Chapter 4 HUMAN ACTIVITY RECOGNITION USING ENLARGED TEMPORAL DIMENSION OF DEPTH MAP SEQUENCES

4.1. Introduction

An activity takes many seconds to complete, which makes it a spatiotemporal structure. Many contemporary techniques tried to learn activity representation using the convolutional neural network from such structures to recognize activities from videos. Nevertheless, these representations failed to learn complete activity because they utilized very few video frames for learning. In this work, we use raw depth sequences considering its capabilities to record geometric information of objects and apply proposed enlarged time dimension convolution to learn features. Due to these properties, depth sequences are more discriminatory and insensitive to lighting changes as compared to RGB video. As we use raw depth data, time to do pre-processing are also saved. The 3 dimensional space-time filters have been used over increased time dimension for feature learning. Experimental results demonstrated that by lengthening the temporal resolution over raw depth data, the accuracy of activity recognition has been improved significantly. We also studied the impact of different spatial resolution and concluded that accuracy stabilizes at larger spatial sizes. Many applications, in particular, intelligent surveillance, human robot interaction, video or image annotations, education, security, clinical applications, digital libraries, and video conferencing, are mainly based on human activity recognition. Human activities can be seen as a sequence of basic motions. For example, activities like brushing hair or hand waving can be described as a sequence of consecutive raising and lowering of the hand. The research in this field is mainly focussed on RGB videos and handcrafted features only even though none of the handcrafted features is considered universally best for all datasets [311].

Many techniques to represent motion dynamics and classify activities in videos have been studied in recent years because of its large number of real world applications. However, to recognize human activities in unconstrained videos is a great challenge due to some real conditions such as occlusions, different viewpoints, different action speeds, light variances, etc. [312]–[314]. In comparison with traditional RGB cameras, depth cameras allow us to obtain the traditional two-dimensional color video sequences as well as the depth sequences which are more insensitive to illumination changes [315]and more discriminating than color and texture features in many computer vision problems like segmentation, object detection, and activity recognition [316]. Many studies have been conducted on activity recognition which employed deep convolutional networks based on either RGB images or depth sequences but in most of these methods networks are feed on handcrafted features calculated from depth maps instead of raw depth maps.

Convolutional networks have been achieved remarkable results in action recognition and classification task from images. The capability of deep convolution network to learn complex representations from large visual data datasets makes them suitable for video based activity recognition. However, there are two significant factors that influence video based activity recognition with deep convolutional network. First, the length of time dimension which helps to understand the dynamics in activity videos [317]. Second, large volume of training data to obtain optimum accuracy. Recent CNN approaches for activity recognition frequently extend CNN architectures for fixed images [318] to learn activity representations for short video intervals varying from 1 to 16 frames [50], [85], [319]. Nevertheless, typical human activities like hand-shaking and eating, as well as series of repetitive actions such as running and swimming often last some seconds and extend over tens or hundreds of video frames, small number of frame may not be enough to learn temporal structure of activity.

Successful state-of-the-art approaches for activity recognition in practice are deep convolution neural networks based motion representation [85] and motion-based video descriptors such as HOG and MBH [150], [320]. Although with large scale training datasets [321], [322]. CNNs demonstrates the power of learning visual representations [320], the existing methods experience the inability to concisely encode long-term motion dependencies because they generally emphasis on appearance and short-term motion.

Unlike existing methods, in this work, we proposed enlarged temporal convolutions using 3D filters to learn spatio-temporal feature from raw masked depth sequence. The 3D deep convolutional neural network based architecture [319] are considered and study framework for proposed method. The network framework works on raw depth sequences i.e no complex computations are required to be done to prepare input. The learned features encapsulate the complete temporal structure information at the cost of reduced spatial structure to manage the complexity of the model. We investigate architecture with temporally extended convolutions. In other words, both spatial and time dimension convolution over the input video are taken by deep convolutional neural network with varied temporal resolution. Our method reports the state-of-the-art performance on NTU RGB-D[289], MSRAction3D[290] and MSRDailyActivity3D[126] dataset without using any handcrafted feature calculation and major preprocessing.

