STIP-GCN: Space-time interest points graph convolutional network for action recognition

Sravani Yenduri, Vishnu Chalavadi, and C Krishna Mohan
Indian Institute of Technology Hyderabad, Kandi, India
{cs18resch02001, cs16m18p000001, ckm}@iith.ac.in

2022 International Joint Conference on Neural Networks (IJCNN) | 978-1-7281-8671-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/IJCNN55064.2022.9891959

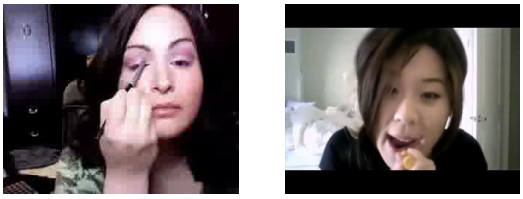
***Abstract*—Action recognition requires modelling the interactions between either human & human or human & objects. Recently, graph convolutional neural networks (GCNs) are exploited to effectively capture the structure of action by modelling the relationship among entities present in a video. However, most of the approaches depend on the effectiveness of object detection frameworks to detect the entities. In this paper, we propose a graph-based framework for action recognition to model the spatio-temporal interactions among the entities in a video without any object-level supervision. First, we obtain the salient space-time interest points (STIP) that contain rich information about the significant local variations in space and time by using the Harris 3D detector. In order to incorporate the local appearance and motion information of the entities, either low-level or deep features are extracted around these STIPs. Next, we build a graph by considering the extracted STIPs as nodes and are connected by spatial edges and temporal edges. These edges are determined based on a membership function that measures the similarity of entities associated with the STIPs. Finally, GCN is employed on the given graph to provide reasoning among different entities present in a video. We evaluate our method on three widely used datasets, namely, UCF-101, HMDB-51, SSV2 to demonstrate the efficacy of the proposed approach.**

***Index Terms*—Action recognition, STIP, GCN, graph representations**

I. Introduction

Action recognition, one of the video understanding tasks is employed in various fields such as smart surveillance, human-computer interaction, autonomous vehicles etc., to learn the spatio-temporal interactions between the human and objects occurring in a given video. However, modeling of these interactions is challenging because of two key factors. Firstly, the typical issues like occlusion, illumination conditions, viewpoint variations etc., cause difficulty in perceiving/identifying the entities present in a video. Secondly, uncertainty among the actions due to the similarity in either spatial arrangement or temporal information. For example, two actions considered from UCF-101 dataset such as, 'apply lipstick' and 'apply makeup' have a similar spatial arrangement i.e., applying makeup/lipstick on the facial regions as shown in Fig. 1. Likewise the similar motion cues between the 'table tennis shot' and 'tennis swing' actions create ambiguity while recognising these actions (see Fig. 2).

In order to overcome these issues, several methods in literature have been explored varying from the traditional hand-crafted to deep learning approaches. The traditional methods [1] extract low-level features such as HOG, HOF, and MBH to obtain the local appearance and motion information. A unique representation of an action is acquired by encoding these features using various aggregation frameworks [2], [3] for action classification. Whereas, the deep learning methods like two-stream networks [4], 3D-CNNs [5], and temporal stream networks (TSNs) [6] exploit the concept of end-to-end learning to capture the spatial and temporal information. Although these frameworks achieve better performance for the actions exhibiting high inter-class variability, (eg. 'Archery', 'basket ball') they fail to distinguish between the actions with low inter-class variability and high intra-class variability (eg. 'playing sitar' and 'playing cello'). This is because, the deep learning methods extract the global features of a video and fail to model the relationship among the entities present in a scene.



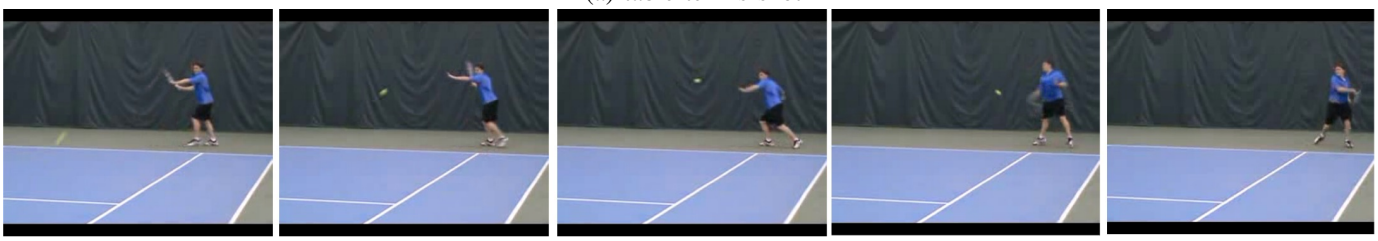
(a) Apply Eyemakeup (b) Apply lipstick

Fig. 1: Examples of actions from UCF-101 exhibiting similar spatial arrangement

Recently, graph convolutional networks (GCN) are exploited to effectively capture the structure of action by modeling the relationship among entities present in a video. The interaction of the entities in a video is formulated as a graph and fed to GCN to classify the actions. The existing works [7], [8] consider the selection or extraction of entities as the crucial part in action recognition. In [9], Yan et al, consider the skeletal joints locations as entities and builds a graph to model the dynamics of the human body skeletons in order to recognize the human actions. However, the contribution of different joints is distinct for various actions. Hence, Ahmad et al. [10] proposed an attention mechanism to attend only to those joints that contribute to the corresponding actions. Whereas, the region-based approaches [7], [8], [11] employ the off-the-shelf object detection framework to detect the entities present in a scene. Wang et al. [7] represent a video as space-time graphs to model the temporal dynamics and contextual relationship between the human and objects. The nodes of a space-time graph are the region proposals extracted from the region proposal network (RPN) and these nodes are connected

978-1-7281-8671-9/22/$31.00 ©2022 IEEE

(a) table tennis shot



(b) tennis swing

Fig. 2: Illustration of actions from UCF-101 having similar motion cues

based on (i) similarity of appearance and (ii) spatio-temporal relations. Along the lines of [7], Mavroudi et al. [12] proposed visual-symbol graphs to capture the visual and semantic cues between the human and objects by constructing two graphs. A visual graph to model the spatio-temporal interactions between the actor & objects and a symbolic graph to capture the semantic relationships between them.

