Smoke Detection Model on the ALERTWildfire Camera Network

Rodrigue de Schaetzen, Raphael Chang Menoni, Yifu Chen, and Drijon Hasani

University of British Columbia, Canada

{deschaet, menoni, yifuch01, drijonha}@student.ubc.ca

Abstract

With climate change effects projected to intensify over the next century, the steady increase in destructive wildfires has been a major cause for concern. While a major advancement in developing large-scale fire monitoring services was made with the establishment of ALERTWildfire, it has yet to adopt an automated smoke detection system to alleviate the task of manually monitoring live-feed footage for wildfires. To address this, we present a multi-label image classifier to predict forest fire smoke based on Pan-Tilt-Zoom (PTZ) image data. Our approach is to divide the image into a KxK grid and predict the cells of the grid containing smoke. We achieve $3-4\%$ improvement over baseline for a $4x4$ mesh and empirically establish upper bounds on grid size resolution. As a secondary contribution, we release the first smoke-annotated video dataset which consists of 139 hours of footage from PTZ cameras across 678 videos. We hope the public release of our dataset will accelerate future research efforts in large-scale automatic wildfire monitoring.

1. Introduction

The climate crisis has had significant impact on the frequency and intensity of forest fires [1]. Extreme heat and prolonged droughts have contributed to increasingly variable and severe wildfire-prone weather. As a result, regions which tend to have dry seasons and combustible vegetation (e.g. dry grass, weed, etc.), have increased spending on wildfire management services. For instance, the California Department of Forestry and Fire Protection increased its spending by more than 60% from 2006-07 to 2017-18 [2]. Of the various challenges managing wildfires, a key component is reliable fire-detection systems, particularly for remote areas.

To address the challenge of monitoring large and remote regions, several universities in western US have collaborated to form the ALERTWildfire organization [3]. It consists of over 300 specialized Pan-Tilt-Zoom (PTZ) cameras providing first responders crucial ground-level information on regions across Washington, Oregon, Idaho, California, and Nevada. Live streams of these cameras are accessible to the public via the ALERTWildfire website. Since 2016, over 1,000 fires have been mitigated as a result of critical information provided by ALERTWildfire cameras. Given its success, the organization intends to continue expanding its fleet of cameras to provide more coverage; 1,000 cameras are projected to be installed in California by 2022 [3], which we estimate generates an expenditure of $2.5 million each year for monitoring alone (using the average per-camera operator cost of $2,500 from the wildfire detection cost review by [4]).

In some districts, recruited volunteers actively monitor these streams during wildfire season to assist state emergency services. However, they likely face common cognitive challenges in human surveillance such as lack of concentration and vulnerability to distractions [5]. Other challenges include less availability on holidays and during night-time, while research [6] shows that more arson and recreational fires occur during holidays. In addition, volunteers may need to monitor concurrent streams or additional operators will be needed to account for the growing number of

installed cameras. For these reasons the current approach is sub-optimal and propose to instead leverage an automated smoke detection system.

In this work, we develop a multi-label image classifier for automatic smoke detection based on PTZ camera data. Each label corresponds to a grid segment of a KxK grid on a given RBG image. We trained our models on our novel video dataset consisting of smoke-annotated daytime footage from ALERTWildfire cameras. This paper is structured as follows: Section 2 provides a brief summary on existing methods for automated smoke detection, Sections 3-4 describe our data generation pipeline and give a detailed discussion on the advantages of a grid approach, Section 5 we experiment with various mesh sizes and CNN architectures, and finally Sections 6-7 we provide a discussion on the limitations of our approach, future work including incorporating temporal modeling, and the broader impact of this paper.

2. Background

In the domain of on-the-ground automated wildfire detection, researchers have developed machine learning-based systems [7]. For example, Inception CNN networks are found to have best performance on smoke detection compared to ResNet and VGG [8]. Commercial systems such as FireWatch [9] and ForestWatch [10] outperform these models [11], with a median detection time of 40 minutes [12]. In a recent collaboration between ALERTWildfire and UC Berkeley, an InceptionV3 model using image segmentation is proposed [13]. Unlike previous CNN-based methods, they proposed a grid approach of segmenting a 3000x2000 image into 299x299 grid cells and perform classification individually on each cell, and outperformed all commercial wildfire detection systems. However, they admitted that there was not an exhaustive investigation of alternative architectures with various grid sizes [13]. Our study is the first to explore the effects of combinations between different CNN architectures and grid sizes to smoke detection. Furthermore, we discuss the relation between grid sizes and the CNN's capacity to learn the shape of smoke, which may inform future research in the detection of similar formless objects.