This chapter evaluate the effect of different spatial and temporal resolution of input sequences to performance of feature learning with 3D based deep convolution neural network architecture. In the remaining part of the chapter we described proposed method and network architectures in Section 4.2 and Section 4.3 presents experimental results and discussion

4.2. The Proposed Method

In this section we present the proposed enlarged time dimension convolution over the network framework given in Fig. 4.1 to learn features from raw depth sequences. Next we specify the varied spatial and temporal input sizes that will be supplied to the network. Finally, we provide details on learning procedures.

4.2.1. Network Architecture

The intuition behind the effectiveness of the network architecture is its *simple* and *compact* design which utilizes the ability of 3D convolutional filters to learn spatio-temporal information more *efficiently.* Learning unique characteristic patterns with long-term temporal structure for each activity with 3D filters from more accurate geometric information provided by depth sequences, makes the model represent high quality features and use them to perform activity recognition task. A 3D deep architecture can be as deep as possible for large dataset provided the machine memory limit and computation affordability. In accordance with currently available computation power and memory, our network architecture contains 5 convolution layers, 5 pooling layers, two fully connected layers, and softmax classifier.

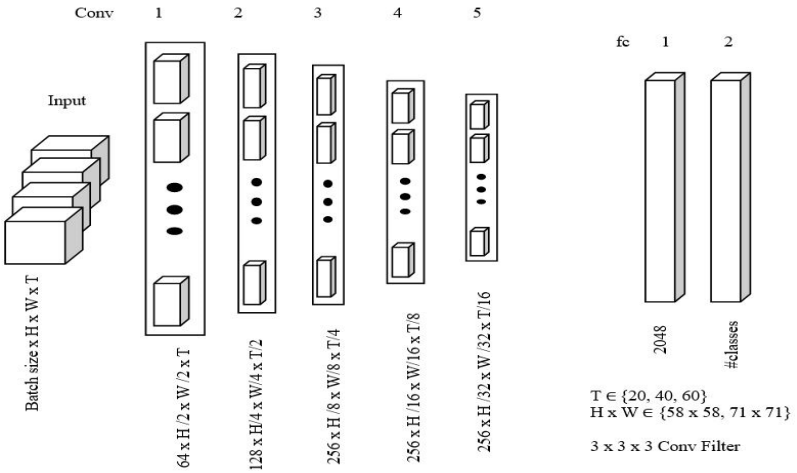


Figure 4.1 3D Deep Convolutional Neural Network framework with 3D filters. (Input is normalized Depth sequences of fixed cuboid size. Convolution filters size is 3 x 3 x 3 in all the 5 layers and after every convolution max pooling is applied for downsampling)

The 3-dimensional deep network is responsible for spatiotemporal feature learning from masked depth maps. It consists of 5 three dimensional convolution layers with 64, 128, 256, 256 and 256 filter response maps, each of which followed by a rectified linear unit (ReLU) activation and a max pooling layer. The rectified linear unit is a non-linear activation function which output 0 for negative input and raw output for positive input. This results in simple gradient computation and hence speed up the training. Max pooling layer helps to reduce the sample size which further reduce computation cost. The receptive fields of convolution kernel size 3 × 3 for deep network frameworks have been found best performer in 2D ConvNet [323]. Further, [319] empirically proves that for 3D ConvNets, 3 × 3 × 3 convolution kernel is the best choice. Hence, we fix the receptive field to 3 × 3 × 3 convolution kernel for our architecture. Size of max pooling filters are 2 × 2 × 1 for first layer and 2 × 2 × 2 for rest of the layers. This means after every convolution layer all the three dimensions will be halved except the first where temporal size preserved. Finally, two fully connected linear transformation layers (FC) are applied and soft-max layer is used as the classifier. The vector dimensions of two fully connected linear transformation are 2048 and number of classes respectively. Padding of 1 pixel are used in all dimensions to keep the size of convolution outcome constant. Stride of filters for all three dimensions is 1 and 2 for convolution and pooling operations respectively. Dropout is used only for the first fully connected layer whereas ReLU layers are added after each fully connected layer. At the end of the network softmax layer is employed to produce class scores.