However, the existing methods discussed above have the following limitations: (i) Skeletal-based approaches are sensitive to view-point variations and fail to model the actions involving interactions between the human and objects (eg. playing violin, playing guitar), (ii) Region-based methods depend on the effectiveness of the object detection framework involved during the extraction of salient regions of entities present in a scene.

In this paper, we propose a graph-based framework for action recognition to model the spatio-temporal interactions among the entities in a video without using any object-level supervision. We first obtain the salient space-time interest points (STIP) that contain the rich information about the significant local variations in space and time by using Harris 3D detector [13]. In order to incorporate the local appearance and motion information of the entities, either low-level or deep features are extracted around these STIPs. Next, we build a graph by considering the extracted STIPs as nodes and are connected by spatial edges and temporal edges. These edges are determined using a membership function that measures the similarity of entities associated with the STIPs. Finally, GCN [14] is applied on the given graph to provide reasoning among different entities present in a video. We evaluate our method on three widely used datasets, namely, UCF-101, HMDB-51, something something v2 (SSV2) to demonstrate the efficacy of the proposed approach.

The main contributions of this paper are:

- We construct a graph whose nodes are space-time interest points (STIPs) obtained from the Harris 3D detector and are connected based on the appearance and motion dynamics of these interest points.
- We investigate the effectiveness of graph convolutional neural networks (GCNs) on the constructed graph to model the interactions among the entities present in a video.
- The efficacy of the proposed approach is demonstrated on three common datasets, namely, UCF-101, HMDB-51, and SSV2. These datasets contain both human-human and human-object interactions with similar spatial and temporal properties.

II. Related Work

Typically, most of the existing works in literature employ either visual models or graphical models to learn the prominent spatio-temporal representations essential for action recognition. In this section, we explore various approaches in visual and graphical models in detail.

A. Visual models for action recognition

Traditional visual models focus on manually designing the features that contribute to the discriminative representation of actions. These hand-engineered features such as improved dense trajectories (IDT) [1], space-time interest points (STIP) [15], etc., contain the descriptors that are rich in appearance and motion information. The derived descriptors are encoded using various aggregation frameworks [2], [3] to obtain a video-based representation for action classification. On the other hand, the deep learning approaches focus on learning the deep features from a video in an end-to-end fashion eliminating the need for hand-engineering. Multi-stream networks [16], one of the popular deep learning methods train the spatial and temporal streams independently to capture the appearance and motion representations. The multi-stream network learns

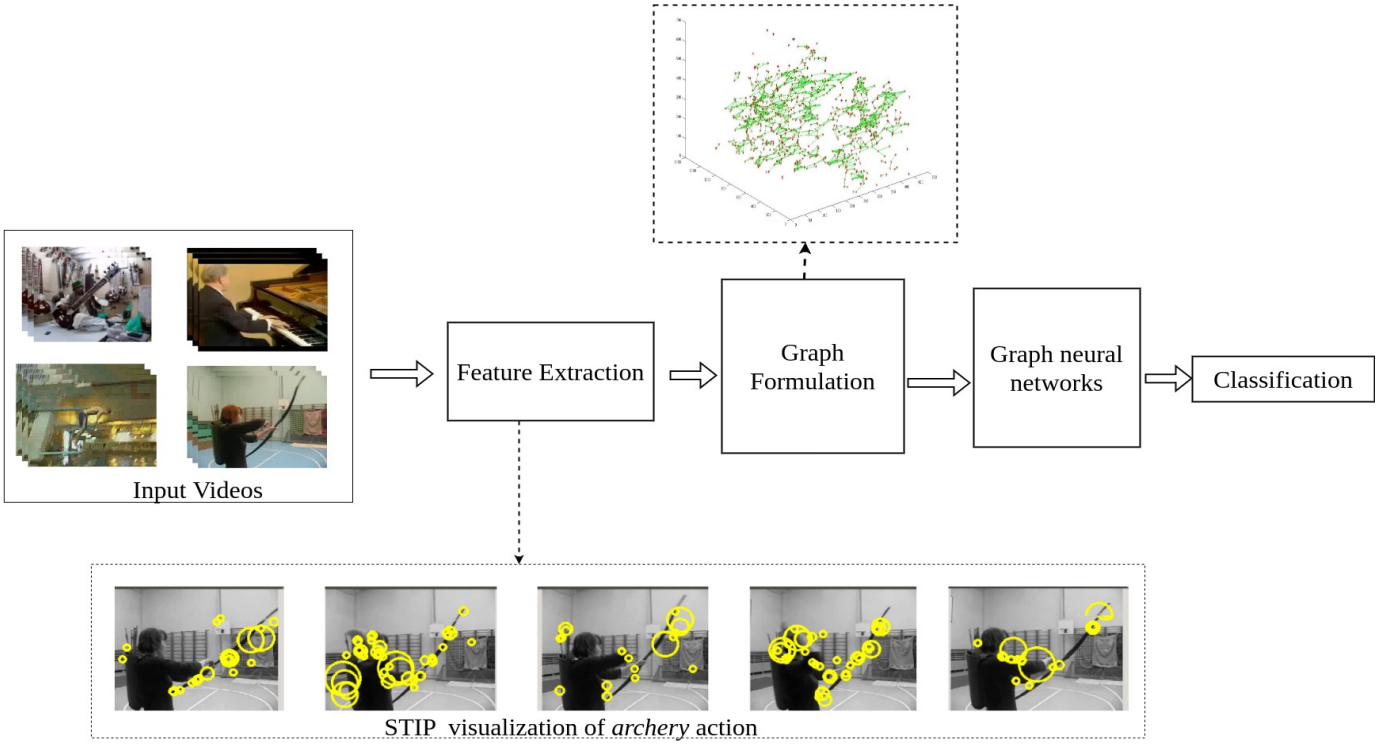


Fig. 3: Block diagram of the proposed approach (best viewed in color).