In 2019, a paper presented a novel Attention Enhanced Bidirectional Long Short-Term Memory Network (ABi-LSTM) for video-based forest fire smoke recognition in China [14]. They emphasize the importance of a spatiotemporal representation of smoke given its highly dynamic features, and show an improvement of 4.4% in accuracy from image-based models. The authors also argue they are the first to introduce an attention mechanism to their network capable of adapting to discriminate frames. Nevertheless, a possible limitation of this paper is that they only tested their model on video footage recorded in China; it remains unclear whether their work is transferable to America due to different vegetation and geology features.

3. Dataset

We found there was a major lack of available smoke-annotated data and zero existing datasets that consisted of labeled videos. There is however massive amounts of raw historical data. We are among the first to make a smoke-annotated dataset publicly available, and the first to do so in such a large scale and on entire videos. We describe our process of generating this dataset in the detailed steps below.

- **Obtain ALERTWildfire data -** We scraped 1,067 videos from a YouTube channel managed by the Nevada Seismological Laboratory [15], a core collaborator of ALERTWildfire. It contains variable length (several minutes to 1 hour) clips of a variety of past wildfire related events including fire ignitions, smoke clouds, and fires being extinguished. The raw resolution of the video sequences were 1080p and took 30 GB of storage in total.
- **Annotate frames -** We configured the Computer Vision Annotation Tool (CVAT) to enable online collaborative data labeling. We recruited volunteers to assist with the labeling. Each

annotator was tasked with tightly bounding smoke with rectangular boxes. Note, multiple boxes were used when necessary to better bound the smoke from non-smoke regions. Annotators were given daytime frames only. In total, our final dataset consisted of 1,385,938 annotated frames from 678 different videos.

3.1 Known weaknesses

We noticed there were a number of frames with inaccurate labels, especially frames that were outsourced for labeling. This was particularly evident for frames containing large smoke, as annotators tended to draw one large box covering most of the image. During the labeling process we randomly reviewed and assessed annotated videos. Feedback was given several times to annotators and the most common comment was to make use of more boxes to tighten the bound of smoke.

Another minor cause for labelling inconsistency was the overuse of the "skip multiple frames" and "interpolation" features. Since most fires only started following the first quarter/third of the video length, annotators were allowed to use the skip multiple frames feature to quickly reach the point of when smoke started to become visible. The interpolation feature, while useful to generate annotations between two key frames (a key frame is a frame that was manually annotated), it also expedited the labeling process and hence slightly degraded the quality of our dataset.

A final weakness we discovered is most of the videos have a third to a half of their frames timestamped after a wildfire has already been flagged by a human operator. Once a fire has been detected, operators use the cameras to pan, tilt, and zoom to the ongoing fire. This creates a bias in our dataset for frames containing centered and large smoke. Figure 3a demonstrates this bias via a heat map.

3.2 Comparison to other datasets

We found only two available annotated datasets for smoke detection. The first one made by the AI For Mankind organization, consists of 744 box-annotated frames gathered from HPWREN cameras (a core ALERTWildfire partner [16]). For the second, the Fuego project annotated 1,740 frames also from HPWREN cameras [17]. Both of these datasets have annotated smoke with the intent of training an object detection model which means smoke is bounded by a single box.

4. Experimental details

In this section we discuss our design decisions for the smoke detection task. We considered 5 broad approaches: 1) segmentation, 2) bounding box detection, 3) binary classification, 4) multi-class and 5) multi-label classification. In contrast to solid objects, smoke rarely has a well defined edge and instead often gradually fades in the background scene. This makes smoke a difficult candidate for image segmentation. Similarly, smoke has a tendency to form irregular cone-like shapes with a trail heading towards one direction making it tough to tightly bound with a single box. For binary image classification, it is tricky to provide an explicit smoke "signal" to the network especially when there are clouds/fog and when the smoke is small in size.

