Abstract

<https://www.springer.com/journal/11263/updates/25233244>

Introduction

Monitoring human behavior requires fine-grained understanding of actions. Free living activities (FLA) may look simple but their recognition is often more challenging than activities present in sport, movie or Youtube videos. FLA often have very low inter-class variance making the task of discriminating them

from one another very challenging. In the recent literature, the main focus is the recognition of actions from internet videos [1] [2] and very few studies have attempted to recognize ADL in indoor scenarios [3][4] [5]

Following a different direction, action recognition for ADL has been dominated by the use of human 3D poses [6]. They provide a strong clue for understanding the visual patterns of an action over time. Action recognition methods based on skeleton data have been widely investigated and attracted considerable attention due to their strong adaptability to the dynamic circumstance and complicated background.

[31, 8, 6, 27, 22, 29, 33, 19, 20, 21, 14, 13, 23, 18, 17, 32, 30, 34].

Conventional deep-learning-based methods manually structure the skeleton as a sequence of joint-coordinate vectors. [6, 27, 22, 29, 33, 19, 20] , which is fed into RNNs or CNNs to generate the prediction.

However, representing the skeleton data as a vector sequence or a 2D grid cannot fully express the dependency between correlated joints. The skeleton is naturally structured as a graph in a non-Euclidean space with the joints as vertexes and their natural connections in the human body as edges. The previous methods cannot exploit the graph structure of the skeleton data and are difficult to generalize to skeletons with arbitrary forms. Recently, graph convolutional networks (GCNs), which generalize convolution from image to graph, have been successfully adopted in many applications[16, 7, 25, 1, 9, 24, 15]. For the skeleton based action recognition task, Yan et al. [32] first apply GCNs to model the skeleton data. They construct a spatial graph based on the natural connections of joints in the human body and add the temporal edges between corresponding joints in consecutive frames. A distance-based sampling function is proposed for constructing the graph convolutional layer, which is then employed as a basic module to build the final spatiotemporal graph convolutional network (ST-GCN).

The skeleton graph employed in ST-GCN is heuristically predefined and represents only the physical structure of the human body which is not the optimal graph to represent the relationship between human body joint. Ett el represented the concept of Adapative graph convolution which suggests a dynamic graph structure which is more flexible and adaptive based on the data. This data-driven method increases the flexibility of the model for graph construction and brings more generality to adapt to various data samples. AGCN is considered as a more improved version of the ST-GCN approach.

However we found some potential drawbacks both these approaches.

1. In ST-GCN based methods spatial convolution and temporal convolution take place parallelly which makes it capturing both spatiality and temporality of the video is captured at the same time. Before capturing the full version of the spatial features, node feature penetrate through the temporsal direction which makes this approach less accurate in capturing the sequential dependency between spatial features at different time points.
2. ST-GCN based methods lack incorporating the other modalities of information which is an essential property in FLA (especially for human-object interaction)
3. Since spatial and temporal convolution occurs concurrently, it doesn’t provision early fusion of other modalities of information

To solve the above problems, a novel adaptive graph convolution based architecture is proposed in this work. Proposed approach handles the spatiality and temporality in a sequential manner.

The main contributions of our work lie in three folds:

1. Adaptive graph convolution based network architecture is proposed to prove the hypothesis that performing spatial and temporal features separately is performing better than them performing them in a parallel manner.
2. At the same time proposed architecture provisions for early fusion of multiple modalities of information. We validated this fusion operation by adding RGB modality to the corresponding spatial features of the graph convolution data but this is not just limited to RGB data and it can be extended for any number of modalities.
3. On three large-scale datasets for skeleton-based action recognition, the proposed MMAGCN- exceeds the state-of-the-art by a significant margin. The code will be released for future work and to facilitate communication

Related work

Below, we discuss the relevant action recognition algorithms w.r.t. their input modalities.

Image based Action recognition

Traditionally, image level features [49,50] have been aggregated over time using encoding techniques like Fisher Vector [34]andNetVLAD[1]. But these video descriptors do not encode long-range temporal information. Then, temporal patterns of actions have been modelled in videos using sequential networks. These sequential networks like LSTMs are fed with convolutional features from images [10] and thus, they model the temporal information based on the evolution of appearance of the human actions. However, these methods first process the image level features and then capture their temporal evolution preventing the computation of joint spatio-temporal patterns over time.