the appearance information from static video frames and temporal information from the local motion vectors derived from the optical flow frames. However, the multi-stream networks train the temporal and spatial streams independently leading to the lack of interaction among the streams and fail to model the long-term temporal dynamics of an action which is essential for efficient recognition of actions. To overcome this issue, recurrent neural networks (RNN) [17], 3D-CNN [18], [19] have been proposed. As these models are difficult to train and computationally expensive, Yang et al [20] proposed an asymmetric 3D convolutional network to reduce the number of parameters and computational complexity. However, the obtained representation from these frameworks is the global representation of a whole scene and fails to capture the spatio-temporal interaction between the entities present in a scene.

B. Graphical models for action recognition

Graph networks have been employed in many domains [21], [22] where the reasoning between different entities is desired. Since a definite structure is not present in video for vision tasks, the selection of atomic elements for representing graph nodes is important. In order to mimic the definite structure, several approaches [23], [24] have considered the human skeleton as an explicit structure to model the dynamics of the human body during the course of action. Yan et al [9] were first to propose a spatio-temporal graphical approach for skeleton-based action recognition. The spatio-temporal graph consists of skeletal joints as nodes and edges are connected based on the natural connectivity of joints. Huang et al [25] incorporated the multi-granularity information by proposing the split-transform-merge concept in graph convolutional networks (GCN). This model aggregates the multi-scale information from spatial and temporal paths to improve the performance of GCNs in action recognition. Later Zhang et al. [26] integrated the contextual information into graphs by providing information of all other nodes of a human skeleton for increasing the receptive field of the local graph convolution operation. This method eliminates the need for stacking multiple layers to incorporate long range dependencies among the nodes. Likewise, to enhance the flexibility of receptive fields of graphs, shift-GCNs [27] incorporating point-wise convolutions and shift operations were introduced. These operations are lightweight and can reduce the computational complexity of GCN-based methods. However, the skeletal-based approaches are sensitive to view-point variations and fail to model the actions involving interactions between the human and objects.

In order to overcome the above limitation, space-time video graphs [7] are introduced to model the interactions between human and objects. These graphs are constructed with the region proposals generated from off-the-shelf object detection frameworks as nodes and are connected based on appearance and motion relationships. Following this, Ji et al. [29] proposed a method to learn the complex interactions between several objects without increasing the computational complexity. These interactions are captured by learning the higher-order relationships among the objects. Similarly, Mavroudi et al. [12] introduced a hybrid graph neural network to capture

the semantic and contextual information for understanding the interactions among the entities in a video. The hybrid graph is built based on the spatio-temporal interactions & semantic representation learnt by a visual graph and symbolic graph, respectively. However, these approaches depend on the effectiveness of the object detection framework involved during the extraction of salient regions of entities present in a scene. To overcome this, Duta et al. [28] presented a method that learns the salient regions responsible for an action dynamically eliminating the need for object detectors. These regions are considered as graph nodes and graph neural networks are applied to model the reasoning for interactions among the objects.

III. Proposed Methodology

To overcome the above limitations, we propose a graph-based framework to model the interactions between human and objects for efficient recognition of actions without any object-level supervision as shown in Fig. 3. The proposed methodology consists of three modules, namely, feature extraction, graph formulation, and graph convolution networks (GCNs) for classification.

A. Feature extraction

We extract spatio-temporal features from each input video containing uniformly sampled T frames to represent an action. These features are extracted at specific locations (also known as interest points) where variations across space and time are significant. The procedure for detecting these interest points and extraction of features are described in the following sub-sections.

1) Space-time interest points (STIP) descriptors: The linear scale representation $F(x, y, \sigma_v^2, \tau_v^2)$ for an input video $V(x, y, t)$ is given by convolving a Gaussian kernel $G(x, y, \sigma_v^2, \tau_v^2)$ with V as

$$F(x, y, \sigma_v^2, \tau_v^2) = V(x, y, t) * G(x, y, \sigma_v^2, \tau_v^2), \quad (1)$$

where, σ_v^2 and τ_v^2 are the variations across space and time and '*' denotes a convolution operator. The aim of Harris 3D corner detector is to find the locations where V has significant variations in three directions. Such locations are determined by convolving the second-moment matrix S with a Gaussian function $g(x, y, \sigma_k^2, \tau_k^2)$ of spatial and temporal variance, σ_k^2 & τ_k^2, respectively.

$$S = g(x, y, \sigma_k^2, \tau_k^2) * \begin{bmatrix} F_x^2 & F_xF_y & F_xF_t \\ F_xF_y & F_y^2 & F_yF_t \\ F_xF_t & F_yF_t & F_t^2 \end{bmatrix}. \quad (2)$$

where the second order derivatives are given by

$$F_x^2 = \frac{\partial^2 F}{\partial x^2} \qquad F_y^2 = \frac{\partial^2 F}{\partial y^2} \qquad F_t^2 = \frac{\partial^2 F}{\partial t^2}$$
$$F_xF_y = \frac{\partial}{\partial x}\left(\frac{\partial F}{\partial y}\right) \quad F_yF_t = \frac{\partial}{\partial y}\left(\frac{\partial F}{\partial t}\right) \quad F_xF_t = \frac{\partial}{\partial x}\left(\frac{\partial F}{\partial t}\right)$$