We argue image classification on gridded images is the optimal solution. Each image is evenly divided up into an KxK grid where each segment of the grid corresponds to a binary class i.e. 0 for no-smoke, and 1 for smoke. This means in total there are 2^{KxK} possible outputs for a multi-class model. For a multi-label approach, the model's objective is to output a probability that a particular grid segment contains smoke. In theory, multi-label is the better approach over multi-class since we have the option to set different thresholds when mapping probabilities to smoke/no-smoke predictions. Below we discuss our steps to convert our raw annotated dataset for a multi-label grid approach. Figure (1) gives a summary of our dataset generation pipeline.

- **Generate grid labels -** We converted the CVAT box annotations to grid labels for various different grid sizes. Any grid segment that was overlaid above a certain threshold by a bounding box was mapped to a 1. In a similar manner, segments that failed to meet this criteria were assigned a 0. For the case of an image containing smoke that failed to meet the criteria for all of the grid segments, only the square with the highest percentage of coverage was set to 1. Note, we left this threshold as a hyperparameter of our model pipeline and tried several different values which we discuss in our results section. Figures (2a), (2b), and (2c) show sample frames with its corresponding CVAT and grid annotations with three different thresholds.
- **Downsample and filter frames -** Each video file was converted to a series of still images. We set our base sampling rate to 100 frames per video. At this stage we also discarded frames that were black and white, too dark (mean HSV value < 39), too orange due to huge fires (mean HSV hue between 0 and 40 degrees), or too blurry (Laplacian variance < 400). We also resized all the images to 224x224 to reduce disk size, resulting in 56,486 images of size $\sim$25KB, totalling 1GB. We will refer to this dataset as **SmokeFrames-50k**.
- **Split dataset -** We divided our dataset into a 60-20-20 train-validation-test split. Frames from the same video were kept in the same split to ensure our model generalizes well to all new cameras.

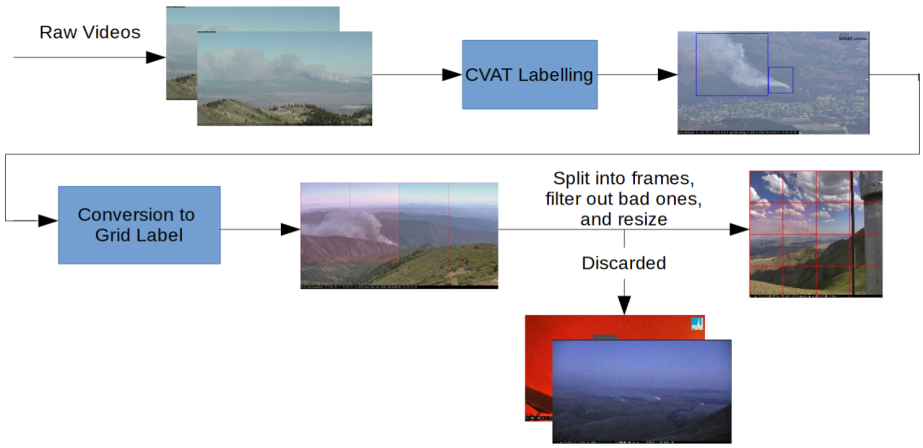
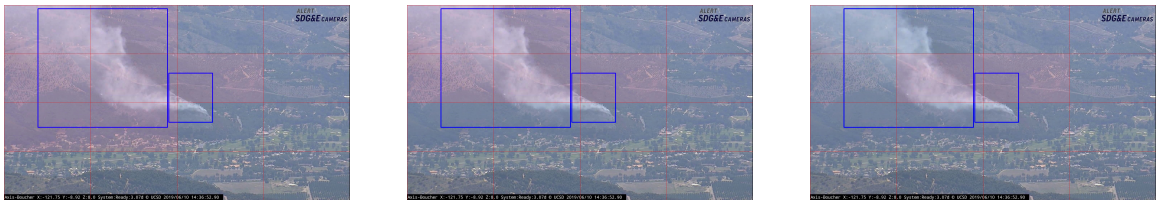


Figure 1: Diagram of the data pipeline. Raw videos were annotated using CVAT and then converted to grid labels. A series of filters were applied to maximize the quality of the dataset. The filtered images are the input to the network.

Dataset Information	
Number of videos	678
Number of frames	56,486
Number of frames with smoke	38,580

Table 1: Basic details of our **SmokeFrames-50k** dataset.

