IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received October 3, 2020, accepted October 16, 2020, date of publication October 22, 2020, date of current version November 2, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3033190

Video Activity Recognition With Varying Rhythms

BULENT AYHAN[1], (Member, IEEE), CHIMAN KWAN[1], (Senior Member, IEEE), BENCE BUDAVARI[1], JUDE LARKIN[1], DAVID GRIBBEN[1], AND BAOXIN LI[2], (Senior Member, IEEE)

[1]Applied Research LLC, Rockville, MD 20850, USA
[2]School of Computing, Informatics, and Decision Systems Engineering, Arizona State University, Tempe, AZ 85281, USA

Corresponding author: Chiman Kwan (chiman.kwan@signalpro.net)

This work was supported by the U.S. Navy under Grant N6833520G2007.

ABSTRACT Recognizing normal and anomalous events in long and complex videos with multiple sub-activities has received considerable attention in recent years. This task is more challenging than traditional action recognition in short and relatively homogeneous video clips. Other than the difficulty in recognizing activities in long videos, one other challenge is the varying activity rhythms. The rhythm of sub-actions in an activity can differ in nature and can pose additional challenges that affect the performance of activity recognition methods. In this article, five video activity recognition methods were evaluated using two publicly available video datasets, Breakfast and VIRAT, which consist of long and complex videos. Extensive experiments and analyses showed that among these methods, VideoGraph, was found to perform distinctly better than the other investigated methods while maintaining high accuracy even if the test videos were exposed to severe rhythm changes. The results indicated that VideoGraph is less sensitive to varying rhythms in contrast to other investigated methods. By changing some of the architecture parameters, we also observed performance improvements in VideoGraph.

INDEX TERMS Activity recognition, human action recognition, varying rhythm, event recognition, video, surveillance.

I. INTRODUCTION

There is an emerging interest in automating 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], elderly care [8], [9], and traffic monitoring [10] and video summarization [11], [12]. One of the easiest acquired input data that can be used for activity recognition are color (RGB) videos captured by cameras. Recognizing activities in videos thus has received significant attention in recent years. The works in this emerging field mostly consist of recognizing human actions using datasets like UCF101 [13], KTH [14], HMDB51 [15], Kinetics [16]. These datasets consist of relatively short and homogeneous video clips, which are generally well-segmented and contain only one action event in which human actions take few seconds to unfold [17]. As an example in [18], the authors used UCF101 and HMDB51 datasets for demonstrating their two-stream 3-D-convNet fusion pipeline, which can recognize human actions in videos of arbitrary size and length using multiple features. In [19], UCF101 and HMDB51 datasets were used and saliency-aware three-dimensional (3-D) CNN with LSTM is introduced for video action recognition. However, it is highly likely that some of these methods using datasets which consist of only short homogenous video clips could face challenges when it comes to recognizing normal and anomalous events in datasets that consist of long and complex videos with multiple sub-actions in it such as Breakfast [20] and VIRAT [21].

Graph-based methods have also found their use for video activity recognition. In [22], the authors proposed a semi-supervised annotation approach by learning an optimized graph from multi-cues (i.e., partial tags and multiple features). There are some other graph-based methods which utilize the sub-action level annotations for human activity recognition in long and complex video datasets [23]–[30]. However, finding datasets with sub-action annotations is not easy and not very practical. Other than the difficulty in recognizing activities in long videos, one other essential challenge is the varying activity rhythms. The rhythm of sub-actions in an activity can differ in nature. As an example, considering "getting into a car" activity in the VIRAT dataset, one can

The associate editor coordinating the review of this manuscript and approving it for publication was Jiachen Yang.

open the door and get in the car immediately, or open the door then take some time before getting in the car. Even though these two sets of actions are both categorized with the same label, their temporal rhythms differ considerably. Varying rhythm of actions in real videos may arise from at least two sources. First, the rhythm of sub-activities in an event can differ in nature such as the different rhythms of getting in a car. Second, the rhythm issue may occur due to non-uniform or different sampling rates between the training and testing stages of the applied recognition method. Ignoring varying rhythms may seriously affect the activity recognition performance. It is quite likely that an event recognition algorithm may fail to accurately classify the activity when trained with one rhythm but tested with another rhythm.

The objective of this article is to investigate the performance of video recognition methods which do not use any sub-action level annotations for long duration and complex videos that are captured with stationary cameras and also to examine the recognition sensitivity of these methods to varying rhythms. Five video activity recognition methods were evaluated using the RGB color videos of two challenging public domain video datasets. These are Breakfast [20] and VIRAT 2.0 dataset [21], which are prepared by Brown University and DARPA, respectively. To simulate varying rhythms in these videos, we manipulated the original test videos in these datasets in three different ways and examined the sensitivity of the trained models with these methods (which were trained using the original rhythm videos in the training set) on the manipulated varying rhythm test videos.

Two of the investigated video activity recognition methods are Convolutional Neural Network - Long Short-Term Memory (CNN-LSTM) of which its source codes were found from [31] and Long Term Recurrent Convolutional Networks (LRCN) [32]. These two methods are considered as benchmark methods. The third method is CNN-IndRNN method [33], 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. The fourth method is called CNN-SkipRNN+ [33], or SkipRNN+ in short, which uses the same framework of IndRNN. However, SkipRNN+ has advantages over IndRNN by alleviating the gradient vanishing problem that occurs because of the many RNN (Recurrent Neural Network) layers used in the frame selection phase of the framework. VideoGraph [34] is the fifth and the last method. VideoGraph is a graph-based method in which the graph nodes are fully inferred from data and it is also extensible to datasets without node-level annotations. Similar to SkipRNN+ and other investigated methods, it also does not need annotations in sub-action level to train a model. VideoGraph learns an undirected graph from the video dataset. The nodes in the formed graph represent the key latent concepts (or the so-called sub-actions) that the human activity is composed of. The edges in the graph are considered to represent the temporal relationship between the latent concepts. VideoGraph is noted to model human activities for up-to thirty-minute videos [34]. It not only learns the graph nodes without any need for node-level annotation but also learns the relationships between graph nodes. The temporal structure of long-range human activities are represented via the constructed graph which is another interesting attribute of VideoGraph that can be utilized for visualization and video understanding. IndRNN, SkipRNN and VideoGraph are included in this work since these three methods were used with long and complex videos in some past works [33], [34].

In our results, the recognition results of VideoGraph were found to be superior to the other investigated methods reaching to close to 60% in the Breakfast dataset (Split-4), 92% for Breakfast 3-grouped class dataset, 92.5% accuracy in the VIRAT 4-event dataset, and over 62% in the VIRAT 6-event dataset. Among the five investigated methods, the varying rhythm sensitivity analysis investigations were conducted for IndRNN, SkipRNN+ and VideoGraph methods. The two conventional methods, CNN-LSTM and LRCN were applied to the original rhythm (R0) videos only. Since these two conventional methods had relatively lower recognition performance in the original rhythm (R0) case in the investigated datasets, no further investigation was conducted for the three varying rhythms. The sensitivity to varying rhythm results indicated that VideoGraph maintained its high recognition accuracy with varying rhythms. Some additional investigations with VideoGraph on the Breakfast dataset by varying some of the design parameters in its architecture also showed some slight performance improvements. Other than superior recognition results, VideoGraph's representation of activities via constructed graphs is demonstrated to bring significant value to the overall video understanding and activity recognition analyses.

