One-Shot Learning for Surveillance Anomaly Recognition using Siamese 3D CNN

Amin Ullah
Intelligent Media Laboratory,
Digital Contents Research
Institute, Sejong University
Seoul, South Korea
aminullah@ieee.org

Khan Muhammad
Intelligent Media Laboratory,
Digital Contents Research
Institute, Sejong University
Seoul, South Korea
khan.muhammad@ieee.org

Killichbek Haydarov
Intelligent Media Laboratory,
Digital Contents Research
Institute, Sejong University
Seoul, South Korea
kilichbek.haydarov@gmail.com

Ijaz Ul Haq
Intelligent Media Laboratory,
Digital Contents Research
Institute, Sejong University
Seoul, South Korea
ijazulhaq@ieee.org

Miyoung Lee
Intelligent Media Laboratory,
Digital Contents Research
Institute, Sejong University
Seoul, South Korea
miylee@gmail.com

Sung Wook Baik*
Intelligent Media Laboratory,
Digital Contents Research
Institute, Sejong University
Seoul, South Korea
sbaik@sejong.ac.kr

***Abstract*—One-shot image recognition has been explored for many applications in computer vision community. However, its applications in video analytics is not deeply investigated yet. For instance, surveillance anomaly recognition is an open challenging problem and one of its hurdles is the lack of accurate temporally annotated data. This paper addresses the lack of data issue using one-shot learning strategy and proposes an anomaly recognition framework which exploits a 3D CNN siamese network that yields the similarity between two anomaly sequences. This paper also investigates the existing 3D CNNs for this task and then proposes a lightweight 3D CNN model that efficiently handles one-shot anomaly recognition. Once our network is trained, then we can use the powerful discriminative 3D CNN features to predict anomalies not only for the new data but also for entirely new classes. The proposed model is trained using temporally annotated test set of UCF Crime dataset. Finally, the trained model is used to recognize the anomalies and produce temporal automatic labels for the video level weakly annotated training set of the dataset.**

Keywords—Artificial intelligence, deep learning, convolutional neural network, anomaly recognition, siamese network, one-shot learning.

I. INTRODUCTION

Surveillance cameras are one of the most reliable sources for the investigation of crime/anomaly scenes. However, advancements in computer vision and artificial intelligence took it one step further by detecting and recognizing the anomaly in real-time, helping in instantaneous reporting systems [1]. Most of these methods with high performance are based on various deep neural network architectures that rely on massive amount of annotated video datasets for training with powerful computational resources [2]. In addition, these models require retraining when there is a need for adding a new class in a classification task [3]. These facts impose problems on training neural networks. In such scenarios, one-shot learning can provide a potential solution which discovers how to perform a classification task by only looking at a single sample of each possible class even if the data is scarce [4]. This kind of learning process under the constraint removes the necessity of retraining models for new classes and facilitates the learning process in dynamically changing data environments [5].

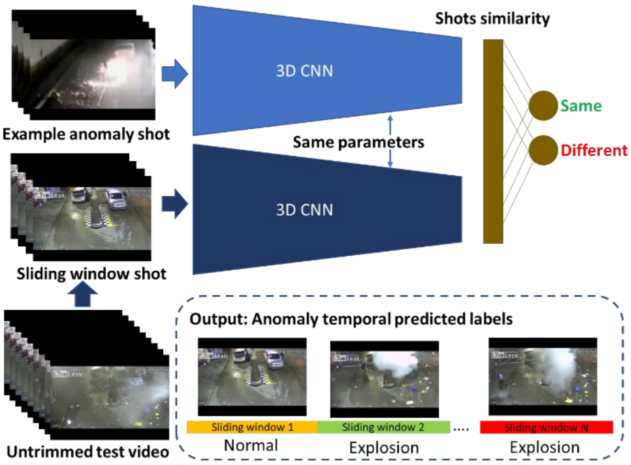


Figure 1: The proposed framework for one-shot anomaly recognition using 3D CNNs siamese network. The sliding window shot is compared with different example anomaly shots and the one outputs as same is considered as recognized anomaly.

Current methods for one-shot learning are inclined towards the meta-learning approaches. The basic idea of meta-learning is to exploit knowledge obtained from prior learning experience to learn more efficiently in future tasks [6, 7]. There exist several approaches in the literature that addressed the one- and few-shot learning. For instance, one category of such approaches is to treat deep neural networks as learners for feature encodings and train a separate meta-learner which learns how to update rules [8-10] or directly generate weights for the inference model [11]. In this way, the meta-learner directs the inference model to swiftly adjust its parameters to each specific task. On the other hand, instead of learning the updated parameters, the MAML [12] focused on finding the optimal initial parameters that can achieve a good generalization across similar tasks and make the task-specific fine-tuning process more efficient. Similarly, some approaches demonstrated that neural networks with augmented memory capabilities can act as a meta-learner. For instance, the method presented in [13] utilized Neural Turing Machine as the base model and trained it in such a way that the memory can encode and retrieve new information quickly.