Due to this reason, Du et al. [48] have proposed 3D convolution to model the spatio-temporal patterns within an action. The 3D kernels provide tight coupling of space and time towards better action classification. Later on, holistic methods like I3D [5], slow-fast network [12], MARS [6] and two-in-one stream network [59] have been fabricated for generic datasets like Kinetics [17] and UCF-101 [45]. But these networks are trained globally over the whole 3D volume of a video and thus, are too rigid to capture salient features for subtle spatio-temporal patterns for ADL.Due to this reason, Du et al. [48] have proposed 3D convolution to model the spatio-temporal patterns within an action. The 3D kernels provide tight coupling of space and time towards better action classification. Later on, holistic methods like I3D [5], slow-fast network [12], MARS [6] and two-in-one stream network [59] have been fabricated for generic datasets like Kinetics [17] and UCF-101 [45]. But these networks are trained globally over the whole 3D volume of a video and thus, are too rigid to capture salient features for subtle spatio-temporal patterns for ADL.

Skeleton based action recognition

Skeleton and joint trajectories of human bodies are robust to illumination change and scene variation, and they are easy to obtain owing to the highly accurate depth sensors or pose estimation algorithms (Shotton et al. 2011; Cao et al. 2017a). There is thus a broad array of skeleton based action recognition approaches. The approaches can be categorized into handcrafted feature based methods and deep learning methods.

The first type of approaches design several handcrafted features to capture the dynamics of joint motion. These could be covariance matrices of joint trajectories (Hussein et al. 2013), relative positions of joints (Wang et al. 2012), or rotations and translations between body parts (Vemulapalli, Arrate, and Chellappa 2014). The recent success of deep learning has lead to the surge of deep learning based skeleton modeling methods. These works have been using recurrent neural networks (Shahroudy et al. 2016; Zhu et al. 2016; Liu et al. 2016; Zhang, Liu, and Xiao 2017) and temporal CNNs (Li et al. 2017; Ke et al. 2017; Kim and Reiter 2017) to learn action recognition models in an end-to-end manner. Recently, Yan et al. [32] propose a spatiotemporal graph convolutional network (STGCN) to directly model the skeleton data as the graph structure. It eliminates the need for designing handcrafted part assignment or traversal rules, thus achieves better performance than previous methods. Furthermore Adaptative Graphs Convolution based action is proposed by Pan et al.[7] has proved better performance over the ST-GCN for action recognition

Multi modal approaches

In order to make use of the pros of both modalities, i.e. image based methods and skeleton based methods it is desirable to fuse these multi-modal information into an integrated set of discriminative features. As these modalities are heterogeneous, they must be processed by different kinds of network to show their effectiveness. There are two types of approaches 1) Fusion based multi modal approaches and 2) Model based Multi modal HAR

Fusion based multi modal approaches creates a joint representation related to model-agnostic approaches that concatenate representations at either the feature or decision level.

Das et al. [4] proposed a attention based fusion model for action recognition while pan et al. [7] proposed a cross-stream network for action recognition. Model-based multimodal methods address multimodal HAR at the model level, which is consistent with the concept of co-learning [8] and represents a kind of fusion method. In these methods multiple models have been trained independently and fused in the later stage of the network to get the results. This is also known as ensemble methods yielded good results in recent literature. [9][10]

We propose a graph convolution based multi modal network to gain the best outcomes of both skeleton based methods and while utilising the complimentary representation power in multi modal networks. We will address the following drawbacks in the existing GCN based methods.

1. Since spatiality and temporality is handled simultaneously , some of salient spatial features which appears after going through multiple spatial convolution layers may not propagate through the temporal axis
2. Since spatiality and temporality is handled simultaneously , difficult to perform a multi modal fusion across spatial domain

[Our model]

Our MM-AGCN adopts the AGCN for capturing the salient features of the human skeleton while harnessing its robustness of view invariance and scene variation. To capture back ground context information we propose a RGB module for capturing the objects and other important informaitiion in the background. It differentiate from other graph based human activity recognition method where it identifies the spatial features first and then fuse multiple modalities together and finally it handles the temporality. As per our understanding this is the first instances which the graph convolution is used with multiple modal fusion for human activity recognition while performing the spatial and temporal aspects sequentially. We have proved that this approach yields better accuracy compared to the simultaneous convolution of spatial and temporal features.

3 Methodology

Proposed Action Recognition Model

To address ADL recognition challenges, we introduce a new architecture based on Adaptive graph convolution [5] The spatial and temporal saliency of human activities can be extracted from the time series representation of pose dynamics, which are described by the 3D joint coordinates of the human body. Furthermore the spatial representation will augmented by the RGB representation which will enable the multi model fusion with AGCN.

