Motion-Aware Feature for Improved Video Anomaly Detection

Yi Zhu
yzhu25@ucmerced.edu

Shawn Newsam
snewsam@ucmerced.edu

Electrical Engineering and Computer Science
University of California, Merced
Merced, US

Abstract

Motivated by our observation that motion information is the key to good anomaly detection performance in video, we propose a temporal augmented network to learn a motion-aware feature. This feature alone can achieve competitive performance with previous state-of-the-art methods, and when combined with them, can achieve significant performance improvements. Furthermore, we incorporate temporal context into the Multiple Instance Learning (MIL) ranking model by using an attention block. The learned attention weights can help to differentiate between anomalous and normal video segments better. With the proposed motion-aware feature and the temporal MIL ranking model, we outperform previous approaches by a large margin on both anomaly detection and anomalous action recognition tasks in the UCF Crime dataset.

1 Introduction

Anomaly detection in video is one of the long standing problems in computer vision and has extensive applications in surveillance monitoring, such as detecting illegal activities, traffic accidents and unusual events etc. Millions of surveillance cameras are being deployed in public places worldwide. However, most of the cameras are just passively recording without actually having any monitoring capability. With petabytes of data generated by the video cameras every minute, it is not possible to understand this large corpus of video data through human effort. We need machine vision to automatically detect anomalies within a video.

Recognizing anomaly in unconstrained videos is extremely hard. The challenges include insufficient annotated data due to the rare occurrence of anomalies, large inter/intra class variations, subjective definition of anomalous events, low resolution of surveillance videos, etc. As humans, we recognize anomalies using our common sense. For example, if multiple people crowd in a street that usually has less traffic, there maybe an anomaly. If violent events such as fighting happen, there maybe an anomaly. For machines, they don't have common sense but only have visual features. In general, the stronger the visual features, the better the anomaly detection performance is expected. In this work, we demonstrate how to obtain strong visual features by incorporating motion information.

Previous work [3, 10, 12, 24, 26] uses either hand crafted features or deep learned features to detect anomalies. Since their performance are reported on different datasets, we conduct an experiment here to make a fair comparison among these features. We evaluate

Features	Network	Dimension	Motion	AUC (%)
VGG16 [14]	deep	4096		68.7
C3D [18]	deep	4096	✓	74.4
Inception [17]	very deep	1024		69.2
I3D [2]	very deep	1024	✓	75.4

Table 1: Evaluation of different features on the UCF Crime dataset [15]. *Motion* indicates whether temporal information is involved. We observe that features incorporating motion information (C3D and I3D) perform much better than features extracted from individual images (VGG16 and Inception), regardless of network depth and feature dimension.

on the UCF Crime dataset [15], a recently released large-scale real world video anomaly benchmark. We adopt the Multiple Instance Learning (MIL) framework proposed in [15] to report the corresponding Area Under the receiver operating characteristic Curve (AUC), while changing only the input features. For volume-based features such as C3D [18] and I3D [2], the input to the network is a 16-frame video clip. For image-based features such as VGG16 [14] and Inception [17], we input the same 16-frame video clip as a mini-batch to the network and average the features. As can be seen in Table 1, we have an important observation that volume-based features that incorporate motion information perform much better than image-based features, regardless of network depth and feature dimension. This intuitively makes sense because most anomalies are irregular abrupt motion patterns, and motion-aware features should be more suitable to detect such events.

Motivated by the above observation, our goal is to learn a strong visual feature by incorporating as much temporal information as we can from the raw video frames. In this work, we propose a temporal augmented network to learn motion-aware features in an unsupervised manner. Our learned feature is efficient to compute, and shown to be competitive with other deep learned features such as C3D [18]. When combined with other features, we obtain a significant performance improvement. Our contributions are as below.

- We propose a temporal augmented network to learn motion-aware features. Such features are shown to be complementary to existing features.

- We introduce an attention-based temporal MIL ranking model, which can take temporal context into the picture and differentiate between anomalous and normal events better.