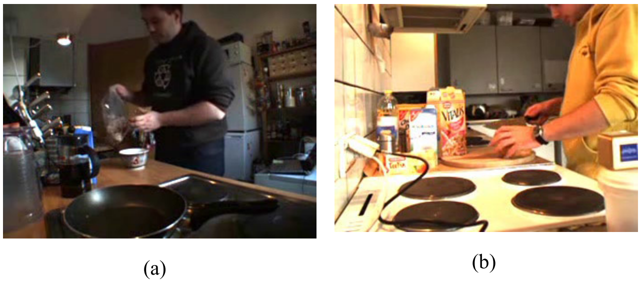
The most significant novelty of this article is providing a comprehensive evaluation of five video recognition algorithms with respect to their sensitivity to varying rhythms when long and complex videos are used. It is our thinking that in the evaluation of activity recognition methods, assessing their robustness to varying rhythms is an important measure which needs to be taken into account. The contributions of this article are as follows:

- We provided a comprehensive evaluation of five video activity recognition methods using two highly challenging activity recognition datasets with long and complex videos.
- We assessed the sensitivity of three of these methods to varying rhythms.
- We demonstrated that if similar activities are grouped in Breakfast dataset, the recognition performances can be improved for the grouped activity classes.
- We showed that by varying some of VideoGraph's design parameters, some performance improvements can be observed.

Our paper is organized as follows. Section 2 provides technical information about the investigated video activity

TABLE 1. Breakfast dataset main events and the number of videos for each event.

Event ID	Event type	Number of videos
1	Coffee	200
2	Orange Juice	187
3	Chocolate Milk	224
4	Tea	223
5	Bowl of Cereal	214
6	Fried Eggs	198
7	Pancakes	173
8	Fruit Salad	185
9	Sandwich	197
10	Scrambled Eggs	188


(a)

(b)

FIGURE 1. Sample images from the Breakfast Dataset.

recognition methods and the datasets used in our experiments. Section 3 contains the performance evaluations for the original videos and the sensitivity to varying rhythm results. Section 4 contains some discussions about the results. Finally, Section 5 concludes the paper with some remarks.

II. DATASETS AND METHODS

A. DATASETS

In the conducted analyses with the five video activity recognition methods, we used the RGB color images of the Breakfast and VIRAT dataset. Information about these datasets and data subsets formed from them are provided in the following.

1) BREAKFAST DATASET

The Breakfast dataset [20] was assembled by Serre Lab of Brown University. The videos in this dataset capture participants preparing breakfast food in many different kitchens at varying camera angles. There are 52 participants where each participant is denoted by *P*. Each participant was filmed in one of 18 different kitchens and with up to five different cameras from different angles and lighting conditions. The videos from these cameras film up to 10 different activities including making coffee, pouring orange juice, making chocolate milk, making tea, preparing a bowl of cereal, frying eggs, cooking pancakes, preparing a fruit salad, making a sandwich, and cooking scrambled eggs. Each video in the dataset is down sampled to 320 × 240 with a frame rate of 15 fps. This dataset was designed to be challenging in that it captured real world conditions with diverse range of lighting and environment. Table 1 shows the Breakfast dataset main events and the number of videos for each event. Fig. 1 shows sample image frames from the Breakfast dataset.

TABLE 2. Four splits in the breakfast dataset.

Split no	Train	Test
Split-1	P16 –P54	P03-P15
Split-2	P03-P15, P29-P54	P16 – P28
Split-3	P03-P28, P42-P54	P29 – P41
Split-4	P03-P41	P42 – P54

TABLE 3. Number of events in the three-group breakfast dataset.

Group Id	Event Description	Number of videos
1	{coffee, milk, tea, juice, cereals}	1048
2	{friedegg, pancake, scrambledegg}	559
3	{salad, sandwich}	382

There are four different splits in the Breakfast dataset for forming the training and testing datasets [20]. Table 2 shows the distributions of the videos which belong to the 52 participants (*P*) (P03-P54) in these four splits.

In the Breakfast dataset investigations, we considered Split-4. One interesting observation from the resultant confusion matrices was that the breakfast activities that are similar to each other like {coffee} and {milk}, or {friedegg} and {scrambled egg} were considerably confused among each other by the classifiers. We considered grouping these 10 breakfast activities into three major classes and formed a three-class version of the breakfast dataset. In addition to using the original 10-event Breakfast dataset, we also used this three-class Breakfast dataset version, and trained models with the five activity recognition methods to examine the recognition accuracy after grouping of similar activities. Among the three groups, the set of five activities, {coffee, milk, tea, juice, cereals} forms the first group. The second group consists of {friedegg, pancake, scrambledegg}. Finally, the third group consists of {salad, sandwich}. Table 3 shows the number of events in the three-group Breakfast dataset. In both Breakfast datasets (10-class and 3-grouped class) we used 65% of videos for training and 35% of videos for testing.

2) VIRAT DATASET

The VIRAT 2.0 dataset [21] is a publicly available video dataset supported by DARPA. The videos in this dataset consist of surveillance footage capturing public areas such as parking lots and college campuses. The VIRAT 2.0 dataset consists of high-definition videos and the original size of the image frames in these videos are 1920 × 1080 in size. Each video contains multiple activities with accompanied labels and bounding boxes. The classified activities in this dataset include: Loading an object, Unloading an object, Opening trunk, Closing trunk, Getting into vehicle, Getting out of a vehicle, Person gesturing, Person carrying an object, Person running, Person entering facility, and Person exiting a facility. Table 4 shows these events and the number of videos for each event. Some of these events, such as person loading an object to a vehicle have very few videos indicating a data

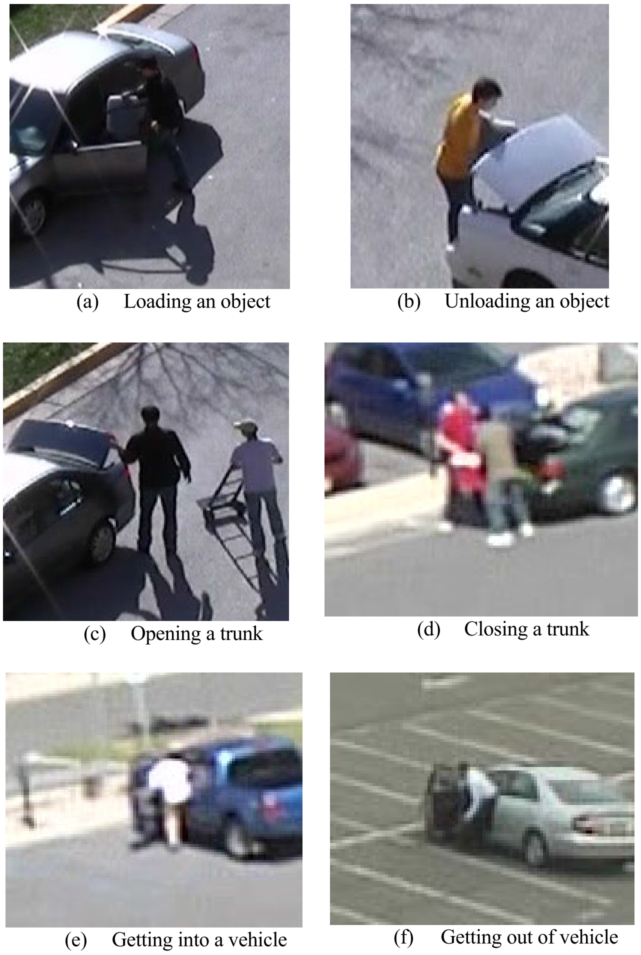


FIGURE 2. Sample images from the VIRAT Dataset.

