electronics

MDPI

Article

Comparing Video Activity Classifiers within a Novel Framework

Chiman Kwan *, Bence Budavari and Bulent Ayhan

Applied Research LLC, Rockville, MD 20850, USA; bencebudavari@gmail.com (B.B.); bulentayhan@gmail.com (B.A.)

* Correspondence: chiman.kwan@signalpro.net

Received: 6 August 2020; Accepted: 18 September 2020; Published: 21 September 2020

Abstract: Video activity classification has many applications. It is challenging because of the diverse characteristics of different events. In this paper, we examined different approaches to event classification within a general framework for video activity detection and classification. In our experiments, we focused on event classification in which we explored a deep learning-based approach, a rule-based approach, and a hybrid combination of the previous two approaches. Experimental results using the well-known Video Image Retrieval and Analysis Tool (VIRAT) database showed that the proposed classification approaches within the framework are promising and more research is needed in this area

Keywords: video event classification; object detection; object tracking; VideoGraph; rule-based approach; hybrid approach

1. Introduction

There is an emerging interest in human activity recognition using intelligent systems. This growing field has a wide range of applications such as human–computer interaction and identity detection [1,2], surveillance and home monitoring [3–5], healthcare [6,7], and elderly care [8,9].

Object detection, tracking, and classification are all essential steps in recognizing activities within a video. Some conventional methods use non-deep learning algorithms, such as optical flow, for object detection and tracking [10–12]. In recent years, there have been some new developments in object detection using deep learning. Representative methods include You Only Look Once (YOLO) [13], Faster Region-based Convolutional Neural Networks (R-CNN) [14], and Single Shot Detector (SSD) [15]. In tracking, the Deep Affinity Network (DAN) [16] has been shown to yield good performance. With respect to video activity classification, there have been many recent works in the literature that implement deep learning methods. For example, References [17–19] used a variety of deep learning methods to recognize complex human activities within videos. In Reference [17], VideoGraph (VG) was proposed for activity recognition. Convolutional Neural Network-Long Short-Term Memory (CNN-LSTM) is another popular technique, and its source codes can be found in Reference [18]. Long-Term Recurrent Convolutional Networks (LRCNs) [19] is another well-known technique which can also address video captioning. Another method is the CNN-Indirect Recurrent Neural Network (IndRNN) method [20], or IndRNN in short, which consists of a two-stage, end-to-end framework and is inspired, in part, by how humans identify events with varying rhythms. In the first stage, the most significant frames are selected, while the second stage recognizes the event using the selected frames. In Reference [20], a method known as CNN-SkipRNN+, or SkipRNN+ in short, uses the same framework, IndRNN. However, SkipRNN+ has advantages over IndRNN by alleviating the gradient vanishing problem that occurs because of the many RNN layers used in the frame selection phase of the framework. In Reference [21], an LSTM-based network was used in order to classify events in

situations where there are multiple viewpoints of an event. In Reference [22], an autoencoder and CNN were utilized in challenging cases where the environment was not stationary. In Reference [23], an activity recognition framework was proposed in which the video stream was first divided into important shots, where shots were selected using CNN-based human saliency features. Next, temporal features of an activity were extracted by utilizing the convolutional layers of another CNN model. Finally, a multilayer long short-term memory was used for activity recognition.

Although deep learning methods are end-to-end algorithms, they need extensive training on large datasets which, in some cases, can be unrealistic. For example, there may be many non-event cases where it is extremely difficult to enumerate the various possibilities, making it extremely difficult to prepare the training data. The vast number of non-events may also create an unbalanced data problem during training. Moreover, there is a more fundamental problem in deep learning. The activity classification in deep learning relies solely on information inside the bounding boxes. The context information around the bounding boxes is completely ignored. For instance, the two events—entering a car and entering a building—have the completely same behavior, except perhaps the last few frames, where parts of the car and building start to show inside the bounding boxes. The above can easily create wrong event classifications.

Although rule-based approaches have not been as popular as deep learning approaches in the past few years, rule-based video surveillance systems have been widely used in commercial applications. For example, the home surveillance camera allows a user to select multiple zones for alerts. That is, if there are motions in some pre-specified zones, then alerts will be issued. In Reference [11], certain regions such as swinging trees in the videos are also masked using rules to reduce false alarms.

In this paper, we propose a video activity detection and classification framework in order to overcome certain shortcomings of deep learning methods in the field of video activity recognition. In particular, our proposed framework encompasses object detection, object tracking, and both rule-based and deep learning event classification modules. For the purpose of being concise, this paper focuses on the object tracking and the classifier modules, except the object detection module. Users can choose deep learning only, rule-based only, or a combination of rule-based and deep learning methods for event classification in videos. For rule-based event classification approaches, some exemplar rules include point of interests, distance from point of interest, etc. For a deep learning-based approach, we adopted the graph-based approach (i.e., VideoGraph). We pay particular attention to eliminate non-events, which are events that are not relevant to some surveillance applications. For example, a person walking on the street is a non-event if our events of interests are defined as entering and exiting buildings and cars. Four distinct events in the well-known VIRAT data [24] were used for illustrating the proposed framework.

The contributions of our paper are as follows:

- We propose a new video activity recognition framework that is flexible, modular, and encompasses object detection, object tracking, and event classification.
- We demonstrated the efficacy of the tracking and classifier modules, except the object detection module, in the proposed framework using the well-known VIRAT dataset. The results are very encouraging.
- We demonstrate the use of logical rules to improve the efficacy of deep learning classifiers.