- We compare with and outperform several state-of-the-art approaches on both anomaly detection and anomalous action recognition tasks in the UCF Crime dataset.

2   Related Work

Here, we discuss additional work related to ours, focusing mainly on the temporal modeling of videos. Video is more than just a stack of images. Modeling the temporal relationship among frames can help understanding the video better. Initial attempts use tracking to design hand crafted features, such as IDT [20]. Recent deep learning based methods use temporal convolution [19], 3D convolution [2, 18], temporal segment networks [21], two-stream networks [13] etc. Among them, two-stream based approaches using optical flow are the top performers on most video benchmarks.
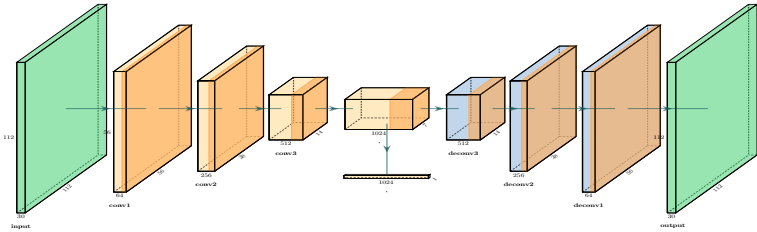
Figure 1: Temporal augmented network. The input (green) is a stack of 15 optical flow maps which this network aims to reconstruct by learning a compact representation. We then use a global average pooling operation to derive our 1024×1 motion-aware feature.

In this work, instead of directly using optical flow, we propose a temporal augmented network as an autoencoder to learn a compact motion-aware feature. This feature is generic, efficient and can be easily integrated with other methods using early fusion. We also incorporate temporal context into classical MIL ranking models by using an attention mechanism. The most similar literature to ours is [15, 23], however, there are several differences. [23] experiments with small scale datasets which does not require MIL formulation, while ours introduces a temporal MIL framework with an attention module. [15] serves as a baseline where we show substantial improvements by our proposed techniques. We fully exploit the temporal constraints within a video for improved anomaly recognition and detection. At the same time, our whole framework runs faster than real-time, which makes it directly applicable to real world problems.

3 Methodology

3.1 Problem Formulation

Given a long untrimmed video, we want to know whether it contains an anomalous event and where the event happens. Due to the massive amount of video recordings and the rare occurrence of anomalies, it is very challenging and costly to obtain precise frame-level annotations to train a powerful neural network. Most video anomaly detection datasets [11, 15] only provide video-level labels. Hence, in this work, we need to develop a weakly supervised approach using such datasets. Our goal is to learn a regressor that can predict the anomaly score for a video clip and detect possible anomalous event within a video.

3.2 Temporal Augmented Network

As we know, volume-based features such as C3D and I3D are computed on multiple video frames using 3D convolutions. They already contain temporal information. That is the reason they outperform image-based features such as VGG16 and Inception. However, as shown in recent action recognition literature [2, 21, 22], volume-based features alone cannot achieve state-of-the-art performance. Combining them with optical flow, as in the popular two-stream network [13], performs the best on most video classification benchmarks. This indicates that learning spatiotemporal features directly from raw video frames is challenging. Extra motion information such as optical flow can help.

Motivated by this observation, we want to learn a motion-aware feature that can complement existing features for improved video anomaly detection. In this work, we propose a
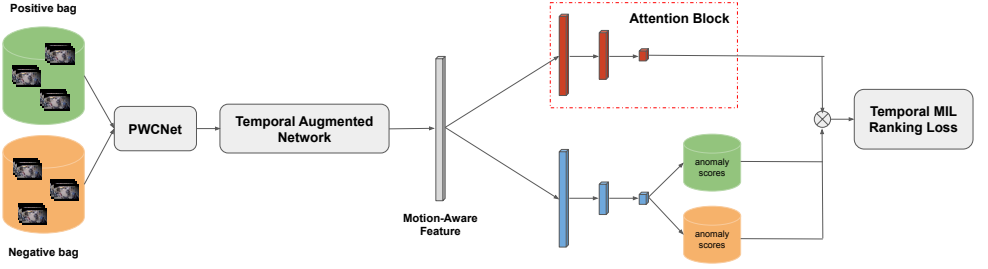
Figure 2: Overall framework. We first obtain the motion-aware feature and then compute the predicted anomaly scores. The attention block is used together with the proposed temporal MIL ranking loss to incorporate temporal context into training for better anomaly detection.