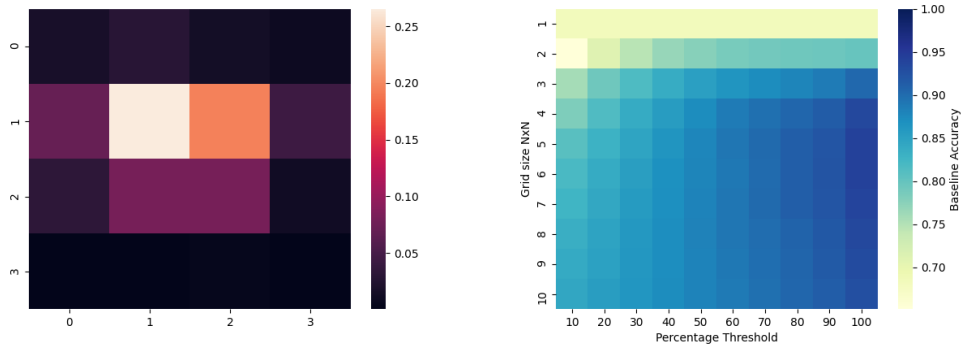


(a) 0.3 threshold (b) 0.5 threshold (c) 1 threshold

Figure 2: A sample annotated frame from our smoke dataset. The two blue-outlined boxes are the raw annotations and the shaded-red segments are the mapped labels for a 4x4 grid. The effect of the threshold is clearly seen across the three images, i.e. the proportion of each segment covered by the raw annotation in order to be mapped to 1, on the resulting grid labels. If a bounding box cannot meet any cell's threshold, it will activate the largest overlapping cell as shown in (c).

4.1 Dataset statistics

Our **SmokeFrames-50k** dataset contains smoke on 68.3% of its frames. This translates to different percentages of positive examples for each square on our grid labels, as well as different baseline accuracies, depending on the grid size and the coverage percentage threshold used to generate the multi-labels. Figure 3a shows the percentage of positive samples on a 4x4 grid with a 100% threshold and figure 3b shows the baseline accuracies for grid sizes 1-10 and thresholds 10-100%.



(a) (b)

Figure 3: a) Percentage of positive samples to total samples for each individual label on the 4x4 grid filtered dataset. b) Baseline accuracy of the dataset for different grid sizes and percentage thresholds.

5. Experiments

The basis of our models were convolutional neural networks (CNNs). The input layer of the network is fed with the preprocessed RGB images from our dataset. The final three layers of the network are 2D adaptive average pooling, a linear layer consisting of KxK nodes, i.e. the number of grid segments, and a sigmoid layer. Each of the CNN backbone architectures we explored share these same components and setup. We trained models both with randomly initialized weights and with pretrained weights. Batch size and number of epochs were set to 16 and 20 respectively. Prior to forward passes of the network, scaling and data augmentations are applied to the loaded training data. Each sample frame from the training data is applied contrast, brightness, and saturation transformations with random intensities.

To account for the imbalance of no-smoke to smoke samples as shown in figure (3a), we used a per-class weighted binary cross-entropy loss function. This weighting was computed for each of the individual grid segment class ratios. Additionally, since accuracy is often a misleading metric for imbalanced datasets (e.g. baseline of 90% accuracy), the metrics of interest were F1-score and ROC AUC which provide better insight to the model's ability to classify smoke vs. no-smoke. These metrics were computed on the validation set at a maximum of 200 evaluation steps per training with early stopping triggered after 10 evaluation steps of no improvement in the F1-score. The following section provides our methods, results, and analysis for the experiments.

5.1 Optimizing grid size and threshold

Two critical hyperparameters of our pipeline are the grid size and label calculation threshold parameters. We wanted to find out which combination of grid size and threshold give optimal results. As seen in figure 2, a lower threshold seems to be more effective at including all the smoke present, at least for a grid size 4x4. We hypothesized that perhaps higher thresholds would yield better results at higher grid sizes.

We experimented with grid sizes 1x1-10x10, and thresholds 10-100% (intervals of 10%) for each grid size, while maintaining all other hyperparameters static. Note that the results do not indicate the highest attainable accuracies, but rather are meant as a relative measure to determine the best grid sizes and thresholds.