Anomaly recognition is literally a video analytics task that can be analyzed by processing the visual and time series information in sequence of video frames. Traditionally, the spatiotemporal and motion-based features have attracted many researchers as they can effectively capture the salient features in visual time series data [14]. For instance, Ullah et al. [15] focused on industrial surveillance videos to automatically recognize normal and abnormal activities. They extracted visual features using optical flow CNN model and applied long short-term memory (LSTM) to learn different activity sequences. Similarly, Ryoo and Matthies [16] investigated local and global motion features to recognize interaction-level first-person activities. Their main goal was to let the first-person know about the ongoing activities in front of him. A TrajectoryNet is presented by Zhao and Xiong [17], which integrated the spatial features with the temporal dimension using a trajectory convolutional operation for anomalous activity recognition. In another study, Chong et al. [18] proposed a novel mehod for abnormal event identification via spatiotemporal autoencoders. The main aim of their method was to learn spatiotemporal features using an efficient autoencoder and to allow it to process up to 140 frames per second. Liu et al. [19] presented an abnormal activity grouping and recognition framework using hierarchical clustering and multi-task learning. They evaluated their method on different realistic datasets and outperformed state-of-the-art accuracies with a high margin. A max-margin based learning framework using soft labelling approach is presented by Hu et al. [20]. In this paper, the authors focused on transition problem between two activities which negatively effects the accuracy of the trained model. This problem is encountered by defining two new labels with 50% weights for each activity under transition. The final decision for the transition part is based on the prediction of the learning algorithm. Most of these methods are based on handcrafted features or CNN models followed by complex process of sequence learning which makes its implementation impracticable in the real-time scenarios.

Metric learning approach become attractive to solve one- or few-shot classification problems due to recent good empirical results [21-23]. The aim of these approaches is to determine similarity or dissimilarity between two input samples based on a distance metric. For instance, Koch et al. [21] used siamese neural network to extract embeddings from two input images and identified whether these images are drawn from the same class, converting the distance between the obtained embeddings to a probability score. A similar work [24] proposed relation network, but the similarity score between a pair of inputs were captured by a CNN classifier instead of computing with absolute distance. A matching network [9] is proposed to learn a classifier for k-shots classification task. They exploited an attention mechanism over learned embeddings from the train samples to predict classes of test samples.

The above discussed techniques are one of the best neural networks assisted approaches for one- and few-shot learning. However, they are limited to process single image and not suitable for processing sequences of frames for video level decision tasks such as anomaly recognition. To the best of our knowledge, the proposed approach is the initial brick to explore video level anomaly recognition using one-shot learning 3D siamese CNN model which is addressed with following major contributions:

- *We proposed one-shot anomaly recognition framework that takes only one or few annotated example videos per class for training. Once the model is trained then its powerful discriminative 3D CNN features can be used to predict anomalies not only for the new data but also for entirely newly added classes without retraining.*

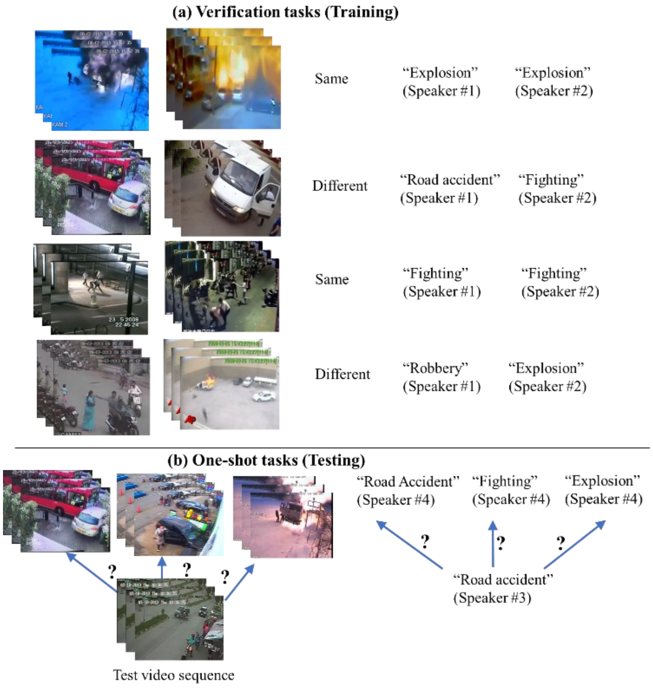
- *We investigated and performed extensive comparative analysis of the existing pretrained 3D CNNs for one-shot learning and proposed a lightweight siamese 3D CNN to extract high-level features from sequence of video frames for anomaly recognition.*

- *Our proposed siamese inference model achieved better competence and performance on fewer number of trainable parameters (2.8 million) with reduced model storage of 33.6 megabytes. These statistics are 90% smaller than the siamese network trained using state-of-the-art 3D CNN model.*

The rest of the paper is structured as follows: Section II contains the explanation and technical details of the proposed one-shot anomaly recognition. Section III provides a detailed discussion on obtained results and comparison with state-of-the-art. Section IV conclude the paper with key findings, limitations and future directions.

II. Proposed Methodlogy

In this section, the proposed one-shot learning for a temporal sequence level anomaly recognition framework is discussed in detail.



(a) Verification tasks (Training)

(b) One-shot tasks (Testing)

Figure 2: A 3-way one-shot learning example for anomaly recognition in video sequences. (a) Verification phase where the model just learns the difference between two sequences. (b) One-shot phase where the model compares the test video with different example anomaly shots to recognize the anomaly.