3.1 Spatio-temporal representation of a video

Taking as input a stack of human cropped images from a video clip,

Input video clip is taken as a stack of frames in the temporal direction. We process these input stacks of frames in two main streams parrrelley as shown in Fig 1. named as Spatial AGCN stream and RGB encoding stream.

As shown in Fig. 4 (a), SAGCN module will proc sl vs australia liveess sequence of input skeletons. Sequence of the input skeletons is modeled by a 3D spatiotemporal matrix, noted as f*in*. For each frame, the 3D body joint coordinates are arranged in a graph within the spatial dimension. fin with the dimensionality tc x jn x c tc denotes the temporal dimension jn denotes the number of joints in the human body and c denotes number of features in each joints

RGB encorder module will process sequence of RGB frames. Fin for the RGB encoder module will have the dimension tc x m x n x c where tc denotes the temporal dimension, m × n the spatial scale and c the channels.

3.1 Graph convolution

Graph construction :

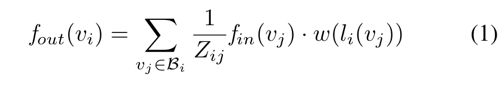
The raw skeleton data in one frame are always provided as a sequence of vectors. Each vector represents the 2D coordinates of the corresponding human joint. A complete action contains multiple frames with different lengths for different samples. We employ a spatiotemporal graph to model the structured information among these joints along both the spatial and temporal dimensions. The structure of the graph follows the work of ST-GCN [11]. The left sketch in Fig. 1 presents an example of the constructed spatiotemporal skeleton graph, where the joints are represented as vertexes and their natural connections in the human body are represented as spatial edges (the orange lines in Fig. 1, left). For the temporal dimension, the corresponding joints between two adjacent frames are connected with temporal edges (the blue lines in Fig. 1, left). The coordinate vector of each joint is set as the attribute of the corresponding vertex.

[Figure 1: Illustration of spatio temporal skelton visualisation]

Graph Convolution :

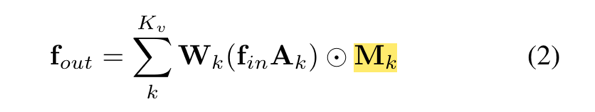
Given the graph defined above, multiple layers of spatial graph convolution operations are applied on the graph to extract the high-level features. Later extracted feature vector is fused with other spatial data to build more stronger spatial encoding. Then LSTM layer to capture the temporal relationship between frames and followed by a softmax classifier are then employed to predict the action categories .

In the spatial dimension, the graph convolution operation on vertex vi is formulated as [11]:



where f denotes the feature map and v denotes the vertex of the graph. Bi denotes the sampling area of the convolution for vi, which is defined as the 1-distance neighbor vertexes (vj) of the target vertex (vi). w is the weighting function similar to the original convolution operation, which provides a weight vector based on the given input. Note that the number of weight vectors of convolution is fixed, while the number of vertexes in Bi is varied. To map each vertex with a unique weight vector, a mapping function li is designed specially in ST-GCN [11]. The right sketch in Fig. 1 shows this strategy, where × represents the center of gravity of the skeleton. Bi is the area enclosed by the curve. In detail, the strategy empirically sets the kernel size as 3 and naturally divides Bi into 3 subsets: Si1 is the vertex itself (the red circle in Fig. 1, right); Si2 is the centripetal subset, which contains the neighboring vertexes that are closer to the center of gravity (the green circle); Si3 is the centrifugal subset, which contains the neighboring vertexes that are farther from the center of gravity (the blue circle). Zij denotes the cardinality of Sik that contains vj. It aims to balance the contribution of each subset.

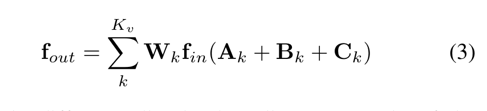
the feature map of the network is actually a C × T × N tensor, where N denotes the number of vertexes, T denotes the temporal length and C denotes the number of channels. To implement this spatial convolution operation , above equation has been transformed in to the following format



where Kv denotes the kernel size of the spatial dimension. With the partition strategy designed above, Kv is set to 3. Ak = Λ− 1 2 k ̄ AkΛ− 1 2 k , where ̄ Ak is similar to the N × N adjacency matrix, and its element ̄ Aij k indicates whether the vertex vj is in the subset Sik of vertex vi. It is used to extract the connected vertexes in a particular subset from fin for the corresponding weight vector. Λii k =∑ j( ̄ Aij k ) + α is the normalized diagonal matrix. α is set to 0.001 to avoid empty rows. Wk is the Cout × Cin × 1 × 1 weight vector of the 1 × 1 convolution operation, which represents the weighting function w in Eq. 1. Mk is an N × N attention map that indicates the importance of each vertex. O denotes the dot product.