temporal augmented network as shown in Figure 1. The network is an autoencoder. Its input is some prior motion information pre-computed from raw video frames, such as optical flow. This forces the network to directly learn complex motion patterns. Then we aim to encode a compact representation so that we can use it to recover the input as closely as possible. This representation is our motion-aware feature and can be used to detect video anomalies.

Since optical flow is the most widely adopted motion representation, we use it as the input to the autoencoder. Specifically, we use the state-of-the-art neural network-based flow estimator PWCNet [16] to compute the optical flow between adjacent frames. We also compare several other motion representations in Section 5. Similar to C3D, we choose 16 frames as a video clip $\mathcal{V}$ and resize them to a resolution of 112×112. We then compute the optical flow on the resized frames. Each optical flow map has two channels, one for horizontal movement and the other for vertical. Hence, the final input to our temporal augmented network is a stack of 15 optical flow maps $\mathcal{F}$ with the dimension of $30 \times 112 \times 112$.

Bearing efficiency in mind, we design the temporal augmented network to have only 7 layers: 3 encoder layers, 1 bottleneck layer and 3 decoder layers. All layers consist of a 2D convolutional layer followed by a ReLU activation. We use a stride of 2 to halve the feature map resolution instead of pooling. The network can be trained in an unsupervised manner on the target dataset using L1 per-pixel reconstruction loss,

$$\text{loss}_{\text{recon}} = |\mathcal{F} - \tilde{\mathcal{F}}| \tag{1}$$

where $\tilde{\mathcal{F}}$ is the reconstructed flow map. Once the training is completed, we can treat it as a feature extractor. For each video clip with 16 frames, we perform a forward pass until the bottleneck layer and conduct a global average pooling operation to derive a 1024×1 feature. This will be our motion-aware feature for anomaly detection. If we want to use it with other features, we can simply concatenate them together. Note that the motion-aware feature is learned from optical flow, hence it contains only motion information without looking at the original frame pixels. We don't perform spatiotemporal feature learning. This will help the network to focus on the moving parts and learn appearance-invariant features.

3.3   Attention-based Temporal MIL Ranking Model

MIL Formulation   Since the precise temporal locations of anomalous events in videos are unknown, we cannot simply learn anomaly patterns like in a standard classification problem. Instead, we can treat it as a Multiple Instance Learning (MIL) problem.

In our scenario, we only have video-level annotations. A video containing anomalies is labeled as positive and a normal video is labeled as negative. Following [15], we represent a positive video as a positive bag $\mathcal{B}_a$, where different temporal segments are individual instances in the bag, $(a^1, a^2, ..., a^m)$, where m is the number of instances in the bag. We assume that at least one of these instances contains the anomaly. Similarly, the negative video is denoted by a negative bag, $\mathcal{B}_n$, where temporal segments in this bag are negative instances $(n^1, n^2, ..., n^m)$. In the negative bag, none of the instances contain an anomaly. In this work, we divide each video into a fixed number of segments (e.g., 32 segments) during training. These segments of a video are the instances in a bag.