TABLE 4. VIRAT 2.0 dataset main events and the number of videos for each event.

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

imbalance problem which poses challenges to applied activity recognition methods. A few image frames for the first six events in VIRAT dataset can be seen in Fig. 2.

We did not use all 13-events of the VIRAT dataset in this work and instead used subsets of it. The reason for this is that the number of videos for each event significantly varies in the VIRAT dataset with some of the events not having enough videos for effective model training as can be seen from Table 4. Because including all 13 events would have resulted in additional challenges such as a significant data imbalance problem with not enough videos for some events, we formed two smaller subsets of the original VIRAT dataset for our investigations. The first subset contains four events with close number of videos for each included event. The four classes in this four-event subset can be seen in Table 5. In the VIRAT dataset annotation files, for all the videos, event ids, event types, start and end frames of the events are provided together with the bounding box locations of the event within these annotation files. Videos in the VIRAT 2.0 database are cropped with respect to the event annotation files. Using the start and end frames in the event annotation files for the four events of interest, these image frames are considered as videos and used for activity recognition. For each of the four VIRAT events, 10 videos are randomly selected for validation purposes while the remaining videos for that event are used for training a model. That is, in the formed subset, there are 40 videos in the validation dataset (10 videos for each of the four events) and there are 457 videos for the four events in the training set (total 497 videos for four events). The high-resolution video image frames are cropped with respect to the bounding box regions.

TABLE 5. Number of events in the small four-event subset of VIRAT 2.0 dataset.

Event ID	Event Description	Number of videos
5	Person getting into a Vehicle	111
6	Person getting out of a Vehicle	97
11	Person entering a facility	156
12	Person exiting a facility	133

TABLE 6. VIRAT 6-event dataset and the number of videos for each event.

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

The second VIRAT subset used in the investigations consists of six events which relate to all six human-vehicle interactions as can be seen in Table 6. This 6-event VIRAT data subset is more challenging than the VIRAT 4-event data subset since the VIRAT 6-event data subset is imbalanced and some activities do not have enough number of videos (such as Event-1, Event-3 and Event-4). This poses additional challenges for the video activity recognition methods. 90% (train)-10% (test) random split is used with this data subset in our investigations.

B. METHODS

1) CNN-LSTM

CNN-LSTM approach [33] first extracts features from the image frames of the video with a Convolutional Neural Network (CNN) and forms features sequences. These feature

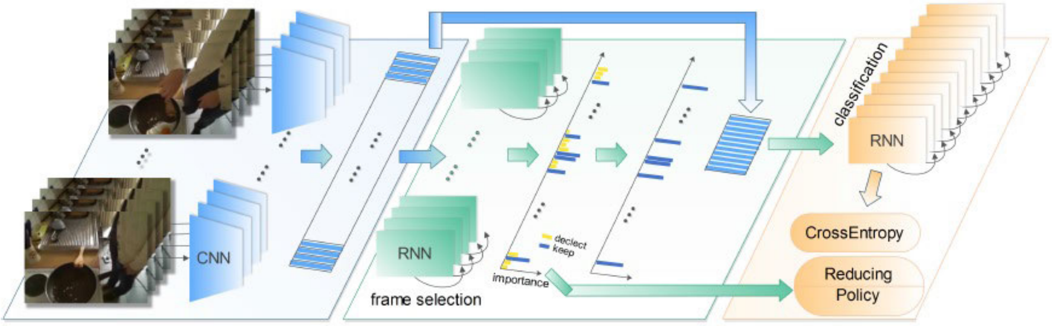


FIGURE 3. SkipRNN+ architecture [33].

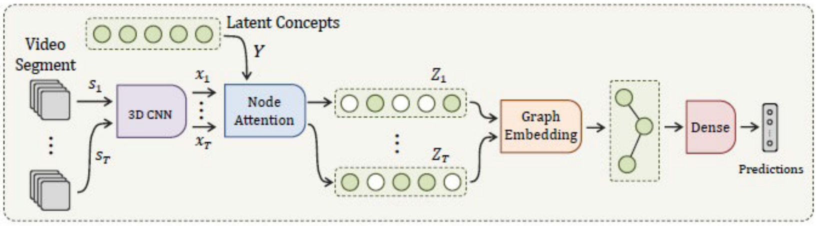


FIGURE 4. VideoGraph block diagram [34].

sequences are passed to a separate LSTM, which is a type of a Recursive Neural Network (RNN) with some additional units [35].

2) LRCN

LRCN makes use of a pretrained CNN in conjunction with a LSTM unit [32]. During the training of a LRCN model, each training frame in a video is individually passed through the CNN where a vector of features is created. These features are then passed on to the LSTM unit. A prediction is generated from the LSTM unit and its state is also passed to the LSTM unit in the next frame until all frames are processed in that video. The predictions across all frames are averaged to get a final prediction for that particular video.

3) CNN-INDEPENDENT RNN (INDRNN)

IndRNN [33] is inspired in part by how humans identify events with varying rhythms by quickly catching frames contributing most to a specific event. The CNN part consists of a VGG16 network and is used to extract visual feature per frame. The RNN part consists of two layers. The most significant frames are selected in the first RNN layer via the use of a regularization term which is included when computing the final loss of the model. The second RNN layer recognizes the event using the selected frames and a cross-entropy based loss is utilized in the recognition part. The sum of regularization term controlled by a parameter and cross entropy loss becomes the final loss of the model. For the classification RNN, Gated Recurrent Units (GRU) [36] is used. In this framework, only activity-level labels are needed in the training stage with no need of sub-action labels.

4) CNN-SKIP RNN (SKIPRNN+)

The details of SkipRNN+ can be found in [33]. In IndRNN method, because the input dimension to the IndRNN layer is high (4096), the output value in the stacked IndRNN layers increase by orders of magnitude resulting in the gradient vanishing problem [33]. In SkipRNN+ method, to mitigate this problem, an improved IndRNN structure is used by skipping state updates to shorten the computation. This idea is originally inspired by [37] which implements skip operation on conventional RNN. Unlike [16], SkipRNN+ structure uses Hadamard's product [38] when computing the gate value. This way the gradient of the SkipRNN+ depends on the weight value instead of the weight matrix product alleviating the gradient vanishing problem. An illustration of SkipRNN+'s architecture is shown in Fig. 3.

5) VIDEOGRAPH

Graph methods, which learn structured representations from videos, are being investigated for human activity recognition in the past [23]–[25]. Even though these graph based methods 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 [34] 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 Fig. 4. The video is first sampled into T segments and each segment, s_i, contains 8 consecutive frames. Using Two-Stream Inflated 3D ConvNet (I3D), which is a 3D CNN model [39], 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 provides 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 feed-forwarded to a classifier to arrive at the activity prediction of the video.