The largest eigen values λ_1, λ_2, and λ_3 of S signify the interest points. These are detected by Harris corner function using

$$H = \lambda_1\lambda_2\lambda_3 - m(\lambda_1 + \lambda_2 + \lambda_3)^3. \quad (3)$$

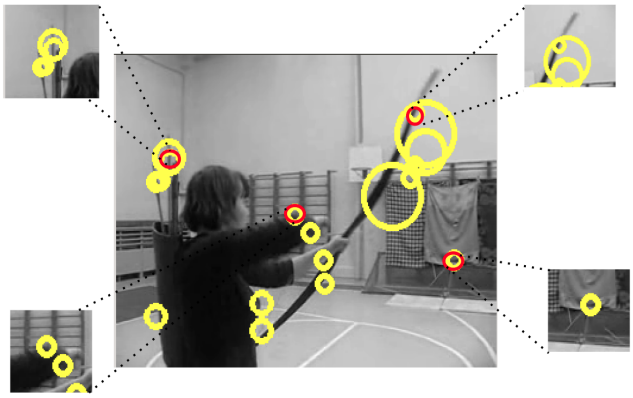


Fig. 4: The $K \times K$ image cropped around the interest points.

Finally, local features such as Histogram of oriented gradients (HoG), Histogram of optical flow (HoF) collectively known as space-time interest points (STIP) descriptors (of 162 dimension) are extracted around these detected interest points to obtain the appearance and motion information, respectively.

2) Deep features: In order to obtain efficient spatio-temporal features, we propose to extract deep features from big transfer (BiT) model [31]. The BiT model is a ResNet architecture with a group normalization layer [30] pre-trained on ImageNet-21K dataset. The input to this model is an $K \times K$ region that is cropped around the interest points as shown in Fig. 4 to extract the local appearance and motion information. Finally, the penultimate layer features (of 2048 dimension) of BiT model are considered to generate a graph which is explained in the next sub-section.

B. Graph formulation

In this sub-section, we construct a graph $G = (N, E)$ from the obtained interest points $R = \{r_i \in (x, y, t)\}_{i=1}^n$ with feature vectors $H = \{h_i\}_{i=1}^n$ to model the relationship among the entities present in a video. Here, N represents the set of STIPs and E gives the set of edges defined by a membership function e_{ij} computed as

$$e_{ij} = \frac{M(h_i, h_j)}{||r_i - r_j||^2}. \quad (4)$$

Here, $M(h_i, h_j)$ is a Gaussian function that measures the similarity between two interest points $r_i \& r_j$. The very high value of e_{ij} represents that either the interest points are extremely close to each other or the features captured by STIPs are the same during feature extraction. Likewise, the value of e_{ij} is very low when the interest points are far from each other. Either of the above cases do not provide any significant information regarding the interaction among entities. Hence, we threshold the edge values by

$$E_{ij} = \begin{cases} 1, & \text{if } e_{ij} > e_{th} \\ 0, & \text{otherwise.} \end{cases} \quad (5)$$

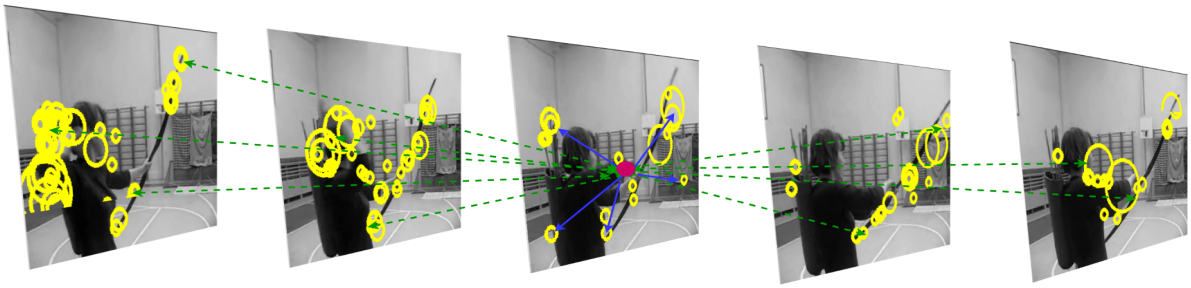


Fig. 5: Spatio-temporal graph depicting possible edges defined by membership function for a vertex (best viewed in color). The *vertex* (in red) is an interest point detected by Harris 3D corner detector. This is connected to other possible nodes through *spatial edges* (blue solid lines) and *temporal edges* (green dotted lines).

Here, e_{th} is usually considered to be mean. And, the nodes within the same frame and across the frames are connected using the above function. Hence, the adjacency matrix E consists of both spatial and temporal edges as shown in Fig. 5.

C. Graph Convolutional networks (GCN)

Given the generated graph from previous step, we employ GCN to classify different actions as shown in Fig. 6. Our GCN model contains series of graph convolutions and pooling layers followed by a readout layer at the end to classify actions present in a video. The graph convolution for l^{th} layer is

$$Q^l = EX^{l-1}W^l, \tag{6}$$

where W^l is a learnable weight matrix, X^l are the features from hidden layer l, and $X^0 = H$ i.e., input spatio-temporal features. The graph convolutions approximated by spectral propogation rule is given as

$$Q^l = \sigma(\tilde{D}^{-1/2}\tilde{E}\tilde{D}^{-1/2}X^{l-1}W^l). \tag{7}$$

Here, $\tilde{D}^{-1}$ is a degree matrix for $\tilde{A} = A + I$, and σ is the ReLu activation function. The convolutional layer is followed by a pooling layer to coarsen the graph. Later, a readout is applied to aggregate the node representations into a graph representation. Any readout operations like sum, mean, max, min are used. However, we consider mean as our readout operation in order to eliminate any outliers. Finally, we predict y by using the multi-layer perceptron and softmax layer on the obtained graph representation.