This paper is organized as follows. Section 2 describes the proposed video activity detection framework and its critical components. Section 3 summarizes the dataset and our experiments. Section 4 discusses the results and future directions. Finally, some concluding remarks are provided in Section 5.

2. Proposed Video Event Classification Framework and Its Critical Modules

In our proposed video activity recognition framework, there are two key components: VideoGraph and DAN. We briefly describe them in Sections 2.1 and 2.2. We then describe the proposed framework in Section 2.3.

2.1. VideoGraph

Graph methods are being investigated for human activity recognition in the past. Although they are noted to learn structured representations from videos, they require the graph nodes and/or edges to be known in advance which limits their practical use since they cannot be used when node or frame-level annotations are not available. In contrast, VideoGraph [17] is a graph-based method in which the graph nodes are fully inferred from data, and it is extensible to datasets without node-level annotations. The block diagram of VideoGraph can be seen in Figure 1. The video is first sampled into T segments and each segment, s_i, contains eight consecutive frames. Using two-stream Inflated 3D ConvNet (I3D), which is a 3D CNN model, features are extracted from s_i, where they are denoted by x_i. An undirected graph with N nodes corresponds to key unit actions in the video, whereas the edges of the graph provide the temporal relationship between these N nodes. The node attention block in VideoGraph learns the latent concept representation. For the initialization of these latent features, the features maps of the last convolutional layer of the I3D backbone are clustered and the resultant centroids are used for initialization. The graph embedding layer learns the graph edges and finalizes the graph structure. VideoGraph extracts two types of relationships and represents them via graph edges. There are the timewise edges indicating how the nodes transition over time and the node-wise edges providing information about the relationships between nodes. The activation output of the first graph embedding layer is used to construct the final graph. Among the two graph embedding layers in VideoGraph, the second one is used for activity prediction. Following a set of pooling operations to the output of the second graph embedding layer both in time and node, the resultant output feature is fed-forward to a classifier to arrive at the activity prediction of the video.

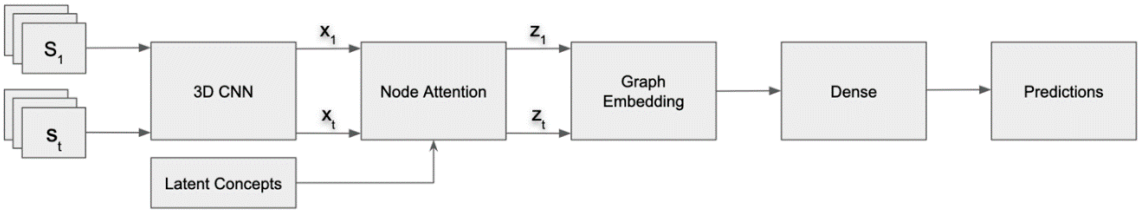


Figure 1. VideoGraph block diagram [17]. A 3D CNN is used to extract features, which are then used to build nodes and graphs.

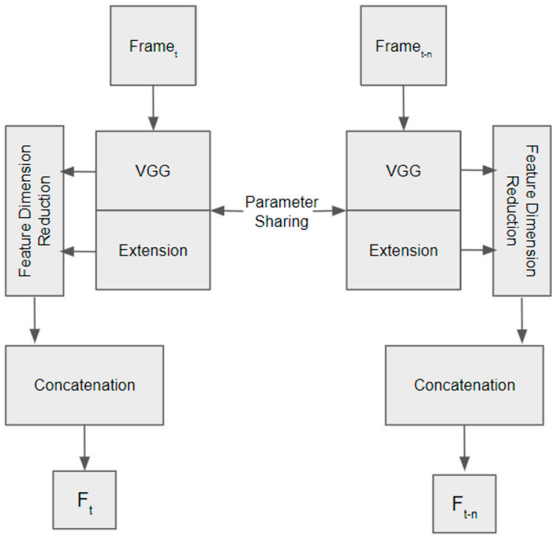
2.2. DAN

Another essential component of the pipeline is an object tracker. Our current pipeline employs DAN, which is a deep learning end-to-end model [16]. The architecture of this model is outlined below in Figure 2. There are two components of DAN: a feature extractor (Figure 2a) and affinity estimator (Figure 2b). The feature extractor is responsible for generating features from the bounding boxes of objects. The affinity estimator takes those object features from one frame and attempts to associate them with the features of subsequent frames. From an implementation standpoint the DAN requires a video as a sequence of frames as well as a text file of all the detected objects within each frame of the video. Once DAN processes the video, it will then output a text file that will include a series of bounding boxes belonging to each of the detected objects in the video.

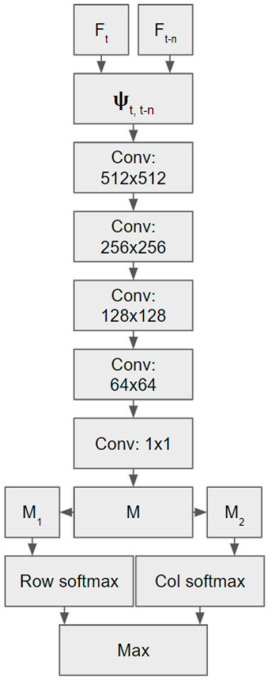
This framework for object tracking is quite robust and has performed very well in the Multiple Object Tracking (MOT) dataset challenges compared to other methods. As seen in Reference [16], the DAN method has better results in all metrics when compared to comparable online methods for

the MOT17 dataset. Even when compared to offline methods, it is at the top or close to the top for all metrics. These metrics are further explained in Reference [16].