III. RESULTS

In addition to applying the investigated methods to the videos with the original rhythm (R0), we also demonstrated the impact of varying rhythm via three other rhythms (R1, R2 and R3) [33]. The testing video sequences have the same sampling rate as the training inputs in the original rhythm (R0). The other three varying rhythm scenarios are designed with different kinds of sampling rates. To prepare the three varying rhythms, the number of frames of each testing video is first divided into three equal intervals and different sampling rates are applied to each interval to form a new testing sequence. To generate the first rhythm (R1), the first and the third intervals are subsampled with every two and five frames respectively to make those two interval periods sparser, while keeping the rhythm intact for the middle interval. The testing inputs of the second rhythm (R2) are similar to R1 except the first and third intervals are subsampled every five and two frames, respectively. As can be noticed this is the reverse of R1. For the last rhythm (R3), half length of the testing video is randomly sampled. All five methods were applied to the original rhythm (R0) videos whereas the varying rhythm sensitivity investigation was conducted only for three methods which are IndRNN, SkipRNN+ and VideoGraph. This is because, overall, the other two methods, LSTM and LRCN, had relatively lower recognition performance in the original rhythm (R0) case and no further investigation was considered for the three varying rhythms.

For performance comparison of the video activity recognition methods, we used the overall accuracy (OA) and Kappa metric [40] measures. Other than these, confusion matrices are also generated to examine which of the activities are generally confused with each other.

In the following, for each dataset and their subsets, we first provide a table that shows the overall accuracy (OA) and Kappa values for the five methods with four rhythms. A bar plot showing the overall accuracies of these methods with four rhythms is provided next. The resultant confusion matrices that belong to the highest overall accuracy for each dataset are also included. The constructed graphs with VideoGraph for the activities in the Breakfast-10 event, VIRAT 4-event and VIRAT 6-event datasets are presented with some brief discussion as well.

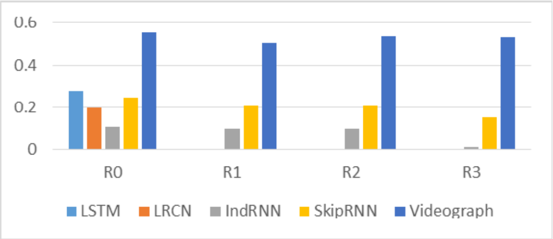


FIGURE 5. Breakfast 10-class dataset overall accuracy comparisons for four rhythms.

A. BREAKFAST 10-EVENT RESULTS

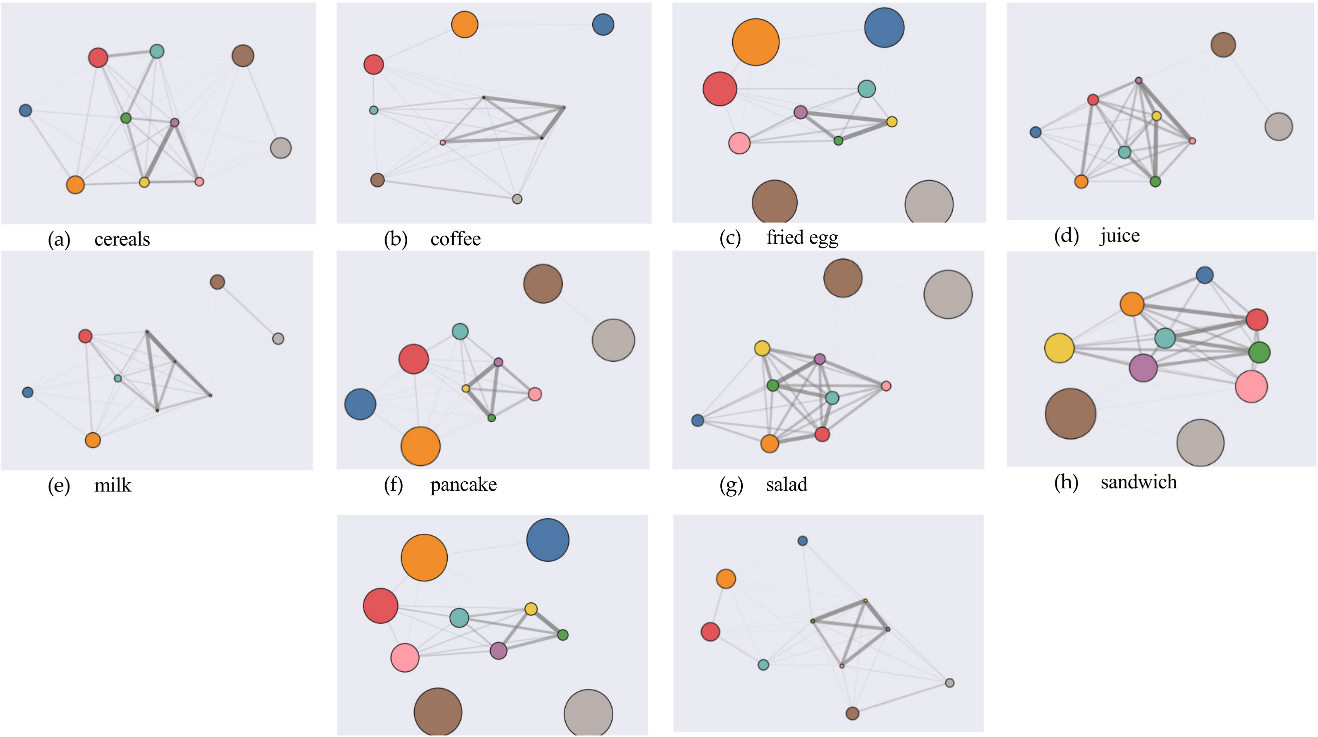
Table 7 shows the 10-event Breakfast dataset results (Split-4) for the original rhythm (R0) and three different rhythms (R1, R2, R3) with five activity recognition methods. For VideoGraph, we used the default '64 segments/8 frames' parameter setting. Figure 5 shows the overall accuracy values for the five methods in a bar plot. From these results, it can be seen that VideoGraph significantly outperforms all other methods and the performance gap between VideoGraph and the next best method is quite wide. VideoGraph is observed to perform well with varying rhythms as well. VideoGraph manages to maintain its original rhythm recognition performance for the varying rhythms and its overall accuracy variation is found to be relatively less in comparison to other three methods. The confusion matrix of the best performing case of VideoGraph is shown in Table 8. From the confusion matrix, it can be observed that breakfast events similar to each other like {cereals} and {milk}, or {fried egg} and {scrambled egg} were confused with each other. Figure 6 shows constructed graphs with VideoGraph for three of the 10 events in the Breakfast dataset. In each constructed graph for an event, the nodes correspond to the latent concepts learned by VideoGraph's graph-attention block. If a node's size is big, it indicates that latent concept is dominant. The edges in the graph emphasize the relationship between these latent concepts represented in the form of nodes. It can be noticed that the node sizes and edge formations are similar in fried egg and scrambled egg events whereas the corresponding graphs of these two events are quite different than the graphs of cereals and milk. Yet, the graphs of cereals and milk also show similarities to each other. We can see more confusions among events when their graphs are similar to each other. The graph representations in VideoGraph can thus add significant value to the recognition and video interpretation analyses.

TABLE 7. Breakfast dataset (10-event) Split-4 results for the original rhythm (R0) and three different rhythms (R1, R2, R3).