IV. Experimental Results

The experiments are executed on 4 Tesla M60 GPUs. The spatial and temporal variances of a Gaussian kernel are considered to be $\sigma_v^2 = 8, \tau_v^2 = 8$ during detection of interest points. Also, m is set to 0.005. We train a 2 layer GCN for 200 epochs with initial learning rate of 0.01 and dropout of 0.5. The size of image cropped to extract deep features is set to 32×32. The increase of cropping size has led to decrease in performance due to addition of noise either by background or irrelevant objects in a video. The comparison with existing state-of-the-art methods and evaluation of the proposed approach on different datasets are discussed in the following sub-sections.

A. Datasets

1) UCF-101: It is a well-known standard action recognition dataset consisting of 101 actions [32]. The 13320 videos are collected from Youtube to recognise actions in a natural real-world environment. It contains actions of human & object interaction, human & human interaction, common body movements, etc. This dataset poses several challenges such as view-point variation, background clutter, occlusion, presence of diverse objects, and large fluctuations in camera movement.

2) HMDB-51: It is another widely used dataset used for recognising actions from realistic movies and YouTube videos [33]. The HMDB-51 dataset contains 6766 videos of 51 action categories. Each video has an individual performing one of 51 actions. It is divided into human & object interaction, body movements, facial actions categories. The dataset contains actions with complex motion cues and is split into 70% training and 30% testing for evaluation.

3) Something something v2 (SSV2): It is a large dataset consisting of 174 fine-grained actions of human object interactions [34]. The dataset has $\sim$220K videos of real-life daily activities with the challenges like presence of diverse objects, view-point variations, occlusion, etc.

B. Analysis of the proposed approach

Table I presents the effect of constructing graphs using STIP features. It is shown that simple multi-layer perceptron trained on STIP descriptors such as HoG and HoF gives 65.4% on UCF-101 and 56.7% on HMDB-51 datasets. Whereas, average pooling of extracted deep features increased the performance by $\sim$5% due to learning of complex information about actions inherently. However, such simple aggregation of features

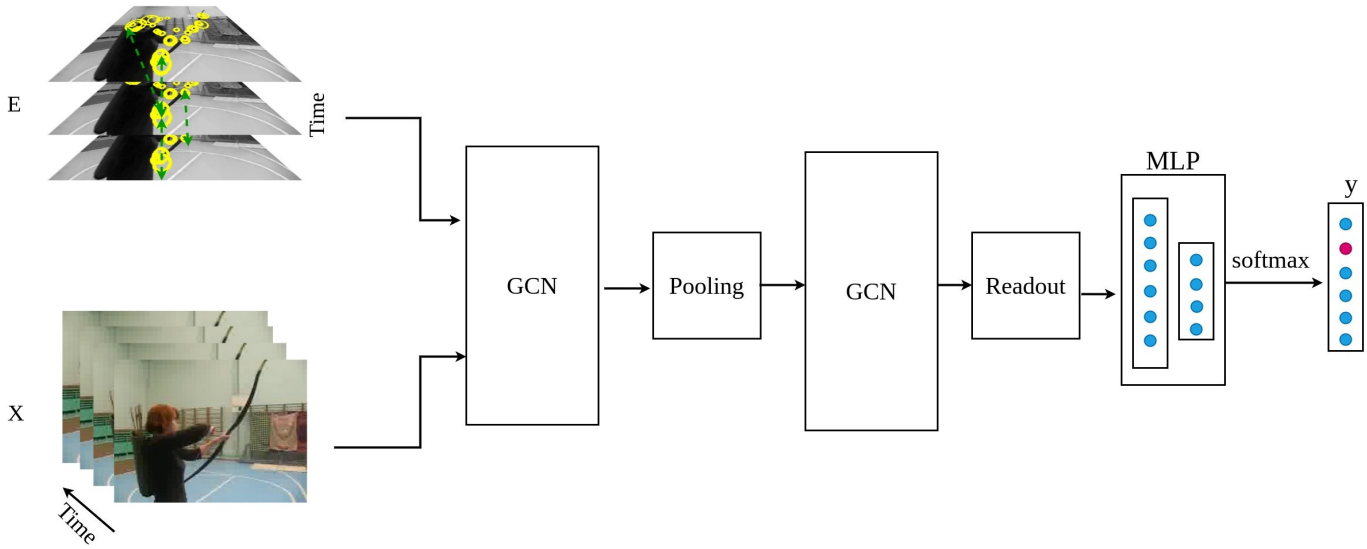
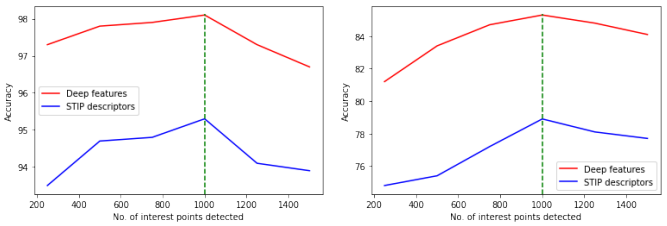


Fig. 6: A graph classification network consisting of a series of graph convolution and pooling layers with readout layer at the end.