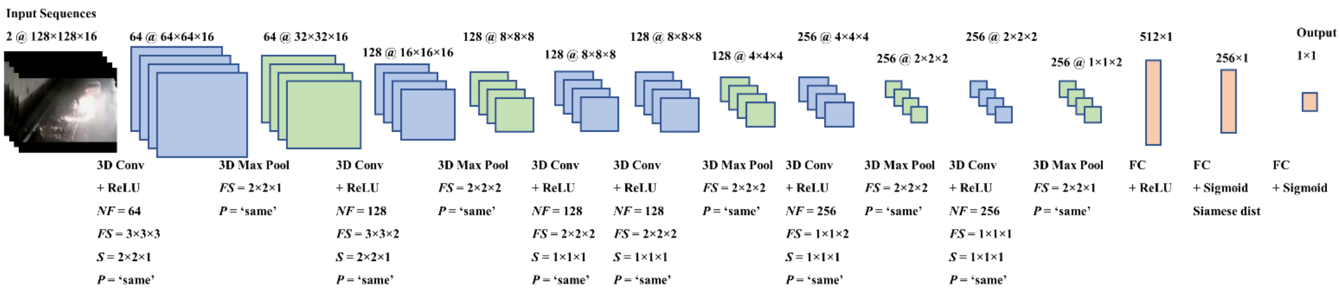

Input Sequences
2 @ 128×128×16
64 @ 64×64×16
64 @ 32×32×16
128 @ 16×16×16
128 @ 8×8×8
128 @ 8×8×8
128 @ 8×8×8
128 @ 4×4×4
256 @ 4×4×4
256 @ 2×2×2
256 @ 2×2×2
256 @ 1×1×2
512×1
256×1
Output
1×1
3D Conv + ReLU NF = 64 FS = 3×3×3 S = 2×2×1 P = 'same'
3D Max Pool FS = 2×2×1 P = 'same'
3D Conv + ReLU NF = 128 FS = 3×3×2 S = 2×2×1 P = 'same'
3D Max Pool FS = 2×2×2 P = 'same'
3D Conv + ReLU NF = 128 FS = 2×2×2 S = 1×1×1 P = 'same'
3D Conv + ReLU NF = 128 FS = 2×2×2 S = 1×1×1 P = 'same'
3D Max Pool FS = 2×2×2 P = 'same'
3D Conv + ReLU NF = 256 FS = 1×1×2 S = 1×1×1 P = 'same'
3D Max Pool FS = 2×2×2 P = 'same'
3D Conv + ReLU NF = 256 FS = 1×1×1 S = 1×1×1 P = 'same'
3D Max Pool FS = 2×2×1 P = 'same'
FC + ReLU
FC + Sigmoid Siamese dist
FC + Sigmoid

Figure 3: The architecture of the proposed 3D CNN for features encoding for our one-shot anomaly recognition framework. The output dimensions are given above and the number of filters NF, filter size FS, stride S, and padding P for each layer are given below the feature maps.

The proposed procedure for one-shot recognition is divided into two main steps. First, we train our siamese 3D CNN model to learn the similarity and dissimilarity of two video sequences. Next, we exploit the one-shot recognition by N-way strategy of matching with example shots. The proposed framework is visualized in Figure 1 and a 3-way one-shot anomaly recognition example is shown in Figure 2.

A. *One-shot recognition overview*

Convolutional neural networks (CNNs) are widely used in computer vision for the image and video representation [25-27], classification, and recognition tasks [28, 29]. But they require a lot of labelled data, which is one of the biggest limitations. There exist several applications which deal with a large number of classes that are dynamically increasing overtime. Furthermore, the samples for each class in such huge database is not enough in order to train the model properly because CNNs require a huge amount of data for each class. Due to the above-mentioned issues CNNs training is very costly and time-consuming process when a new class is inserted to the database. This problem can be tackled by using one-shot learning that requires only one or few training example for each class [23]. In one-shot learning the network does not classify the input (test image or sequence of frames) directly, but it requires two inputs in the form of a pair i.e., one is the test and second is the reference sample and the network tells us the similarity or difference between them [30]. Typically, sigmoid function is used to squeeze the similarity score in the range of 0 and 1. When the input test and reference samples are similar then the similarity score is 1 or closer to 1 and vice versa. In one-shot learning, the network learns a similarity function instead of learning classification.

The anomaly recognition problem can be considered as a binary supervised learning problem. The training dataset consists of (x^i, x^j) pairs, where 'x^i' shows an example shot and 'x^j' sliding window shot and for each pair we have a labeled similarity score 'Y' (0 or 1). The ideal case similarity function of one- shot siamese learning is mathematically given in Eq. 1 and a better visual understanding is drawn the Figure 2.

$$f(x^i, x^j) = \begin{cases} 1 & if \quad i = j \\ 0 & else \end{cases} \tag{1}$$

N-way one-shot learning method is normally used to check the performance of the trained network [21]. An example of 3-way one-shot learning is shown in Figure 2. In this setup, three example shots are created to test the sliding window shots. In a 3-way one-shot learning, sliding window shot of a particular class and a set of 3 example shots from 3 random classes are compared, but the sliding window will be similar with only one example shot in the set of 3 classes. In Figure 2, we are comparing a sliding window shot from a particular class with the set of example shots which are from 3 random classes, as a result we obtain 3 different similarity scores S_1, S_2 and S_3, respectively. If the model is correctly trained, it will return the maximum similarity score for the shots belonging to the same class and return the minimum similarity score for different shots.

In this paper, we have used siamese network, the term siamese means clones or twins. It consists of two CNNs, which are exactly the same networks, having the equal number of layers, and share the same parameters [31]. It takes pairs of 16 frames shots as an input, then after feeding a pair to the network, it creates two feature vectors using its 3D CNNs for each sequence. Then, it finds a distance between feature vectors using its matching network which uses the absolute distance and sigmoid activation for similarity score.