Some key features of DAN that made it ideal for our study include its ability to track even after certain objects are obscured and later reappear. Many instances in the VIRAT dataset include objects that are obscured by buildings or by vehicles, so a tracker that has the ability to overcome obscuration was necessary. In addition, DAN can be used in real-time, so it gives our pipeline the flexibility to use both offline and online processing to meet the needs of a variety of applications.



(**a**)



(**b**)

Figure 2. Deep Affinity Network (DAN) architecture [16]: (**a**) Stage 1 of DAN; (**b**) Stage 2 of DAN. VGG stands for Visual Geometry Group.

2.3. Proposed Video Activity Recognition Pipeline

2.3.1. General Architecture

The architecture of our proposed pipeline is displayed in Figure 3. A video is fed forward to the "Detector" module which performs object detection. Many algorithms such as YOLO, Faster R-CNN, and SSD can be used for object detection. The output of detected objects in the form of bounding boxes is fed into DAN which is responsible for object tracking. The output of DAN is then fed to both the "Rule-Based System" and "Bounding Box Extractor" modules. The "Bounding Box Extractor" simply takes the tracked object locations and extracts the bounding boxes of each tracked individual in order to be used in the "VideoGraph" and "Rule-Based System" modules. The "Rule-based System", which will be discussed in further detail below, takes the tracking information consisting of sequences of bounding boxes (BBs) from DAN, any regions of interest manually selected from the video, and the sequence of BBs for each individual in order to classify the various events within the video. The "VideoGraph" module will simply take the sequences of bounding boxes for each tracked individual to generate event classification results. Most of our interest in this paper was focused on the event classification portion of the pipeline, but we will briefly discuss how the object detection and tracking modules function in our current implementation.

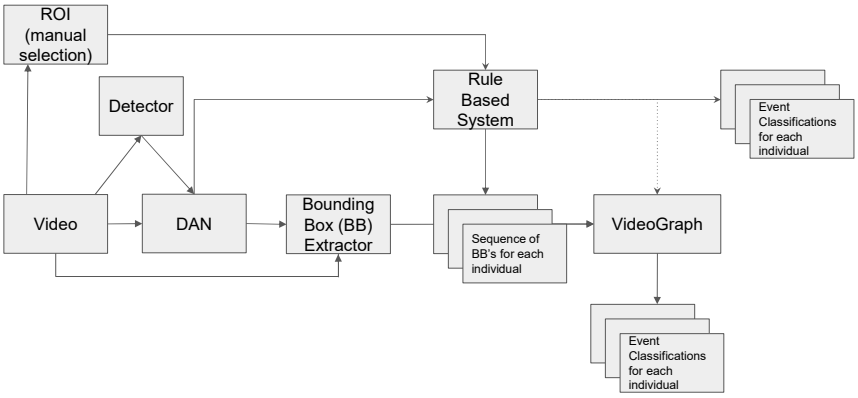


Figure 3. Our proposed video event classification pipeline architecture. Both rules-based and deep learning-based algorithms are included for activity recognition. ROI stands for Region of Interest; BB stands for Bounding Box.

In this paper, the "Detector" module has been replaced with VIRAT's ground truth detection annotations. This was done in order to test the efficacy of the tracker, classifier, and rules-based system. For object tracking, the pipeline utilizes DAN, which is discussed in further detail above. These two components have been used in all of our experiments throughout our study so that we are able to effectively examine the differences between a rule-based and deep learning event classifiers. If these event classifiers are not able to effectively differentiate between the events in the best-case scenario, then there is no need to further examine the previous components of the pipeline.

2.3.2. Rule-Based System

For the VIRAT 4-event case, we developed a set of rules that can effectively classify the four events of interest in addition to the non-event cases. Given a sequence of frames of a particular person, there are several rules that can be used to determine if they performed an event or not. For the events "entering facility" and "exiting facility", we need to annotate the facility and place a point of interest along the entrance of a facility. In a particular scene, there can be multiple facilities and multiple entrances for a particular building. At this point, this was done manually for the VIRAT videos since most of the videos for these events were taken in only a handful of different environments.

Additionally, in a surveillance application with a steady camera, it would not be difficult to manually label a particular region of interest.

Once the pipeline receives both the sequence of frames for a particular point and the points of interest, the first and last frame of each sequence need to be analyzed. We use the following two rules to determine whether a sequence of bounding boxes is exiting or entering a facility.

- Rule 1: Exiting a facility

If the first frame of an object sequence is within a certain distance of any of the points of interest, it can be assumed that the person has exited from a particular facility.

- Rule 2: Entering a facility

If the last frame is within a certain distance of any of the points of interest, it can be assumed that person has entered the facility. This distance threshold has been manually determined depending on the scene.

The above simple rules work quite effectively for the VIRAT events involving facilities, because the person cannot be tracked through the building as the facilities are all brick and not glass. So, in essence, if a person emerges or disappears in the scene near these points of interest, it can be assumed they either entered or exited that particular facility.

Now, we describe the rule for recognizing entering or exiting a vehicle.