As shown in figure 4, grid sizes 2x2 and 3x3 perform well at almost any threshold, grid sizes 4x4 through 8x8 generally get far better results at thresholds 10-30%, whereas grid sizes 9x9 and 10x10 do slightly better at high thresholds. This reinforces our initial hypothesis regarding higher thresholds, with mid-range thresholds yielding the worst results after grid size 3x3.

The model which outputted the best metrics in comparison to its respective baseline was the 2x2 grid. However, the model performed poorly on images containing only 1 true positive i.e. small smoke, mislabeling 3018 out of 3440 instances. These results are not too surprising since the coarser the grid mesh, the more noisy the signal will be and hence more difficult for the model to differentiate between background and smoke.

On the other hand, a super fine mesh will in essence draw out the raw bounding box annotations, as shown in 5, rendering the grid approach useless. We believe a fine balance can be reached between resolution of the mesh which determines the noise in the labels and adequate model performance for a smoke detection model. Also, we found the finer the mesh, the higher the baseline of predicting all 0s. This implies a higher class imbalance which generally translate to a bigger challenge for the network. For instance, a mesh size of 8 and threshold of 100% gives a baseline accuracy of 99% which means even if our model achieves 98% we cannot conclude we have successfully trained a smoke detection model.

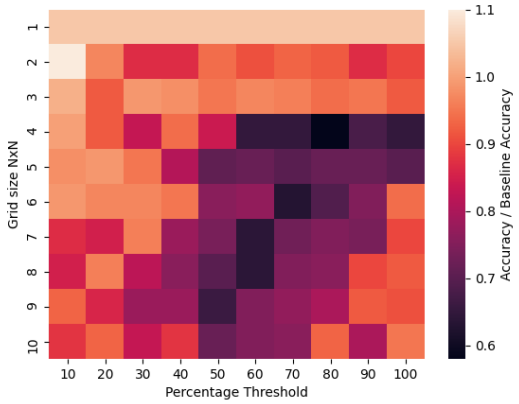


Figure 4: Improvement over baseline accuracy for models trained at 100 combinations of label grid sizes and grid calculation thresholds

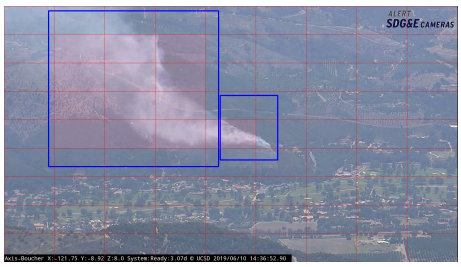
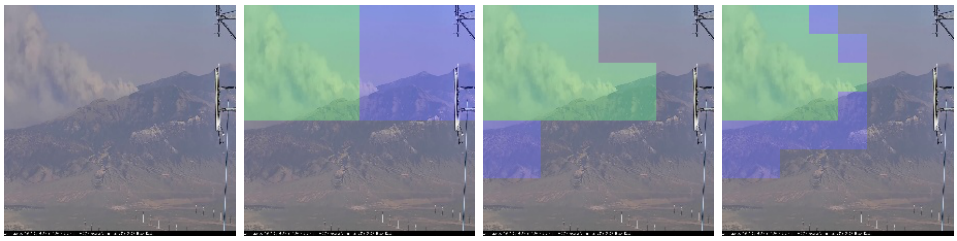


Figure 5: Labeled frame with grid size 9x9, threshold 100%



(a) Raw image (b) Grid size 2x2 (c) Grid size 4x4 (d) Grid size 8x8

Figure 6: Sample image from our test dataset comparing predictions from grid sizes 2x2, 4x4, and 8x8. True positives are shown as green shading, false positives are shown in blue, and true negatives are left as is. Note, there are no false negatives in these sample predictions. This is a great example where smoke is better bounded by a non-box annotation.

5.2 Model Architecture Comparisons

Following our investigation of mesh size and threshold, we explored several common CNN architectures for image classification tasks. A typical trade-off is between the discriminative power and the size of the architecture. While a bigger architecture often implies better model performance, inference

time grows as the number of model parameters increases. Below is a list discussing the various architectures we investigated. All models were initialized with weights pre-trained on the ImageNet dataset. Table 2 provides a summary of the evaluation results on the validation set.

