Single Modality-Based Event Detection Framework for Complex Videos

Sheeraz Arif[1], Adnan Ahmed Siddiqui[2], Rajesh Kumar[3]
Avinash Maheshwari[4], Komal Maheshwari[5], Muhammad Imran Saeed[6]
Department of IT, Barrett Hodgson University, Karachi, Pakistan[1]
Department of Computing, Hamdard University, Karachi, Pakistan[2, 3, 5, 6]
Department Applied Mathematics and Data science, Hochschule Mitteida University, Mitteida, Germany[4]

***Abstract*—Event detection of rare and complex events in large video datasets or in unconstrained user-uploaded videos on internet is a challenging task. The presence of irregular camera movement, viewpoint changes, illumination variations and significant changes in the background make extremely difficult to capture underlying motion in videos. In addition, extraction of features using different modalities (single streams) may offer computational complexities and cause abstraction of confusing and irrelevant spatial and semantic features. To address this problem, we present a single stream (RGB only) based on feature of spatial and semantic features extracted by modified 3D Residual Convulsion Network. We combine the spatial and semantic features based on this assumption that difference between both types of features can discover the accurate and relevant features. Moreover, introduction of temporal encoding builds the relationship in consecutive video frames to explore discriminative long-term motion patterns. We conduct extensive experiments on prominent publically available datasets. The obtained results demonstrate the great power of our proposed model and improved accuracy compared with existing state-of-the-art methods.**

***Keywords*—Event detection; single-stream; feature fusion; temporal encoding**

I. INTRODUCTION

Detection of events in complex and untrimmed videos has been the topic of great concern for many years. Furthermore, it is imperative for many real-world applications such as video indexing, video retrieval, and video surveillance. However, event detection in videos became very challenging due to the different environmental and video recording factors. Video captured from different devices show lots of variations such as variations in environment and variations in recording setting. Variations in environment are due to the occlusion, confusing background, rapid changes in background in video scene, camera motion, noise and viewpoint changes. Variations in video recording also cause different kinds of noise in different lighting conditions. In addition, video low resolution and its high dimensionality may also degrade accurate detection of complex events. Moreover, existing available event detection datasets are too complex and large amounts of uploaded videos on internet are captured in unconstrained conditions. To combat these challenges, there is an immense need for effective and robust activity recognition system to achieve best performance.

In contrast to the simple human action recognition, event detection is a semantic composition of many atomic concepts and there may be involvement of various objects and actors with their different locations and appearances. In addition, videos related to event detection may be of longer duration with multiple scenes and mostly focus on real-world scenarios. For example, the event of "wedding ceremony" in which there are so many related sub-activities with different actors and objects, which can infer the event with a high probability.

Over the past decade, several low level and high-level representations have been proposed to address the issues in context of video event detection. Early attempts are an extension of static image-based representations and pattern recognition. Initially, trajectory-based representations [1-3] have been introduced and obtained satisfactory results. These models utilized Gaussian mixture and Hidden Markov models for the extraction of trajectories and work well for detection of deviant trajectories in less crowded scenes. However, these trajectory-based methods are occlusion-sensitive and not ideal for crowded scenes. These issues are well addressed by hand-crafted methods for example, Histogram of arranged angle (HOG) [4], Histogram of Optical Flow (HOF) [5] and Motion Boundary Histogram (MBH) [6]. These models construct the template behavior and model the background, shape, appearance, and motion and yielded remarkable results. However, these models are only specific to the simple events and do not link between local patterns. Many methods followed Bag of Visual Words (BoVW) by applying dense sampling or detecting spatiotemporal interest points. However, these methods ignore the intrinsic difference between video volumes.

Recently, deep learning achieved a remarkable breakthrough in the image domain and many researchers start applying those learning Spatio-temporal clues by extending deep 2D Convolution Network with 3D Convolution Network [7-13]. These deep learning methods providing high discriminative power and have produced promising results for action recognition. However, CNN based strategy just concentrates visual appearance highlights and comes up short on the capacity to long-run worldly displaying. Most of the researchers implement temporal modeling by introducing two stream-based CNN learning models by applying an extra input stream known as stacked multi-frame dense optical flow along with raw RGB stream.

However, these two stream approaches are not able to capture the motion and semantic changes accurately and only limited to short-term temporal modeling. In light of the above discussion, this research paper proposes lightweight event detection framework by considering only RGB data and address the disadvantages of optical flow in complex and unconstrained videos. Based on assumption that motion can be represented in series of video frames and temporal dynamics of an actor/object can be observed by computing the difference between appearance and semantics. First, we extract two kinds of video features mainly spatial and semantic features by using convolutional and fully connected layers respectively by utilizing by modified 3D Residual Conv Network. Next, we join both low-level (spatial) and high-level information (semantic). To weaken the effect of semantic gap, we add extra learnable filters on the output of different layers. Then, frame-level representation is achieved by employing global average pooling. We also design attention model to take deep insight within the neural network to find important parts in video and ignoring the redundant features and background noise effect for finding the temporal discriminative patterns, temporal encoder is introduced to achieve clip-level representation. Finally, the specific classifier is used to identify the event. The main contributions of this research are listed as under:

1) We only consider RGB stream to extract both spatial and semantic information and extract the motion of the object via the changing of both features and take aside the use of optical flow.