- Rule 3: Entering or exiting a vehicle

The events that involve entering and exiting a vehicle are tested in a very similar way, except with one caveat. Since the vehicles can move, the detection and tracking portions of the pipeline provide the bounding box information of the vehicles. So, this is not a manual process but rather automated. But the exiting and entering vehicles events are classified in the same way as the facility events. That is, we check the position of the first and final bounding box in relation to the point of interest, which is the center of the vehicle closest to the person.

There are a few limitations if only the above rules are used for both events with facilities and vehicles. A rule to address those limitations is shown below.

- Rule 4: Special case

If a person's last frame was within the distance threshold of a vehicle or facility but the video ends, then the person would be classified as having entered into the facility. In these cases, the person's speed information can be quite valuable for determining their particular classification. Speed in this case is simply determined by the pixel distance from the center of one bounding box to the next bounding box. For example, when somebody enters a vehicle, their speed will drastically decelerate. We averaged the speed of each object over eight frames and compared the acceleration information between those eight frame segments to better assess the potential trajectory of the object.

There are some cases where the bounding boxes are far away from any of the points of interest. So, we have the following rule for classifying the above event sequence as a non-event.

- Rule 5: Non-event determination based on distance from points of interest

Given a sequence of bounding boxes, if the distances of the centroids of the bounding boxes to those points of interest are larger than a certain threshold, then this sequence can be considered as non-events (not entering or exiting a facility or a vehicle).

In some surveillance scenes, there is prior knowledge about the scene contents in which the following rule can be utilized:

- Rule 6: Elimination of certain events based on the scene contents.

The majority of the VIRAT videos do not have both a vehicle and facility event. In other words, most videos will only include entering and exiting of facility events or entering and exiting vehicles events.

This simple rule could significantly improve the recognition performance in quite a few videos.

When using the VideoGraph approach, it is also possible to come up with one or two simple rules. For example, the following rule can help eliminate some false alarms.

- Rule 7: VideoGraph Probability Thresholding.

We applied a 0.6 threshold for the prediction probability in the Softmax output of VideoGraph. That is, if the prediction probability is less than 0.6, we ignore the decision.

3. Results

We conducted extensive tests to demonstrate the effectiveness of classification approaches in our proposed framework. In these experiments, the ground truth locations of bounding boxes were used as input to the DAN tracker. Afterwards, the DAN tracker generated sequences of bounding boxes associated with a particular target. The bounding box extractor module cropped the video frames according to the bounding box information for each target and fed the temporal sequence for each object into one of the three possible classifiers. We focused on event recognition by comparing the VideoGraph approach, the rule-based approach, and a hybrid approach in order to examine their strengths and weaknesses.

3.1. Dataset

The VIRAT 2.0 dataset [21] is a publicly available video dataset supported by Defense Advanced Research Projects Agency (DARPA). The videos in this dataset consist of surveillance footage capturing public areas such as parking lots and college campuses. The VIRAT 2.0 videos are in high definition, and the original size of the image frames in these videos are 1920×1080 in size. Each video contained multiple activities, but using accompanied labels and bounding boxes, we were able to crop each activity from the footage to be used for model training. The classified activities we used from this dataset included: a person getting into a vehicle, a person getting out of a vehicle, a person entering a facility, and a person exiting a facility. Table 1 shows these events and the number of videos for each event. The events of interest for this study are bolded. In addition, some example frames for these four events are shown in Figure 4.

Some of these events, such as a person loading an object into a vehicle, have very few videos indicating a data imbalance problem which poses challenges to applied deep learning methods. To prevent the negative effects of imbalanced data, we took an equal amount of data from each event for the training of our models in this study.

It should be noted that these videos were annotated using Internet Mechanical Turk. This sort of annotation method essentially outsources the task of annotating videos to a variety of individuals which may result in inconsistencies from video to video. For the VIRAT dataset specifically, this means that some objects outside of the events of interest may not be annotated. Other issues, like segmented tracked objects, also occurred. This means that a single person may be annotated as two different people. These are important to keep in mind, as they pose problems for our proposed video activity recognition pipeline.

3.2. Comparison of Three Event Recognition Approaches

We assembled a collection of 20 videos from the VIRAT dataset in order to test at least 10 cases for each of the events. There are three approaches: VideoGraph approach, rule-based approach, and the hybrid approach.

Table 1. The VIRAT 2.0 Dataset main events and the number of videos for each event. The bold letters indicate events of interest in this paper. There are 12 events in total in the VIRAT dataset. Bold characters indicate the events of interest in this paper.

Event ID	Event Type	Number of Videos
1	Person loading an Object to a Vehicle	21
2	Person Unloading an Object from a Car/Vehicle	59
3	Person Opening a Vehicle/Car Trunk	42
4	Person Closing a Vehicle/Car Trunk	41
5	**Person getting into a Vehicle**	**111**
6	**Person getting out of a Vehicle**	**97**
7	Person gesturing	51
8	Person digging	0
9	Person carrying an object	822
10	Person running	22
11	**Person entering a facility**	**156**
12	**Person exiting a facility**	**133**

Figure 4. Sample image frames for the four investigated events from the VIRAT dataset: (**a**) a person getting into a vehicle; (**b**) a person getting out of a vehicle; (**c**) a person entering a facility; (**d**) a person exiting a facility.