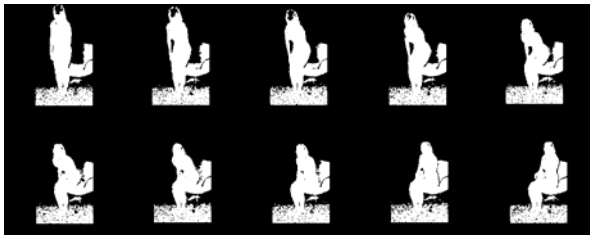
4.2.2. Network Input

The network are feed with the cuboids of depth sequences which are first pre-processed and normalized to the same size for a particular dataset. The pre-processing of input

cuboid involves center crop, rescaling to normalize spatial and temporal dimension, and depth value normalization. In center crop each frame is cropped to center region. After that resize them to size 200 x 200 pixels. Next, randomly sample a cuboid of specific dimension from depth sequences in a dataset to normalize spatial and temporal dimension. Finally depth values of all pixels are normalized to 0-1 range by applying min-max normalization. The raw depth sequences from NTU-RGB+D dataset for sitting activity is depicted in Fig. 4.2(a). Fig. 4.2(b) show the frame after cropping to center region of size 200 × 200 pixels. Further, on rescaling of the raw depth sequence to 58 × 58 pixels, it will look as shown in Fig. 4.2(c).

To evaluate the effect of enlarged temporal convolutions, the network inputs with varied time resolution. The study has been conducted using orderly increased temporal resolution T ∈ {20, 40, 60} frames and spatial resolution {58 x 58, 71 x 71} pixels. The spatial resolution is kept small to manage the network complexity. As illustrated in Table 4.1, for each of the five convolutional layers in 60f network the temporal resolution corresponds to 60, 30, 15, 7 and 3 frames.

(a)



(b)

(c)

Figure 4.2 (a) Raw depth map of size 512 x 424 pixels of sitting activity from NTU-RGB+D dataset [289] (b) Crop depth sequence of sitting activity to center region of size 200 x 200 pixels (c) Resize center crop sequence to 58 x 58 pixels.

In contrast, the temporal dimension of the 20f network shrinks more drastically to 20, 10, 5, 2 and 1 frame at every convolutional layer. We believe that more complicated temporal structures can be learned by retaining the temporal resolution at higher convolutional layers. The spatiotemporal resolution of the outcomes of the last convolutional layers in case of {58 x 58} pixels is 1 × 1 × 1, 1 × 1 × 2 and 1 × 1 × 3 for the 20f, 40f and 60f networks respectively and for {71 x 71} pixels is 2 × 2 × 1, 2 × 2 × 2 and 2 × 2 × 3 for the 20f, 40f and 60f networks respectively.

Table 4.1 Max pool filter sizes and spatial (HxW) and temporal (T) size of output corresponding to each convolution layer

Layer	*Max Pool Filter size*	*Size of Output*				
		H x W (58 x 58)	*H x W (71 x 71)*	*T (20)*	*T (40)*	*T (60)*
1	2 x 2 x 1	29 x 29	35 x 35	**20**	**40**	**60**
2	2 x 2 x 2	14 x 14	18 x 18	**10**	**20**	**30**
3	2 x 2 x 2	7 x 7	9 x 9	**5**	**10**	**15**
4	2 x 2 x 2	3 x 3	4 x 4	**2**	**5**	**7**
5	2 x 2 x 2	1 x 1	2 x 2	**1**	**2**	**3**

4.2.3. Learning

The network is trained on the training set of 20 subjects of NTU RGB-D dataset, which contain 40,320 videos. The stochastic gradient descent optimizer is used on mini-batches with negative log likelihood criterion to optimize the parameters. The negative log likelihood criterion requires normalized log-probabilities model output which is achieved with a softmax function. The mini-batch of 30 videos is used in case of 20f networks. However, the batch size is reduced to 15 videos and 10 videos for 40f and 60f networks respectively because of limitations of computing power. The model is trained with initial learning rate of 5×10^{-4} for learning from scratch. The learning rate is divided by the factor of 10^{-1} when testing accuracy stops increasing. The experimental setup uses 0.5 dropout ratio and initialize weight decay with 5×10^{-4}.