MIL Ranking Model Following previous work [15], we formulate anomaly detection as an anomaly score regression problem. We hope the segments from an anomalous video to have higher anomaly scores than the segments from a normal video. If we have the segment-level annotations, we can simply use a ranking loss as

$$f(\mathcal{V}_a) > f(\mathcal{V}_n), \qquad (2)$$

where $\mathcal{V}_a$ and $\mathcal{V}_n$ are anomalous and normal video segments. f is the function that maps a video segment to its corresponding predicted anomaly scores ranging from 0 to 1. Here, f is designed to be a 3-layer fully-connected neural network. The first fully-connected layer has 512 units followed by 32 unit and 1 unit fully-connected layers. Dropout regularization is used between these layers. We use ReLU activation and Sigmoid activation for the first and the last fully-connected layers, respectively. However, we only have access to video-level annotations. [15] thus proposed a MIL ranking loss

$$\max_{i \in \mathcal{B}_a} f(\mathcal{V}_a^i) > \max_{i \in \mathcal{B}_n} f(\mathcal{V}_n^i). \qquad (3)$$

Here, *max* is taken over all video segments in each bag. The intuition behind this ranking objective is that the segment with highest anomaly score in the positive bag should rank higher than the segment with highest anomaly score in the negative bag because a negative bag does not contain any anomaly. In order to keep a large margin between the positive and negative instances, [15] introduced a hinge-based ranking loss

$$l(\mathcal{B}_a, \mathcal{B}_n) = \max(0, 1 - \max_{i \in \mathcal{B}_a} f(\mathcal{V}_a^i) + \max_{i \in \mathcal{B}_n} f(\mathcal{V}_n^i)). \qquad (4)$$

However, there are at least two limitations to this ranking loss. First, we note that Equation 3 ignores the underlying temporal structure of the anomalous video. A single max operation is not expressive. There maybe an anomalous video that contains multiple anomaly events. For a normal video, some segments could also look anomalous. Reasoning on temporal context should be useful to differentiate anomalous and normal video segments better. Second, the hinge-based ranking loss in Equation 4 can easily lead to a degenerate solution where we predict most video segments to be normal.

Temporal MIL Ranking Model In this section, we introduce our temporal MIL ranking model by using temporal context information. Motivated by the above limitations, we turn to an attention-based framework which can capture the total anomaly score of a video,

$$\sum_{i \in \mathcal{B}_a} w_i f(\mathcal{V}_a^i) > \sum_{i \in \mathcal{B}_n} w_i f(\mathcal{V}_n^i). \qquad (5)$$

subfigure

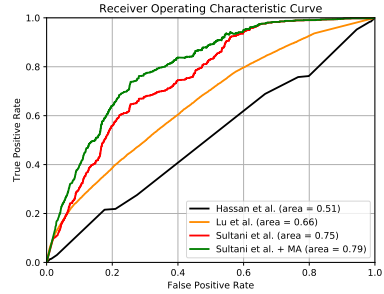| Method	AUC (%)
Hasan et al. [4]	50.6
Lu et al. [9]	65.5
Sultani et al. [15]	75.4
MA	72.1
Hasan et al. [4] + MA	62.7
Lu et al. [9] + MA	73.4
Sultani et al. [15] + MA	**79.0**

Table 2: Performance comparison on the UCF Crime dataset. *MA* indicates our motion-aware feature learned from temporal augmented network. Left: Comparison to state-of-the-art approaches. Our motion-aware feature complements existing methods for better anomaly detection. Right: Visual comparison in terms of ROC and AUC. [15] with motion-aware feature (green) achieves higher true positive rates than without (red).

where w_i indicates the learned attention weights. The intuition is, the overall anomaly score of an anomaly video should be larger than that of a normal video. We should include temporal context into consideration and compute the anomaly score video-wise, not segment-wise.