3.2.1. VideoGraph Only

From Table 2, it can be seen that the VideoGraph only approach was excellent in identifying across the four events of interest. VideoGraph had an average of 95% accuracy across the four events, but had a 0% classification for the non-events. The VideoGraph model had issues with correctly classifying the non-event cases in the videos because VideoGraph utilizes only information inside the bounding boxes and no information from the surroundings. The VideoGraph had a tendency to have very confident prediction probabilities for one of the four events when trying to classify non-events. This is because VideoGraph cannot effectively handle events that are not present in the training data. Although the model was not trained on any non-event cases, theoretically the prediction probabilities should be quite

low for each event when the model is trying to predict a non-event. However, this was not the case. In Table 2, there are 28 non-events and VideoGraph wrongly categorized all of them into one of those four events. In particular, 13 non-events were classified as Event 5 (entering vehicle); 10 non-events were classified as Event 6 (exiting vehicle); 2 non-events were classified as Event 11 (entering facility); 3 non-events were classified as Event 12 (exiting facility). This study clearly demonstrates the limitation of the VideoGraph as well as other deep learning approaches which rely only on information inside the bounding boxes and non-contextual information from the environment was taken into account. The overall accuracy (OA) was 0.56 due to the mis-classifications of non-events.

Table 2. Event Classification performance using VideoGraph only for 20 VIRAT videos. In addition to a confusion matrix, the overall accuracy (OA) was also included.

		Classified							
		Event 5	Event 6	Event 11	Event 12	Non-Event	Total	Accuracy	OA
Ground Truth (GT)	Event 5	9	1	0	0	0	10	0.9	0.56
	Event 6	0	10	0	0	0	10	1	-
	Event 11	0	0	10	0	0	10	1	-
	Event 12	0	0	1	9	0	10	0.9	-
	Non-Event	13	10	2	3	0	28	0	-

3.2.2. Rule-Based Approach

We applied Rules 1 to 5 mentioned in Section 2.3.2 in our rule-based approach. In Table 3, we can see that the averaged classification accuracy for the four events was 77.5%. For those four events, the rule-based approach did not perform as well as VideoGraph. However, for those non-events, the rule-based approach correctly classified 67% or 19 out of 28 non-events. This clearly demonstrates that rule-based approach is a viable alternative to the deep learning approaches such as VideoGraph. We can see that the OA was improved to 0.74 which is better than Videograph.

Table 3. Event classification performance using a rule-based approach for only 20 VIRAT videos. Confusion matrix and OA are included.

		Classified							
		Event 5	Event 6	Event 11	Event 12	Non-Event	Total	Accuracy	OA
GT	Event 5	8	1	0	0	1	10	0.8	0.74
	Event 6	0	8	0	0	2	10	0.8	-
	Event 11	0	0	7	1	2	10	0.7	-
	Event 12	0	0	1	8	1	10	0.8	-
	Non-Event	3	2	3	1	19	28	0.68	-

3.2.3. Hybrid Approach

One way to improve on the VideoGraph's classification of non-events is to implement a simple rule (Rule 6 in Section 2.3.2) that checks whether a vehicle or facility of interest is present within the scene. The majority of the videos do not have both a vehicle and facility event, so this could significantly improve the results. An additional rule (Rule 7 in Section 2.3.2) of having a 0.6 threshold for the prediction probability alongside the previous rule allows for much better classification of non-events. With the combination of the above simple rules with VideoGraph, one can clearly see from Table 4 that a significant reduction of wrong classifications. In particular, non-event classification accuracy improved to 61%. That is, 17 out of 28 non-events were correctly classified as non-events. This is a big improvement over the pure VideoGraph approach. It should be noted that the four events were still classified quite accurately like the VideoGraph only approach. It can be seen that the hybrid approach has achieved an OA of 0.81, which is better than both the VideoGraph and Rule-based approaches.

Table 4. Event classification performance using a hybrid approach only for 20 VIRAT videos. Confusion matrix and OA are included.

		Classified							
		Event 5	Event 6	Event 11	Event 12	Non-Event	Total	Accuracy	OA
GT	Event 5	9	1	0	0	0	10	0.9	0.81
	Event 6	0	10	0	0	0	10	1	-
	Event 11	0	0	10	0	0	10	1	-
	Event 12	0	0	1	9	0	10	0.9	-
	Non-Event	5	2	2	2	17	28	0.61	-

3.3. Three Detailed Videos

This section takes a closer look at how the various approaches within the pipeline deal with three videos. The three approaches for the classifier module are VideoGraph only, VideoGraph with two rules, and rule-based only. The two rules (Rule 6 and Rule 7) used with the VideoGraph check whether a facility or building is present in the scene (if not present, the two correlated events would be removed) and a confidence threshold for the predicted probabilities of the remaining available classes.

3.3.1. Building Entrance Scene: VIRAT_S_010000_04_000530_000605

As shown in Figure 5, this video scene is located on a college campus and includes several entrances to facilities. The camera is stationary. There are two "exiting facility" and two "entering facility" events within this video. There are no "entering vehicle" and "exiting vehicle" events. The remaining nine individuals are all non-event cases. The results using three separate approaches of our proposed pipeline are summarized in Tables 5–7. From Table 5, one can see that the VideoGraph only approach did a good job in classifying Event 11 and Event 12, but wrongly classified all of the nine non-events. From Table 6, we can see that the rule-based approach correctly classified all the events and non-events. Finally, from Table 7, it can be seen that the hybrid approach wrongly classified one of the nine non-events.

Table 5. Event classification using the VideoGraph only approach for one VIRAT video: VIRAT_S_010000_04_000530_000605. The confusion matrix and OA are included.

