Segmentation of Low-Level Temporal Plume Patterns from IR Video

Rajeev Bhatt[1] M. Gokhan Uzunbas[2] Thai Hoang[2] Ozge C. Whiting[2]
[1]Pennsylvania State University, [2]Avitas Systems Inc.
[1]rza67@psu.edu [2]{mustafa.uzunbas, thai.hoang, ozge.whiting}@avitas-systems.com
www.avitas-systems.com

Abstract

In this paper, a method to segment out gas or steam plumes in IR videos collected from fixed cameras is presented. We propose a spatio-temporal U-Net architecture that captures deforming blobs of gas/steam plumes that have a unique temporal signature. In this task, the blob shapes are not semantically meaningful and change from frame to frame with no consistency across different exemplar plumes; however, there is spatial and temporal continuity in the way blobs deform suggesting a need for a low-level spatio-temporal segmentation network. The proposed method is compared to an LSTM-based segmentation network on a challenging IR video dataset collected in a controlled environment. In the controlled dataset there is motion due to steam plumes with deforming blob patterns as well as due to walking people with more structured high-level patterns. The experiments show that plume patterns are successfully segmented out with no confusion to moving people and the proposed spatio-temporal U-Net outperforms LSTM-based network in terms of pixelwise accuracy of output masks.

1. Introduction

Thermal monitoring and inspection of industrial assets using Long-wavelet Infrared (LWIR) or Mid-wavelet Infrared (MWIR) cameras is an important application with several use cases. Infrared modality is preferred over visible range as it provides more robustness against illumination changes, shadows and has obvious advantages when observed phenomena has a distinct thermal signature. Plume detection is one such application where methane and other types of gas leaks present a distinguishable pattern with good thermal contrast to background depending on the leak rate. There are several cameras that are specifically tailored to leak inspections [1]; however, the concept of operations still involves a human inspector observing the camera feed and manually segmenting out plumes if present. This is a labor-intensive procedure for the inspector and also prone to human errors. In this paper, a generic spatio-temporal segmentation framework is proposed to automate plume detection process where a plume segmentation mask is provided as output for each frame in an input video sequence. Accurate and detailed frame by frame segmentation masks can robustly signal existence of a leak as well as enabling quantification of leak rate by further analysis [2].

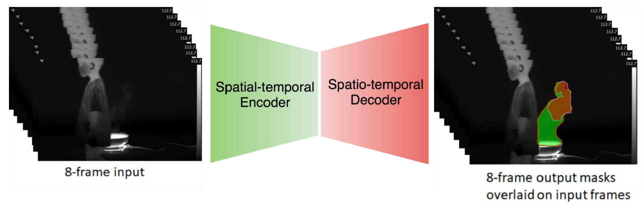

Figure 1: A sequence-to-sequence encoder-decoder model with the proposed temporal extension to U-Net. The output masks are different in each frame capturing the deformation of the plume in the input sequence. Note that the plume is water vapor in this case and it is barely visible to human eye. On the output mask, green pixels show the ground-truth plume annotation and red pixels show the plume prediction.

In this paper, a generic end-to-end deep learning framework is proposed to process a sequence of IR images captured from a fixed camera that delineates low level spatial patterns with a temporal coherence from other types of movement or from background. This network is designed for use cases where the camera is observing a portion of the facility where humans/vehicles can be occluding the scene from time to time or there are other types of spatio-temporal motion such as background thermal fluctuations due to vegetation, wind, reflections, etc. More specifically, a spatio-temporal encoder-decoder network extending a spatial U-Net [3] to temporal domain is proposed to create a pixel-level mask for each frame in an input sequence, see Figure 1. Note that the proposed algorithm is generic and would apply to segmentation of plumes from different types of gas or steam in LWIR or MWIR video in both hot-foreground or cold-foreground scenarios. The proposed algorithm is a supervised technique which requires frame by frame ground-truth annotations in a set of training sequences.

An alternative network architecture is to restrict the decoder and encoder to spatial domain while capturing a temporal signature in the encoding space, see Figure 2. Such an architecture can be accomplished as a combination of standard LSTM cells with spatial encoders and decoders which are commonly employed in semantic segmentation tasks such as DeConvNet [5] and SegNet [6]. With this architecture, temporal relations in the final high-level encoding space are captured by the *recurrent* structure. Such a network would cover use cases where the nature of temporal phenomena builds upon high abstractive concepts. For example, concepts such as head, limbs etc. which are observed in human pose space for purposes of human action recognition and segmentation on a per-frame basis. However, in our use case, the shape of the plume is not semantically meaningful, has low abstraction, high texture and changes from frame to frame with no consistency across different exemplar plumes. This phenomenon suggests using a spatio-temporal encoder-decoder scheme as proposed in this paper which can capture and isolate lower-level patterns.