4.3. Experimental Results and Discussion

In this section, we evaluate the impact of increased spatial and temporal resolution on the feature learning. Afterwards, the results will be compared to the-state-of-art on three depth datasets for activity recognition: NTU-RGB+D, MSRAction3D and MSRDailyActivity3D dataset.

The network is trained and tested on NTU-RGB+D, MSRAction3D and MSRDailyActivity3D dataset for activity recognition. The NTU-RGB+D dataset [289] were captured using Microsoft Kinect v2 sensors and provide depth sequences of two dimensional depth values in millimetres. It contains 57K videos for 60 activities classes performed by 40 distinct subjects and 80 viewpoints. The resolution of each depth frame is 512 × 424 pixels and 30 fps frame rate. Each subject performed each activity twice. Cross subject evaluation metric as mentioned in [324] is used in which the 40 subjects

are split into training and testing groups. Each group consists of 20 subjects. For this evaluation, the training and testing sets have 40,320 and 16,560 samples, respectively.

The MSRAction3D dataset was collected with depth sensor. It consists of 20 activity types performed by 10 subjects. Every activity is performed by every subject 2 or 3 times. There are 567 depth sequences of resolution 640 x 240 pixels in total. The cross subject setting is used to evaluate the network where five subjects 1,3,5,7,9 are used for training and rest of the five 2,4,6,8,10 for testing.

The third dataset is MSRDailyActivity3D dataset which was also recorded with an MS Kinect-V1 sensor. It contain 16 activities: drink, eat, walking, read book, write on paper, use laptop, cheer up, use vacuum cleaner, sit still, toss paper, play guitar, play game, lay down on sofa, sit down, stand up, and call cell phone. In this dataset ten different subjects perform each activity two times, once in standing and the other in sitting position. There are a total of 320 depth sequences in the dataset.

The pre-processing of each video sequence is done in order to make network less insensitive to different subjects. NTU-RGB+D dataset provide masked depth sequences which means foreground information is removed but in case of MSRAction3D and MSRDailyActivity3d dataset foreground extraction has been done as a part of preprocessing.

4.3.1. Performance of Varied Spatial and Temporal Sizes on NTU-RGB+D Dataset

We evaluate the efficiency of the network and compare the recognition results to the state-of-the-art approaches on the NTU-RGB+D dataset in Table 4.2. We adopt cross-subject method to evaluate performance on all activities in which we train the network against training subjects 1, 2, 4, 5, 8, 9, 13, 14, 15, 16, 17, 18, 19, 25, 27, 28, 31, 34, 35, 38 and

remaining subjects are reserved for testing. In order to evaluate the performance on different spatiotemporal size to performance of 3D deep CNN, we normalized the depth sequences of each activity to different temporal dimension sizes. The input cuboid of size 71 x 71 x 20, 58 x 58 x 20, 58 x 58 x 40, 71 x 71 x 40, 58 x 58 x 60 and 71 x 71 x 40 are evaluated.

Table 4.2 Performance of proposed approach on NTU-RGB+D dataset for different spatial and temporal dimensions

NTU-RGB+D dataset		
	Accuracy	
	{58 x 58}	*{71 x 71}*
T = 20	52.63 %	56.31 %
T = 40	59.48 %	61.23 %
T = 60	64.03 %	64.57 %

Table 4.2 reports the performance comparison at varying spatial and temporal resolution for depth sequences. From this table, we observe that the classification accuracy is improved with increased temporal sizes and believed that it can be further improved with more increase in temporal resolution. However, the gain from increasing spatial resolution is noticeable at lower values of temporal resolution, but accuracy stabilizes with larger spatial sizes since high temporal and spatial sizes learn the structure information of sequences completely. In Fig. 4.3, we plot the accuracy corresponding to different temporal size T ∈ {20, 40, 60} frames and at different spatial resolution {58 x 58, 71 x 71} pixels. The possible reason for the trend can be seen in Fig. 4.4 which clearly depicts that short frame interval are not sufficient to distinguish among activities because of similarities in motions before actual activity takes place.