		Classified						
		Event 5	**Event 6**	**Event 11**	**Event 12**	**Non-Event**	**Accuracy**	**OA**
GT	Event 5	-	-	-	-	-	-	0.31
	Event 6	-	-	-	-	-	-	-
	Event 11	0	0	2	0	0	1	-
	Event 12	0	0	0	2	0	1	-
	Non-Event	1	6	1	1	0	0	-

Table 6. Event classification using the rule-based only approach for one VIRAT video: VIRAT_S_010000_04_000530_000605. The confusion matrix and OA are included.

		Classified						
		Event 5	**Event 6**	**Event 11**	**Event 12**	**Non-Event**	**Accuracy**	**OA**
GT	Event 5	-	-	-	-	-	-	0.92
	Event 6	-	-	-	-	-	-	-
	Event 11	0	0	2	0	0	1	-
	Event 12	0	0	0	2	0	1	-
	Non-Event	0	0	1	0	8	0.89	-

Table 7. Event classification using the hybrid approach for one VIRAT video: VIRAT_S_010000_04_000530_000605. The confusion matrix and OA are included.

		Classified						
		Event 5	**Event 6**	**Event 11**	**Event 12**	**Non-Event**	**Accuracy**	**OA**
	Event 5	-	-	-	-	-	-	0.85
	Event 6	-	-	-	-	-	-	-
GT	Event 11	0	0	2	0	0	1	-
	Event 12	0	0	0	2	0	1	-
	Non-Event	0	0	1	1	7	0.78	-

Figure 5. VIRAT_S_010000_04_000530_000605 Scene. This building scene has a lot of human activity.

3.3.2. Parking Lot Scene: VIRAT_S_010111_08_000920_000954

This scenario shown in Figure 6 is opposite to the previous case in Section 3.2.1. Here, there are only entering and exiting vehicles events and no entering and exiting facility events. This video is located in a busy parking lot and includes over 40 vehicles. Some of the vehicles are stationary while others move. There is a single entering vehicle event and a single exiting vehicle event. There are an additional three people who are all non-event cases. From Table 8, the VideoGraph only approach correctly classified Event 5 and Event 6 but wrongly classified all the non-events. This is understandable because VideoGraph was not trained with any non-events. The results of the rule-based only approach are summarized in Table 9 which shows that it correctly classified the two events and two out of three non-events. Finally, the results of the hybrid approach are shown in Table 10. We can see that the two non-events are wrongly classified. It should be emphasized that the rules in the rule-based approach and the hybrid approach are different. In the rule-based approach, Rule 1 to Rule 5 were used, and in the hybrid approach, Rules 6 and 7 were used.

Table 8. Event classification using the VideoGraph only approach for one VIRAT video: VIRAT_S_010111_08_000920_000954. The confusion matrix and OA are included.

		Classified						
		Event 5	**Event 6**	**Event 11**	**Event 12**	**Non-Event**	**Accuracy**	**OA**
	Event 5	1	0	0	0	0	1	0.4
	Event 6	0	1	0	0	0	1	-
GT	Event 11	-	-	-	-	-	-	-
	Event 12	-	-	-	-	-	-	-
	Non-Event	1	1	1	0	-	0	-

Table 9. Event classification using the rule-based only approach for one VIRAT video: VIRAT_S_010111_08_000920_000954. The confusion matrix and OA are included.

		Classified						
		Event 5	**Event 6**	**Event 11**	**Event 12**	**Non-Event**	**Accuracy**	**OA**
GT	Event 5	1	0	0	0	0	1	0.8
	Event 6	0	1	0	0	0	1	-
	Event 11	-	-	-	-	-	-	-
	Event 12	-	-	-	-	-	-	-
	Non-Event	0	1	0	0	2	0.67	-

Table 10. Event classification using the hybrid approach for one VIRAT video: VIRAT_S_010111_08_000920_000954. Confusion matrix and OA are included.

		Classified						
		Event 5	**Event 6**	**Event 11**	**Event 12**	**Non-Event**	**Accuracy**	**OA**
GT	Event 5	1	0	0	0	0	1	0.6
	Event 6	0	1	0	0	0	1	-
	Event 11	-	-	-	-	-	-	-
	Event 12	-	-	-	-	-	-	-
	Non-Event	1	1	0	0	1	0.33	-

Figure 6. VIRAT_S_010111_08_000920_000954 Scene. This parking lot scene is quite challenging due to the presence of multiple events.

3.3.3. Mixed Parking Lot and Building Entrance Scene: VIRAT_S_000102

As seen in Figure 7, this video is located on a campus where there is a parking lot and a facility. This video contains both facility events and vehicle events. Since this video is quite long, only the last quarter of the video is used. This section alone contains three exiting facilities, four entering facilities, 1 exiting vehicle, and 2 entering vehicle events. In this particular case, using the VideoGraph with the rules-based approach was not feasible, since it relies on scenes with only vehicles or only facilities. Because of this, only the VideoGraph and rule-based approaches were compared. From Table 11, similar to earlier cases, the VideoGraph only approach performed quite well for classifying the four events, but failed completely for the non-events. In contrast, the rule-based approach results in Table 12 performed slightly worse in classifying the four events but correctly classified all the six non-events.

Figure 7. VIRAT_S_000102 Scene. This scene has both parking lot and building entrance events.

Table 11. Event classification using the VideoGraph only approach for one VIRAT video: VIRAT_S_000102. The confusion matrix and OA are included.