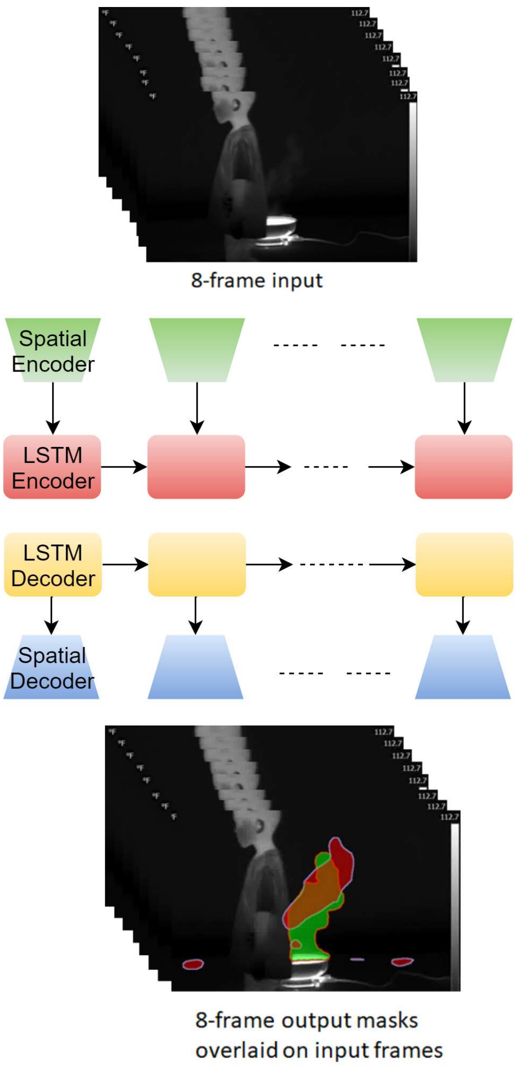

Figure 2: A sequence-to-sequence encoder-decoder with an alternative LSTM-based recurrent neural network which performs only spatial encoding and decoding on a per-frame basis and learns temporal patterns only in the encoding space. Note that the hidden state of decoder LSTM is initialized with the final state of encoder LSTM. The output masks are the ground-truth plume annotations and red pixels show the plume predictions.

In Section 3, we explain the network architecture of the proposed technique as well as the alternative LSTM-based segmentation network in more details. In Section 4, we explain details of our dataset: a set of 12 IR video sequences which are collected in a controlled environment where water vapor is escaping from a steamer in varying rates. This IR steam video database will be made publicly available in support of further research. In Section 5, we provide comparison of the methods proposed in this work and show that the spatio-temporal U-Net successfully segments out water vapor patterns without confusion to more structured moving human patterns that occlude the steam from time to time while outperforming LSTM-based technique.

In summary, the contributions of the paper are:

- Extending spatial U-net architecture to temporal domain for purposes of creating pixelwise segmentation masks from IR image sequences that capture low-level temporal phenomena such as deforming gas/steam plumes.
- A database of 12 IR video sequences captured in controlled environment with varying steamer settings and occlusion conditions.
- Comparative performance analysis of the proposed network on the IR steam video database where an LSTM-based temporal segmentation network is used as an alternative.

2. Related Work

Due to uncertainty and complexity of gas flow in leaky regions, early plume detection algorithms that use linear classifiers fail to capture the spatio-temporal patterns and result in a poor performance. In [23], temporal redundancy is used to estimate the location of plume followed by a Binary Partition Tree which is finally pruned according to the previous estimate to identify the extent of plume. The temporal redundancy assumption is no longer valid if the plume trajectory is non-linear.

Principal Component Analysis (PCA) is used in [24] to reduce the dimensions of multi-channel IR video to retain the most information. This is followed by image processing and clustering techniques to remove flicker for maintaining a consistent pixel signature. However, plumes that arise from small cross section or those that move slowly compared to the median velocity will be treated as noise by PCA which otherwise could be very important. For gases that travel at higher velocities or that have an abrupt acceleration in the initial frame, the pixel signature consistency can no longer be maintained.

Different image processing techniques such as Discrete Cosine Transform (DCT), Continuous Wavelet Transform (CWT), Hidden Markov Models (HMM) etc. were used to detect flame and smoke from videos [25-30]. [30] uses a Hidden-Markov model to analyze the temporal behavior of smoke but the stationarity assumption is no longer valid when spatio-temporal correlation is at play. Such statistical methods frequently fail when dealing with non-linear data because of underlying linearity assumptions in mathematical models and they suffer from the curse of dimensionality [24]. This motivates use of a CNN based technique which can capture non-linear relationships and learn complex problem-specific feature spaces to separate out background phenomena from plume patterns.

Abnormal plume patterns would be reflected in the fine-detail texture features that are present in the lower layers of a CNN. However, we should notice that an unknown pixel activation in the lower level layers can also be categorized as abnormal if it is present in the blob of abnormal plume patterns in higher levels of the abstraction hierarchy, i.e. in cases where the receptive field of the first layer is not large enough to capture the whole blob. So, in order to create a fine grained and accurate segmentation map, a U-Net [3] type architecture is needed which can exploit the lower layers for texture and higher layers for context. This idea has been explored in other tasks such as image restoration [9].

In plume detection use case, temporal low-level patterns also need to be learned from a stack of input frames. Action recognition literature provides examples of spatiotemporal feature extraction techniques. For example, [33] exploits 3D convolutions over spatial and temporal dimensions proving the value of the technique in comparison to 2D-only feature extraction on multiple frames followed by various temporal fusion techniques. Similar 3D convolution idea is also applied in IR domain [32] for action recognition task.