2) We introduce global averaging pooling to represent frame-level representation along with attention mechanisms to learn temporal focus of action.

3) The temporal encoder is applied to detect motion in series of frames.

4) The proposed model experimentally demonstrates the super performance when evaluated on publically benchmark datasets and obtained state-of-the-art results.

The rest of article is organized as follows: Section 2 provides the high level of related works. In Section 3, we present our approach in detail. In Section 4, we demonstrate the experimental evaluation. Finally, conclusion is drawn in Section 5.

II. Related Work

For video analysis, many previous methods adopted a similar approach to image analysis. The video domain is different and complex from the image domain due to the ever-changing motion patterns with target actors/objects and their appearances in different scenes. For the accurate and robust video event detection motion resides in temporal dimension plays crucial role. Many spatiotemporal representation methods such as HOG, HOF, HOG3D [14] and SIFT3D [15] have been proposed to present the motion in a video sequence. In these models extracted features are encoded or pooled in hierarchal structure before feeding to the Support Vector Machine (SVM) classifier. To take the full advantage of motion features dense-point trajectory model [16] has been proposed. These all hand-crafted features models have shown remarkable performance, however, there are several weaknesses are present. These models are computationally expensive and do not consider the changes in semantic clues along the temporal dimension. In addition, features extracted by these schemes are not very discriminative and limited to only simple event detection. Recently, deep convolutional neural networks (DCNN) have achieved great success in many research areas such as object/action detection, classification, recognition. These networks have great potential to learn features automatically from a large datasets. Most of these networks are the natural extension from 2D CNN which are now using in time dimension to represent motion using 3D sensitive filters. More recently, Kinetic 3D networks such as ResNet-3D [16] and I3D [17] obtained great success in the area of action analysis and event detection. However, simultaneously learning appearance and motion brings complexity in the process. Most of the models adopted optical stream as an additional information methodology to catch movement portrayal, for example, [9]. In this model author utilized stacked of dense optical flow as extra stream along with RGB stream to extract static and motion features respectively. The phenomena of optical flow introduced computational complexity and also optical flow may not very robust and accurate capturing semantic and motion changes. In addition, this practice is not ideal for real-world untrimmed and unconstraint videos due to the irregular camera movement. Extra computation by optical flow may degrade the efficiency of event detection framework. Based on this analysis, it is required to re-think the capturing process of motion for complex event detection. This research study represents a single stream model for spatial and temporal feature extraction by considering only RGB frames. RGB frame represented by high dimensional features such as background, objects, and actors. RGB single frame usually encodes static information; however, this study observes the object motion by analyzing the difference in both extracted features i.e. appearance and semantic features. We follow the modified version of ResNet and utilized convolution and fully-connected layers to extract spatial and semantic features respectively. We combine these two features to obtain frame-level representation by using global average pooling. We also introduce a special type of encoding scheme i.e. temporal encoding to achieve clip-level representation.

III. Proposed Framework

Before this section provides a detailed description of our framework, which inputs untrimmed RGB frames of video and detects the event accordingly. The overall flowchart of our method is demonstrated in Fig. 1. Our framework mainly comprises of feature extractor, feature encoder, and classifier. We explain the detailed description of each step in the following sub-section.

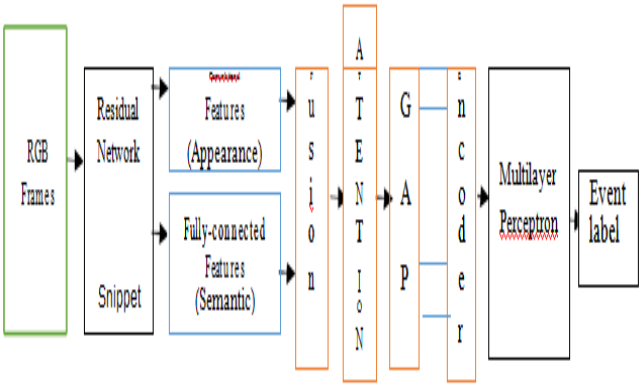

Fig. 1. Feature Extraction and Encoding.

A. Frame-level Representation

For the extraction of appearance and semantic feature representation the pre-processed RGB data is used as input. Deep Residual networks [18, 19] with deep layers are followed as our base network. We combine the low level and higher-level features by using fusion scheme and global average pooling (GAP) is adopted to achieve frame-level representation of video.

