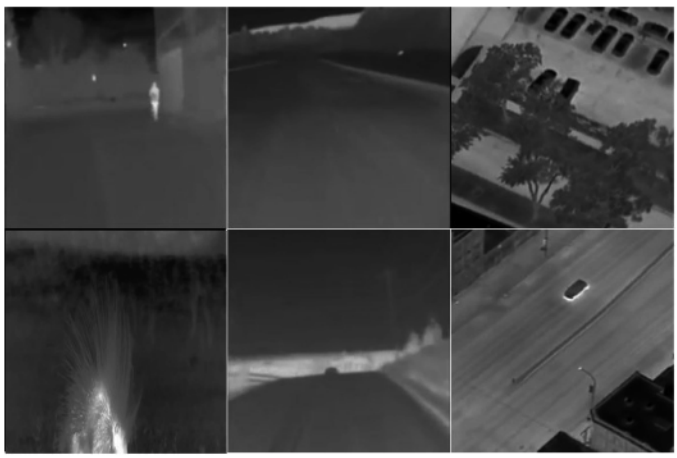


Fig. 9. Example ordinary image frames from IR videos used in training the neural networks.

signals come from VOC leaks, the lowest score is 0.26. We can set the threshold for the confidence score defined in Eq. (1) as 0.16. All the other scores of video scenes 1–7 are much lower than 0.16. Example 1-D signals and some intermediate layers are shown in Fig. 10.

On the other hand, the LSTM-based classifier does not produce as good results as the CNN-based classifier does. The confidence scores of positive videos are consistently higher compared to negative (no-leakage) videos. We can still set a threshold separating the two classes in our data set. However, the score margin between the two classes is small when compared to that of the CNN-based classification (0.35–0.31 vs 0.26–0.16). This suggests that the LSTM-based classifier has a higher chance of producing more false alarms in comparison with the CNN based classifier.

We also subtracted the mean values of 1-D signals before feeding them to CNN (column 6 of Table III). This strategy also produced good results except for one video clip (Scene 4).

B. IR Video Dataset for 2-D Spatio-Temporal Processing

We extracted infrared image frames from 17 publicly available videos, which account for 48 different scenes, and constructed a training data set for the spatial classifier stage.[2] For validation, we utilized the data set we used earlier for establishing the confidence score for the 1-D classifier as mentioned in Section III-A (Scene 1–20).

We used entire frames and we tried different temporal depths. In particular, we set our temporal depth to 3, 4 and 5 image frames, respectively. In our data set we normalized the image input size to 112×112. We gathered a total of 8,400 frames of VOC scenes and 10,246 frames of ordinary scenes for our training data set. As for the validation data set, we used the data set that we test upon the 1-D neural network as detailed in Section III-A.

In order to enhance the capabilities of the network to detect VOCs, we randomly rotated the frames in the spatial domain during training so that the network is exposed to the textures in all different locations. This mitigates the risk of having the classifier over-fit the background scenes. Furthermore, We employed the early stopping criterion based on the recognition rates of the validation data set. Our validation dataset consists of 7 normal scenes and 12 VOC-leakage scenes. This is the same dataset we used for test the 1-D temporal signals. It should be noted that both training and validation sets are disjoint. We implemented an ordinary 2-D convolutional neural network and an AddNet, both of which have the architecture shown in Table II. We investigated 3, 4 and 5 consecutive image frames for the temporal depth of the input. We used Adam optimizer with a learning rate of 10^{-4}. We used Tensorflow-Keras in our implementation. We compared our method with a regular smoke detection algorithm that we developed with Mobilenet-V2 [30]. We utilized transfer learning using Mobilenet-V2 due to the relatively small data set size. We trained (fine-tune) only the last dense layers while keeping the weights of the convolutional layers intact. Our VOC image frame recognition results over the validation data set were 91–95% for regular networks with different number of input frames (3–5) and 93% for AddNet. We were able to identify the events of VOC leak in all of the videos. Smoke detection algorithm developed using Mobilenet-V2 did not produce as good results as our algorithms.