	R0 (OA/Kappa)		R1		R2		R3	
	OA	Kappa	OA	Kappa	OA	Kappa	OA	Kappa
LSTM	0.2754	0.1933						
LRCN	0.1982	0.042						
IndRNN	0.1094	0.0087	0.101	-0.00029	0.101	-0.0000249	0.0145	0
SkipRNN	0.248	0.166	0.207	0.140	0.205	0.113	0.155	0.099
VideoGraph	0.5549	0.5053	0.5072	0.4526	0.5383	0.4868	0.5342	0.4823

TABLE 8. Confusion matrix for Breakfast dataset's best overall accuracy in original rhythm (R0): VideoGraph, 55.49%.

	cereals	coffee	friedegg	juice	milk	pancake	salad	sandwich	scr_egg	tea
cereals	16	4	0	0	27	1	0	0	0	1
coffee	1	22	0	0	15	4	1	0	0	3
friedegg	0	0	21	0	1	12	1	0	16	0
juice	0	0	0	24	7	0	7	2	9	0
milk	4	2	0	0	41	0	1	1	0	0
pancake	0	0	3	0	0	24	2	3	15	0
salad	0	0	2	0	1	0	36	9	0	0
sandwich	0	1	0	0	2	1	8	31	1	0
scr_egg	0	0	2	0	2	8	0	4	35	0
tea	1	3	3	2	19	0	1	2	0	18



(a) cereals (b) coffee (c) fried egg (d) juice

(e) milk (f) pancake (g) salad (h) sandwich

(i) scrambled egg (g) tea

FIGURE 6. Constructed graphs with VideoGraph for 10 breakfast events.

B. BREAKFAST 3-CLASS RESULTS

Figure 7 and Table 9 correspond to the three-class Breakfast dataset results (Split-4) for the original rhythm (R0) and three different rhythms (R1, R2, R3). The default setting of '64 segments/8 frames' is used in VideoGraph. A similar performance trend is observed and VideoGraph performs significantly better, reaching to an overall accuracy of ~92 % in the original rhythm (R0). We also included the confusion matrix for VideoGraph with the original rhythm (R0) in Table 10. The recognitions are also found to be extremely good with VideoGraph for the three varying rhythms. Although this can be considered as an imbalanced dataset, VideoGraph's performance reaching to ∼ 92 % overall accuracy is quite significant.

TABLE 9. Three-class Breakfast dataset Split-4 results for the original rhythm (R0) and three different rhythms (R1, R2, R3).

	R0 (OA/Kappa)		R1		R2		R3	
	OA	Kappa	OA	Kappa	OA	Kappa	OA	Kappa
LSTM	0.7578	0.5948						
LRCN	0.49	0						
IndRNN	0.582	0.221	0.516	0.0344	0.516	0.0344	0.509	0.0197
SkipRNN	0.768	0.597	0.673	0.406	0.669	0.399	0.615	0.290
VideoGraph	0.9193	0.8678	0.8861	0.8099	0.9089	0.8494	0.9006	0.8347

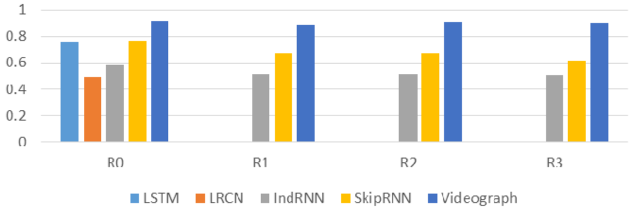


FIGURE 7. Three-class Breakfast dataset overall accuracy comparison using bar charts.

TABLE 10. Confusion matrix for three-class Breakfast dataset best overall accuracy in original rhythm (R0): VideoGraph, 91.93%.

	Group1	Group2	Group3
Group1	233	2	7
Group2	3	142	4
Group3	16	7	69

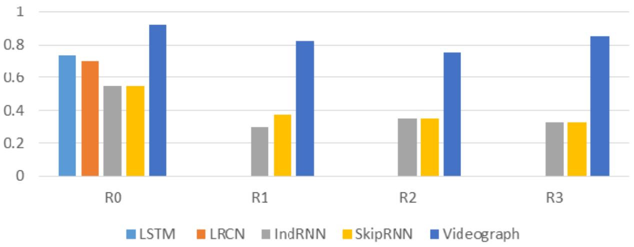
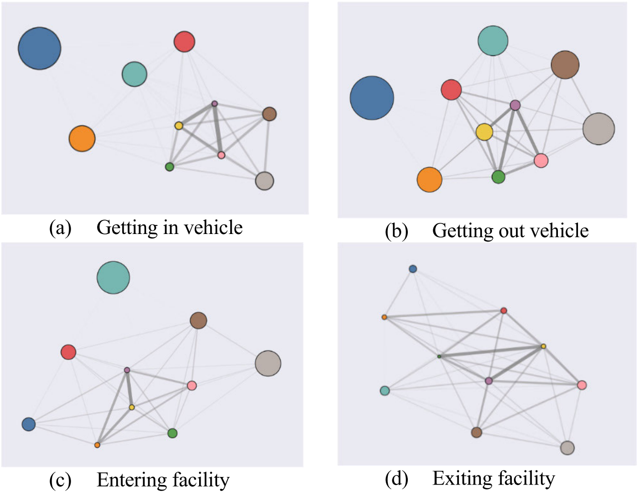


FIGURE 8. VIRAT 4-event dataset overall accuracy comparison using bar charts.


(a) Getting in vehicle

(b) Getting out vehicle

(c) Entering facility

(d) Exiting facility

FIGURE 9. Constructed graphs with VideoGraph for all four events of VIRAT-4 event dataset.

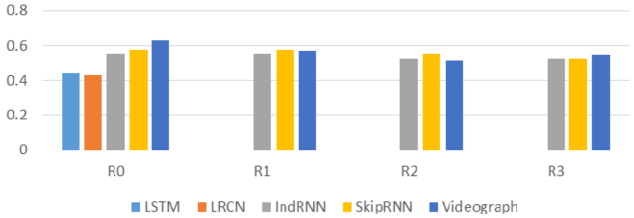


FIGURE 10. VIRAT 6-event dataset overall accuracy comparison using bar charts.