- **MobileNet:** MobileNets [18] are lightweight, efficient networks targeted for mobile and embedded vision applications. The basis of the network are depthwise separable filters which replaces standard convolution filters, reducing computation cost by 8-9 times while minimally affecting model performance. We experimented with the MobileNetV1 architecture to consider the case of having embedded GPUs instead of a centralized model running on cloud computing resources.
- **Inception:** The Inception [19] architecture is an efficient deep neural network. Its release was a major advancement in the computer vision world due its highly optimized structure reducing memory and compute resources.
- **ResNet:** ResNet [20] was the winner of the 2015 ILSVRC classification task. The novelty of this architecture is the "skip" or "shortcut connections" connections to address the common accuracy degradation problem with deep networks.
- **ResNeXt:** ResNeXt [21] is a highly modularized network and exhibits a simple design for an image classification network.

Model	# of Parameters	Accuracy	F1-score	ROC-AUC
MobileNetV2	2 mill.	0.85	0.69	0.90
InceptionV3	21 mill.	0.84	0.67	0.89
ResNet-50	23 mill.	0.85	0.68	0.90
ResNeXt-101-32x8d	86 mill.	0.86	0.70	0.89

Table 2: Results from exploring various CNN backbone architectures. We used a mesh size of $4x4$ for the data and the baseline accuracy was 0.82. The best performing model was the ResNeXt network.

The model with the highest F1-score in the architecture search was ResNeXt, though the smaller MobileNetV2 network performed relatively well in comparison. Looking at the other metrics accuracy, and ROC-AUC it becomes less clear which model is truly better performing.

6. Future work/Impact

6.1 Future work

A major limitation of our model owes to the fact temporal information is completely discounted. Smoke has a particular dynamic behavior which differs from the common misclassified instances like clouds and fog. A study done in a Chinese province showed a combination of spatial and temporal modelling improved model performance for early smoke detection [14]. To account for temporal information, many more frames than standard image classification tasks need to be labeled, which makes acquiring the necessary labelled video data one of the hurdles for temporal modelling. Fortunately, we have already made this jump with our 678 smoke-annotated videos. In future experiments we plan to compare and quantify model improvement from incorporating temporal modelling layers such as LSTMs into our model.

Moving forward, we also aim to run further experiments to measure our models generalizability to other PTZ camera datasets including the AI for Mankind dataset. In section 3 we briefly discussed our dataset having a bias towards centered and zoomed-in smoke images. For future work, we plan to run several experiments to test our model's ability to detect smoke of smaller size and located in corners of images.

6.2 Impact discussion

We believe our work will prove to be highly beneficial for both researchers and government officials considering options for automated smoke detection systems. Our results demonstrated the feasibility of a multi-label approach to detecting smoke from PTZ camera data and should be highly regarded as a viable method. While we are not the first group of researchers to apply a gridded image approach for smoke detection, we are however the first to determine the relationship between model performance and parameters mesh size and label calculation threshold. On top of this, our results reinforced the notion that PTZ camera data is sufficient for a smoke detection system and does not necessarily require expensive multi-spectral sensors. Overall, this part of our contribution provides both a convincing motivation and a clear methodology for state services to develop a cheap, simple, and effective automated smoke detection system to enhance the current ALERTWildfire framework, with the potential to reduce or possibly eliminate extended volunteer hours and improve response time to wildfires.

Our secondary contribution is the release of the first publicly available large-scale smoke-detection video dataset consisting of 678 annotated clips captured with ALERTWildfire cameras. Since there is currently a lack of open-source annotated datasets for smoke detection tasks, this release will complement existing image datasets and will reduce the overhead associated with building an end-to-end smoke detection pipeline.

7. Conclusion

In this paper, we demonstrated the feasibility of smoke detection in PTZ camera data using a multi-label image classification approach. We applied a gridded image approach where the model predicts the grid segments containing smoke. We experimented with various mesh sizes to determine the effect on model performance and found even a coarse grid resolution of 4x4 results in accurate predictions. In addition, comparisons to different CNN architectures showed smaller models such as MobileNetV2 perform comparably to massive networks. Finally, we provided in-depth details to our data generation pipeline and have released our smoke-annotated video dataset to the open source community.