We also tested the neural networks over a set of videos we gathered using a bolometer type FLIR IR camera. These 12 videos contain butane leakage.[3] We report the results per video/scene in Table IV and Table V. Images generated by low cost bolometer IR cameras are corrupted by noise as shown in Fig. 11.

As it can be seen from Table IV and V, the true positive rates are high in the case of CNN and AddNet, in contrast to Mobilenet-V2, which misses out more VOC-positive videos. This can be attributed to the fact that the earlier layers in Mobilenet-V2, which are used for feature extraction and are

[2]See Table X in Appendix A for details about the data sources.

[3]These videos are available on YouTube given the following link [Online]. Available: https://www.youtube.com/playlist?list=PL9_9ATqpfzwPcHBnq2UdxHJ96aSVgVTF-

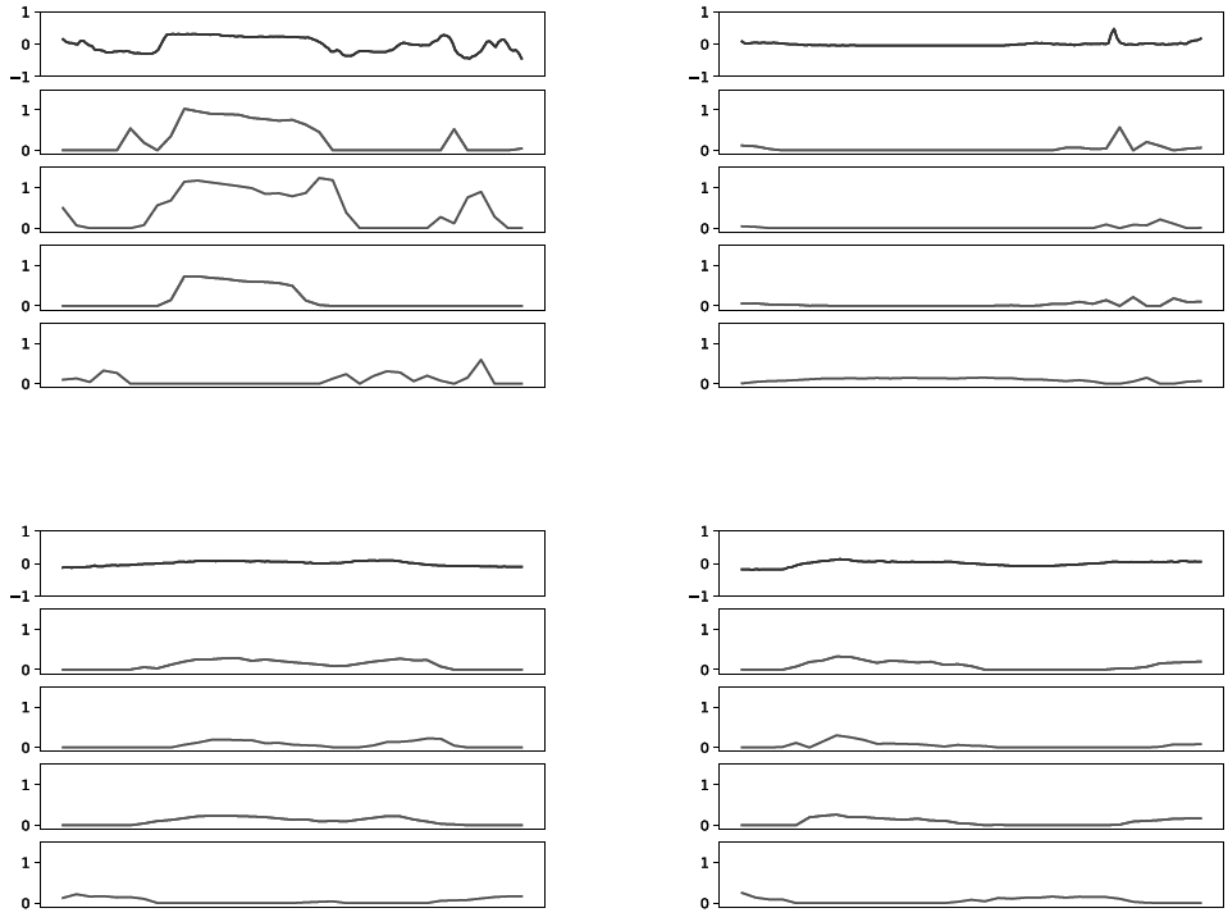


Fig. 10. Intermediate feature maps for 4 different example time-series signals. The signals in blue are the input signals and the signals in green and red are four different features maps from the second convolutional layer. Example (a) is taken from a VOC-positive video and classified as VOC-positive. Example (b) is taken from a VOC-positive video and classified as VOC-negative. Example (c) is taken from a normal (VOC negative video) and classified as negative, whereas Example (d) is taken from a VOC negative video and classified erroneously as VOC-positive. Each feature signal has a length of 35 samples and is scaled to match the length of the original signal (160 samples) for demonstration purposes.

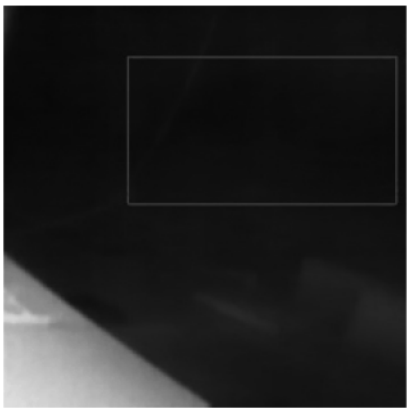


Fig. 11. An example of thermal image obtained by a low-cost bolometer-type IR camera. The darker region corresponds to butane leakage.