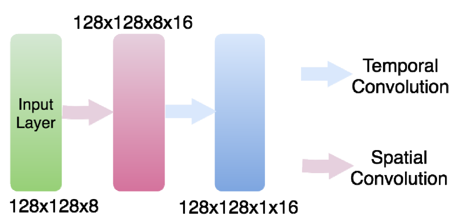

Figure 3: Input layer of proposed network with 8 single channel frames with size 128x128 where spatial convolutions of kernel size 3x3 and stride 1x1 are applied to all frames. Convolution kernels applied to all eight frames have shared weights but there are 16 sets of kernels producing an output tensor of size 128x128x8x16.

In the field of spatio-temporal data, videos are well-studied, for tasks such as video-prediction, human action recognition and pose estimation etc. Almost all of these propose a hybrid of two powerful neural network models the CNN, for capturing spatial information and the RNN, for capturing temporal information. Such a CONV-LSTM model was first proposed for precipitation nowcasting [10]. A CNN is first used to extract spatial features followed by an LSTM to learn temporal features of variation. Following their work, several attempts were made to combine a CNN and an LSTM for spatio-temporal feature extraction especially for combining computer vision and natural language processing tasks such as image/video description generation [11-13], visual activity recognition [14], video classification [15] and, traffic flow forecasting [16]. For predicting citywide crowd flows, [17] proposed a deep spatio-temporal residual network in which the input frames are grouped into distant, near and recent frames that are eventually fed through a CNN to obtain features from 3 different times.

Inspired from their success, spatio-temporal models are also applied to graph structured data, while a few are a direct extension of the models used for structured data, others are modified with graph convolutions. Structural-RNN [18], a spatio-temporal graph is modelled into a mixture of recurrent neural networks for human pose estimation. [19] uses a diffused convolutional layer for extracting spatial features from each timestamp of a graph structured input followed by an encoder-decoder LSTM network. A Graph Convolutional Network (GCN) is used in [20-21] to extract spatial features. While a gated CNN is used to extract temporal features in [20], spatial features are fed to a Gated recurrent unit (GRU) in ST-UNet of [21]. The U-net structure processes the data through multiple scales in order to understand the local and global properties in the input data for pixel-level tasks as introduced in [22]. It has the capacity to represent both the global information and also the locally distributed features. ST-UNet achieves the required multi-granularity using spatio-temporal Pool and Unpool operations. Note that though we picked the same abbreviation as ST-UNet for the architecture proposed in this paper, our ST-UNet is significantly different from [21]. The architecture proposed in this paper uses 3D convolutions to capture non-linear temporal features at multiple scales as opposed to a GRU.

3. Spatio-temporal Segmentation Networks

In this section, we will provide more detailed descriptions of proposed temporal extension to U-Net, Section 3.1, and the alternative LSTM-based encoder-decoder network, Section 3.2. Note that all the figures in these sections will be provided for input frames of size 128 by 128 but the input layers can be extended to any size since both networks entail only convolutions or deconvolutions and no fully connected networks. Also, we chose the length of image sequence as 8 frames without loss of generality.

3.1. Spatio-temporal U-Net (ST-UNET)

The proposed network consists of an encoder followed by a decoder unit that are comprised of spatial and temporal convolutions during the encoding phase and transpose convolutions during the decoding phase. The input layer consists of spatial convolutions as shown in Figure 3, generating eight activation maps $A^1, A^2, ..., A^8$ with shared weights applied to each of the eight frames. If the plume patterns exhibited linear temperature profiles as in Figure 4, then plume could be separated from other phenomena by just calculating the slope:

$$m = \frac{T_2 - T_1}{t_2 - t_1} = w_2 T_2 + w_1 T_1 \qquad (1)$$

Using any gradient descent optimizer, there exists weights w_1, w_2 which can be trained to represent $w_1 A^1 + w_2 A^2 + \cdots + w_8 A^8$ as the plume heat transfer pattern. A single 3D convolutional layer as in Figure 4 would be able to represent this sum and learn the necessary slopes in temporal dimension.

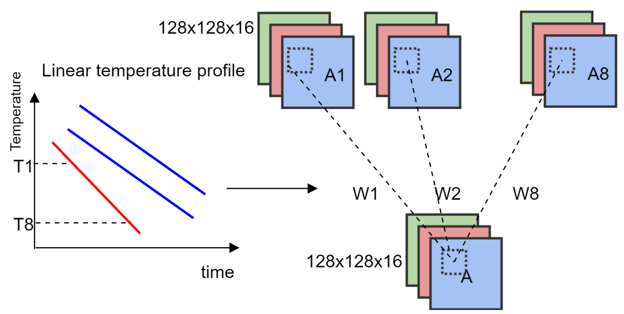

Figure 4: 3D convolution with one layer capable of capturing differences between patterns of linear temperature variation profiles.

However, there are more complex non-linear variations in temperature profile of plume blobs over time as illustrated in Figure 5. To mitigate this, we propose to add a hidden layer and introduce nonlinearity using nonlinear activation functions like sigmoid, ReLU etc.