3.2 Adpative graph convoilution

In our implementation we have adopted the adaptive graph convolution (AGCN) [5] which is a improved version of the convolution introduced in ST-GCN implementation [11] . In contrast to the fixed predefined grph topology , AGCN utilises a learnable graph topology to define graph convolution. The graph topology is unique for different layers and samples, which greatly increases the flexibility of the model. To make the graph structure adaptive, we change Eq. 2 into the following form:



In this equation , Adjecency matrix has three componenets. Ak, Bk and Ck. Ak is a fixed normalised NxN adjacency matrix represents the physical structure of the human body. Bk is also a NxN adjacency matrix and its elements are parameterized and optimized together with the other parameters in the training process. Bk is completely learned according to the training data. The third part (Ck) is a data-dependent graph which learn a unique graph for each sample. Wk is the weight function.

Spatial Adaptive Graph Convolutional block

SAGCN block defines the steps involved in a single spatial convolution iteration. Spatial convolution is followed by a batch normalization (BN) layer and a ReLU layer. Additional dropout layer with the drop rate set as 0.5 to reduce the oiverfitting. In the proposed architecture temporal convolution does not take place in this stage where it focuses only on caturing the spatial features . Temporal dependencies are captured I a later part of the model. This one of the key difference compared to the original ST-GCN implementation.

[Figure 4 : Illustration of the Spatial AGCN

Convs represents the spatial GCN, which is are followed by a BN layer and a ReLU layer and a dropout layer

Spatial adaptive graph convolutional network (SAGCN) is the stack of these basic blocks, as shown in Fig. 4. There are a total of 7 blocks. The numbers of output channels for each block are 64, 64, 64,64, 128, 128 and 128. A data BN layer is added at the beginning to normalize the input data. Output of the SAGCN (fs) will be feed to the Multi modal coupler to fuse with other modalities of spatial data.

3.2 RGB encoder

Skelton based representation in SAGCN is really good in capturing behaviourr in the human body in view invariant manner, but at the same time it loses some of the background context information which might be critical when identifying some of the action classes specially the ones having object interactions. Primary objective of RGB encoder is to augment the spatial feature vector derived in SAGCN module.

Fin for the RGB encoder module will have the dimension tc x m x n x c where tc denotes the temporal dimension, m × n the spatial scale and c the channels.

The input of RGB encoder module are successive RGB frames of human body along the video. Using ResNet50 image encoding arrchitectuer we get 1000D image encoding for each frame. Starting from the input of 4000 human-cropped frames from a video V , the spatio-temporal representation g is the feature map extracted from an intermediate layer of the ResNet50 [12] . RGB encoder module consists with ResNet50 encoder and three standard convolution layers as shown in Fig 5. Each convolutional layer is followed by Relu activation. Here we use a pretrained ResNet50 module to get the initial encoding of the image. Dropout layers have been added to reduce the overfitting. Number of output channels frm each sub unit is 1000, 512, 256 and 128 respectively. Output of the RGB encoder module (frgb) will be feed to the Multi modal coupler to fuse with other modalities of spatial data.

3.3 Multi modal coupler

Multiple modalities of information are fused in this module. In our case, we are focusing on the pose information extracted from the SAGCN module and the feature vector extracted from the RGB encoder module. But functionality of multi modal coupler is not limited to two modalities. It can be expanded to a higher number of modalities. Intermediate fusion [Ref] is taking place here where representations from different modalities are merged together to build up an even more rich representation of the input video.

Input from the SAGCN module (f1) will have the dimensionality tc x m x c where tc denotes the temporal dimension, and m represents the number of nodes in the pose and c denotes the number of channels

Similarly the input from the RGB (f2) will have the dimensionality tc x c2 where tc denotes the temporal dimension and c2 denotes the number of output channels.

Multiple modalities are fused by merging operation

F= f1+f2 … + fn

This merging operation works as booth modalities have the same temporal frequency. Output of the multimodal couple will have the dimensionality of tc x (dim(f1) + dim (f2) + .. + dim(fn)) . Each modality is normalised before performing the merging operation.