not trained during fine-tuning, were originally optimized using a dataset, namely ImageNet, which is radically different from IR-thermal images. Furthermore, despite the fact that AddNet and CNN do suffer from false positive rates in some videos as demonstrated in Table V, what matters the most is not missing out any VOC-positive leak. It should be pointed out that we use image analysis to verify the results of 1-D network which detects all the VOC leaks in our validation data set. Two feature maps corresponding to the same filter from the first convolutional layers are shown in Fig. 13 for a VOC leakage example and a normal example. As it can be seen from Fig. 13, feature maps have high response in the regions of darker areas. This is expected since the gas leak will generate dark spots in white-hot mode thermal IR video. Nevertheless, we can see that the response is zero for the animal appearing in the negative example as in Fig. 13(a). We can interpret this feature extraction process to be sensitive to darker areas, while not responsive to other patterns.

Furthermore, in order to verify that we have a small false negative rate, we extracted successive frames from each video. The total number of videos is 80. All of these frames have VOC leaks. The videos were obtained from The Oil & Gas Threat Map website [20], [6]. We tested our spatio-temporal neural networks over these frames and we could recognize VOC gas leaks in 77 out of 80 scenes in the case of a regular 2-D spatio-temporal network. In the case of AddNet, we were able to recognize 78 out of 80 scenes, i.e., AddNet and the regular convolutional neural networks are on par with their event based recognition results.