The attention weights are learned end-to-end within the network. As can be seen in Figure 2, we add an attention block after the input features. The block consists of three fully-connected layers and two tanh activations in between. The first fully-connected layer has 256 units followed by 64 unit and 1 unit fully-connected layers. For each video with m segments, we will learn a $1 \times m$ attention score for all the segments. Similar to Equation 4, our hinge-based temporal ranking loss is defined as

$$l(\mathcal{B}_a, \mathcal{B}_n) = \max(0, 1 - \sum_{i \in \mathcal{B}_a} w_i f(\mathcal{V}_a^i) + \sum_{i \in \mathcal{B}_n} w_i f(\mathcal{V}_n^i)). \tag{6}$$

We also employ the sparsity constraints [15, 27] because anomalies occur only rarely. There should be only a few segments that have high anomaly score. In the end, our final loss function becomes

$$\text{Loss} = l(\mathcal{B}_a, \mathcal{B}_n) + \lambda_1 \sum_{i \in \mathcal{B}_a} w_i f(\mathcal{V}_a^i). \tag{7}$$

λ_1 is the loss weight for the sparsity constraint. Note that we do not use the temporal smoothness constraint introduced in [15]. We empirically find it harmful for model training. Our overall framework can be seen in Figure 2.

4 Experiments

4.1 Dataset

Previous datasets [8, 9, 11] for video anomaly detection are either small in terms of the number of videos or have limited anomaly classes. Since we are doing comparisons among multiple features, we need a large, diverse and balanced dataset to reach a convincing conclusion. We use a recently released large-scale real world anomaly detection benchmark, UCF Crime [15], to evaluate our model and design choices. This dataset consists of 1900
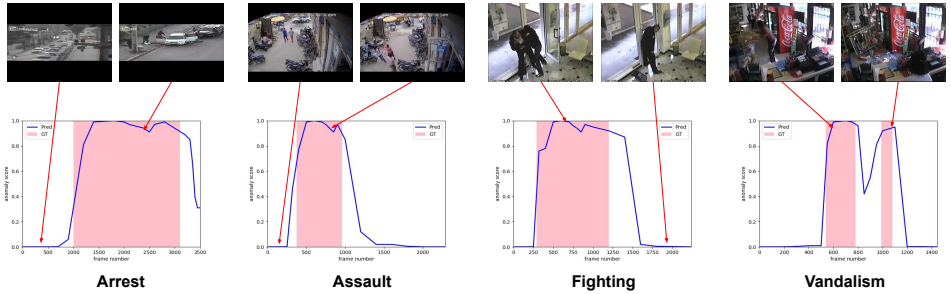
Figure 3: Visual examples of prediction results. For the anomalous frames, our model is able to provide accurate detection by generating high anomaly scores. For the normal frames, our model consistently produces low anomaly scores.

real-world surveillance videos, half of which contain anomalous events and the other half normal activities. For the anomalous videos, there are 13 different classes, including *Abuse, Arrest, Arson, Assault, Accident, Burglary, Explosion, Fighting, Robbery, Shooting, Stealing, Shoplifting, and Vandalism*. The official training split divides the dataset into two parts: the training set consisting of 800 normal and 810 anomalous videos and the testing set including the remaining 150 normal and 140 anomalous videos. Following previous works [15], we use frame based receiver operating characteristic (ROC) curves and corresponding area under the curve (AUC) to evaluate the performance of our method.

4.2 Implementation Details

We use the PyTorch framework to train our model. For the temporal augmented network, we randomly select video clips of 16 frames and use PWCNet [16] to compute the optical flow. The batch size is set to 50. We use the Adagrad optimizer with an initial learning rate of 0.005. We train the model for a total of 50K iterations, and decrease the learning rate by half at 25K, 40K and stop at 50K. For the MIL ranking model, we first divide each video into 32 non-overlapping segments. If the video has less than 32 frames, we duplicate its frames. Within each segment, we compute our motion-aware feature for every non-overlapping 16-frame video clip. If the segment has multiple 16-frame video clips, we take the average of all features followed by a L2 normalization. Hence, for each video, we have a 32×1024 feature. To train the MIL ranking model, we randomly select 30 positive and 30 negative bags as a mini-batch. We use the Adagrad optimizer with an initial learning rate of 0.001. We train the model for a total of 10K iterations, and decrease the learning rate by half at 4K, 8K and stop at 10K. λ_1 is set to 8×10^{-5}. For all other features used in this paper such as C3D and I3D, we adopt the implementation kindly provided by the original authors [2, 18].