1) Residual network: Residual networks are developed similar to VGG networks [20] in which layers are arranged to learn residual function with respect to the given input. Residual networks play an important role to avoid the information bottleneck by introducing residual units. This practice allows skipping connection which permits direct signal propagation between the first and final layers of the network. This network is pre-trained on ImageNet as basic architecture. Residual networks comprise of small 3 x 3 spatial filters with 1 x 1 filters for learned dimensionality expansion and reduction. The network takes input of size 224 x 224 which is the reduced size by stride 2. To prevent the direct fitting of underlying mapping P(l), a residual mapping G(l) = P(l) – l is introduced by training deep network. We can represent the residual unit as follows:

$$l_{n\,1} \quad (l_n \quad G(l_{\,n}\,; Wn)) \qquad (1)$$

Where l_n and $l_{n\,1}$ are the input and output of the n^{th} layer of the network, $G(l_n\,; Wn)$ is a nonlinear mapping based on residual given by Weights of convolutional filters.

$Wn\ \{Wn,\ s\ |1\ s\ S\}$ with $^{S}\ \{2,3\}$, and represents the ReLU function. This practice can achieve direct propagation across all layers of network. In addition, the problem of gradient explosion and disappearance can be avoided. Another advantage is that short connection does not introduce extra computational complexity and parameters. Moreover, ResNet follows the batch normalization (BN) before the activation layer which not only addresses the issues of covariant shift but also speeds up the performance of network.

2) Extraction of appearance and semantic information: We utilize the RGB data as input to the ResNet and to perform the sampling on the data, we adopt two different sampling strategies i.e. dense sampling and sparse sampling. For thick testing, every video is partitioned into T cuts with length of 1-2 seconds and afterward we haphazardly select a picture/outline from each clasp and organize them in an arrangement {N1, N2… NT}. For meager examining, we select three edges of equivalent span from video arrangement and receive setting rules given in [21]. As referenced before, movement of any article/entertainer can be investigated by means of the distinction of both appearance and semantic highlights. The yield of the profundity layers for the most part gives the elevated level (semantic) highlights. In our base organization, the yield highlights of both convolution and the completely associated layers are extraordinary. The output of convolution layer is appearance features (outline, shapes, etc.), while fully-connected layer provides semantic features (rotation invariance, and location invariance). Our baseline CNN generates two feature maps for the nth frame. The last pooling layer of the network generates feature maps fcln and fully-connected layers outputs feature maps f fcln. Both feature maps having the dimension (W x H x D, C.), representing width, height, temporal depth and number of feature channels respectively. The matrix representation of both feature maps for the video length of duration T can be given as:

$$\text{fcln} \quad [\ \text{fcl1}\ ,\ldots,\ \text{fclt}\ ,\ldots,\ \text{fclT}\ ]\ ¡ \qquad \text{WHTC} \qquad (2)$$

$$\text{f fcln} \quad [\ \text{f fcl1}\ ,\ \text{f fcl2}\ ,\ldots,\ \text{f fclt}\ ,\ldots,\ \text{f fclT} \qquad]\ ¡\text{DC} \qquad (3)$$

3) Fusion and attention mechanism: In the previous sub-section, we obtain two feature maps produced by the pooling layer and FC layer. Next, we perform weighted linear fusion scheme to integrate both appearance and semantic features by employing pixel-wise operation. After fusion, we again obtain a frame blended with both spatial and semantic properties. Then, we apply attention mechanism by computing the weights of both appearance and semantic features. The purpose of attention model is to decide important frames in a video for event recognition. The attention mechanism is very close to human visual model as humans always concentrate and focus on moving objects instead of whole frame or static background. In addition, it plays important role to eliminate the effect of background noise and adds a dimension of interpretation ability. If we assume that W is the weight mapping of both appearance and semantic information of tth element of frame and N is the number of frames then the probability of informative frame can be represented as follow:

$$= \frac{\exp(W^{T})}{(W_{t}^{T})^{\,\mathrm{N}}} \qquad (4)$$

Where t is the probability with which the corresponding frame is considered an informative frame. Finally, we get the vector representation of each selected (attention mechanism) frame by using global average pooling. The function of the global average pooling (GAP) layer is to average the feature values of the respective pixels in each chosen frame, and the average value is taken as the probability value of each feature. After applying this pooling scheme, a video can be represented as a sequence of vector V = {v1, … vM }of M clip of input video. Each vm¡ d is the expression of M video sequence i.e. Sm.