B. *Encoding shots via 3D CNN*

The features encoding from images using deep CNNs reached human level accuracy, however, from video it is still truly a difficult task because of the diverse nature of video data [32]. For instance, an apple category in image classification may have different colors, shapes, and sizes in training sample. However, for a road accident in video classification, everything in each training sample is totally changed because of the sequence of frames being not still images. In this paper, we are preforming anomaly recognition which occurs in the sequence of frames. In the video analytics literature, researchers from computer vision community have encoded video frames via 2D [33] or 3D [34] CNNs for various applications like human action and activity recognition. The siamese network has two encoding nets which provide feature vectors for matching the similarity of the inputs. Therefore, utilizing the 2D CNNs for frame level representation and then providing features to the matching net for the similarity score is not a good solution. Because each frame will give a 1000-dimensional feature vector and collectively it is a very high dimensional features which may require extreme computations. The 3D CNN is well suited in our case because it processes the sequence of frames at once instead of processing individual frames. It is end-to-end connected with the matching net that helps the optimizer to precisely update the parameters of 3D CNN and matching net during training.

In this study, we investigated two state-of-the-art 3D CNNs i.e., C3D [34] and I3D [35] for one-shot anomaly recognition using siamese network. The C3D is first popular 3D CNN model which employs a homogenous structure of kernels throughout its convolutional layers. It has eight convolutional layers with 3×3×3 filters, five down sampling layers with 2×2×2 filters, and two fully connected layers followed by a SoftMax classifier. The large filters size and more convolutional layers make C3D a very large model of 305 megabytes and has 79 million trainable parameters. The I3D [35] is an inflated 2D pretrained CNN which is expanded to the 3D model. On the backend, it utilized the pretrained parameters of Inception-V1 [36], that contains nine blocks and each block has multiple convolutional and pooling layers. It consists of total 22 weighted layers, five down sampling, and a SoftMax layer. I3D is also a very large model of 138 megabytes and 9 million trainable parameters. Since, siamese network has two features encoding networks running in parallel to processes two sequences at a time. Therefore, the parameters become double and models with millions of parameters are not efficient. Furthermore, both models process the three-channel video frames because they are trained for action recognition and three-channel processing is also computationally expensive. However, the samples we have for anomaly recognition are totally different from each other where the color and background information have no role for its recognition task and only the motion and temporal information are important. So, keeping all these facts in our minds we design a lightweight 3D CNN model features encoding for our one-shot anomaly recognition.

Our siamese network consist of twin 3D CNNs which has different settings on each layer. It has a total of six 3D convolutional, four 3D downsampling, and two fully connected layers. The overall architecture of our proposed model is visualized in Figure 3. We exploited 64 filters of size 3×3×3 in the first layer and down sampled the features maps using 2×2×1 in only spatial dimension while keep the temporal data unchanged. This helps our model to easily learn the temporal patterns in the next layer where we applied 128 3×3×2 filters followed by 2×2×2 max pooling. In the middle of the network, we established two consecutive convolutional layers of 2×2×2 kernels to learn the representations without quickly losing the spatiotemporal information. The hierarchy of convolutional layers ends with the output of 256×1×1×2 followed by two fully connected layers with 512 and 256 dimensions, respectively. The absolute distance of feature vectors from example and sliding window shots are calculated and processed via sigmoid activation to calculate the similarity score. The rectified linear (ReLU) activation unit is used in all convolutional and first fully connected layers. Furthermore, the Adam optimizer [37] is utilized for the parameter optimization. We processed gray level 16 frames sequences via small kernels of different sizes instead of large and homogenous nature kernels to reduce the model size and to capture the tiny patterns in small respective field.

C. Shots similarity and recognition decision

Once our model learns how to encode a pair of video sequences into feature vectors through a non-linear function f_α with parameters α, then we compute absolute distance *D* between these vectors using Eq. 2.

$$D = \left|f_\alpha(x_i) - f_\alpha(x_j)\right| \tag{2}$$

The obtained distance vector tells us how close two samples are to each other in feature space. However, the network must be able to decide whether a pair of video sequences comes from the same category based on the computed distance [38]. Therefore, the last output layer consisting of a single neuron employs the sigmoid function to obtain the probability of two video sequences belonging to the same class or not which can be computed using Eq. 3.

$$p(x_i, x_j) = \sigma(WD) \tag{3}$$

where W corresponds to the weights of last layer. However, it is still challenging to make a prediction relying merely on the similarity score p between two inputs. We take a test sequence and randomly select single example sequence from all classes to match the similarity score between them. Finally, the decision of anomaly class is based on the highly similar example shot.

III. Expremental Evaluation

The experiments for evaluating our proposed one-shot anomaly recognition framework are discussed in this section. We performed training and testing on UCF-Crime dataset [39] and then an extensive comparisons with state-of-the-art is performed including overall accuracy, receiver operating characteristic curve (ROC), aera under the curve (AUC), number of parameters in model, size of inference model, and time complexity. The proposed siamese network for one-shot anomaly recognition is implemented and tested in Python 3.6 on Windows 10 operating system. The deep learning toolbox known as "Keras" [40] is used for 3D CNN siamese network implementation. The hardware is equipped with Corei7 CPU, 32-GB RAM, and GeForce-RTX 2080 with 12-GB graphics processing unit (GPU).

A. UCF crime dataset

UCF-Crime is one of the most recently released anomaly recognition benchmark dataset which consists of 128 hours videos. This dataset comprises 1900 uncut surveillance videos of 14 different categories including Assault, Abuse, Arson, Arrest, Burglary, Explosion, Road Accident, Stealing, Fighting, Robbery, Shoplifting, Vandalism, and Normal. This dataset is well-balanced i.e., half of the videos are anomalous events and half of the videos have normal events. This dataset is split into training and testing set by the publishing authors where the training set consists of 800 normal and 810 anomalous events while the testing set includes 150 normal events and 140 anomalous events. The challenging part of this dataset is that its training part is not temporally annotated, and the training videos contain a lot of shots having no anomalies. However, its testing videos are entirely temporally annotated. The one-shot learning requires a fewer number of samples for recognition task therefore, we utilized its testing set for verification task in one-shot learning.