4.3 Results

We present our results in Table 2. We compare our method with a state-of-the-art approach [15] and two other baselines [4, 9] for anomaly detection. In order to make fair comparisons, we keep the model training setting the same.

As we can see in Table 2 left, our motion-aware feature learned from the temporal augmented network achieves competitive performance with the previous best [15] in terms of anomaly detection AUC score (72.1 VS 75.4), but has smaller size (1024-dim VS 4096-dim)

Method	AUC (%)
MA (max)	70.6
MA (attention)	72.1
[15] (max)	75.4
[15] (attention)	76.2
[15] + MA (max)	77.1
[15] + MA (attention)	79.0

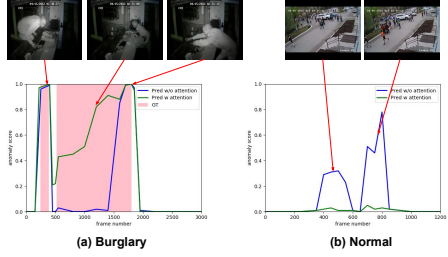(a) Burglary   (b) Normal

Table 3: Attention is useful. Left: Quantitative results. Right: Two visual examples. Temporal context can help to differentiate between anomalous and normal events better.

and faster speed (400+ fps VS 300+ fps). When combined with [15], we can achieve a performance improvement of 3.6% (75.4 → 79.0). As for per-class breakdown, we observe that classes with fast motion benefit a lot from our motion-aware feature. For example, Arrest (46.0 → 57.4), Assault (20.8 → 41.2) and Fighting (32.4 → 47.1). Similarly, when combined with [4, 9], we are able to get significant performance boosts of 12.1% and 7.9%, respectively. This demonstrates the effectiveness of our learned motion-aware feature.

In terms of visualization, we show the comparison of ROC curves in Table 2 right. We can see that [15] with our motion-aware feature (green) achieves higher true positive rates than without (red) at low false positive rates. This will help to reduce the false alarm rate.

We also combine our motion-aware feature with other widely adopted features such as VGG16, Inception and I3D. We observe consistent improvements: VGG16 (68.7 → 74.2), Inception (69.2 → 74.9) and I3D (75.4 → 79.8). The large improvements indicate the strong complementarity of our feature. At the same time, we may conclude that motion patterns are strong indicators for detecting anomalies. The more motion information captured in the visual feature, the better performance we will have.

In Figure 3, we show several visual examples of our qualitative results. We can see that for the anomalous frames, our model is able to provide successful and timely detection by generating high anomaly scores. For the normal frames in which no anomaly occurs, our model consistently produces low (almost zero) anomaly scores.

5   Discussion

5.1   Effectiveness of Attention Mechanism

In this section, we investigate the effectiveness of the attention mechanism in temporal MIL ranking model, to see the benefit of using Equation 6 over Equation 4. As can be seen in Table 3 left, adding attention consistently brings us 1% to 2% AUC improvement. We believe temporal context plays a key role here to differentiate anomalous and normal events. In terms of visualization, we show two examples in Table 3 right. The first video contains burglary events. Without attention, the model (blue) fails to report the anomaly from frame 500 to 1600 which looks normal. After adding attention, our method can detect the anomalous event there (green) because it has knowledge of the temporal context. The second video doesn't contain any anomalies but there are people grouping and running in the middle. Without attention, the model classifies the middle two parts as anomalies (high spike of blue curve). After incorporating attention, the model stops producing high anomaly scores for those parts (green).