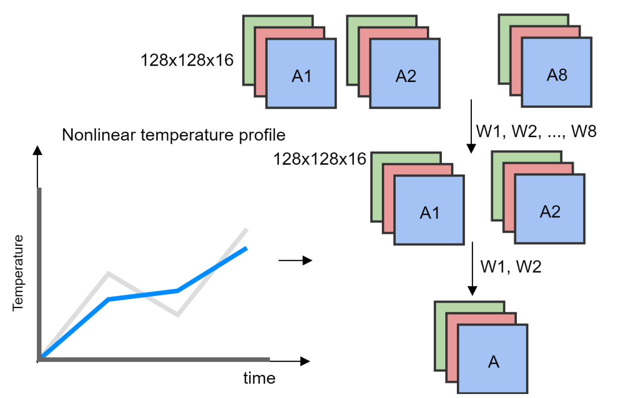

Figure 5: Neither target plume patterns nor other background phenomena do not necessarily follow linear temperature profiles over time. A 2-stage 3D convolution is applied to capture nonlinear patterns in temporal dimension.

For our architecture, we are using a 3D convolutional block with a hidden layer as shown in Figure 6. First stage consists of 16 filters with a kernel of size 3x1x1 and stride 2x1x1. Second 3D convolutional layer consists of 16 filters each with a kernel of size 3x1x1 and stride 2x1x1.

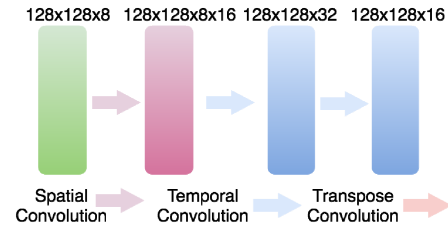

Figure 6: 3D convolutions applied in two stages to extract 16 temporal feature maps with a final tensor size of 128x128x1x16.

From here onwards, all the 2D convolutional layers have kernel size 3 x 3 and stride 2 x 2 to halve the resolution and the number of filters is doubled after each stage, see Figure 7.

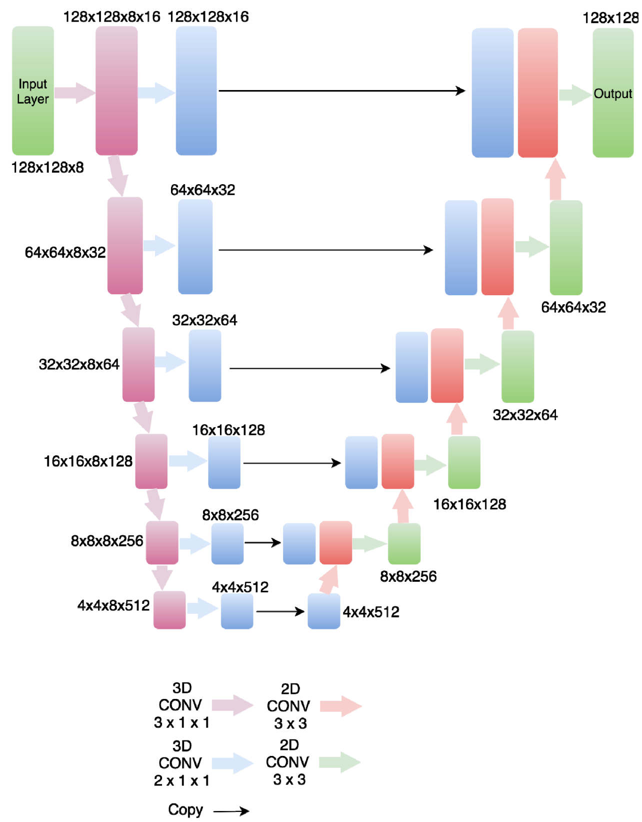

Figure 7: Full network architecture of proposed spatio-temporal U-Net (ST-UNet). Left side is the encoder network and right side is the decoder network. The arrows denote spatial, temporal and transpose convolution operations following the same color palette in Figure 6.

For the decoding stage, we have feature maps from 6 stages each varying in degree of texture and context they capture. Starting from the 4 x 4 x 512 feature map of stage-6 which captures the highest context (largest receptive field), we increase the resolution of the segmentation map incrementally by incorporating higher context from i^{th} stage with fine-detailed texture from $(i-1)^{th}$ stage via the operations shown in Figure 8. During the decoding stage, the 2D transpose convolutional layer has a kernel size of 3 x 3 and stride 2 x 2 to double the resolution and the 2D convolutional layer has a kernel size of 1 x 1 and stride 1 x 1.

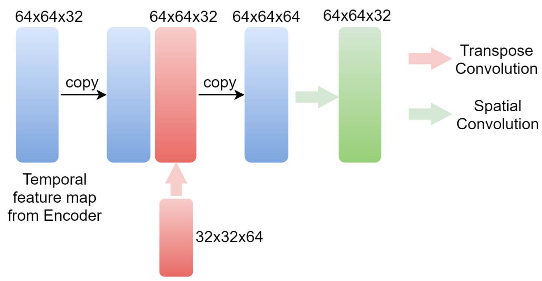

Figure 8: Decoding stage operations on a given layer where feature map size from previous layer is doubled via transpose convolutions and the temporal feature map of encoding layer is concatenated to the feature map of the current layer.