C. VIRAT 4-EVENT RESULTS

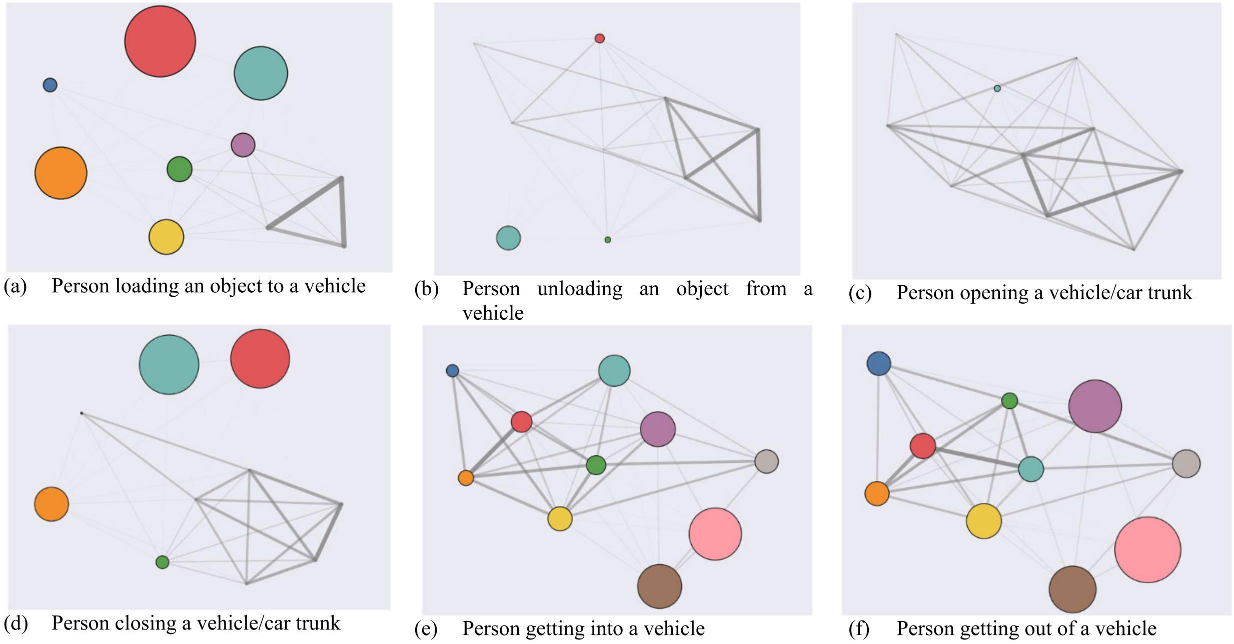
Figure 8 and Table 11 correspond to VIRAT 4-event dataset results. The default parameter setting (64 segments/8 frames) is used in VideoGraph. Similarly, VideoGraph performs superior to other methods, reaching to an overall accuracy of 92.5% in the original rhythm. The corresponding confusion matrix for the best VideoGraph case is shown in Table 12. The recognitions are also found to be considerably well with VideoGraph for the three varying rhythms. The constructed graphs with VideoGraph for VIRAT 4-event dataset can be seen in Figure 9. From Figure 9, it is interesting to observe that the nodes in the graphs for ‘Getting in vehicle’ and ‘Getting out vehicle’ events significantly differ from the graphs of the two other events which are ‘Getting in facility’ and ‘Getting out facility’. That is, the differences between the graphs of human-car and human-facility interaction events can be clearly observed.

D. VIRAT 6-EVENT RESULTS

Figure 10 and Table 13 correspond to the six-event VIRAT dataset results. The default parameter setting of ‘64 segments/8 frames’ is used for VideoGraph. This is not only a highly imbalanced dataset but also contains very small number of videos for some of the events. From the results, we can see that VideoGraph performs better than others especially in the original rhythm and reaches to an overall accuracy of ∼63%. The confusion matrix of the best performing case (VideoGraph) is shown in Table 14. However, there is not a wide performance gap between VideoGraph and the other methods as was previously observed in the former three datasets. It is thought that being an imbalanced dataset and containing not enough number of videos for some of the activities could be contributing to this result. In any event, overall, VideoGraph still performs considerably better than the others especially in the original rhythm. Figure 11 shows

TABLE 11. VIRAT 4-Event Dataset results for the original rhythm (R0) and three different rhythms (R1, R2, R3).

	R0 (OA/Kappa)		R1		R2		R3	
	OA	Kappa	OA	Kappa	OA	Kappa	OA	Kappa
LSTM (50 frames)	0.7333	0.6308						
LRCN	0.7	0.6						
IndRNN	0.550	0.400	0.300	0.067	0.350	0.133	0.325	0.100
SkipRNN	0.550	0.400	0.375	0.167	0.350	0.133	0.325	0.100
VideoGraph	0.9250	0.8944	0.825	0.8241	0.75	0.6667	0.85	0.80



(a) Person loading an object to a vehicle

(b) Person unloading an object from a vehicle

(c) Person opening a vehicle/car trunk

(d) Person closing a vehicle/car trunk

(e) Person getting into a vehicle

(f) Person getting out of a vehicle

FIGURE 11. Constructed graphs with VideoGraph for all six events in VIRAT-6 event dataset.

TABLE 12. Confusion matrix for VIRAT 4-Event dataset best overall accuracy in original rhythm (R0): VideoGraph, 92.50%.

	Getting in	Getting out	Entering facility	Exiting facility
Getting in (vehicle)	8	2	0	0
Getting out (vehicle	0	10	0	0
Entering facility	0	0	10	0
Exiting facility	0	0	1	9

the constructed graphs for the six events in VIRAT-6-event dataset.

E. CHANGING SEGMENT AND FRAME NUMBER PARAMETERS IN VIDEOGRAPH

For VideoGraph, in addition to the default ‘64 segments/8 frames’ parameter setting, two other segment/frame combinations are considered as well. This investigation was conducted using the Breakfast 10-event dataset. Table 15 shows the resultant performance metrics for three parameter combinations of VideoGraph including the default setting of ‘64 segments and 8 frames’. It can be noticed that when using ‘16 segments/32 frames’, relatively a higher recognition accuracy is achieved in the original rhythm and also in two of the three simulated varying rhythms. Table 16 shows the confusion matrix for the ‘16 segments/32 frames’ case which provided the highest overall accuracy in the original rhythm.

F. COMPUTATION TIME COMPARISON

The computation time comparison of the five investigated methods using the Split-4 of the Breakfast 10-event dataset can be seen in Table 17. The comparisons are with respect to feature extraction time, training time and test times. The computer platforms used for retrieving these times are also provided.

IV. DISCUSSIONS

The investigations with Breakfast and VIRAT datasets which contain long and complex videos clearly showed that among the five investigated activity recognition methods, VideoGraph performs significantly better than the others. Especially in the 10-event Breakfast dataset, VideoGraph’s classification performance is distinctively better than SkipRNN+ (VideoGraph: 59.21% vs SkipRNN+: 24.8%). Similarly, in VIRAT-4 event dataset, the performance gap between VideoGraph and SkipRNN+ is quite wide

TABLE 13. VIRAT 6-Event dataset results for the original rhythm (R0) and three different rhythms (R1, R2, R3).

	R0 (OA/Kappa)		R1		R2		R3	
	OA	Kappa	OA	Kappa	OA	Kappa	OA	Kappa
LSTM	0.4412	0.2651						
LRCN	0.4286	0.2568						
IndRNN	0.5526	0.4268	0.5526	0.4278	0.5263	0.3989	0.5263	0.3936
SkipRNN	0.5789	0.4532	0.5789	0.4614	0.5526	0.4328	0.5263	0.3947
VideoGraph	0.6286	0.5206	0.5714	0.4826	0.514	0.4297	0.5428	0.4615

TABLE 14. Confusion matrix for VIRAT 6-event dataset' best overall accuracy in original rhythm (R0): VideoGraph, 62.86%.

	Loading an object	Unloading an object	Opening trunk	Closing trunk	Getting in vehicle	Getting out vehicle
Loading an object	0	1	0	1	0	0
Unloading an object	0	3	0	1	1	0
Opening trunk	0	1	0	3	0	0
Closing trunk	0	0	1	3	0	0
Getting in vehicle	0	1	0	0	10	0
Getting out vehicle	0	0	1	2	0	6

TABLE 15. Breakfast dataset (10-event) Split-4 results for the original rhythm (R0) and three different rhythms (R1, R2, R3) with different segment and frame number combinations in VideoGraph.