B. Data preparation and evaluation metrics

To train the network, we generated a batch consisting of *K* (less than or equal to the total number of classes) pairs of video clips with their target labels at each iteration. The first element of all pairs is the same while the second element of the pair is randomly sampled without replacement from the dataset.

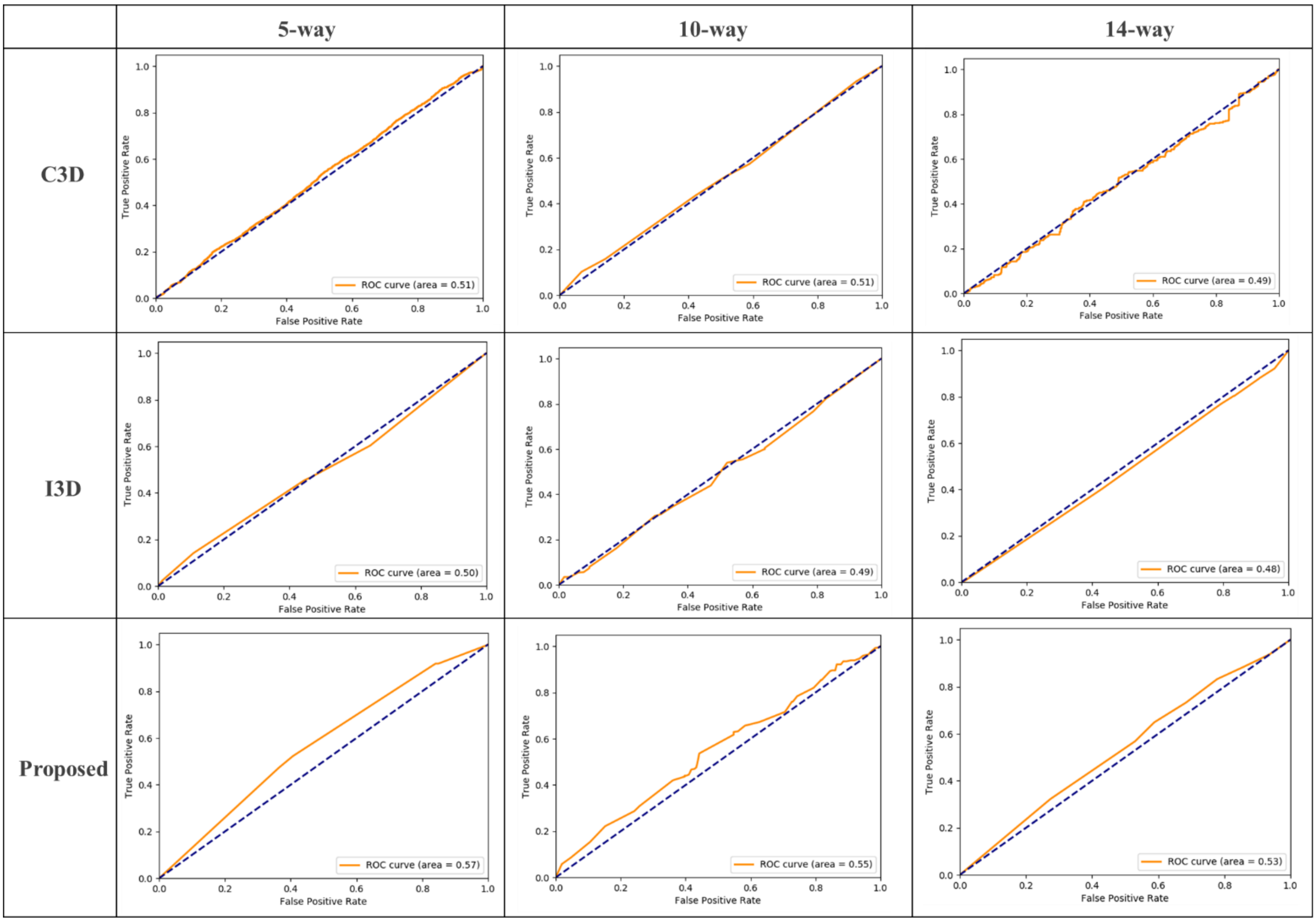


Figure 4: The comparison using ROC and AUC of the proposed 3D CNN siamese network with the networks constructed using state-of-the-art 3D CNNs including C3D and I3D. The experimental strategies are presented in columns i.e., 5-way, 10-way, and 14-way. The AUC values for all comparisons are given at the bottom of each graph.

Table 1: The comparison of proposed 3D CNN siamese network with state-of-the-art models trained for one-shot anomaly recognition using overall accuracy, number of trainable parameters, and trained model size.

Methods	**5-way Accuracy (%)**	**10-way Accuracy (%)**	**14-way Accuracy (%)**	**Trainable Parameters (million)**	**Model Size (MB)**
C3D [34]	35.7	**31.5**	27.3	96.9	1200
I3D [35]	36.1	29.7	28.8	15	180.7
Proposed	**38.4**	30.8	**29.1**	**2.8**	**33.6**

The target scores are set to either 1 if the elements of a pair are from the same class and 0 otherwise. The half of the batch contained pairs of videos from the same class and the other half with different classes. In validation task, we exploited N-way strategy. The batch generation process was almost identical to the one in training phase with some minor differences. Here, we set the number of video pairs to N. In our experiments we chose 5, 10 and 14 to validate our model. Moreover, the validation batch contained only a single pair of elements with the same class label. We expected that pair would give us the highest similarity score, treating it as a correct prediction. In order to calculate the accuracy of the model, we generated batches of random pairs for T trials and then find the ratio between the number of correct predictions c for batch i and the number of T as given in Eq. 4.

$$accuracy = \left(\frac{1}{T}\sum_{i=1}^{T} c_i\right) \times 100 \tag{4}$$

For visualization of the performance of our classifier, we utilized ROC curves and AUC [41]. ROC represents the trade-off between the true positive rate (TPR) and false positive rate (FPR) whereas AUC indicates a degree or measure of separability. The ROC curve is plotted with TPR against the FPR, being on y-axis and x-axis, respectively. An ideal model has AUC near to the 1.