B. Video Level Representation

As we mentioned earlier that temporal information is presented in sequence of video frames. To model the relationship between video frames we introduce a temporal encoder E. We can combine the different features from the entire video sequence into powerful and compact clip level representation. If V = {v1, … vM }is the input to the encoder then clip-level representation can be obtained by applying simple function or neural network. This research work, apply and compare the three different encoders mainly Average Encoder, Max Encoder, and LSTM Encoder. All these encoders take the sequence of vectors of M video clips and generate video representation as a single vector Z such as Z ¡ d. Where Z is the vector representation of video integrating the high-level semantic and low-level appearance features along with temporal relationship. We can define working of each the encoder as follows: Average encoder: This encoder performs the element-wise addition on the feature vectors and compute the single feature vector using the length of M video clips as:

$$Z \quad \frac{1}{M} \quad (v \quad v \quad v \ldots v \quad) \qquad (5)$$

1) Max encoder: This encoder represents a video by a single vector using maximum feature value (highly weighted) from the list of finite values and can be given as under:

$$Z \max(v1, v2, v3 \ldots vM) \qquad (6)$$

2) LSTM encoder: This encoder outputs the feature vector Z using the hidden state of the LSTM hj at time step j and feature vector vj.

$$h_j \quad LSTM(v_j, h_{j-1}), 0 \quad j \quad M \qquad (7)$$

C. Event Classification

We require a prediction function F(Z) to detect the event category for the given video. We adopt a multi-layer perceptron as classifier which comprises of FC-Dropout-FC pipeline. The dropout option is used to prevent the framework from overfitting. If yˆ is the prediction of classier and y is the ground-truth label of the video then final loss can be formulated as:

$$L(\hat{y}, y) \sum_{i=1}^{C} y_i (\hat{y} \quad \log \quad \exp y_j) \qquad (8)$$

In addition, if our temporal encoder E is differentiable, so our network can also be differentiable. We can utilize the multiple frames to jointly optimize the model parameter W with the standard back-propagation scheme. We can compute the loss P by using the chain rule using gradient W as follows:

$$\frac{P(\hat{y}, y)}{W} \quad \frac{P}{\hat{y}} \frac{\hat{y}}{Z} \sum_{i=1}^{M} \frac{Z}{v_i} \frac{v_i}{W} \qquad (9)$$

IV. Experiments and Analysis

Evaluate the performance of our proposed framework, we carried out several experiments on challenging publically available datasets. We analyze our introduced method using different aspects. The portrayal of datasets with their approval plans, trial arrangement, results, and similar examination are introduced in ensuing segments.

A. Datasets

For our experiments, we use well-known event detection datasets, namely YLI-MED [22], MEDTest-14 [23] and Columbia Consumer Video (CCV) [24].

YLI-MED: This dataset comprises of 1823 videos and each of them classified is classified into 10 event categories. These videos are divided into training (1000 videos) and testing (823 videos). The Videos length duration is variable, which makes event detection more challenging. We measure the accuracy of the test set for all experiments. Columbia Consumer Video (CCV): The Columbia Consumer Video dataset contains 9,317 videos in total from 20 semantic categories, including events like “parade” and “baseball”.

B. Implementation Setup and Details

We extract the RGB frames from the original video by using the guidelines given in FFMPEG [25]. For the training of our model, we augment the extracted images to reduce the effect of overfitting. We horizontally flip input images with 55% probability and then crop them by resizing of 320 x 240. We scale the height and width of cropping rectangle by a randomly selected factor of in the range of 0.8 ~1. We utilize ResNet50 which network weights are initialized by pre-training on ImageNet. We replace the final classifier with a two-layer perceptron. The unit number of FC layer is set to 512. The dropout ratio is set to 0.8. For LSTM encoders, we set up one hidden layer with 512 units. The momentum of stochastic gradient is selected as 0.9 for optimizing the model. All experiments are conducted on a single GPU with weight decay of 1x10-4 and mini-batch of size 16. The initial learning rate is set as 0.003 and decreased to 12% at 150 epochs. The whole training procedure is stopped at 300 epochs.

C. Experiments and Discussion

We direct broad tests to assess the exhibition of our proposed technique. In this part, we introduced significant trial results and execution investigation. We direct broad tests to assess the exhibition of our proposed technique. In this part, we introduced significant trial results and execution investigation.

1) Exploration results: First, we tested our model by employing different exploration aspects. We conducted our experiments on YLI-MED and (CCV) datasets and use all videos associated with 10 event categories of YLI-MED dataset and 20 event categories of CCV. We explore the performance of our proposed method by using convolution features, semantic features and fusion of both features with and without using attention mechanism and obtained results are demonstrated in Table I and Table II.

TABLE I. MAP(%) ON YLI-MED DATASET ON DIFFERENT FEATURE INFORMATION

Feature Information	With attention	Without attention
Convolutional	76.2	74.5
Fully-Connected	78.9	77.1
Fusion (Both)	**82.2**	81.1

TABLE II. MAP(%) ON CCV DATASET ON DIFFERENT FEATURE INFORMATION

Feature Information	With attention		Without attention	
Convolutional	71.2			69.5
Fully-Connected	74.9			72.1
Fusion (Both)	**78.2**			75.1

It can be viewed from the obtained results, the introduced method (Fusion of appearance and semantic features) performs better than using convolution (appearance) and fully-connected (semantic) features separately. This illustrates that it is necessary to combine both appearance and semantic features in the temporal domain. This practice can discover more useful information for robust and accurate event detection. It can be also observed that introduction of attention mechanism yields improved performance especially in YLI-MED dataset and performs better than without attention model in both datasets. This attention mechanism provides insight for finding important parts of the video and prevents the background noise, thus, play important role to achieve better event recognition accuracy.