	R0 (OA/Kappa)		R1		R2		R3	
	OA	Kappa	OA	Kappa	OA	Kappa	OA	Kappa
VideoGraph(64seg/8 fr)	0.5549	0.5053	0.5072	0.4526	0.5383	0.4868	0.5342	0.4823
VideoGraph(32seg/16 fr)	0.5859	0.5396	0.5797	0.5330	0.5693	0.5214	0.5652	0.5168
VideoGraph(16seg/32 fr)	0.5921	0.5466	0.5528	0.5030	0.5963	0.5512	0.6128	0.5711

TABLE 16. Confusion matrix for Breakfast dataset's best overall accuracy in original rhythm (R0) using '16 segments/32 frames': VideoGraph, 59.21%.

	cereals	coffee	friedegg	juice	milk	pancake	salad	sandwich	scr_egg	tea
cereals	38	1	0	0	8	0	0	0	0	1
coffee	9	22	0	0	10	0	0	0	0	5
friedegg	1	0	34	0	0	2	3	3	8	0
juice	0	1	0	43	3	0	0	2	0	0
milk	11	4	0	0	32	0	1	0	0	1
pancake	1	0	10	2	0	19	0	2	13	0
salad	0	0	0	7	5	0	20	16	0	0
sandwich	1	0	0	4	4	1	6	27	1	0
scr_egg	0	0	3	2	0	4	0	2	40	0
tea	3	5	0	3	13	0	1	3	0	21

(VideoGraph: 92.5% vs SkipRNN+: 55.0%). The same performance trend can be also observed in the other two datasets. VideoGraph is also found to be less sensitive to varying rhythms because it provided accuracy values close to the accuracy value with the actual rhythm for all three varying rhythms. One other analysis with VideoGraph on the 10-event Breakfast dataset was to examine the recognition performance when the segment and frame number parameters are varied. We observed that some parameter combinations provided better results than VideoGraph's default parameter setting and this showed that there could be more room to further improve the accuracy values by varying these parameters and some other parameters such as kernel sizes used in graph embedding layer in VideoGraph's architecture. The constructed graphs with VideoGraph demonstrated that these graphs have the potential to add significant value to the overall video understanding and activity recognition analyses which could be further tapped into and exploited. The results for the Breakfast-3-grouped class dataset also provided some potential future investigation ideas with VideoGraph and other classifiers in the sense that if a set of additional classifiers trained specifically for the activities within each of the three groups are applied, this second layer of classifiers could perhaps further boost VideoGraph's performance for the 10-event case via a potential two-step activity recognition framework (three-grouped class classification followed by individual classifications for each group). Another future investigation idea is to examine VideoGraph's recognition performance on video datasets that consist of videos with varying image resolutions and various actors in the scene that are captured with moving cameras such as the UCF-Crime dataset [41].

TABLE 17. Computation Time Comparison of the Investigated Methods Using Breakfast dataset (10-event) Split-4.

Method	Computer specs	Feature extraction time (min)	Training (min)	Test (min)
CNN-LSTM	Windows 10, 16GB RAM, i7-9700K, GPU RTX 2070	25	20	1
LRCN	Windows 10, 16GB RAM, i7-9700K, GPU GTX Titan Black	30	35	3
IndRNN	Ubuntu 18.04, 16GB RAM, i7-4790K, GPU GTX Titan Black	195	105	3
SkipRNN	Ubuntu 14.06, 24GB RAM, i7-4790, GPU GTX Titan X	2796	3900	6
VideoGraph	Windows 10, 16GB RAM, i7-9700K, GPU RTX 2070	81	76	2

V. CONCLUSION

Robustness to varying rhythms can be a discerning measure when comparing the performance of activity recognition methods since the rhythm of sub-actions in an activity can differ in nature and pose challenges for the activity recognition methods due to the fact not all rhythm variations can be included in training dataset for model learning. This article contained comprehensive investigations of five video activity recognition methods with two datasets that consist of long and complex videos in consideration of varying rhythms. The results showed that among them, VideoGraph performs significantly better than others and is found to be less sensitive to varying rhythms since it provided accuracy values for varying rhythms close to the accuracy values observed with the original rhythm. Having noted some performance improvements after varying some of VideoGraph's parameters also indicated that there could be more room for improvement in VideoGraph by searching optimal hyperparameters.

ACKNOWLEDGMENT

The views, opinions and/or findings expressed are those of the author and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.

REFERENCES

[1] S. S. Rautaray and A. Agrawal, "Vision based hand gesture recognition for human computer interaction: A survey," *Artif. Intell. Rev.*, vol. 43, no. 1, pp. 1–54, Jan. 2015.

[2] S. N. Paul and Y. J. Singh, "Survey on video analysis of human walking motion," *Int. J. Signal Process., Image Process. Pattern Recognit.*, vol. 7, no. 3, pp. 99–122, Jun. 2014.

[3] S. Vishwakarma and A. Agrawal, "A survey on activity recognition and behavior understanding in video surveillance," *Vis. Comput.*, vol. 29, no. 10, pp. 983–1009, Oct. 2013.

[4] W. Lin, M.-T. Sun, R. Poovandran, and Z. Zhang, "Human activity recognition for video surveillance," in *Proc. IEEE Int. Symp. Circuits Syst.*, May 2008, pp. 2737–2740.

[5] T. V. Duong, H. H. Bui, D. Q. Phung, and S. Venkatesh, "Activity recognition and abnormality detection with the switching hidden semi-Markov model," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, San Diego, CA, USA, vol. 1, Jun. 2005, pp. 838–845.

[6] Y.-M. Kuo, J.-S. Lee, and P.-C. Chung, "A visual Context-Awareness-Based sleeping-respiration measurement system," *IEEE Trans. Inf. Technol. Biomed.*, vol. 14, no. 2, pp. 255–265, Mar. 2010.

[7] H. H. Huynh, J. Meunier, J. Sequeira, and M. Daniel, "Real time detection, tracking and recognition of medication intake," *World Acad. Sci. Eng. Technol.*, vol. 3, no. 12, pp. 2801–2808, Dec. 2009.

[8] H. Foroughi, B. S. Aski, and H. Pourreza, "Intelligent video surveillance for monitoring fall detection of elderly in home environments," in *Proc. 11th Int. Conf. Comput. Inf. Technol.*, Khulna, Bangladesh, Dec. 2008, pp. 219–224.

[9] C. Liu, P. Chung, Y. Chung, and M. Thonnat, "Understanding of human behaviors from videos in nursing care monitoring systems," *J. High Speed Netw.*, vol. 16, no. 1, pp. 91–103, 2007.

[10] J. Zhou and C. Kwan, "Anomaly detection in low quality traffic monitoring videos using optical flow," *Proc. SPIE*, vol. 10649, Apr. 2018, Art. no. 106490F.

[11] C. Kwan, J. Zhou, Z. Wang, and B. Li, "Efficient anomaly detection algorithms for summarizing low quality videos," *Proc. SPIE*, vol. 10649, Apr. 2018, Art. no. 1064906.

[12] C. Kwan, J. Yin, and J. Zhou, "The development of a video browsing and video summary review tool," *Proc. SPIE*, vol. 10649, Apr. 2018, Art. no. 1064907.