C. Comparison with state-of-the-art

We investigated two state-of-the-art 3D CNN models including C3D [34] and I3D [35] for sequence of video frames features encoding before modeling our own architecture. The comparison using overall accuracy, number of trainable parameters, and model size is given in Table 1. For 5-way evaluation setting, the C3D achieved 35.7%, I3D and proposed method reached 36.1% and 38.4% accuracy score, respectively. The 5-way is easy evaluation because inferencing with only five examples and the probability of

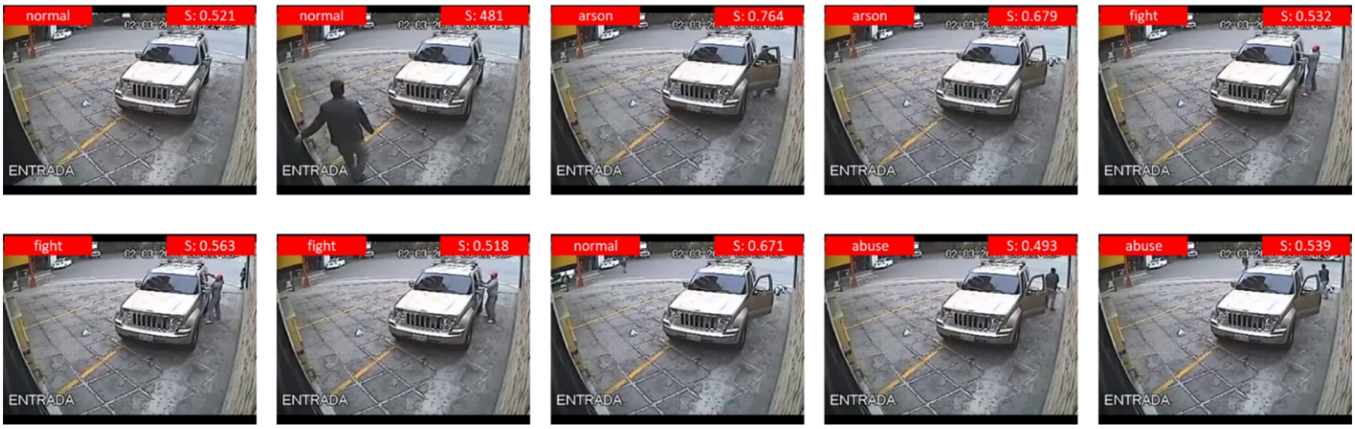


Figure 5: The visual results of the proposed one-shot anomaly recognition with predicted class and its probability scores for sequences of 16 video frames. The test video is taken from the UCF-Crime training videos which is weakly labeled as shooting class. However, we have generated automatic temporal annotation for each 16 frames sequence.

making wrong prediction is very low as compare to 10-way and 14-way. In 5-way we have choose only five anomaly classes for evaluation, in 10-way only ten anomalies, and in 14-way evaluation we selected thirteen anomaly classes and one normal class. For 10-way evaluation the C3D model achieved highest accuracy of 31.5%, the I3D and our method achieved 29.7% and 30.8%, respectively. The 14-way is very important but challenging due to comparison of one sliding window shot with 14 examples and then deciding to which category it has higher similarity. For this challenging setting, we achieved 29.1% accuracy which is 1% greater than I3D results and 2% greater than C3D results. The comparison using ROC curve and AUC scores are shown in Figure 4. The proposed method achieved higher AUC of 0.57, 0.55, and 0.53 for 5-way, 10-way, and 14-way anomaly recognition evaluation using 5-way, 10-way, and 14-way, respectively. The AUC for C3D and I3D are approximately 0.50 for all three N-way settings. For instance, the C3D achieved AUC of 0.51, 0.51, and 0.49 for 5-way, 10-way, and 14-way anomaly recognition, respectively. The I3D achieved AUC of 0.5, 0.49, and 0.48 for 5-way, 10-way, and 14-way anomaly recognition, respectively. The overall accuracies of the aforementioned 3D CNNs are not deficient because the proposed method achieved better performance on 5-way and 14-way and the C3D perform well on 10-way anomaly recognition. However, comparing them using number of trainable parameters in the inference model then the C3D is more expensive because it has 96.9 million parameters and required 1.2 GB storage capacity. Similarly, the I3D is built from 15 million parameters and required 180.7 MB size in memory. The original C3D and I3D has fewer number of parameters as compare to the statistics we shown in Table 1. Because of some additional weights of the matching network in the model, therefore, parameters increased from the original C3D. In contrast with this the proposed model contains only 2.8 million parameters and required only 33.6 MB space in memory. This make our proposed inference model superior from both state-of-the-art 3D CNNs. Furthermore, our model can be easily implemented over the resource contained devices for real-time anomaly recognition in surveillance.