2) Effect of sampling strategies and encoders: We also carried out some experiments to analyze the effect of sampling strategies and encoders on our proposed model. Our model takes a series of frame sequence and we use two different sampling strategies: dense sampling and sparse sampling. For dense sampling, each video is divided into T clips with duration of 1-2 seconds and then we randomly select an image/frame from each clip and arrange them in a sequence. For sparse sampling, we select 3 frames of equal duration from video sequence and adopt setting guidelines given in [21]. We use YLI-MED and Columbia Consumer Video (CCV) dataset for these experiments and consider all events categories in both datasets. We explore the capacity of both sampling strategies using three encoders i.e. Max encoder, Average encoder, and LSTM encoder. The obtained results can be shown in Fig. 2 and Fig. 3. It can be seen from results that dense sampling achieves better performance than sparse sampling in the presence of max encoder. The possible reason is that sparse sampling may miss more crucial and important frames of the video sequence as compared to dense sampling and there may be loss of some important semantic features. We also analyze the performance of three different encoders in this experiment in the presence of both sampling strategies. According to the result, max encoder obtains the best performance against average and LSTM encoder. The underlying reason is that max encoder strengthens the features which are useful for specific event over a long-range. Both average and LSTM encoders perform similarly on both datasets. We can observe that performance of both encoders is relatively moderate. One reason is that both encoders are complex encoders as they have more parameters may lead to over fitting problem. We will adopt dense sampling with max encoder for our framework for all remaining experiments.

3) Class-Wise accuracy for event classification: We further investigate the event classification accuracy of our method by constructing the confusion matrix of two datasets i.e. YLI-MED and CCV datasets. The confusion matrixes of our introduced approach on both datasets can be depicted in Table III and Fig. 4. The confusion matrix indicating the accuracy of each action and correspondence between the target classes along x-axis (true label) and output classes (predicted label) along y-axis. We consider 10 event categories from YLI-MED dataset and 16 event categories from CCV dataset to conduct our experiment. Table III demonstrates the accuracy of each action category in the form of confusion matrix. The intensity of the true score is high (diagonal) for each category, and our method achieves 83% for all 10 event categories. It is interesting to note that some of categories with similar actions are more easily confused with each other, such as Birthday Party (Event-1), Wedding Ceremony (Event-9) and grooming an animal (Event-6), hand-feeding an animal (Event-7); these classifications meddle with one another and yield low scores. A potential purpose behind this is the comparability of the highlights and portrayals among activities. Be that as it may, our proposed approach actually performs well with the majority of function classes.

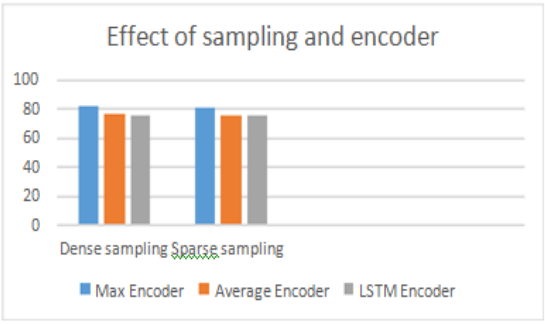

Fig. 2. Comparison of different Sampling Strategies and Encoders on YLI-MED Dataset.

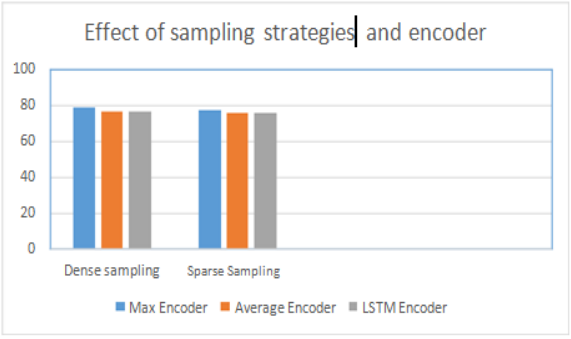

Fig. 3. Comparison of different Sampling Strategies and Encoders on CCV Dataset.