Comparison to state-of-the-art methods: In Table 4.3 and Figure 4.5, classification accuracy on NTU-RGB+D dataset is compared with other existing techniques. It includes methods based on handcrafted feature extraction from depth maps, skeleton features and deep convolution neural network.

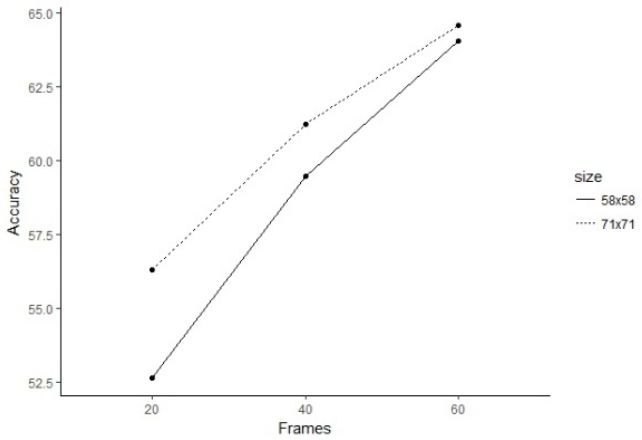


Figure 4.3 Results for NTU-RGB+D using network of varying temporal and spatial resolution

Depth-map based features (HOG [325], Super normal vector [316], and HON4D [313]) have been evaluated on NTU-RGB+D dataset in [289]. These techniques performed poorly due to their susceptibility towards learning low level appearances and view-dependent motion patterns. The performance of Skeletal-based features (Lie group [326], Skeletal Quads [327], and FTP Dynamic Skeletons [328]) are better because of view-independent nature of 3D skeletal representation, but they are vulnerable to faults of the body tracker. In comparison to our method which involved nearly zero pre-processing, all the above mentioned methods can model short range activity only and even after pre-processing depth data using extensive calculations and computations they performed poorly.

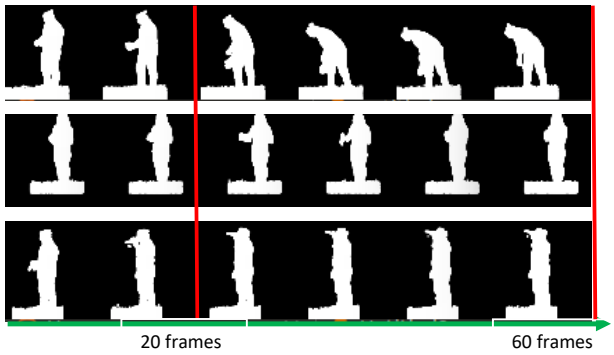


Figure 4.4 Visualization of 6 frames extracted at every 10 frames of activity Drink (row 1), Drop (row 2) and Tear up paper (row 3). Our network can capture the long interval, whereas 20-frame networks fail to recognize such long-term activities.

In comparison to handcrafted features based methods deep learning method performs better. In paper [289], the inability of recurrent neural networks (RNN) model in finding long-term mutual dependencies of input makes it inefficient for long-range activities even after utilizing two layers of RNN and a lot of handcrafted engineering which is computationally expensive in comparison to our raw depth based enlarged temporal convolutions. Further, the long short term memory (LSTM) model has been employed in long-term context in [289]. The significant improvement in performance has been recorded but storage capacity is the major drawback. Another HBRNN-L [324] model had used five networks for the task of activity recognition whereas our method employed only one 3D network for the same task and outperforms. In other words, number of parameters are comparatively much less in number in our proposed model than HBRNN-L.