D. Performance analysis and discussion

The visual results for one weakly labeled video are shown in Figure 5. The video is annotated as shooting class in the dataset. However, inside the video there are multiple shots representing different actions. For instance, the first and second 16 frames sequences are recognized as normal category and truly it is normal, but in the dataset, it is mentioned as shooting. In the third and fourth sequence a man is approaching the vehicle which are recognized as arson, a wrong prediction, however, in the arson category of the dataset, there exist several sample videos where man approaches the car. In the next three sequences a man come to shoot the driver in the car and the driver try to resist which is recognized as fighting class and semantically it is not wrong. In the final two frames there is casualty which is recognized as abuse while abuse category usually contains similar scenarios of injures in human and animal. Overall, the accuracy of the proposed model is less but it has produced semantically very correct results because the dataset is weakly labeled and need temporal annotations for better performance of anomaly recognition.

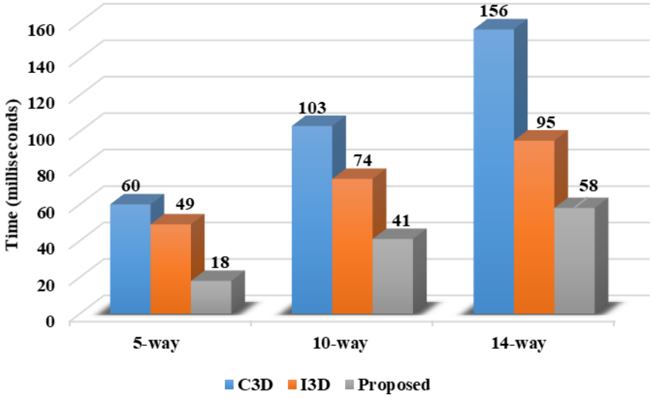


Figure 6: Total time required to process a single sequence in three N-way settings including 5-way, 10-way, and 14-way.

The time complexity comparison of the proposed one-shot anomaly recognition with C3D and I3D models are given in Figure 6. The processing time of surveillance anomaly recognition is very important for which our proposed model achieved better performance taking only 18, 41, and 58 milliseconds to process 5-way, 10-way, and 14-way one-shot recognition, respectively. On the other hand, the C3D and I3D take almost triple and double time from the proposed model.

As we increased the parameter N, the accuracy of our model dropped accordingly because it needs to decide in large number of similarity scores. The proposed method achieved better performance in terms of accuracy and processing time from the state-of-the-art and can be deployed for surveillance anomaly recognition.

IV. Conculsion

This paper presents a real-time anomaly recognition framework for smart surveillance. The proposed method is based on one-shot learning strategy, where we exploited a 3D CNN siamese network that yields the similarity between two anomaly sequences that gives free hand to add new data as well as entirely new classes. We also investigated the performance of existing two famous state-of-the-art 3D CNNs by implementing it as siamese network. Our proposed siamese inference model shows better performance on fewer number of trainable parameters (2.8 million) with reduced storage size of 33.6 megabytes that make it 90% smaller than state-of-the-art models. Further, we perform extensive experiments for comparative analysis of the existing 3D CNNs on UCF crime dataset.

In future, we aim to explore few-shot recognition and try to address the shortcomings of one-shot anomaly recognition. Furthermore, spiking neural networks and fuzzy neural networks [42] are very hot topics these days and very much suitable to be investigated for anomaly recognition task. In addition, we intend to optimize and deploy our model on resource constrained devices such as nano-Jetson and raspberry Pi with neural kit. Furthermore, the current work is based on full frame analysis, we plan to analyze multiple anomalies by detection and tracking different targets in the video. In addition, we are motivated to use multi-view surveillance data for precise anomaly recognition in the visual sensor network.

Acknowledgment

This work was supported by Institute of Information & communications Technology Planning &Evaluation (IITP) grant funded by the Korea government (MSIT) (2019-0-00136, Development of AI-Convergence Technologies for Smart City Industry Productivity Innovation)

References

[1] J. Kim, T. Kim, S. Kim, and C. D. Yoo, "Edge-Labeling Graph Neural Network for Few-shot Learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11-20.

[2] Z. Lu, S. Qin, L. Li, D. Zhang, K. Xu, and Z. Hu, "One-Shot Learning Hand Gesture Recognition Based on Lightweight 3D Convolutional Neural Networks for Portable Applications on Mobile Systems," *IEEE Access,* vol. 7, pp. 131732-131748, 2019.

[3] J. Zhang, C. Zhao, B. Ni, M. Xu, and X. Yang, "Variational Few-Shot Learning," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 1685-1694.

[4] A. T. Chen, M. Biglari-Abhari, and K. I. Wang, "Fast One-Shot Learning for Identity Classification in Person Re-identification and Tracking," in *2018 15th International Conference on Control, Automation, Robotics and Vision (ICARCV)*, 2018, pp. 1197-1203.

[5] H. Li, D. Eigen, S. Dodge, M. Zeiler, and X. Wang, "Finding Task-Relevant Features for Few-Shot Learning by Category Traversal," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1-10.

[6] B. Liu, X. Yu, A. Yu, P. Zhang, G. Wan, and R. Wang, "Deep Few-Shot Learning for Hyperspectral Image Classification," *IEEE Transactions on Geoscience and Remote Sensing,* vol. 57, no. 4, pp. 2290-2304, 2019.

[7] X. Wei, P. Wang, L. Liu, C. Shen, and J. Wu, "Piecewise Classifier Mappings: Learning Fine-Grained Learners for Novel Categories With Few Examples," *IEEE Transactions on Image Processing,* vol. 28, no. 12, pp. 6116-6125, 2019.

[8] M. Andrychowicz *et al.*, "Learning to learn by gradient descent by gradient descent," in *Advances in neural information processing systems*, 2016, pp. 3981-3989.

[9] K. Li and J. Malik, "Learning to optimize," *arXiv preprint arXiv:1606.01885,* 2016.