TABLE III. CONFUSION MATRIX ON THE YLI-MED DATASET USING OUR MODEL

Categories	Event-1	Event-2	Event-3	Event-4	Event-5	Event-6	Event-7	Event-8	Event-9
Event-1		0	0.01	0	0	0.01	0.02	0	0.04
Event-2	0		0	0.17	0	0	0.02	0	0.04
Event-3	0	0.17		0.02	0.05	0	0.02	0.04	0
Event-4	0	0.16	0.01		0	0	0	0	0
Event-5	0.01	0.01	0.02	0.01		0	0	0.02	0
Event-6	0.02	0	0	0	0		0.09	0	0
Event-7	0.03	0.02	0.01	0	0.01	0.16		0.04	0
Event-8	0	0	0.03	0	0.05	0	0.03		0
Event-9	0.11	0.02	0	0	0	0	0.03	0	
Event-10	0	0.01	0.02	0	0.02	0.02	0	0.03	0.05
Average									
Accuracy									

In addition, we also investigate the class-wise recognition accuracy of our method by constructing confusion matrices of CCV datasets. We consider 16 event classes from this dataset. The confusion matrices are given in Fig. 4. In this figure, the x-axis represents the classified labels of action classes whereas y-axis denotes the ground truth label. The accuracies in the diagonal cells are indicated by different colors and yellow cells show the 100% accuracy achieved for the particular action class. From the results, it can be seen that both of the confusion matrices are well diagonal zed. However, some of the action classes are giving low prediction scores by giving different colors of cells other than yellow it means few categories are mixed up when classifying. The possible reasons for interfering and misclassification are the motion similarity in actions or the same background, objects and appearance and motion-based features. However, most of the scores are well diagonal zed.

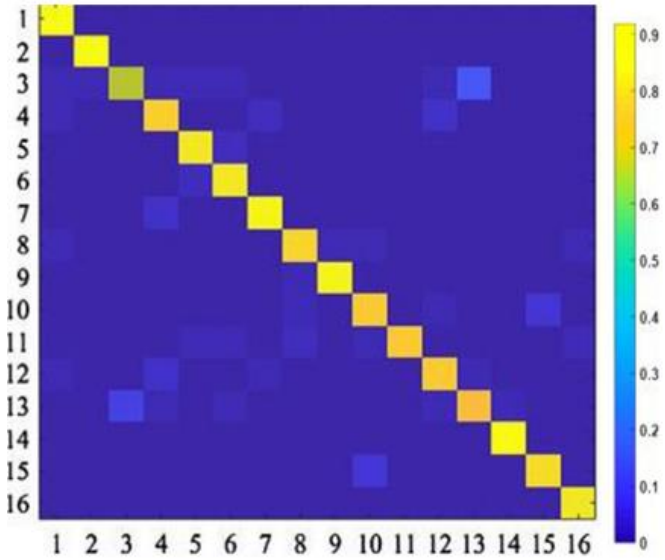

Fig. 4. Confusion Matrix on the CCV Dataset using our Model.

4) Visualization of feature embedding: Furthermore, we investigate the discriminative power of our learned fused features for human activity recognition. We consider the 10 different event categories (Birthday Party, Flash Mob, Vehicle unstuck, Parade, Board trick, grooming animal, Feeding animal, Landing a fish, Wedding Ceremony, Woodworking Project) from YLI-MED datasets. For each of the event category, we utilize 30 video clips of each event class for our experiment. Each video clip can be viewed by a single color point and we used the same color for al videos related same action class. For successful recognition of these action classes, an action recognition framework must possess high discriminative power. We adopt the method of t-SNE visualization [34] and show the visualization of feature representation embedding extracted by our introduced approach in Fig. 5. It can be observed from results that our method provides the better-separated clusters and clip-level features are semantically well separated as compared to the other existing prominent methods (two-stream model, and C3D). Thus, we can conclude that our proposed method can integrate both appearance and semantic features and possesses high discriminative information.

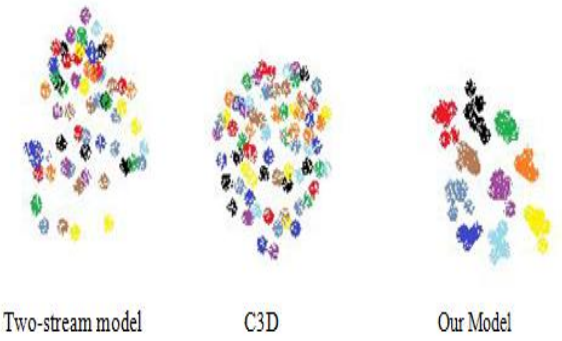

Fig. 5. Visualization of Clip-Level Features Embedding Learned by Two-Stream Model, C3D and Proposed Model.