Table 4.3 Performance comparison of proposed method with other state-of-the-art methods on NTU-RGB+D dataset [289]

Input Modality	*Method*	*Accuracy*
Depth map based baseline method	HOG [325]	32.24 %
	Super Normal Vector[316]	31.82 %
	HON4D [313]	30.56 %
Skeleton-based baseline method	Lie Group [326]	50.08 %
	Skeletal Quads [327]	38.62 %
	FTP Dynamic Skeletons [328]	60.23 %
Skeleton-based deep learning method	HBRNN-L [324]	59.07 %
	2 Layer RNN [289]	56.29 %
	2 Layer P-LSTM [289]	62.93 %
Depth based deep learning method	*Ours (3D Deep CNN + 60f)*	*64.57 %*

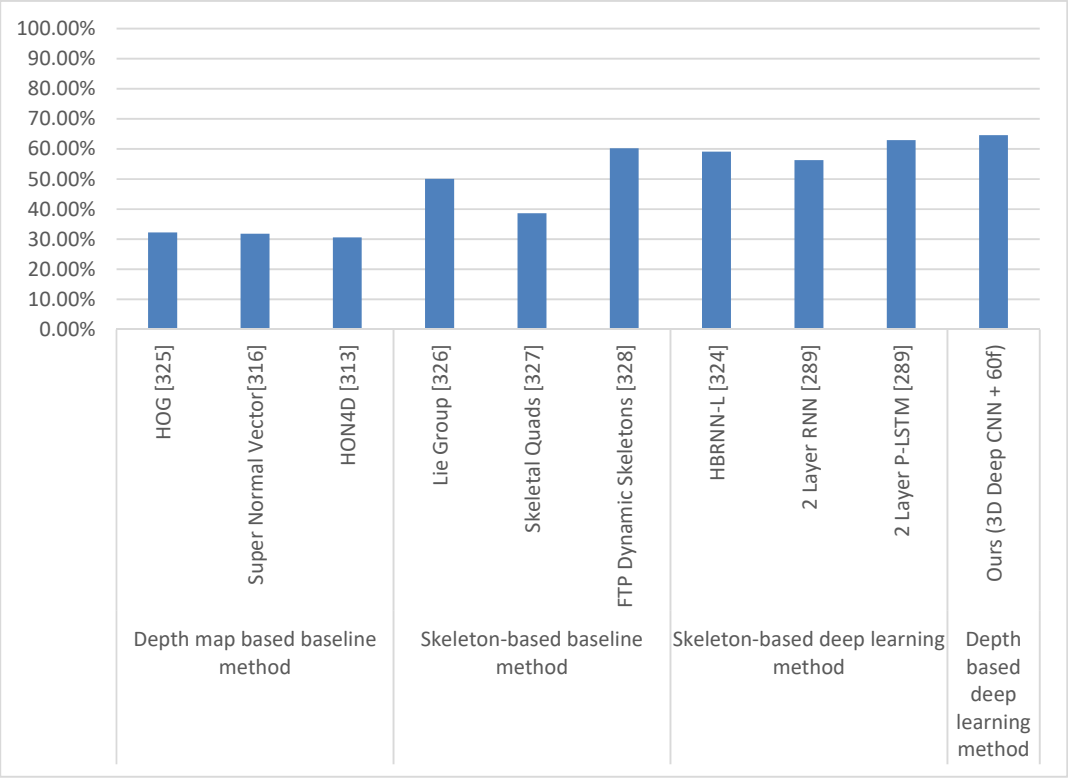


Figure 4.5 Result comparison chart of proposed method with other state-of-the-art methods on NTU-RGB+D dataset

A 3D deep architecture can be as deep as possible for large dataset provided the machine memory limit and computation affordability. Our proposed network architecture has been designed in accordance with currently available computation power and memory. In the proposed enlarged time dimension 3D deep model, though we reduce the spatial resolution, we are able to achieve the-state-of-art results on NTU-RGB+D dataset by lengthening the network to 60 frames. Our results are obtained when input to 3D deep CNN are cuboid of size 71 x 71 x 60. The activity recognition accuracy is 64.57%.