TABLE IV
TRUE POSITIVE RATES OVER BUTANE-POSITIVE IR-THERMAL VIDEOS WE GATHERED USING A LOW RESOLUTION BOLOMETER-TYPE IR CAMERA. THE ABBREVIATION "FRM." REFERS TO THE NUMBER OF INPUT CHANNELS (FRAMES) OF THE 2D CNN

Video ID	# of Frames	True Positive Rate (%)				
		CNN			Add-Net	Mob-NetV2
		3 frm.	4 frm.	5 frm.		
104414	101	100.0	60.5	72.2	98.6	56.3
104703	56	84.0	82.0	65.6	70.7	27.6
103421	104	22.7	20.3	100.0	0.0	0.0
104741	95	100.0	80.2	100.0	100.0	0.0
103934	103	100.0	100.0	100.0	98.2	3.0
104903	100	3.9	70.6	100.0	100.0	93.0
104955	97	100.0	98.9	100.0	100.0	100.0
104255	99	100.0	100.0	100.0	100.0	30.2
103732	78	100.0	80.0	100.0	100.0	10.0
104534	88	100.0	100.0	100.0	100.0	0.0
103236	113	74.8	20.2	10.1	2.6	67.8
103126	97	100.0	94.5	100.0	100.0	100.0
Total	1131	81.1	74.8	86.6	79.3	42.1

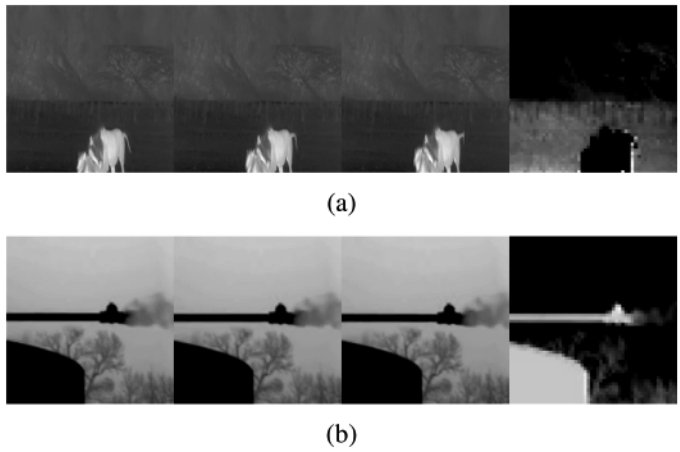
Fig. 12. Example image frames from various videos that we used in testing our deep neural networks. All the frames except the bottom-left frame are correctly recognized by the CNN and AddNet.

As for the missed scenes, we notice that the VOC leak is very dim. This is probably because the camera is placed too far away from the VOC leak sources. We show some example IR image frames in Fig. 12. We also note that bolometer type low-cost IR cameras are not as reliable as regular LWIR or MWIR cameras due to the noisy nature of bolometer images.

C. Joint Performance Evaluation and Discussion

We tested the entire VOC detection system over a data set consisting of 7 normal and 5 VOC-positive scenes extracted from 9 videos.[4] The recognition results are reported in Table VI. As one can see from Table VI, the 1D CNN classifier with zero-mean input (CNN 0) recognizes all the VOC leaks in positive video clips. However, it produces false alarms in aerial Scene 26 (aerial scene) and the wildlife scene 27. The 2-D spatio-temporal network with 3 or 5 image inputs can correct the false-alarm in Scene 26 as shown in the 6-th row of Table VI.

[4]For information about the data sources see Table IX in the following link [Online]. Available: https://github.com/Diaa0/Volatile-Organic-Compound-VOC-Leakage-Detection/blob/master/VOC_leak_appendix.pdf



(a)

(b)

Fig. 13. Example feature maps of the first convolutional layer for two examples: (a) A wildlife scene with no VOC and (b) a scene with VOC gas leak. The values are re-scaled for demonstration purposes.

Fig. 14. Example image from Scene 27 (pump gas leak).

The 1-D network without mean subtraction (CNN 1) recognizes all the VOC-leaks except the gas-pump (scene 28) and produces a false alarm in the Scene 27. The gas pump leak is relatively faint compared to other VOC leak video clips. An example frame from Scene 27 is shown in Fig. 14. The 2-D spatio-temporal network recognizes the leak in more than 80% of the image frames of the IR video consisting of 177 frames.