5) Comparison with the state-of-the-art models: In this part, we further check the viability and plausibility of our model, we contrast our proposed approach with various existing cutting edge human activity acknowledgment draws near on both YLI-MED and Columbia Consumer Video (CCV) datasets for all videos of each event category. The comparison results are reported in Table IV in which average detection accuracies (%) are reported for both datasets. We consider two-stream models such as two-stream ConvNets [9], Two Stream 3D-Nets [26] and two-stream Fusion [27]. We also consider existing state-of-the-art hybrid model-based techniques such as TDD-IDT [28] and MTC3D-IDT [29] and P3D+IDT [30]. All of these models follow improved dense trajectories (iDT) for trajectories extraction and adopt higher-order encoding scheme i.e. Fisher Vector (FV) to encode the hand-crafted features.

TABLE IV. COMPARISON OF PROPOSED METHOD WITH THE STATE-OF-THE-ARTS APPROACHES ON YLI-MED AND CCV

Modality	Method	Input	YLI-MED (%)	ccv (CCV) (%)
Two-Streams	Two-stream ConNet [9]	RGB	69.9	58.9
	Two-stream ConNet [9]	O.F	53.4	49.2
	Two-Stream 3D-Nets	RGB	71.9	66.4
	[26]	O.F	64.3	61.7
	Two-Stream 3D-Nets	RGB	75.2	66.9
	[26]	O.F	67.3	61.5
	Two-stream Fusion [27]			
	Two-stream Fusion [27]			
Hybrid	TDD-IDT [28]	RGB	77.2	74.3
	MTC3D-IDT [29]	RGB	76.2	73.9
	P3D+IDT [30]	RGB	79.3	72.7
Very deep ConvNet	C3D [31]	RGB	65.6	63.2
	3D-ResNet [32]	RGB	72.6	69.0
	TSN [33]	RGB	74.5	70.0
Ours	SM-AB	RGB	**82.2**	78.2
	SM-AB	O.F	69.1	66.8

For the two-stream models, we analyze their performance on both stream i.e. RGB and optical flow and we can notice that performance of the optical flow is worst against the RGB images. This phenomenon verifies the assumption that the optical flow (O.F) in less flexible and inaccurate to capture motion of object due to the movement of camera and large-scale perspective transformation in complex videos. We also analyze the performance of some hybrid-features model in which features from both domains i.e. deep learning and hand-crafted features (improved dense trajectories) are incorporated and obtained competitive results, however, our approach outperforms them by fair margin on both datasets. We also compare our model with existing prominent and successful 3D convolution based methods such as C3D [31] and 3D-ResNet [32] and Temporal Segment Network model TSN [33]. Our approach possesses higher discriminative power and our system to be on par with the state-of-the-art. We compare the performance of our proposed model on both modalities i.e. RGB and Optical flow data and we achieve far better results when using only RGB frames so obtained results suggest that temporal long term dynamics can be capture from RGB frames. Thus, from results we can say that our model in the presence of only RGB data explores more relationships between video clips and semantic features and introduction of max encoder works well by capturing the long-term dependencies and successful for the detection of complex events.

V. CONCLUSION

This paper proposes a new lightweight framework for video event detection, which comprises CNN, features fusion, attention mechanism, and global average pooling. This framework obtains high representational power and finds the discriminative patterns in complex videos for event detection. We just use the RGB data to extract appearance and semantic features for each frame of video using convolution and fully-connected layers. This practice avoids the additional computational power required by optical flow. We explore the motion by computing the difference between both semantic and appearance features. We also employ the attention mechanism to concentrate and focus on key frames keeping motion information and avoiding the redundant effect of static background. Furthermore, we utilize temporal encoder to establish temporal relationships between frames and explore discriminative long-term motion patterns. The introduced model achieved promising performance when tested on two widely used challenging datasets. In future work, we will try to improve the sampling strategy or may modify the pooling or fusion layers in the network.

ACKNOWLEDGMENT

The authors would like to thank the editor and reviewers for their work on this manuscript.

REFERENCES

[1] F. Jiang, J. Yuan, S.A, Tsaftaris and A.K Katsaggelos. “Anomalous video event detection using spatiotemporal context”. Comput. Vis. Image Underst. pp. 323–333, vol. 115, 201.

[2] B.T Morris and M.M Trivedi.”Trajectory learning for activity understanding: Unsupervised, multilevel, and long-term adaptive approach”. IEEE Trans. Pattern Anal. Mach. Intell. pp. 2287–2301, vol. 33, 2011.

[3] S. Calderara, U. Heinemann, A. Prati, R. Cucchiara and N. Tishby.”Detecting anomalies in people’strajectories using spectral graph analysis”. Comput. Vis. Image Underst. pp. 1099–1111. Vol. 115, 2011.

[4] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in proceeding of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), pp. 886–893, June 20-25, 2005.

[5] H. Wang, A. Klaser A and C. Schimid, “Dense trajectories and motion boundary descriptor for action recognition,” in proceeding international journal of computer vision, vol. 103, pp. 60-79, March, 2013.

[6] Wang H, Ullah MM, Kl¨aser A, Laptev I, Schmid C (2009) Evaluation of local spatio-temporal features for action recognition. In: British Machine Vision Conference.