References

[1] M. D. Flannigan et al. "Forest Fires and Climate Change in the 21ST Century". In: *Mitigation and Adaptation Strategies for Global Change* 11.4 (July 2006), pp. 847–859. ISSN: 1573-1596. DOI: 10.1007/s11027-005-9020-7. URL: https://doi.org/10.1007/s11027-005-9020-7.

[2] *Governor's Wildfire-Related Proposals*. 2020. URL: https://lao.ca.gov/Publications/Report/4172.

[3] *About: Nevada Seismological Lab*. URL: http://www.alertwildfire.org/about.html.

[4] A. J. MacAuley et al. "A strategic review of the wildfire detection programme in Saskatchewan". In: (2004).

[5] M. Helen et al. "See No Evil: Cognitive Challenges of Security Surveillance and Monitoring". In: *Journal of Applied Research in Memory and Cognition* 6.3 (2017), pp. 230 –243. ISSN: 2211-3681. DOI: https://doi.org/10.1016/j.jarmac.2017.05.001. URL: http://www.sciencedirect.com/science/article/pii/S2211368117300207.

[6] M. P. Plucinski. "The timing of vegetation fire occurrence in a human landscape". en. In: *Fire Safety Journal* 67 (July 2014), pp. 42–52. ISSN: 0379-7112. DOI: 10.1016/j.firesaf.2014.05.012. URL: http://www.sciencedirect.com/science/article/pii/S0379711214000678 (visited on 11/24/2020).

[7] A. Filonenko et al. "Comparative study of modern convolutional neural networks for smoke detection on image data". In: *2017 10th International Conference on Human System Interactions (HSI)*. July 2017, pp. 64–68. DOI: 10.1109/HSI.2017.8004998.

[8] S. Frizzi et al. "Convolutional neural network for video fire and smoke detection". In: *IECON 2016 - 42nd Annual Conference of the IEEE Industrial Electronics Society*. Oct. 2016, pp. 877–882. DOI: 10.1109/IECON.2016.7793196.

[9] *IQ FireWatch: A Global Success Story*. 2020. URL: https://www.iq-firewatch.com/references.

[10] Vizzuality. "Forest Monitoring, Land Use Deforestation Trends". In: *Global Forest Watch* (2020). URL: https://www.globalforestwatch.org/.

[11] Simon Hohberg. "Wildfire smoke detection using convolutional neural networks". In: *21 Technical report, Freie Universitt Berlin, Berlin, Germany* (2015).

[12] Stuart Matthews et al. "Field evaluation of two image-based wildland fire detection systems". en. In: *Fire Safety Journal* 47 (2012), pp. 54–61. ISSN: 0379-7112. DOI: 10.1016/j.firesaf.2011.11.001. URL: http://www.sciencedirect.com/science/article/pii/S0379711211001457 (visited on 12/01/2020).

[13] Kinshuk Govil et al. "Preliminary Results from a Wildfire Detection System Using Deep Learning on Remote Camera Images". en. In: *Remote Sensing* 12.1 (Jan. 2020). Number: 1 Publisher: Multidisciplinary Digital Publishing Institute, p. 166. DOI: 10.3390/rs12010166. URL: https://www.mdpi.com/2072-4292/12/1/166 (visited on 11/29/2020).

[14] Yichao Cao et al. "An attention enhanced bidirectional LSTM for early forest fire smoke recognition". In: *IEEE Access* 7 (2019), pp. 154732–154742.

[15] NV Seismolab. *NV Seismolab*. URL: https://www.youtube.com/user/nvseismolab.

[16] *High Performance Wireless Research and Education Network (HPWREN)*. URL: http://hpwren.ucsd.edu/ (visited on 11/23/2020).

[17] FUEGO. *fuego*. URL: https://github.com/fuego-dev/firecam.

[18] Andrew G. Howard et al. *MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications*. 2017. arXiv: 1704.04861 [cs.CV].

[19] Christian Szegedy et al. *Going Deeper with Convolutions*. 2014. arXiv: 1409.4842 [cs.CV].

[20] Kaiming He et al. *Deep Residual Learning for Image Recognition*. 2015. arXiv: 1512.03385 [cs.CV].

[21] Saining Xie et al. *Aggregated Residual Transformations for Deep Neural Networks*. 2017. arXiv: 1611.05431 [cs.CV].