The 2D spatio-temporal networks with 3, 4 or 5 inputs recognize the VOC leaks in all the video clips. They have a low recognition rate in Scene 29 but it is enough to recognize the VOC leak even once to alert the security officer who will be verifying the results of the 2-D spatio-temporal system. The 2D network with 3 or 5 images produces a false alarm only in Scene 27.

D. Computational Efficiency of AddNet

We did time analysis over inference passes for a regular CNN and AddNet on a PC equipped with a CPU of type Intel Core I7-7700HQ. We measured the inference time for mini-batches of different sizes. The averaged results are presented in Table VII. As it can be seen from Table VII, computational efficiency of AddNet is not significant in the case of a single-example batch.

TABLE V
RECOGNITION RATES OVER THE POSITIVE AND THE NEGATIVE SCENES WE USED IN OUR VALIDATION DATA SET

Scene ID	Scene Description	Contains leak?	# Frames	CNN			AddNet	MobileNet-V2
				3 frames	4 frames	5 frames		
				False Positive Rate (%)				
Scene 1	wildlife	No	48	0.0	0.0	0.0	0.0	41.7
Scene 2	wildlife	No	48	0.0	0.0	0.0	0.0	14.6
Scene 3	wildlife	No	48	0.0	0.0	0.0	100.0	0.0
Scene 4	wildlife	No	82	7.2	0.0	0.0	1.2	7.3
Scene 5	road	No	143	26.6	55.2	11.8	0.0	11.9
Scene 6	road	No	485	4.1	0.0	0.0	0.0	2.4
Scene 7	road	No	67	28.4	0.0	0.0	0.0	9.0
Total	-	-	921	8.3	8.6	1.8	5.3	7.4
				True Positive Rate (%)				
Scene 8	pipe leak	Yes	157	95.6	80.2	62.3	80.3	87.3
Scene 9	pipe leak	Yes	289	99.7	100.0	91.9	91.0	82.0
Scene 10	jet engine	Yes	51	60.8	100.0	38.2	100.0	31.4
Scene 11	chimneys	Yes	72	100.0	100.0	100.0	100.0	100.0
Scene 12	pipe gas leak	Yes	156	100.0	100.0	100.0	54.5	94.9
Scene 13	natural gas leak	Yes	48	100.0	100.0	100.0	100.0	100.0
Scene 14	natural gas leak	Yes	58	100.0	49.1	100.0	12.7	100.0
Scene 15	chimneys	Yes	50	98.0	0.0	100.0	100.0	100.0
Scene 16	VOC leak	Yes	41	100.0	100.0	100.0	0.0	100.0
Scene 17	VOC leak	Yes	629	75.8	100.0	99.8	92.9	4.3
Scene 18	chimneys	Yes	406	100.0	99.7	100.0	69.7	55.4
Scene 19	natural gas leak	Yes	3544	99.4	100.0	93.2	98.9	97.7
Scene 20	natural gas leak	Yes	528	41.5	96.8	100.0	100.0	0.0
Total	-	-	6029	91.6	97.7	94.1	92.9	75.0

TABLE VI
THE PERFORMANCE RESULTS OF THE SPATIO-TEMPORAL VOC DETECTION SYSTEM OVER A TEST DATA SET

Scene ID	Scene Description	Contains leak?	# of frames	confidence Score (1D)		VOC Recognition rates (2D CNN)		
				CNN 1	CNN 0	3 frames	4 frames	5 frames
Scene 21	building	No	417	0.08	0.0	0.0	0.0	0.0
Scene 22	pedestrians	No	237	0.14	0.04	0.0	0.0	0.0
Scene 23	pedestrians	No	507	0.08	0.05	0.0	16.9	0.0
Scene 24	building	No	717	0.08	0.07	0.0	0.0	0. 0
Scene 25	aerial scene	No	230	0.06	0.04	0.0	10.0.0	0.0
Scene 26	aerial scene	No	117	0.11	0.41	0.0	89.0	0.0
Scene 27	wildlife	No	148	0.25	0.23	89.0	87.8	87.5
Scene 28	gas pump	Yes	177	0.10	0.57	88.0	90.0	85.1
Scene 29	oil well	Yes	177	0.27	0.44	12.4	13.6	18.2
Scene 30	pipe gas leak	Yes	177	0.18	0.47	100.0	100.0	100.0
Scene 31	pipe gas leak	Yes	217	0.19	0.26	32.5	100.0	100.0
Scene 32	pipe gas leak	Yes	207	0.19	0.40	99.5	97.2	98.0