TABLE I: Analysis of the proposed approach

Methods	UCF-101	HMDB-51	SSV2
STIP descriptors	65.4	56.7	50.1
Deep features + Avg pooling	77.4	60.5	61.4
STIP descriptors + 1-layer GCN	93.7	72.6	57.9
Deep features + 1-layer GCN	95.4	81.7	67.2
STIP descriptors + 2-layer GCN	**95.3**	**78.9**	**59.7**
Deep features + 2-layer GCN	**98.1**	**85.3**	**68.2**
STIP descriptors + 3-layer GCN	94.9	77.6	58.4
Deep features + 3-layer GCN	97.7	84.6	67.9

cannot model the interactions among the entities present in a video. Hence, to incorporate the reasoning between various entities in a video, GCNs are employed to classify actions which are depicted in the form of graph. It can be seen from the Table I that the incorporation of GCN improves the performance on all the datasets. Also, as the number of GCN layers increases, the performance also improves. However, we can observe that there is no significant improvement with more than 2 layers. Fig. 7 shows the optimum number of interest



(a) UCF-101 (b) HMDB-51

Fig. 7: Plot of accuracy vs number of interest points detected

points to be detected from each video in order to obtain better accuracy. It can be seen that the performance of our method increases with the number of interest points detected (up to 1000). That is, a 1000 node graph best represents the action occurring in a video. Rather than selecting the nodes randomly, we choose the nodes based on score given by a SVM classifier trained using the STIP descriptors. The illustration of graphs generated for a few actions from UCF-101, HMDB-51, and SSV2 datasets are presented in Fig. 8. The graphs effectively depict the significant interaction among the detected while eliminating the irrelevant background noise. That is, 'hand to eye' interaction in *ApplyEyeMakeup* action, the 'spinning motion' of an individual during *diving* action, and 'movement of an object' in *moving object from left to right* action.

TABLE II: Comparison of the proposed approach with the state-of-the-arts on UCF-101, HMDB-51, and SSV2 datasets

Methods	UCF-101	HMDB-51	SSV2
C3D + IDT [35]	90.4	-	-
Two-stream fusion [36]	92.5	65.4	-
R(2+1)D - RGB [37]	96.8	74.5	-
I3D - RGB [38]	98	80.7	-
P3D [41]	88.6	-	-
TSN [6]	94.2	69.4	-
TSM [39]	94.5	70.7	63.4
STG [7]	-	-	46.1
Slowfast [46]	-	-	61.7
STM [40]	96.2	72.2	64.2
R(2+1)D + BERT [47]	-	84.77	-
STC [45]	93.7	66.8	-
MViT [44]	-	-	67.1
ViViT [43]	-	-	65.4
Proposed approach (STIP features)	95.3	78.9	59.7
Proposed approach (Deep features)	**98.1**	**85.3**	**68.2**

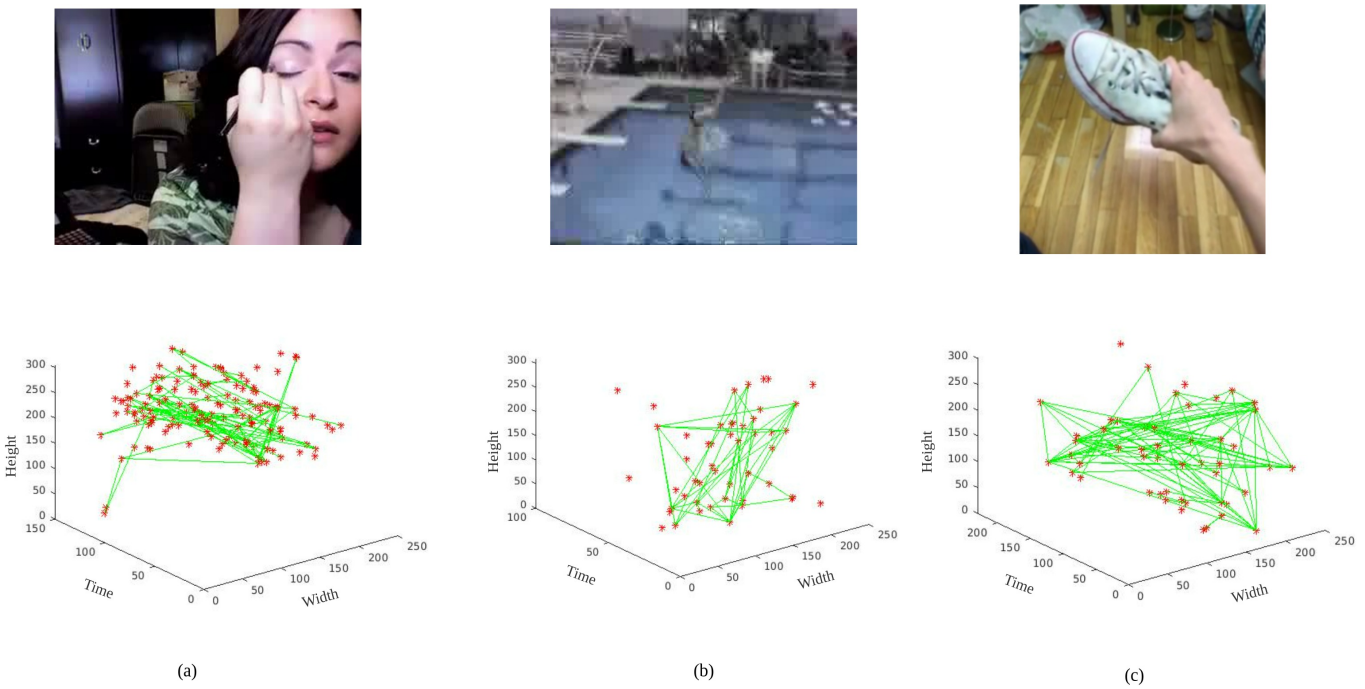


Fig. 8: 3D Illustration of generated graphs for (a) *ApplyEyeMakeup* from UCF-101 dataset, (b) *Diving* from HMDB-51 dataset, and (c) *Moving object from left to right* from SSV2 dataset.