Method	Speed (fps)	AUC (%)
TVL1 [25]*	102	71.7
FlowFields [1]	3	70.6
FlowNet2 [6]*	185	73.2
PWCNet [16]*	767	72.1
Motion Vector	> 1000	51.8
Video Saliency [7]*	523	56.9

Method	Accuracy (%)
Motion-Aware	20.2
C3D [18]	23.0
C3D + Motion-Aware	**26.1**
TCNN [5]	28.4
TCNN + Motion-Aware	**31.0**

Table 4: Left: Ablation study on motion representations, the input to our temporal augmented network. All speeds are evaluated on an image of resolution 112×112. The speed only includes the time to compute the motion representation. * indicates the method uses a GPU for inference. Right: Anomalous activity recognition experiments. Our motion-aware feature can complement state-of-the-art video features and lead to large performance improvements.

5.2 Ablation Study on Motion Representations

Recall from Section 3 that any motion representation can be fed to our temporal augmented network as input. There are many flow estimators, such as TVL1, FlowFields, FlowNet2 and PWCNet etc. Besides optical flow, we also have other motion representations, such as motion vectors and video saliency etc. Here, we perform an ablation study among these representations to see which one is the most effective.

First, we compare different flow estimators. TVL1 [25] and FlowFields [1] are classical methods, while FlowNet2 [6] and PWCNet [16] are neural network based methods. As can be seen in Table 4 left, FlowNet2 achieves the best AUC score due to its accurate and sharp flow predictions. However, it is relatively slow to compute. PWCNet is a good trade-off, performing competitively to FlowNet2 but running significantly faster.

Second, we compare different types of motion representations: motion vectors, optical flow and video saliency. Here, we use PWCNet [16] to compute optical flow, and we use a state-of-the-art method [7] to obtain video saliency. As can be seen in Table 4 left, PWCNet achieves the best performance. Both the motion vectors and video saliency perform badly. We find that the resolution of motion vectors is too coarse to extract useful motion information. For video saliency, we observe that the predictions are not consistent across frames thus may complicate the learning process of temporal augmented network.

5.3 Anomalous Activity Recognition Experiments

To further demonstrate the generalizability of our motion-aware feature, we use the same dataset to conduct anomalous action recognition experiments. Following the official setting, there are 4 splits and we report the average recognition accuracy. As can be seen in Table 4 right, our motion-aware feature alone can achieve reasonable performance. When combined with other state-of-the-art video features, we can obtain large performance improvements, 3.1% for C3D and 2.6% for TCNN respectively.

6 Conclusion

In this work, we propose a temporal augmented network to learn a motion-aware feature. This feature alone can achieve competitive performance with previous state-of-the-art methods, and when combined with them, can achieve significant performance improvements. We

also incorporate temporal context into the MIL ranking model by using an attention block. The learned attention weights can help to differentiate anomalous and normal video segments better. With the proposed motion-aware feature and temporal MIL ranking model, we achieve new state-of-the-art results for both anomaly detection and anomalous action recognition tasks in the UCF Crime dataset. Note that, our model still has difficulties in some known challenging scenarios, including fast motion, people grouping, low resolution, dark images, etc. In the future, we want to investigate other MIL formulations as in the recent zero-shot learning literature [28]. We also want to make our pipeline have just one stage with end-to-end learning to obtain more robustness.

References

[1] Christian Bailer, Bertram Taetz, and Didier Stricker. Flow Fields: Dense Correspondence Fields for Highly Accurate Large Displacement Optical Flow Estimation. In *International Conference on Computer Vision (ICCV)*, 2015.

[2] Joao Carreira and Andrew Zisserman. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[3] Kai-Wen Cheng, Yie-Tarng Chen, and Wen-Hsien Fang. Video Anomaly Detection and Localization Using Hierarchical Feature Representation and Gaussian Process Regression. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[4] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K. Roy-Chowdhury, and Larry S. Davis. Learning Temporal Regularity in Video Sequences. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[5] Rui Hou, Chen Chen, and Mubarak Shah. Tube Convolutional Neural Network (T-CNN) for Action Detection in Videos. In *The IEEE International Conference on Computer Vision (ICCV)*, 2017.