However, AddNet was able to process batches of 3 images faster than CNN by 5% in a regular PC. It can achieve 15% efficiency when the batch size increases to 20 images. Cameras output compressed video and decoders generate batches of image frames in practice. The time saving results that we achieved using AddNet gives more room for processing mini batches of spatio-temporal frames in order to increase the recognition capacity of the system. Energy efficiency of AddNet depends on the type of the processor that is being used in video analysis but it is related to the computational time savings.

TABLE VII
EXECUTION TIME RESULTS OF CNN AND ADDNET MINI-BATCH INFERENCE FOR DIFFERENT MINI-BATCH SIZES

Mini-Batch Size	CNN Average (ms)	AddNet Average (ms)	AddNet Saving Rate
1	2.532	2.489	1.70%
3	1.175	1.106	5.88%
5	1.147	1.051	8.37%
10	1.074	0.968	9.87%
20	0.994	0.839	15.59%

IV. CONCLUSION

In this paper, we presented a computationally efficient VOC gas leak detection method, which is based on two neural networks connected in series. The first neural network analyzes the time-series data generated by some of the moving dark pixels of the thermal IR camera. If such pixels exhibit an erratic behavior it is possible that the scene may contain a dark cloud-like region due to a VOC gas leak. In such cases, three or more consecutive frames of the video are fed to the 2D spatio-temporal neural network. The overall system has reached high recognition rates in our dataset.

The VOC-leak detection structure that we propose is scalable in the sense that one can use only 1D temporal history signals, if the processor of the IR camera is a simple one such as the Raspberry PI or Arduino. If more processing power is available, the 2D spatio-temporal network can be also used for more reliable VOC-leak detection results. The spatio-temporal network will verify the results of the 1D temporal neural network.

We also used a novel neural network, AddNet, which is based on what we call the "additive-correlation" operation. The AddNet can be used in mobile applications including drones because it is more energy-efficient than corresponding regular neural networks. Recognition results of the AddNet is slightly inferior to the regular deep 2-D convolutional neural network.

ACKNOWLEDGMENT

The authors, Badawi and Cetin, would like to thank Nvidia for an equipment grant.

REFERENCES

[1] United States Environmental Protection Agency, *Leak Detection and Repair Compliance Assistance Guidance Best Practices Guide - ldarguide.pdf*, United States Environmental Protection Agency. Feb 2014. [Online]. Available: https://www.epa.gov/sites/production/files/2014-02/documents/ldarguide.pdf, Accessed on: Mar. 10, 2020.

[2] United States Environmental Protection Agency, Enforcement and Compliance, *Inspection Manual: Federal Equipment Leak Regulations For the Chemical Manufacturing Industry*, United States Environmental Protection Agency, Enforcement and Compliance Assurance, 1998. [Online]. Available: https://archive.epa.gov/compliance/resources/publications/assistance/sectors/web/pdf/insmanvol1.pdf, Accessed on: Mar. 10, 2020.

[3] National Institute of Standards and Technology, *Toluene*, National Institute of Standards and Technology, United States. [Online]. Available: https://webbook.nist.gov/cgi/cbook.cgi?ID=C108883&Type=IRSPEC&Index=2#IR-SPEC, Accessed on: Mar. 10, 2020.