[10] S. Ravi and H. Larochelle, "Optimization as a model for few-shot learning," 2016.

[11] D. Ha, A. Dai, and Q. V. Le, "Hypernetworks," *arXiv preprint arXiv:1609.09106,* 2016.

[12] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 2017, pp. 1126-1135: JMLR. org.

[13] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. Lillicrap, "One-shot learning with memory-augmented neural networks," *arXiv preprint arXiv:1605.06065,* 2016.

[14] K.-P. Chou *et al.*, "Robust feature-based automated multi-view human action recognition system," vol. 6, pp. 15283-15296, 2018.

[15] A. Ullah, K. Muhammad, J. D. Ser, S. W. Baik, and V. H. C. d. Albuquerque, "Activity Recognition Using Temporal Optical Flow Convolutional Features and Multilayer LSTM," *IEEE Transactions on Industrial Electronics,* vol. 66, no. 12, pp. 9692-9702, 2019.

[16] M. S. Ryoo and L. Matthies, "First-person activity recognition: Feature, temporal structure, and prediction," *International Journal of Computer Vision,* vol. 119, no. 3, pp. 307-328, 2016.

[17] Y. Zhao, Y. Xiong, and D. Lin, "Trajectory Convolution for Action Recognition," in *Advances in Neural Information Processing Systems*, 2018, pp. 2204-2215.

[18] Y. S. Chong and Y. H. Tay, "Abnormal event detection in videos using spatiotemporal autoencoder," in *International Symposium on Neural Networks*, 2017, pp. 189-196: Springer.

[19] A.-A. Liu, Y.-T. Su, W.-Z. Nie, and M. Kankanhalli, "Hierarchical clustering multi-task learning for joint human action grouping and recognition," *IEEE*

transactions on pattern analysis and machine intelligence, vol. 39, no. 1, pp. 102-114, 2017.

[20] N. Hu, G. Englebienne, Z. Lou, and B. Kröse, "Learning to recognize human activities using soft labels," *IEEE transactions on pattern analysis and machine intelligence,* vol. 39, no. 10, pp. 1973-1984, 2016.

[21] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese neural networks for one-shot image recognition," in *ICML deep learning workshop*, 2015, vol. 2.

[22] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Advances in Neural Information Processing Systems*, 2017, pp. 4077-4087.

[23] O. Vinyals, C. Blundell, T. Lillicrap, and D. Wierstra, "Matching networks for one shot learning," in *Advances in neural information processing systems*, 2016, pp. 3630-3638.

[24] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1199-1208.

[25] I. Mehmood *et al.*, "Efficient Image Recognition and Retrieval on IoT-Assisted Energy-Constrained Platforms From Big Data Repositories," *IEEE Internet of Things Journal,* vol. 6, no. 6, pp. 9246-9255, 2019.

[26] I. U. Haq, K. Muhammad, A. Ullah, and S. W. Baik, "DeepStar: Detecting Starring Characters in Movies," *IEEE Access,* vol. 7, pp. 9265-9272, 2019.

[27] K. Muhammad, T. Hussain, M. Tanveer, G. Sannino, and V. H. C. J. I. I. o. T. J. de Albuquerque, "Cost-Effective Video Summarization using Deep CNN with Hierarchical Weighted Fusion for IoT Surveillance Networks," 2019.

[28] F. U. M. Ullah, A. Ullah, K. Muhammad, I. U. Haq, and S. W. J. S. Baik, "Violence detection using spatiotemporal features with 3D convolutional neural network," vol. 19, no. 11, p. 2472, 2019.

[29] E.-J. Cheng *et al.*, "Deep sparse representation classifier for facial recognition and detection system," vol. 125, pp. 71-77, 2019.

[30] F.-F. Li, R. Fergus, and P. Perona, "One-shot learning of object categories," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 28, no. 4, pp. 594-611, 2006.

[31] M. H. Abdelpakey and M. S. Shehata, "DP-Siam: Dynamic Policy Siamese Network for Robust Object Tracking," *IEEE Transactions on Image Processing,* vol. 29, pp. 1479-1492, 2020.

[32] T. Hussain, K. Muhammad, A. Ullah, Z. Cao, S. W. Baik, and V. H. C. d. Albuquerque, "Cloud-Assisted Multiview Video Summarization Using CNN and Bidirectional LSTM," *IEEE Transactions on Industrial Informatics,* vol. 16, no. 1, pp. 77-86, 2020.

[33] A. Ullah, K. Muhammad, I. U. Haq, and S. W. Baik, "Action recognition using optimized deep autoencoder and CNN for surveillance data streams of non-stationary environments," *Future Generation Computer Systems,* vol. 96, pp. 386-397, 2019/07/01/ 2019.

[34] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4489-4497.

[35] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299-6308.

[36] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1-9.

[37] D. P. Kingma and J. J. a. p. a. Ba, "Adam: A method for stochastic optimization," 2014.

[38] K. Thanikasalam, C. Fookes, S. Sridharan, A. Ramanan, and A. Pinidiyaarachchi, "Target-Specific Siamese Attention Network for Real-Time Object Tracking," *IEEE Transactions on Information Forensics and Security,* vol. 15, pp. 1276-1289, 2020.

[39] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6479-6488.

[40] F. Chollet, "Keras," ed, 2015.

[41] S. J. T. D. S. Narkhede, "Understanding AUC-ROC Curve," vol. 26, 2018.

[42] O. P. Patel, N. Bharill, A. Tiwari, and M. Prasad, "A Novel Quantum-inspired Fuzzy Based Neural Network for Data Classification," *IEEE Transactions on Emerging Topics in Computing,* pp. 1-1, 2019.