Figure 7 shows full architecture of the proposed network where the output is a binary mask per 8-frame sequence separating plume pattern from all other phenomena. The ground-truth mask of the last frame in an 8-frame sequence is used to train the network. During training, the energy function of U-Net is used where a pixelwise soft-max over the final feature map is combined with the cross-entropy loss function. Since this is a relatively shallow network, no weight initialization schemes with other tasks such as compression-decompression are used but rather all weights are initialized randomly.

Requiring one ground-truth mask per 8-frame sequence is advantageous as it reduces the amount of annotation required to create training datasets. However, an alternative network, ST-UNet-Full is also implemented with a slight variation to the decoder network architecture where the network outputs one mask for each frame in a given 8-frame input sequence. With this variation, all 8 ground-truth masks are used by the loss function to train the network. Although this change increases the network size considerably, the experiments in Section 5 show that pixelwise precision improves over ST-UNet.

3.2. LSTM-based Encoder-Decoder

In this section, the details of the recurrent model that will be compared to our proposed temporal U-Net architecture will be explained. This model utilizes convolutional LSTM cells as the bottle neck structure between the image feature encoder and predicting decoder that works as a memory unit between multiple consecutive input frames. Here, an RNN architecture is trained end-to-end in order to capture the dependency between image frames similar to Markov Random Processes. This helps the model in learning the conditional dependency rules for final decisions on the thermal images.

Each module of recurrent architecture is composed of an encoder, an LSTM unit, and a decoder, as shown in Figure 2. The encoder part of the model is a classical convolutional segmentation network encoder which is a very common backbone architecture called VGG-8 [31]. It is composed of 15 convolutional, max pooling and rectified linear unit (RELU) layers as shown below in the figure. We also use drop-out technique after the 7^{th} conv layer in order to reduce the overfitting to the training data.

Between the encoder-decoder network, the convolutional LSTM units are connected by operating on the low dimensional (8x8x2) embedded space where these embeddings store the useful spatio-temporal features. The LSTM unit outputs are connected to a decoding structure which is an end-to-end trainable network composed of deconvolutional layers that learns to propagate the encoded activations back to original input signal resolution in addition to inferencing pixel level prediction scores for the target labels. This architecture is also inherited from the decoder part of VGG8 [31] network. Since VGG is a relatively deeper architecture, a version of it with weights pre-trained on PASCAL VOC object segmentation challenge is used. Weights of LSTM layers were initialized randomly.

3.3. Training

All networks are trained on 3 GPU's with an effective batch size of 12 examples (each example has 8 frames) using TensorFlow ADAM optimizer with a learning rate of 2e-5. Batch normalization is applied after each layer for faster training and to reduce overfitting.

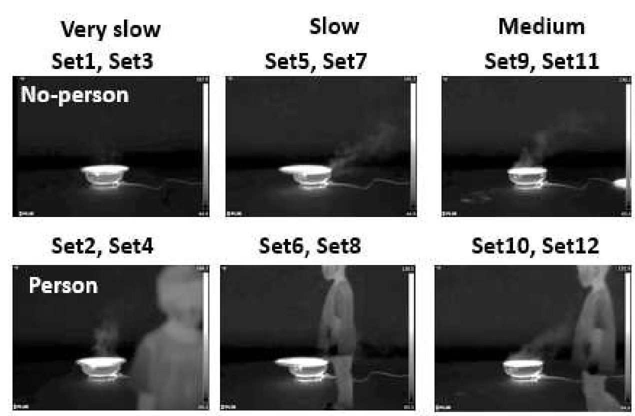

Figure 9: IR steam video dataset collected for the experiments in this paper. A total of 12 sets are collected with varying humidifier speeds and with a person walking and occluding the steam or not.

4. IR Steam Video Database

To support comparative study in this paper, a controlled environment is set up to collect a series of IR videos using a