[4] A. E. Cetin and B. U. Toreyin, "Method, device and system for determining the presence of volatile organic compounds (voc) in video," U.S. Patent 8 432 451, Apr. 30, 2013.

[5] F. Erden, E. B. Soyer, B. U. Toreyin, and A. E. Cetin, "VOC gas leak detection using pyro-electric infrared sensors," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2010, pp. 1682–1685.

[6] The New York Times, *It's a Vast, Invisible Climate Menace. We Made It Visible*. The New York Times, NY, United States, Dec. 2019. [Online]. Available: https://www.nytimes.com/interactive/2019/12/12/climate/texas-methane-super-emitters.html, Accessed on: Mar. 10, 2020.

[7] B. U. Toreyin and A. E. Çetin, "Volatile organic compound plume detection using wavelet analysis of video," in *Proc. 15th IEEE Int. Conf. Image Process.*, 2008, pp. 1836–1839.

[8] B. U. Töreyin, Y. Dedeoğlu, U. Güdükbay, and A. E. Cetin, "Computer vision based method for real-time fire and flame detection," *Pattern Recognit. Lett.*, vol. 27, no. 1, pp. 49–58, 2006.

[9] Y. H. Habiboglu, O. Gunay, and A. E. Cetin, "Real-time wildfire detection using correlation descriptors," in *Proc. IEEE 19th Eur. Signal Process. Conf.*, 2011, pp. 894–898.

[10] O. Gunay, B. U. Toreyin, K. Kose, and A. E. Cetin, "Entropy-functional-based online adaptive decision fusion framework with application to wildfire detection in video," *IEEE Trans. Image Process.*, vol. 21, no. 5, pp. 2853–2865, May 2012.

[11] B. U. Töreyin, Y. Dedeoğlu, and A. E. Cetin, "Wavelet based real-time smoke detection in video," in *Proc. IEEE 13th Eur. Signal Process. Conf.*, 2005, pp. 1–4.

[12] T. Celik and H. Demirel, "Fire detection in video sequences using a generic color model," *Fire Saf. J.*, vol. 44, no. 2, pp. 147–158, 2009.

[13] Y. H. Habiboğlu, O. Günay, and A. E. Çetin, "Covariance matrix-based fire and flame detection method in video," *Mach. Vision Appl.*, vol. 23, no. 6, pp. 1103–1113, 2012.

[14] B. Toreyin *et al.*, "Dynamic texture detection, segmentation and analysis," in *Proc. Conf. Image Video Retrieval: Proc. 6 th ACM Int. Conf. Image Video Retrieval*, 2007, vol. 9, no. 11, pp. 131–134.

[15] R. Péteri, S. Fazekas, and M. J. Huiskes, "Dyntex: A comprehensive database of dynamic textures," *Pattern Recognit. Lett.*, vol. 31, no. 12, pp. 1627–1632, 2010.

[16] A. Enis Cetin and F. Porikli, "Special issue on dynamic textures in video," *Mach. Vision Appl.*, vol. 22, no. 5, pp. 739–740, 2011.

[17] B. U. Toreyin and A. E. Cetin, "Computer vision based forest fire detection," in *Proc. IEEE 16th Signal Process., Commun. Appl. Conf.*, 2008, pp. 1–4.

[18] T. Celik, H. Demirel, H. Ozkaramanli, and M. Uyguroglu, "Fire detection using statistical color model in video sequences," *J. Visual Commun. Image Representation*, vol. 18, no. 2, pp. 176–185, 2007.

[19] H. Pan, D. Badawi, X. Zhang, and A. E. Cetin, "Additive neural network for forest fire detection," *Signal, Image Video Process.*, pp. 1–8, 2019.

[20] The Oil and Gas Threat Map, Infrared Video, The Oil and Gas Threat Map, United States, Apr. 2016. [Online]. Available: https://oilandgasthreatmap.com/about/infrared/, Accessed on: Mar. 10, 2020.