[7] G.W Taylor, R. Fergus and Y. LeCun, "Convolutional learning of spatio-temporal features," in proceeding of 11th European conference on Computer vision, pp. 140-153, September 5-11, 2010. Article (CrossRef Link).

[8] Ji Si, Xu W, Yang M, et al., "3d convolutional neural networks for human action recognition," IEEE Trans. Pattern Anal. Mach. Intell., vol.35, no.1, pp.221–231, January, 2013.

[9] D. Tran, L. Bourdev and Fergus, "Learning spatiotemporal features with 3d convolutional networks," In proceeding of IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, pp. 4489–4497, December 7-13, 2015. Article (CrossRef Link).

[10] Limin Wang, Yuanjun, Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaou Tang. Temporal segment network: Towards good practices for deep action recognition," in ECCV, 2016.

[11] Julia bernd, Damian Borth, Benjamin Elizade, et al. YL1-Med corpus: characteristics, procedures, and plans, in arXiV: 1503.04250, 2015.

[12] A. Diba, A. M. Pazandeh, and L. V. Gool, "Efficient two-stream motion and ap-pearance 3D CNN for video classification," in Proceedings of European Conferenceon Computer Vision, 2016, pp. 1- 4.

[13] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition", in Poc. IEEE Conf. Comp. Vis. Pattern Recognit., Jun. 2016, pp-1933-1941.

[14] L. Wang, Y. Qiao, and X. Tang, "Action recognition with trajectory-pooled deep-convolutional descriptors," in Proceeding of IEEE international conference on Computer Vision and Pattern Recognition, 2015, pp. 4305-4314.

[15] Y.-G. Jiang, S. Bhattacharya, S.-F. Chang and M. Shah. High-Level Event Recognition in Unconstrained Videos. In IJMIR, 2012.

[16] X. Lu, H. Yao, and S. Zhao, "Action recognition with multi-scale trajectory-pooled 3D convolutional descriptors," Multimedia Tools and Applications, 2017 pp.1-17.

[17] Z. Qiu, t. Yao, and T. Mei, "Learning spatio-temporal representation with pseudo-3D residual networks, "in proc. IEEE Intl. Conf. Comut. Vis., Oct. 2017, pp. 5533-5541.

[18] A. Karpathy, G. Toderici, S. Shetty and T. Leung, "Large-scale video classification with convolutional neural networks," in proceeding IEEE conference on computer vision and pattern recognition, pp. 1725 – 1732, June 23-28, 2014.

[19] L. Wang, Y. Xiong, Z. Wang, and Y. Qiao. Towards good practices for very deep two-stream convnets. arXiv preprint arXiv:1507.02159, 2015

[20] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In Proc. ECCV, 2014.

[21] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler. Efficient object localization using convolutional networks. In Proc. CVPR, 2015.

[22] S. Venugopalan, M. Rohrbach, R. Mooney, T. Darrell, and K. Saenko. Sequence to sequence video to text. In Proc. ICCV, 2015.

[23] N. Srivastava, and R. Salakhutdinov. Multimodal Learning with Deep Boltzmann Machines. In NIPS'12, 25, pages 2231-2239.

[24] L. Sun, K. Jia, and D. Yeung, "Human action recognition using factorized spatio-temporal convolutional networks," in proceeding of IEEE International Conference on computer vision (ICCV), pp. 4597 – 4605, December 7-13, 2015.

[25] A. Klaser, M. Marszalek, and C. Schmid, "A Spatio-Temporal Descriptor Based on 3D-Gradients," in proceeding of 19th British Machine Vision Conference, British Machine Vision Association: Leeds, United Kingdom, pp.1–10, September, 2008. Article (CrossRef Link).

[26] P. Scovanner, S. Ali and M. Shah, "A 3-Dimensional SIFT Descriptor and its Application to Action recognition," in Proceedings of the 15th International Conference on Multimedia, pp. 357–360,September 25 – 29, 2007.

[27] H. Wang and C. Schmid, "Action recognition with improved trajectories," in proceeding of IEEE International conference on computer vision, pp. 3551-3558, December 1-8, 2013.

[28] Joao Carreira and Andrew Zisserman, action recognition? A new model and the kinetics dataset, in CVPR, 2017.

[29] Kaming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual for image recognition. arXiv preprint arXiv: 1512/03385, 2015.

[30] Kaming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mapping in deep.

[31] Karen Simonyan and Andrew Zisserman. Very deep convolutional networksfor large-scale image recognition. In proc. ICLR, 2014.

[32] Limin Wang, Yuanjun, Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaou Tang.Temporal segment network: Towards good practices for deep action recognition," in ECCV, 2016.

[33] N. Srivastava, and K. Salakhutdinov. Multimodal Learning with Deep Boltzmann Machines. In NIPS'12, 25, pages 2231-2239.