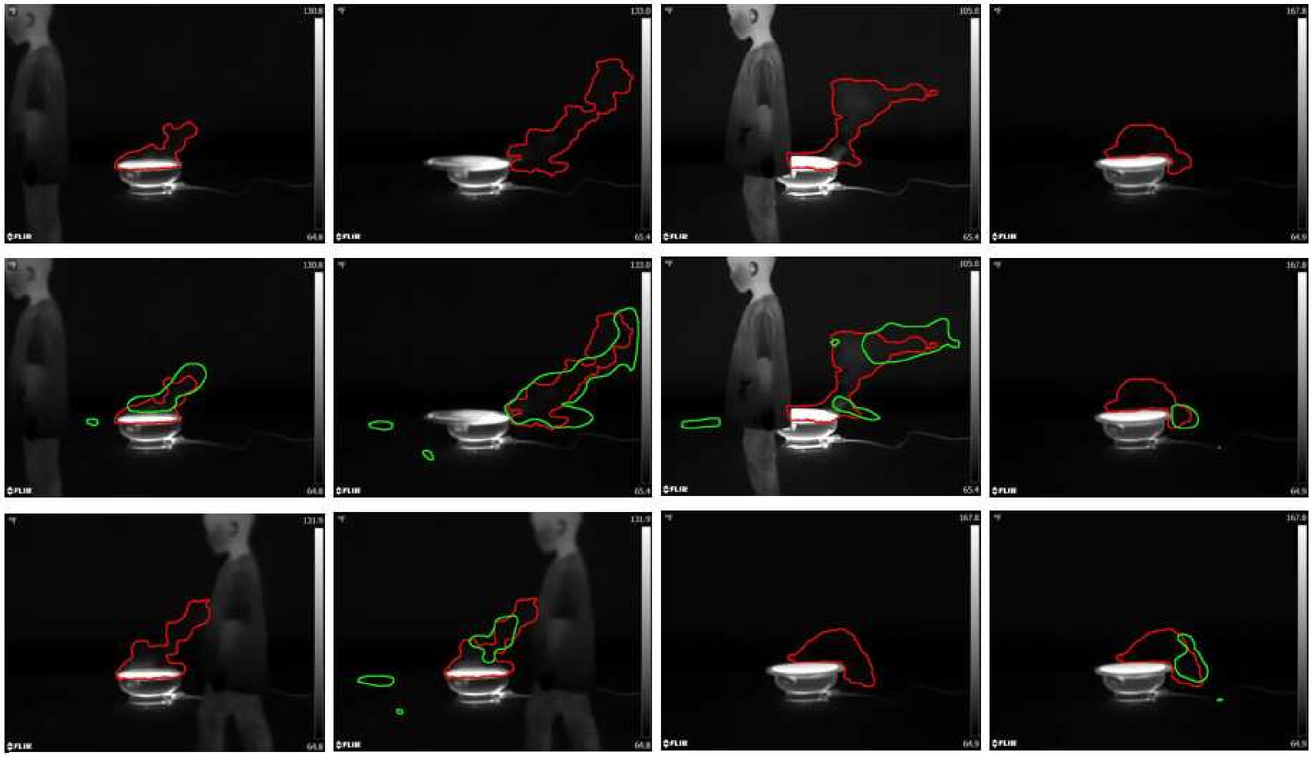

Figure 10: Results on selected frames for LSTM network. Red outlines are ground-truth plume annotations and green outlines are predictions of the network.

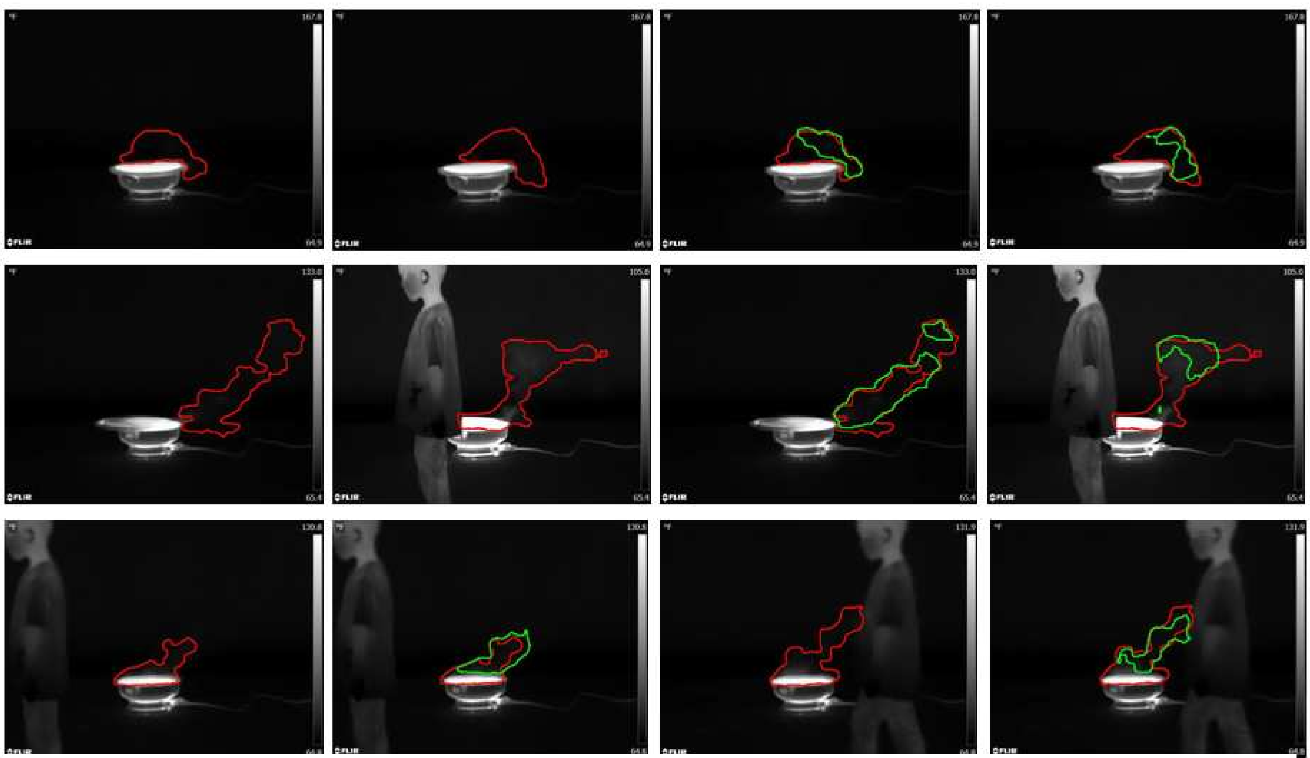

Figure 11: Results on selected frames for ST-UNet-Full network. Red outlines are ground-truth plume annotations and green outlines are predictions of the network.