[21] A. E. Çetin *et al.*, "Video fire detection–review," *Digit. Signal Process.*, vol. 23, no. 6, pp. 1827–1843, 2013.

[22] A. E. Cetin, B. Merci, O. Günay, B. U. Töreyin, and S. Verstockt, *Methods and techniques for fire detection: Signal, image and video processing perspectives*. Academic Press, 2016.

[23] S. Verstockt, P. Lambert, R. Van de Walle, B. Merci, and B. Sette, "State of the art in vision-based fire and smoke dectection," in *Proc. 14th Int. Conf. Automat. Fire Detection*, Dept. Commun. Syst., Univ. Duisburg-Essen., vol. 2, 2009, pp. 285–292.

[24] P. Barmpoutis, K. Dimitropoulos, K. Kaza, and N. Grammalidis, "Fire detection from images using faster R-CNN and multidimensional texture analysis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 8301–8305.

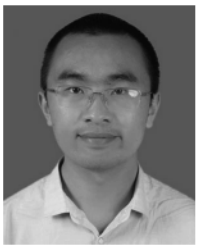
[25] B. C. Ko, K.-H. Cheong, and J.-Y. Nam, "Fire detection based on vision sensor and support vector machines," *Fire Saf. J.*, vol. 44, no. 3, pp. 322–329, 2009.

[26] F. Yuan, "A fast accumulative motion orientation model based on integral image for video smoke detection," *Pattern Recognit. Lett.*, vol. 29, no. 7, pp. 925–932, 2008.

[27] S. Aslan, U. Güdükbay, B. U. Töreyin, and A. E. Çetin, "Early wildfire smoke detection based on motion-based geometric image transformation and deep convolutional generative adversarial networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 8315–8319.

[28] A. Afrasiyabi, D. Badawi, B. Nasir, O. Yıldız, F. T. Yarman-Vural, and A. E. Çetin, "Non-Euclidean vector product for neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 6862–6866.
[29] D. Badawi, S. Ozev, J. B. Christen, C. Yang, A. Orailoglu, and A. E. Çetin, "Detecting gas vapor leaks through uncalibrated sensor based CPS," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 8296–8300.
[30] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 4510–4520.

Diaa Badawi received the B.Sc. degree in communications engineering from An-Najah National University, Nablus, Palestine, in 2015. He received the M.Sc. degree in electrical engineering from Bilkent University, Ankara, Turkey, in 2018. He is currently working toward the Ph.D. degree in electrical and computer engineering with the University of Illinois at Chicago, Chicago, IL, USA. His current research interests are in gas leak detection using machine learning and signal processing.



Hongyi Pan received the B.S. degree in automation from Chang'an University, Xi'an, China, in 2018 and the M.S degree in electrical and computer engineering from the University of Illinois at Chicago, Chicago, IL, USA, in 2019. He is currently working toward the doctoral degree in electrical and computer engineering with the University of Illinois at Chicago. His research interests are signal processing and deep learning.

Sinan Cem Cetin received the B.Sc. degree from the Department of Computer Science, University of California, Los Angeles, Los Angeles, CA, USA, in 2019. He is currently a Software Engineer with Flexera, Ithasca, IL, USA.

A. Enis Cetin (Fellow, IEEE) received the B.Sc. degree from METU, Ankara, Turkey and the M.S.E and Ph.D. degrees in systems engineering from the University of Pennsylvania, Philadelphia, PA, USA, in 1986 and 1987, respectively. He was an Assistant Professor of Electrical Engineering with the University of Toronto from 1987 to 1989. He was on the Faculty of Bilkent University, Ankara, Turkey from 1989 to 2017. He is currently a Research Professor with the University of Illinois at Chicago, Chicago, IL, USA. He is the Co-Author of the book *Methods and Techniques for Fire Detection: Signal, Image and Video Processing Perspectives*. Academic Press, 2016. He is an Editor-in-Chief of Signal, Image and Video Processing, Springer-Nature.