C. Comparison with existing approaches

Table II gives comparison of the proposed method with existing state-of-the-art approaches on UCF-101, HMDB-51, and SSV2 datasets. Conventional two-stream network [36] explored various ways to fuse the appearance and motion cues from independent streams without loss of information and achieved 92.5% on UCF-101 and 65.4% on HMDB-51 dataset. In order to capture the spatial and temporal information simultaneously, several 3D-CNN architectures [37], [41], [46] pre-trained on large datasets have been investigated. Among these, I3D network pre-trained on Kinetics-400 has shown the prominent performance in action recognition. Later, some methods [39], [41] are introduced to reduce the computational complexity of 3D-CNN architectures without compromising its performance. Inspired by the success of transformers in NLP tasks, some of the methods [43], [44], [47] incorporated variants of transformers to attend to spatial and temporal information. However, the above mentioned approaches are limited in differentiating actions involving human object interactions. It can be seen from the Table II that our proposed method performs on par with the existing approaches. It achieves 98.1% on UCF-101, 85.3% on HMDB-51, and 68.2% on SSV2 datasets. This is due to modeling of relationships among the entities by the GCNs from the graphs constructed based on the STIP feature descriptors.

V. Conclusion

In this paper, we propose a graph-based approach to model the interactions between human and objects occurring in an action. The graphs are constructed to provide the local relationships and long-range temporal dependencies among the entities present in a video. These entities are represented by space-time interest points (STIP) where there is significant local variations in appearance and motion. These entities are connected based on appearance, motion, and position using a novel membership function. The graphs are classified using graph convolutional networks (GCNs) to effectively aggregate the node information for recognition of actions. We demonstrate the effectiveness of the proposed approach by evaluating on UCF-101, HMDB-51, and something-something v2 datasets. The experiments show that our approach performs on par with existing state-of-the-arts due to efficient modelling of reasoning between human and objects by the generated graphs.

References

[1] Wang, Heng, and Cordelia Schmid. "Action recognition with improved trajectories." In Proceedings of the IEEE International Conference on Computer Vision, pp. 3551-3558. 2013.

[2] Li, Yingwei, Weixin Li, Vijay Mahadevan, and Nuno Vasconcelos. "Vlad3: Encoding dynamics of deep features for action recognition." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1951-1960. 2016.

[3] Wang, Sen, Zhigang Ma, Yi Yang, Xue Li, Chaoyi Pang, and Alexander G. Hauptmann. "Semi-supervised multiple feature analysis for action recognition." IEEE Transactions on Multimedia 16, no. 2 (2013): 289-298.

[4] Simonyan, Karen, and Andrew Zisserman. "Two-stream convolutional networks for action recognition in videos." arXiv preprint arXiv:1406.2199 (2014).

[5] Hara, Kensho, Hirokatsu Kataoka, and Yutaka Satoh. "Learning spatio-temporal features with 3d residual networks for action recognition." In Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 3154-3160. 2017.

[6] Wang, Limin, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. "Temporal segment networks for action recognition in videos." IEEE Transactions on Pattern Analysis and Machine Intelligence 41, no. 11 (2018): 2740-2755.

[7] Wang, Xiaolong, and Abhinav Gupta. "Videos as space-time region graphs." In Proceedings of the European Conference on Computer Vision (ECCV), pp. 399-417. 2018.

[8] Yang, Jianwei, Jiasen Lu, Stefan Lee, Dhruv Batra, and Devi Parikh. "Graph r-cnn for scene graph generation." In Proceedings of the European Conference on Computer Vision (ECCV), pp. 670-685. 2018.

[9] Sijie Yan, Yuanjun Xiong, and Dahua Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in AAAI Conference on Artificial Intelligence, 2018, pp. 7444–7452.

[10] Ahmad, Tasweer, Huiyun Mao, Luojun Lin, and Guozhi Tang. "Action Recognition Using Attention-Joints Graph Convolutional Neural Networks." IEEE Access 8 (2019): 305-313.

[11] Zeng, Runhao, Wenbing Huang, Mingkui Tan, Yu Rong, Peilin Zhao, Junzhou Huang, and Chuang Gan. "Graph convolutional networks for temporal action localization." In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7094-7103. 2019.

[12] Mavroudi, Effrosyni, Benjamín Béjar Haro, and René Vidal. "Representation learning on visual-symbolic graphs for video understanding." European Conference on Computer Vision, pp. 71-90, 2020.

[13] Harris, Chris, and Mike Stephens. "A combined corner and edge detector." In Alvey vision conference, vol. 15, no. 50, pp. 10-5244. 1988.

[14] Kipf, Thomas N., and Max Welling. "Semi-supervised classification with graph convolutional networks." arXiv preprint arXiv:1609.02907 (2016).

[15] Laptev, Ivan. "On space-time interest points." International Journal of Computer Vision 64, no. 2 (2005): 107-123.

[16] Concha, Darwin Ttito, Helena De Almeida Maia, Helio Pedrini, Hemerson Tacon, André De Souza Brito, Hugo De Lima Chaves, and Marcelo Bernardes Vieira. "Multi-stream convolutional neural networks for action recognition in video sequences based on adaptive visual rhythms." In 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), pp. 473-480. IEEE, 2018.

[17] Ullah, A., Ahmad, J., Muhammad, K., Sajjad, M. and Baik, S.W., 2017. Action recognition in video sequences using deep bi-directional LSTM with CNN features. IEEE access, 6, pp.1155-1166.

[18] Zong, Ming, Ruili Wang, Zhe Chen, Maoli Wang, Xun Wang, and Johan Potgieter. "Multi-cue based 3D residual network for action recognition." Neural Computing and Applications 33, no. 10 (2021): 5167-5181.

[19] Li, Xinyu, Bing Shuai, and Joseph Tighe. "Directional temporal modeling for action recognition." In European Conference on Computer Vision, pp. 275-291. Springer, Cham, 2020.

[20] Yang, Hao, Chunfeng Yuan, Bing Li, Yang Du, Junliang Xing, Weiming Hu, and Stephen J. Maybank. "Asymmetric 3d convolutional neural networks for action recognition." Pattern Recognition 85 (2019): 1-12.