FLIR T640 camera with a 41 mm lens. In this setup, a humidifier with adjustable speeds is used to generate varying rates and density of steam plumes. To introduce varying foreground conditions, a person walked in front of the steamer in some of the videos. A total of 12 videos are generated with 3 humidifier settings: very slow, slow and medium where 4 videos per setting are collected. For each setting, two of the videos contained a human walking in front of the camera occluding the humidifier and the plume. Figure 9 shows the names of videos as set1, 2, etc and the combinations of environment conditions. 20 to 30 second video clips are collected at 30 fps rate. The videos are scaled, and contrast enhanced using FLIR SDK with emissivity 0.95 and reflective temperature 77°F settings.

To support training and performance analysis, a total of 900 frames are manually annotated with the outline of the steam in Sets 1, 3, 4, 5, 6, 7, 9, 10 and 11. A total of 450 frames are manually annotated with the outline of the steam plume in Sets 2, 8 and 12 as a separate test set.

5. Experiments and Results

We verify the effectiveness of the proposed spatio-temporal U-Net on the IR steam video database. All three networks are trained on the annotated training frames by generating random exemplars of 8-frame sequences. Standard precision and recall measures at pixel level are calculated on 8-frame test sequences by comparing ground-truth manual annotations to network predictions. For fairness, only the last frame's ground-truth mask is compared to last frame's prediction for all three networks even though ST-UNet-Full generates a mask for all eight frames in the sequence. Figures 10 and 11 show prediction results as overlaid on ground-truth plume annotations for LSTM and ST-UNet-Full networks respectively. Table 1 summarizes prediction performance of all three networks on test sequences from three sets of IR videos. ST-UNet variations clearly outperform LSTM-based network especially in terms of precision. Even if full plume is not segmented out fully (around 70% recall), ST-UNet does a better job at overlapping with the ground-truth polygons (better than 80% precision). LSTM seems to be more vulnerable to some random perturbations in the background heat patterns generating false alarms at irrelevant parts of the frames. The results are especially impressive for Set2 where the plume is barely visible to human eye.

Table 1: Comparison of pixelwise mask prediction performance for proposed ST-UNet and ST-UNet-Full architectures with LSTM-based network on test video sequences.

Method	Set12		Set8		Set2	
	Precision	Recall	Precision	Recall	Precision	Recall
LSTM	0.70	0.62	0.73	0.76	0.60	0.57
ST-UNET	0.63	**0.68**	0.72	**0.85**	0.67	**0.86**
ST-UNET-Full	**0.82**	0.62	**0.84**	0.77	**0.83**	0.71

6. Conclusions

In this paper, a deep learning based low-level segmentation framework is proposed to detect gas/vapor plumes in IR videos. A new IR video dataset is generated by collecting sequences in a controlled environment where there is a combination of both low-level and high-level spatio-temporal motion patterns. The proposed technique is shown to isolate low-level plume patterns from high-level ones successfully on this dataset as well as outperforming an alternative LSTM-based spatio-temporal segmentation network.

References

[1] 1. *Infrared camera for methane and voc detection,* FLIR GF320. [Online]. Available: https://www.flir.com/products/gf320/.

[2] 2. Abdel-Moati, H., Morris, J., Zeng, Y., Kangas, P., & McGregor, D., "New Optical Gas Imaging Technology for Quantifying Fugitive Emission Rates," in *Proc. 2015 Int. Petroleum Tech. Conf.*

[3] 3. O. Ronneberger, P. Fischer, T. Brox, "U-net: Convolutional networks for biomedical image segmentation", *Proc. Med. Image Comput. Comput.-Assisted Intervention*, pp. 234-241, 2015.

[4] 4. J. Donahue, L.A. Hendricks, M. Rohrbach, S. Venugopalan, S. Guadarrama, K. Saenko, T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 677-691, April 2017.

[5] 5. H. Noh, S. Hong, B. Han, "Learning deconvolution network for semantic segmentation", *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 1520-1528, 2015.

[6] 6. V. Badrinarayanan, A. Kendall, R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for scene segmentation", *TPAMI*, 2017.

[7] 7. D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, "Learning spatiotemporal features with 3d convolutional networks", *Computer Vision (ICCV) 2015 IEEE International Conference on*, pp. 4489-4497, 2015.

[8] 8. T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, "Feature pyramid networks for object detection", *Proc. 2017 IEEE Conf. Comput. Vision Pattern Recog.*, pp. 936-944, 2017.

[9] 9. X.-J. Mao, C. Shen, Y.-B. Yang, "Image restoration using convolutional auto-encoders with symmetric skip connections", *Adv. Neural Inform. Process. Syst.*, vol. 29, pp. 2802-2810, 2016.