[6] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. FlowNet 2.0: Evolution of Optical Flow Estimation with Deep Networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[7] Lai Jiang, Mai Xu, Tie Liu, Minglang Qiao, and Zulin Wang. Deepvs: A deep learning based video saliency prediction approach. In *The European Conference on Computer Vision (ECCV)*, 2018.

[8] Weixin Li, Vijay Mahadevan, and Nuno Vasconcelos. Anomaly Detection and Localization in Crowded Scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 36(1):18–32, 2014.

[9] Cewu Lu, Jianping Shi, and Jiaya Jia. Abnormal Event Detection at 150 FPS in MAT-LAB. In *IEEE International Conference on Computer Vision (ICCV)*, 2013.

[10] Medhini G. Narasimhan and Sowmya Kamath S. Dynamic Video Anomaly Detection and Localization Using Sparse Denoising Autoencoders. *Multimedia Tools and Applications*, 77(11):13173–13195, 2018.

[11] Hamidreza Rabiee, Javad Haddadnia, Hossein Mousavi, Maziyar Kalantarzadeh, Moin Nabi, and Vittorio Murino. Novel Dataset for Fine-Grained Abnormal Behavior Understanding in Crowd. In *IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 2016.

[12] Jing Shao, Chen Change Loy, Kai Kang, and Xiaogang Wang. Slicing Convolutional Neural Network for Crowd Video Understanding. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[13] Karen Simonyan and Andrew Zisserman. Two-Stream Convolutional Networks for Action Recognition in Videos. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2014.

[14] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *International Conference on Learning Representations (ICLR)*, 2015.

[15] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-World Anomaly Detection in Surveillance Videos. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[16] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. PWC-Net: CNNs for Optical Flow Using Pyramid, Warping, and Cost Volume. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[17] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going Deeper with Convolutions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[18] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning Spatiotemporal Features with 3D Convolutional Networks. In *IEEE International Conference on Computer Vision (ICCV)*, 2015.

[19] Gul Varol, Ivan Laptev, and Cordelia Schmid. Long-term Temporal Convolutions for Action Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2017.

[20] Heng Wang and Cordelia Schmid. Action Recognition with Improved Trajectories. In *IEEE International Conference on Computer Vision (ICCV)*, 2013.

[21] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal Segment Networks: Towards Good Practices for Deep Action Recognition. In *European Conference on Computer Vision (ECCV)*, 2016.

[22] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking Spatiotemporal Feature Learning: Speed-Accuracy Trade-offs in Video Classification. In *European Conference on Computer Vision (ECCV)*, 2018.

[23] Dan Xu, Elisa Ricci, Yan Yan, Jingkuan Song, and Nicu Sebe. Learning Deep Representations of Appearance and Motion for Anomalous Event Detection. In *British Machine Vision Conference (BMVC)*, 2015.

[24] Huan Yang, Baoyuan Wang, Stephen Lin, David Wipf, Minyi Guo, and Baining Guo. Unsupervised Extraction of Video Highlights Via Robust Recurrent Auto-Encoders. In *IEEE International Conference on Computer Vision (ICCV)*, 2015.

[25] Christopher Zach, Thomas Pock, and Horst Bischof. A Duality Based Approach for Realtime TV-L1 Optical Flow. In *DAGM Conference on Pattern Recognition*, 2014.

[26] Ji Zhang, Yannis Kalantidis, Marcus Rohrbach, Manohar Paluri, Ahmed Elgammal, and Mohamed Elhoseiny. Large-Scale Visual Relationship Understanding. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2019.

[27] Ji Zhang, Kevin J. Shih, Ahmed Elgammal, Andrew Tao, and Bryan Catanzaro. Graphical Contrastive Losses for Scene Graph Parsing. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[28] Yi Zhu, Yang Long, Yu Guan, Shawn Newsam, and Ling Shao. Towards Universal Representation for Unseen Action Recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.