[21] Zhang, Ziqi, Yaya Shi, Chunfeng Yuan, Bing Li, Peijin Wang, Weiming Hu, and Zheng-Jun Zha. "Object relational graph with teacher-recommended learning for video captioning." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 13278-13288. 2020.

[22] Jiang, Bo, Xixi Wang, and Bin Luo. "PH-GCN: Person re-identification with part-based hierarchical graph convolutional network." arXiv preprint arXiv:1907.08822 (2019).

[23] Peng, Wei, Jingang Shi, Tuomas Varanka, and Guoying Zhao. "Rethinking the ST-GCNs for 3D skeleton-based human action recognition." Neurocomputing 454 (2021): 45-53.

[24] Huang, Zhen, Xu Shen, Xinmei Tian, Houqiang Li, Jianqiang Huang, and Xian-Sheng Hua. "Spatio-temporal inception graph convolutional networks for skeleton-based action recognition." ACM International Conference on Multimedia, pp. 2122-2130, 2020.

[25] Huang, Qingqing, Fengyu Zhou, Jiakai He, Yang Zhao, and Runze Qin. "Spatial–temporal graph attention networks for skeleton-based action recognition." Journal of Electronic Imaging 29, no. 5 (2020): 053003.

[26] Zhu, Guangming, Liang Zhang, Hongsheng Li, Peiyi Shen, Syed Afaq Ali Shah, and Mohammed Bennamoun. "Topology-learnable graph convolution for skeleton-based action recognition." Pattern Recognition Letters 135 (2020): 286-292.

[27] Cheng, Ke, Yifan Zhang, Xiangyu He, Weihan Chen, Jian Cheng, and Hanqing Lu. "Skeleton-based action recognition with shift graph convolutional network." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 183-192. 2020.

[28] Duta, Iulia, Andrei Nicolicioiu, and Marius Leordeanu. "Discovering Dynamic Salient Regions with Spatio-Temporal Graph Neural Networks." arXiv preprint arXiv:2009.08427 (2020).

[29] Ji, Jingwei, Ranjay Krishna, Li Fei-Fei, and Juan Carlos Niebles. "Action genome: Actions as compositions of spatio-temporal scene graphs." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10236-10247. 2020.

[30] Wu, Yuxin, and Kaiming He. "Group normalization." In Proceedings of the European Conference on Computer Vision (ECCV), pp. 3-19. 2018.

[31] Kolesnikov, Alexander, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. "Big transfer (bit): General visual representation learning." In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16, pp. 491-507. Springer International Publishing, 2020.

[32] Soomro, Khurram, Amir Roshan Zamir, and Mubarak Shah. "UCF101: A dataset of 101 human actions classes from videos in the wild." arXiv preprint arXiv:1212.0402 (2012).

[33] Kuehne, Hildegard, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. "HMDB: a large video database for human motion recognition." In 2011 International Conference on Computer Vision, pp. 2556-2563. IEEE, 2011.

[34] Goyal, Raghav, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel et al. "The" something something" video database for learning and evaluating visual common sense." In Proceedings of the IEEE International Conference on Computer Vision, pp. 5842-5850. 2017.

[35] Tran, Du, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. "Learning spatiotemporal features with 3d convolutional networks." In Proceedings of the IEEE International Conference on Computer Vision, pp. 4489-4497. 2015.

[36] Feichtenhofer, Christoph, Axel Pinz, and Andrew Zisserman. "Convolutional two-stream network fusion for video action recognition." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1933-1941. 2016.

[37] Tran, Du, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. "A closer look at spatiotemporal convolutions for action recognition." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6450-6459. 2018.

[38] Carreira, Joao, and Andrew Zisserman. "Quo vadis, action recognition? a new model and the kinetics dataset." In proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6299-6308. 2017.

[39] Ji Lin, Chuang Gan, and Song Han. Temporal shift module for efficient video understanding. arXiv preprint arXiv:1811.08383, 2018

[40] Jiang, Boyuan, MengMeng Wang, Weihao Gan, Wei Wu, and Junjie Yan. "Stm: Spatiotemporal and motion encoding for action recognition." In Proceedings of the International Conference on Computer Vision, pp. 2000-2009. 2019.

[41] Qiu, Zhaofan, Ting Yao, and Tao Mei. "Learning spatio-temporal representation with pseudo-3d residual networks." In proceedings of the IEEE International Conference on Computer Vision, pp. 5533-5541. 2017.

[42] Lin, Ji, Chuang Gan, and Song Han. "Tsm: Temporal shift module for efficient video understanding." In Proceedings of the International Conference on Computer Vision, pp. 7083-7093. 2019.

[43] Arnab, Anurag, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. "Vivit: A video vision transformer." arXiv preprint arXiv:2103.15691 (2021).

[44] Fan, Haoqi, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. "Multiscale vision transformers." arXiv preprint arXiv:2104.11227 (2021).

[45] Diba, Ali, Mohsen Fayyaz, Vivek Sharma, M. Mahdi Arzani, Rahman Yousefzadeh, Juergen Gall, and Luc Van Gool. "Spatio-temporal channel correlation networks for action classification." In Proceedings of the European Conference on Computer Vision (ECCV), pp. 284-299. 2018.

[46] Feichtenhofer, Christoph, Haoqi Fan, Jitendra Malik, and Kaiming He. "Slowfast networks for video recognition." In Proceedings of the International Conference on Computer Vision, pp. 6202-6211. 2019.

[47] Kalfaoglu, M. Esat, Sinan Kalkan, and A. Aydin Alatan. "Late temporal modeling in 3d cnn architectures with bert for action recognition." In European Conference on Computer Vision, pp. 731-747. Springer, Cham, 2020.