[10] 10. X. Shi, Z. Chen, H. Wang, D. Y. Yeung, W. K. Wong, W. C. Woo, "Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting", *NIPS*, 2015.

[11] 11. O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in Proc. IEEE Conf. Comput. Vis. Pattern Recog., 2015, pp. 3156–3164.

[12] 12. L. Yao et al., "Describing videos by exploiting temporal structure", *Proc. Int. Conf. Comput. Vis.*, Dec. 2015.

[13] 13. A. Peris, M. Bola˜nos, P. Radeva, and F. Casacuberta, "Video description using bidirectional recurrent neural networks," 2016. [Online]. Available: arXiv:1604.03390.

[14] 14. J. Donahue et al., Long-term recurrent convolutional networks for visual recognition and description, Nov. 2014, [online] Available: http://arxiv.org/abs/1411.4389.

[15] 15. Y.-G. Jiang, Z. Wu, Z. Li, X. Xue, S.-F. Chang, J. Tang, "Modeling multimodal clues in a hybrid deep learning framework for video classification", *IEEE Trans. Multimedia*, vol. 20, no. 11, pp. 3137-3147, Nov. 2018.

[16] 16. Y. Wu, H. Tan, "*Short-term traffic flow forecasting with spatial-temporal correlation in a hybrid deep learning framework*," 2016. [Online]. Available: arXiv:1612.01022.

[17] 17. J. B. Zhang, Y. Zheng, D. K. Qi, "Deep spatio-temporal residual networks for citywide crowd flows prediction", *Proc. 31st AAAI Conf. Artif. Intell.*, pp. 1655-1661, 2016.

[18] 18. A. Jain, A. R. Zamir, S. Savarese, A. Saxena, "Structural-RNN: Deep learning on spatio-temporal graphs", *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 5308-5317, 2016.

[19] 19. Y. Li, R. Yu, C. Shahabi, Y. Liu, "Diffusion convolutional recurrent neural network: Data-driven traffic forecasting", *Proc. Int. Learning Representations (ICLR '18)*, Jul. 2017.

[20] 20. B. Yu, H. Yin, Z. Zhu, "Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting", *Proc. 27th Int. Joint Conf. Artif. Intell. (IGCAI)*, pp. 3634-3640, 2018.

[21] 21. B. Yu, H. Yin, Z. Zhu, "ST-UNet: A Spatio-Temporal U-Network for Graph-structured Time Series Modeling," 2019. [Online]. Available: arXiv:1903.05631.

[22] 22. O. Ronneberger, P. Fischer, and T. Brox, "*U-net: Convolutional networks for biomedical image segmentation",* In Proc. Med. Image Comput. Comput.-Assisted Intervention, pp. 234–241, 2015.

[23] 23. Guillaume Tochon, Jocelyn Chanussot, Jérōme Gilles, Mauro Dalla Mura, Jen-Mei Chang, Andrea Bertozzi, "Gas Plume Detection and Tracking in Hyperspectral Video Sequences using Binary Partition Trees", *IEEE Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS 2014)*, 2014.

[24] 24. T. Gerhart, J. Sunu, L. Lieu, E. Merkurjev, J.-M. Chang, J. Gilles, A. L. Bertozzi, "Detection and tracking of gas plumes in LWIR hyperspectral video sequence data", *SPIE Conference on Defense Security and Sensing*, 2013.

[25] 25. CELIK Turgay, "Fast and efficient method for fire detection using image processing," ETRI journal, vol. 32, no. 6, pp. 881-890, 2010, eng.

[26] 26. P. Borges, E. Izquierdo, "A probabilistic approach for vision-based fire detection in videos", *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, pp. 721-731, May 2010.

[27] 27. S. Razmi, N. Saad, and V. Asirvadam, "Vision-Based Flame Detection: Motion Detection & Fire Analysis," IEEE Student Conference on Research and Development, pp. 187-191,2010.

[28] 28. Y. Habiboglu, O. Gunay, and A. Cetin, "Flame detection method in video using covariance descriptors," IEEE International Conference on Speech and Signal Processing, pp. 1817-1820, 2011.

[29] 29. B. Toreyin, A. Cetin, "Online Detection of Fire in Video," IEEE Conference on Computer Vision and Pattern Recognition, 2007, pp. 1-7.

[30] 30. B. Toreyin, Y. Dedeoglu, A. Cetin, "Contour based smoke detection in video using wavelets," European Signal Processing Conference (EUSIPCO), 2006.

[31] 31. K. Simonyan, A. Zisserman, "Very deep convolutional networks for large-scale image recognition", *Proc. Int. Conf. Learn. Representations*, 2015.

[32] 32. Z. Jiang, V. Rozgic, S. Adali, "Learning spatiotemporal features for infrared action recognition with 3-D convolutional neural networks", *Proc. Comput. Vis. Pattern Recognit. Workshops*, pp. 309-317, Jul. 2017.

[33] 33. D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, "Learning spatiotemporal features with 3D convolutional networks", *Proc. IEEE Conf. Comput. Vis.*, pp. 4489-4497, 2015.