		Classified						
		Event 5	**Event 6**	**Event 11**	**Event 12**	**Non-Event**	**Accuracy**	**OA**
GT	Event 5	1	0	0	0	0	1	0.6
	Event 6	0	2	0	0	0	1	-
	Event 11	0	0	3	1	0	0.75	-
	Event 12	0	0	0	3	0	1	-
	Non-Event	4	0	1	0	0	0	-

Table 12. Event classification using the rule-based approach for one VIRAT video: VIRAT_S_000102. The confusion matrix and OA are included.

		Classified						
		Event 5	**Event 6**	**Event 11**	**Event 12**	**Non-Event**	**Accuracy**	**OA**
GT	Event 5	1	0	0	0	0	1	0.93
	Event 6	0	1	0	0	0	0.5	-
	Event 11	0	0	4	0	0	1	-
	Event 12	0	0	0	3	0	1	-
	Non-Event	0	0	0	0	6	1	-

4. Discussions and Future Work

4.1. Discussions

Although the rule-based approach performed well in all cases for the VIRAT 4 events, there are a few underlying issues with rule-based approaches. Many of these rules developed for a specific dataset are not scalable or adaptable to new events or datasets. If we expanded the number of events from four to include the other events present in the VIRAT dataset, the rules used for entering and exiting would not be effective at detecting any other classes. Instead, new rules would have to be developed to work alongside the previous rules.

The same could be said of deep learning methods to a certain extent. With the addition of new classes or use of different datasets, a model will have to be retrained. In addition, certain parameters will have to be tweaked to better fit the dataset at hand. But even in those cases, the framework of the model generally stays the same. Whereas in rule-based approaches, the workflow will have to be fundamentally changed in order to incorporate new data. However, our results demonstrate that

when deep learning methods have shortcomings, logical rules can be used with the model to improve its effectiveness. In our case, the VideoGraph approach was able to be improved significantly with the addition of two simple logical rules instead of having to retrain the entire model. One could argue that such a broad class like "non-event" would not even be realistic to train a model to effectively classify.

4.2. Limitations

Although only a single dataset was used in this study, there are 20 videos in our case studies. It should be mentioned that rule-based classifiers are tailored to the VIRAT data. Admittedly, the rules discussed are very specific to the events within the VIRAT dataset and would not necessarily be transferrable to a different dataset. Different rules will need to be generated for different datasets. These rules merely highlight the potential of using logical rules in conjunction with deep learning classifiers.

4.3. Future Work

In order to compare the VideoGraph with a rule-based approach, the object detection and tracking stayed consistent (unchanged) between the two classification modules. Researching the effectiveness of the last portion of the pipeline, the classification module, was a necessary first step. If we work backwards through the pipeline, it will be easier to understand the "best case" scenario for each of the modules. Future work will include comparing various methods for each of the modules outlined in the pipeline.

There are quite a few directions we would like to continue research with this particular pipeline. First of all, it will be necessary to examine how the pipeline results function when the ground truth detection results are not used. In other words, how much degradation in event classification will result if an object detection algorithm is utilized to classify the various events? We predict some of the more challenging videos with multiple overlapping objects, like the scenes with dense parking lots (Section 3.3.2), may see significant degradation due to the fact of misguided detection. This will be an area of further study.

Second, we are interested in turning this pipeline into a real-time application rather than one that uses pre-recorded data. In order for this to happen, a few changes will have to be made to the architecture of the pipeline as well as the internal algorithms. For example, the rule-based approach can no longer rely on the first and last frames in an object's entire sequence of frames to determine exiting and entering. Rather, a buffer of frames will need to be updated and the classification will need to be generated for each buffer. Additional rules will need to be implemented in order to classify these buffers at steady time intervals.

Another important area for future research would be the expansion of the number of events in the VIRAT dataset. In this case, the rules would have to be reworked in order to incorporate the logic of these new events. For example, additional rules would need to be implemented for "unloading object from vehicle". This sort of study would help further identify strengths and weaknesses, especially in regard to scalability, between rules-based approaches and deep learning approaches.

5. Conclusions

Deep learning methods have been effective in dealing with a variety of image and video processing problems, but they also have difficulty in dealing with unexpected inputs such as new classes. These new classes may be falsely identified as existing classes. This may be even more problematic when these new classes are too broad to effectively train a model with. Simple rules can be utilized in conjunction with deep learning models in order to produce better classification results. In some cases, rule-based approaches may even be more effective at classifying events even though these rules may not be as scalable and adaptable to new datasets as deep learning methods. In this paper, we present a video event classification framework that contains many modules such as object detection, object tracking, rule database, event classification, etc. Experiments using representative videos in the VIRAT

database demonstrated the limitation of deep learning event classification approach and potential advantages of using a rule-based approach in conjunction with a deep learning approach for event classification. It was demonstrated that the rule-based approach can effectively handle non-events that the deep learning approach may be difficult to model.

We believe the field of video event classification is still in its early stage of development and more research is needed in order to be practical. We also highlighted some future research directions, which were mentioned earlier.