4.3.2. Performance of Varied Spatial and Temporal Sizes on MSRAction3D Dataset.

Here, again cross-subject evaluation metric is utilized to examine the performance of network on all activities. Subject 1, 3, 5, 7, 9 are taken as training subjects and remaining subjects are used for testing purpose. The length of all the videos in this dataset is not long enough to perform temporal convolution of 60f. Therefore, the input depth

sequences of sizes 58 x 58 x 20, 71 x 71 x 20, 58 x 58 x 40, and 71 x 71 x 40 are only analysed. The performance comparison at different spatial and temporal resolution is presented in Table 4.4 and observed improvement in classification accuracy with increased temporal sizes.

Table 4.4 Performance of proposed method on MSRAction3D dataset for different spatial and temporal dimensions

MSRAction 3D dataset		
	Accuracy	
	{58 x 58}	*{71 x 71}*
T = 20	83.29 %	86.5 %
T = 40	88.47 %	90 %

Comparison to state-of-the-art methods: Existing methods [126], [313], [323], [329]–[331] involved extensive hand engineered feature computation to fulfil the task of activity recognition. However, our method feed network with raw depth maps and as shown in Table 4.5 and Figure 4.6, the accuracy is comparable to the state-of-the-art methods without using any handcrafted feature. [94] performed inferior to our method may be due to few convolutions layers that too with convolution kernel size 6 x 6 x 7 and 5 x 5 x 5 for layer 1 and layer 2 respectively which are inefficient than 3 x 3 x 3 filter size [319]

Table 4.5 Performance comparison of proposed method with other state-of-the-art methods on MSRAction3D dataset [290]

MSRAction3D	
Method	*Accuracy*
DCSF [325]	89.30 %
Bag of 3D points [126]	74.70 %
ROPs [323]	86.50 %
HON4D [313]	85.85 %
DMA+DMH+HOG [330]	90.45 %
DMMs-based GLAC [331]	90.48 %
3DDCNN [94]	88%
Ours (3D deep CNN+40f)	*90 %*

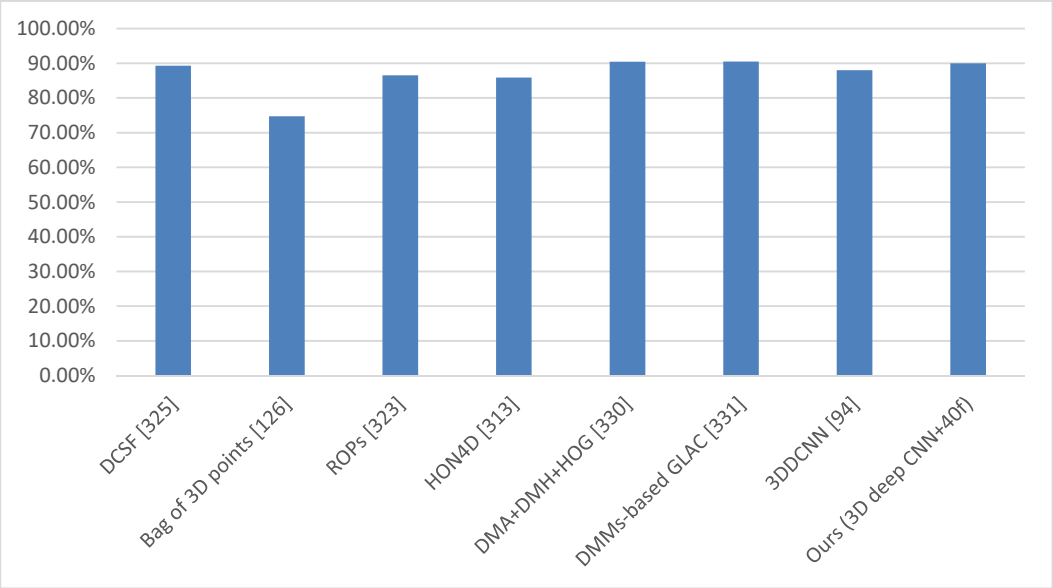


Figure 4.6 Performace comparison chart of proposed method with other state-of-the-art methods on MSRAction3D dataset

4.3.3. Performance of Varied Spatial and Temporal Sizes on MSRDailyActivity3D Dataset

Network as discussed in section 4.2.1 is evaluated for cross subject metric where subject 1, 3, 5, 7, 9 are used for training and testing is done on subject 2, 4, 6, 8, 10. Again due to insufficient depth sequence length for some videos only 58 x 58 x 20, 71 x 71 x 20, 58 x 58 x 40, and 71 x 71 x 40 input resolutions are used. Table 4.6 demonstrate the results and show significant increase in accuracy with increase in the spatial and temporal resolution.