3.4 Temporality handling and Activity classification

Temporal depencey between frames are identified using a RNN on the output feature vector from the multi modal coupler. F is sent through a LSTM and it will be followed by a softmax layer. LSTM is capable of identifying long term dependencies between individual frames in the video. Softmax layer will predict the action label of the input video.

Experiments

In this section we evaluate the performance of SAGCN in skeleton based action recognition experiments. We experiment on two large-scale action recognition datasets : Toyota-Smarthome [3] a recent a daily living action dataset and NTURGB+D [13] the largest in-house captured action recognition dataset. In particular, we first perform detailed ablation study on the Toyota-Smarthome dataset to examine the contributions of the proposed model components to the recognition performance. Then we compare the recognition results of ASGCN with other state-of-the-art methods.

**Toyota-Smarthome** (Smarthome) is a recent daily living activity dataset recorded in an apartment where 18 older subjects carry out tasks of daily living during a day. The dataset contains 16.1k video clips, 7 different camera views and 31 complex activities performed in a natural way without strong prior instructions. This dataset provides RGB data and 3D skeletons which are extracted from LCRNet [14] . For evaluation on this dataset, we follow cross-subject (CS)and cross-view (CV1 and CV2) protocols proposed in [3].

**NTU RGB+D** (NTU-60 & NTU-120) NTU-60 is acquired with a Kinect v2 camera and consists of 56880 video samples with 60 activity classes. The activities were performed by 40 subjects and recorded from 80 viewpoints. For each frame, the dataset provides RGB, depth and a 25-joint skeleton of each subject in the frame. For evaluation, we follow the two protocols proposed in [13] : cross-subject (CS) and cross-view (CV). NTU-120 is a super-set of NTU-60 adding a lot of new similar actions. NTU-120 dataset contains 114k video clips of 106 distinct subjects performing 120 actions in a laboratory environment with 155 camera views. For evaluation, we follow a cross-subject (CS1) protocol and a cross-setting (CS2) protocol proposed in [15].

Training details:

All experiments are conducted on the PyTorch deep learning framework. Stochastic gradient descent (SGD) with Nesterov momentum (0.9) is applied as the optimization strategy. The batch size is 16. Cross-entropy is selected as the loss function to backpropagate gradients. The weight decay is set to 0.0001. In RGB module , each CNN layer is followed by a dropout layer of 0.5 . Base learning rate is set to 0.1. The learning rate is set as 0.1 and is divided by 10 at the 30th epoch and 40th epoch. The training process is ended at the 50th epoch.

Toyota-smarthome dataset , there was only one person appeared in each frame and max number of frames were set to 4000. If there is less number of frames we padded the rest with 0 till 4000 frames.

For the NTU-RGBD dataset, there are at most two people in each sample of the dataset. If the number of bodies in the sample is less than 2, we pad the second body with 0. The max number of frames in each sample is 300. For samples with less than 300 frames, we repeat the samples until it reaches 300 frames.

Runtime :

Training the separable SAGCN model end-to-end takes 10h over 2x Nvidia Tesla V100 GPUs on Smarthome in CS settings. At test time, a single forward pass for a video takes 338ms on 2 GPUs.

5.3 Ablation study

We examined the effectiveness of the proposed architecture of spatial Adaptive Graph Convolution in this section with the Cross-subject benchmark on the Toyota-smarthome dataset. The performance is compared in three differrents aspects.

5.3.1 Why RGB module and LSTM unit

In this section we highlight the importance of temporality factor handling using a LSTM instead of average pooling approach followed in ST-GCN [11]. Furthermore we highlight the iimportance of using the RGB module to capture the context information from RGB image frames.

First we conducted the two types of experiments one using the LSTM for aggregating the image encoding in the temporal direction and the other type using the average pooling for aggregating image encoding. Under each type we conducted two other experiments where one with RGB module and the other one without RGB module input. This is to validate the effect of using the RGB modality as a input to augment the feature extraction from SAGCN component.

|  |  |  |
| --- | --- | --- |
|  |  | Toyota-Smarthome (CS)  Accuracy % |
| Average Pooling | SAGCN + Average pooling | 58.53 |
| SAGCN + RGB + Average pooling | 60.52 |
| LSTM | SAGCN + LSTM | 59.56 |
| SAGCN + RGB + LSTM | 61.46 |

Table 1: Comparisons of the accuracy when using RGB modality and using AVG pooling vs LSTM modules

It is clear that with the introduction of LSTM instread of Average pooling it increases the classification oaccuracy and also with the introiduction of the RGB module, it utilises the new modality of data to enhance its latent representation of each image. So the version with SAGCN in combination with RGB module and the LSTM can be considered as the best out of given four configurations.