Author Contributions: Conceptualization, C.K.; methodology, C.K., B.B.; validation, B.B., C.K.; writing—original draft preparation, C.K., B.B.; investigation, C.K., B.B., B.A.; supervision, C.K.; project administration, C.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Rautaray, S.S.; Agrawal, A. Vision based hand gesture recognition for human computer interaction: A survey. *Artif. Intell. Rev.* **2012**, *43*, 1–54. [CrossRef]
2. Paul, S.N.; Singh, Y.J. Survey on Video Analysis of Human Walking Motion. *Int. J. Signal Process. Image Process. Pattern Recognit.* **2014**, *7*, 99–122. [CrossRef]
3. Vishwakarma, S.; Agrawal, A. A survey on activity recognition and behavior understanding in video surveillance. *Vis. Comput.* **2012**, *29*, 983–1009. [CrossRef]
4. Lin, W.; Sun, M.-T.; Poovandran, R.; Zhang, Z. Human activity recognition for video surveillance. In Proceedings of the 2008 IEEE International Symposium on Circuits and Systems, Seattle, WA, USA, 18–21 May 2008.
5. Duong, T.V.; Bui, H.H.; Phung, D.; Venkatesh, S. Activity Recognition and Abnormality Detection with the Switching Hidden Semi-Markov Model. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 05), San Diego, CA, USA, 20–25 June 2005; Volume 1, pp. 838–845.
6. Kuo, Y.-M.; Lee, J.-S.; Chung, P.-C. A Visual Context-Awareness-Based Sleeping-Respiration Measurement System. *IEEE Trans. Inf. Technol. Biomed.* **2009**, *14*, 255–265. [PubMed]
7. Huynh, H.H.; Meunier, J.; Sequeira, J.; Daniel, M. Real time detection, tracking and recognition of medication intake. *World Acad. Sci. Eng. Technol.* **2009**, *60*, 280–287.
8. Foroughi, H.; Aski, B.S.; Pourreza, H. Intelligent Video Surveillance for Monitoring Fall Detection of Elderly in Home Environments. In Proceedings of the IEEE 11th International Conference on Computer and Information Technology (ICCIT), Khulna, Bangladesh, 24–27 December 2008; pp. 219–224.
9. Liu, C.D.; Chung, P.C.; Chung, Y.N.; Thonnat, M. Understanding of human behaviors from videos in nursing care monitoring systems. *J. High Speed Netw.* **2007**, *16*, 91–103.
10. Kwan, C.; Zhou, J. Anomaly detection in low quality traffic monitoring videos using optical flow. In Proceedings of the Pattern Recognition and Tracking XXIX, Orlando, FL, USA, 30 April 2018; Volume 10649.
11. Kwan, C.; Zhou, J.; Wang, Z.; Li, B. Efficient anomaly detection algorithms for summarizing low quality videos. In Proceedings of the Pattern Recognition and Tracking XXIX, Orlando, FL, USA, 27 April 2018; Volume 10649.
12. Kwan, C.; Zhou, J.; Yin, J. The development of a video browsing and video summary review tool. In Proceedings of the Pattern Recognition and Tracking XXIX, Orlando, FL, USA, 27 April 2018; Volume 10649.
13. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. 2018. Available online: https://arxiv.org/abs/1804.02767 (accessed on 8 April 2018).
14. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef] [PubMed]
15. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. SSD: Single Shot MultiBox Detector. In Proceedings of the 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; pp. 21–37.

16. Sun, S.; Akhtar, N.; Song, H.; Mian, A.S.; Shah, M. Deep affinity network for multiple object tracking. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *1*. [CrossRef] [PubMed]
17. Hussein, N.; Gavves, E.; Smeulders, A.W.M. VideoGraph: Recognizing Minutes-Long Human Activities in Videos. *arXiv* **2019**, arXiv:1905.05143.
18. Five Video Classification Methods Implemented in Keras and TensorFlow. 2017. Available online: https://blog.coast.ai/five-video-classification-methods-implemented-in-keras-and-tensorflow-99cad29cc0b5? (accessed on 29 April 2020).
19. Donahue, J.; Hendricks, L.A.; Guadarrama, S.; Rohrbach, M.; Venugopalan, S.; Darrell, T.; Saenko, K. Long-term recurrent convolutional networks for visual recognition and description. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 2625–2634.
20. Li, Y.; Yu, T.; Li, B. Recognizing Video Events with Varying Rhythms. *arXiv* **2020**, arXiv:2001.05060.
21. Ullah, A.; Muhammad, K.; Hussain, T.; Baik, S.W. Conflux LSTMsNetwork: A Novel Approach for Multi-View Action Recognition. Available online: https://www.researchgate.net/profile/Amin_Ullah3/publication/339133153_Conflux_LSTMs_Network_A_Novel_Approach_for_Multi-View_Action_Recognition/links/5e3f79ce299bf1cdb918f8e4/Conflux-LSTMs-Network-A-Novel-Approach-for-Multi-View-Action-Recognition.pdf (accessed on 20 September 2020).
22. Ullah, A.; Muhammad, K.; Haq, I.U.; Baik, S.W. Action recognition using optimized deep autoencoder and CNN for surveillance data streams of non-stationary environments. *Futur. Gener. Comput. Syst.* **2019**, *96*, 386–397. [CrossRef]
23. Ullah, A.; Muhammad, K.; Del Ser, J.; Baik, S.W.; De Albuquerque, V.H.C. Activity Recognition Using Temporal Optical Flow Convolutional Features and Multilayer LSTM. *IEEE Trans. Ind. Electron.* **2019**, *66*, 9692–9702. [CrossRef]
24. The VIRAT Video Datase. Available online: https://viratdata.org/ (accessed on 11 March 2020).