Table 4.6 Performance MSRDailyActivity3D dataset on different spatial and temporal dimensions

MSRDailyActivity3D dataset		
	Accuracy	
	{58 x 58}	*{71 x 71}*
T = 20	73.61 %	77.5 %
T = 40	81.3 %	83.45 %

Comparison to state-of-the-art methods: Again the [313], [327], [332], [91] perform poorly despite of sophisticated feature calculations may be due to incomplete feature learning from few video frames. Also, in comparison with three ConvNets of [31] very

few parameters are there in the proposed network, which makes it computationally efficient. Table 4.7 and Figure 4.7 shows that the proposed method outperforms state-of-the-art methods. It has been seen that the proposed network has been trained on raw depth sequences instead of handcrafted features calculated from depth maps and still perform better than other methods. It was found that retaining high temporal and spatial resolution helped in learning the complex dynamics of the video more accurately.

Table 4.7 Performance comparison of the proposed method with other state-of-the-art methods on MSRDailyActivity3D dataset [126]

MSRDailyActivity3D Dataset	
Method	*Accuracy*
LOP [327]	42.50 %
Depth Motion Maps [332]	43.13 %
Local HON4D [313]	80.75 %
WHDMM + 3Convnet [91]	75.62 %
Proposed (3D deep CNN+60f)	*83.45 %*

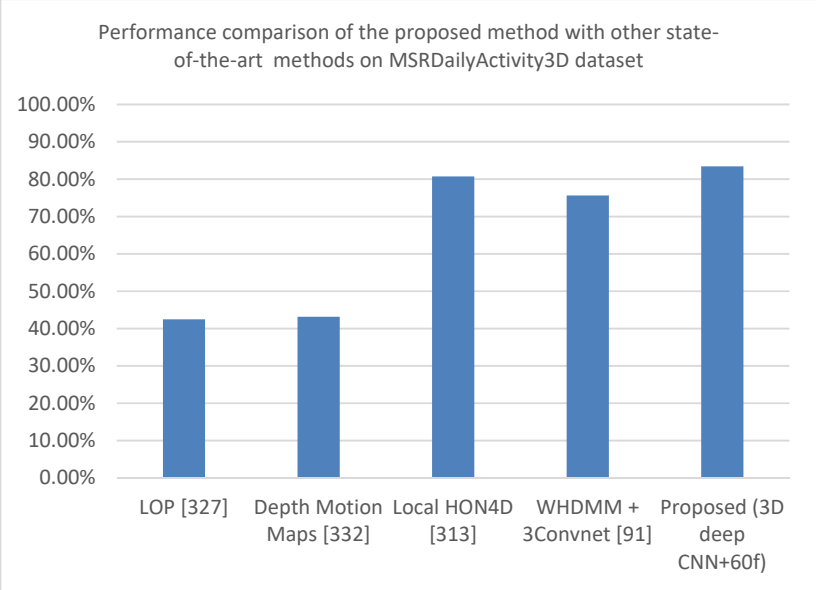


Figure 4.7 Performace comparison chart of proposed method with other state-of-the-art methods on MSRDailyActivity3D dataset

4.4. Conclusion

In this chapter, we investigate the influence of enlarged time dimension convolution to the performance of feature learning with 3D deep convolution neural networkss from raw

depth sequences. The results show that it improves the performance significantly by learning complete activity representation from geometric information recorded in depth sequences. We also obtain state-of-the-art performance using space time filters over a large number of depth maps from a depth sequence on NTU-RGB+D, MSRAction3D and MSRDailyActivity3D activity recognition datasets. In future, we will further increase the temporal length to 100f for large dataset and fuse the results with the results obtained from different handcrafted feature which has been already shown best result for the corresponding dataset such as depth motion maps and skeleton features to further increase the performance.