5.3.2 Number of AGCN layers

Another important consideration is the number of adaptive convolutional layers appear in the SAGCN unit. Here, we compare the performance by having different number of layers in SAGCN unit. In this case we tried with 1, 4 7 and 10 layers. Adding a layer always added a extra complexity to the model by introducing additional set of parameters to adjust.

|  |  |
| --- | --- |
| Number of AGCN layers | Toyota-Smarthome (CS)  Accuracy % |
| 1 | 54.7 |
| 4 | 59.51 |
| 7 | 61.46 |
| 10 | 57.08 |

Table 2: Comparisons of the accuracy when using different number of AGCN layers

It is clear that having moderate level of AGCN layers yield the best acurayc. Having smaller number of spatial convolutions means that the model does not have capacity to capture the salient features of the input data while having too many convolution means that model is getting over complex and ultimately lead to overfitting to the training data set.

5.3.3 Graph convolution

In this section we validate the usage of Adaptive graph convolution instead of the standard graph convolution operation introduced in ST-GCN [11] in the spatial domain.

|  |  |
| --- | --- |
|  | Toyota-Smarthome (CS)  Accuracy % |
| Standard Graph Convolution (ST-GCN implementation) | 59.97 |
| Adaptive Graph Convolution | 61.46 |

Table 3: Comparisons of the accuracy when using different graph convolution operations

In the original literature it shows that adaptive graph convolution outperforms over ST-GCN. But in both implementations spatial and temporal convolution occur simultaneously. But here prove that same performance holds even we Performa spatiality and temporality separately in two stages. Adaptive graph convolution is better in capturing the dynamic nature in human skeleton.

Comparison with the state of the art

We compare SAGCN to the state-of-the-art (SoA) on Smarthome, NTU-60 & NTU-120. Proposed SAGCN outperform each of them. The methods used for comparison include the handcraft-feature-based methods [16], RNNbased methods [17,18], CNN-based methods [1,3] and GCN-based methods [4]. Our model achieves state-of-the-art performance with a considerable margin margin on both of the datasets, which verifies the significance of our model.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Method | Pose | RGB | CS | CV1 | CV2 |
| DT[16] | x | Y | 41.9 | 20.9 | 23.7 |
| LSTM [17] | Y | x | 42.5 | 13.4 | 17.2 |
| I3D [1] | x | Y | 53.4 | 34.9 | 45.1 |
| Separable STA [3] | Y | Y | 54.2 | 34.3 | 43.9 |
| UNIK [18] | x | Y | 58.9 | 21.9 | 58.7 |
| VPN [4] | Y | Y | 60.8 | 43.8 | 53.5 |
| SAGCN (Ours) | Y | Y | 61.4 | 37.27 | 65.53 |

Table 6. Results on smarthome dataset with cross-subject (CS) and cross-view (CV1 and CV2) settings (accuracies in %).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Method | NTU-60 | | NTU-120 | |
|  | CS | CV | CS1 | CS2 |
| 2s-AGCN [5] | 84.2 | 93 | 78.2 | 82.9 |
| UNIK [18] | 85.1 | 93.6 | 79.1 | 83.5 |
| Separable STA [3] | 92.2 | 94.6 | 83.8 | 82.5 |
| VPN [4] | 93.5 | 96.2 | 86.3 | 87.8 |
| SAGCN (Ours) |  |  |  |  |

Table 7. Results (accuracies in %) on NTU-60 with cross-subject (CS) and cross-view (CV) settings and NTU-120 with cross-subject (CS1) and cross-setup (CS2) settings

Conclusion

In this work we proposed a novel articture for processing the spatial and temporal aspects of videos by leveraging the convolutional power of adaptive graph convolution. Our proposed architecture allows to fuse multiple modalities of data together to get a moer accurate latent reprresentttion of the videos. The final model is evaluated on two large-scale action recognition datasets, NTU-RGBD and Smarthome, and it outperforms most of the existing state-of-the art action recognition methods in terms of accuracy. As future work we plan to extract more discriminative features of the human movements which will give more insights by going beyond the simple activity classification.

Acknowledgement

*This work was supported by the Australasian Leadership Computing Grants scheme, with computational resources provided by NCI Australia, an NCRIS enabled capability supported by the Australian Government.*