[13] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," 2012, *arXiv:1212.0402*. [Online]. Available: http://arxiv.org/abs/1212.0402

[14] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local SVM approach," in *Proc. 17th Int. Conf. Pattern Recognit. (ICPR)*, Cambridge, U.K., vol. 3, Aug. 2004, pp. 32–36.

[15] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: A large video database for human motion recognition," in *Proc. Int. Conf. Comput. Vis.*, Barcelona, Spain, Nov. 2011, pp. 2556–2563.

[16] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman, "The kinetics human action video dataset," 2017, *arXiv:1705.06950*. [Online]. Available: http://arxiv.org/abs/1705.06950

[17] H.-B. Zhang, Y.-X. Zhang, B. Zhong, Q. Lei, L. Yang, J.-X. Du, and D.-S. Chen, "A comprehensive survey of vision-based human action recognition methods," *Sensors*, vol. 19, no. 5, p. 1005, Feb. 2019.

[18] X. Wang, L. Gao, P. Wang, X. Sun, and X. Liu, "Two-stream 3-D convNet fusion for action recognition in videos with arbitrary size and length," *IEEE Trans. Multimedia*, vol. 20, no. 3, pp. 634–644, Mar. 2018.

[19] X. Wang, L. Gao, J. Song, and H. Shen, "Beyond frame-level CNN: Saliency-aware 3-D CNN with LSTM for video action recognition," *IEEE Signal Process. Lett.*, vol. 24, no. 4, pp. 510–514, Apr. 2017.

[20] Serre Lab, A Brown University Research Group. *The Breakfast Actions Dataset*. Accessed: Apr. 29, 2020. [Online]. Available: http://serre-lab.clps.brown.edu/resource/breakfast-actions-dataset/

[21] *The VIRAT Video Dataset*. Accessed: Apr. 29,2020. [Online]. Available: https://viratdata.org/

[22] J. Song, L. Gao, F. Nie, H. T. Shen, Y. Yan, and N. Sebe, "Optimized graph learning using partial tags and multiple features for image and video annotation," *IEEE Trans. Image Process.*, vol. 25, no. 11, pp. 4999–5011, Nov. 2016.

[23] X. Wang and A. Gupta, "Videos as space-time region graphs," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, 2018, pp. 399–417.

[24] D. A. Huang, S. Buch, L. Dery, A. Garg, L. Fei-Fei, and J. C. Niebles, "Finding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 5948–5957.

[25] R. Girdhar, J. Joao Carreira, C. Doersch, and A. Zisserman, "Video action transformer network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 244–253.

[26] H. Kuehne, J. Gall, and T. Serre, "An end-to-end generative framework for video segmentation and recognition," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Lake Placid, NY, USA, Mar. 2016, pp. 1–8.

[27] L. Ding and C. Xu, "TricorNet: A hybrid temporal convolutional and recurrent network for video action segmentation," 2017, *arXiv:1705.07818*. [Online]. Available: http://arxiv.org/abs/1705.07818

[28] C. Fermüller, F. Wang, Y. Yang, K. Zampogiannis, Y. Zhang, F. Barranco, and M. Pfeiffer, "Prediction of manipulation actions," *Int. J. Comput. Vis.*, vol. 126, nos. 2–4, pp. 358–374, Apr. 2018.

[29] C. Xu and L. Ding, "Weakly-supervised action segmentation with iterative soft boundary assignment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 6508–6516.

[30] H. Kuehne, A. Arslan, and T. Serre, "The language of actions: Recovering the syntax and semantics of goal-directed human activities," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, Jun. 2014, pp. 780–787.

[31] *Five Video Classification Methods Implemented in Keras and TensorFlow*. Accessed: Apr. 29, 2020. [Online]. Available: https://blog.coast.ai/five-video-classification-methods-implemented-in-keras-and-tensorflow-99cad29cc0b5?

[32] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, T. Darrell, and K. Saenko, "Long-term recurrent convolutional networks for visual recognition and description," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 2625–2634.

[33] Y. Li, T. Yu, and B. Li, "Recognizing video events with varying rhythms," 2020, *arXiv:2001.05060*. [Online]. Available: http://arxiv.org/abs/2001.05060

[34] N. Hussein, E. Gavves, and A. W. M. Smeulders, "VideoGraph: Recognizing minutes-long human activities in videos," 2019, *arXiv:1905.05143*. [Online]. Available: http://arxiv.org/abs/1905.05143

[35] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[36] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," 2014, *arXiv:1412.3555*. [Online]. Available: http://arxiv.org/abs/1412.3555

[37] V. Campos, B. Jou, X. Giro-i-Nieto, J. Torres, and S.-F. Chang, "Skip RNN: Learning to skip state updates in recurrent neural networks," 2017, *arXiv:1708.06834*. [Online]. Available: http://arxiv.org/abs/1708.06834

[38] G. Krishnamurthy, N. Majumder, S. Poria, and E. Cambria, "A deep learning approach for multimodal deception detection," 2018, *arXiv:1803.00344*. [Online]. Available: http://arxiv.org/abs/1803.00344

[39] J. Carreira and A. Zisserman, "Quo Vadis, action recognition. A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, Jul. 2017, pp. 6299–6308.

[40] G. Cardillo. *Cohen's Kappa: Compute the Cohen's Kappa Ratio on a 2 × 2 Matrix*. Accessed: Apr. 29, 2020. [Online]. Available: https://www.github.com/dnafinder/Cohen

[41] UCF-Crime Dataset. *Real-World Anomaly Detection In Surveillance Videos*. Accessed: Sep. 24, 2020. [Online]. Available: https://www.crcv.ucf.edu/projects/real-world/

BULENT AYHAN (Member, IEEE) received the Ph.D. degree in electrical engineering from North Carolina State University, Raleigh, in 2006. He is currently working as a Principal Research Engineer with Applied Research LLC. He has more than 100 journal and conference papers in prestigious journals and conferences. His research interests include machine learning, deep learning, artificial intelligence, signal processing, image processing, computer vision, pattern recognition, remote sensing, and condition monitoring.

CHIMAN KWAN (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from The University of Texas at Arlington, in 1993. He is currently the Chief Technology Officer of Signal Processing, Inc., and Applied Research LLC. He has published one book, four book chapters, over 375 journal and conference papers, and over 550 technical reports. He has served as the Principal Investigator and the Program Manager of more than 120 diverse projects in the past 25 years. He also received numerous awards from IEEE, including the Finalist for Kayamori Best Paper Award in 2003 IEEE International Conference on Robotics and Automation, the IEEE Mikio Takagi Best Student Paper Award in 2016 IEEE International Geoscience and Remote Sensing Symposium, and the Best Paper Award in 2019 IEEE Ubiquitous Computing, Electronics and Mobile Communication Conference.

BENCE BUDAVARI received the B.S. degree in audio engineering from Belmont University, in 2015. He is currently working as a Software Developer with Applied Research LLC.

JUDE LARKIN received the B.S. degree in computer science from the Franciscan University of Steubenville, in 2015. He is currently working as a Software Engineer with Applied Research LLC.

DAVID GRIBBEN received the B.S. degree in computer science and physics from the McDaniel College. He is currently working as a Software Developer with Applied Research LLC.

BAOXIN LI (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from the University of Maryland, College Park, in 2000. He is currently a Professor and the Chair of the Computer Science and Engineering Program with the School of Computing, Informatics, and Decision Systems Engineering, Arizona State University.