D²MM-CNN: DIFFERENCE DEPTH MOTION MAP AND CONVOLUTIONAL NEURAL NETWORKS FOR HUMAN ACTION RECOGNITION

S. Sandhya Rani[1], Dr. G. Appa Rao Naidu[2], Dr.V. Usha Shree[3]
[1]Research Scholar, Department of CSE, JNTUH, India
[2]Professor, Department of CSE, JBIET, India
[3]Principal, Department of ECE, JBREC, India

ABSTRACT

Human Action Recognition has become the most significant research area for several applications like robotics, healthcare, gaming, smart houses, etc. However, in computer vision, action recognition from videos is one of the most challenging issues, due to some extraneous aspects like Occlusions, backgrounds, noises and so on. One solution to overcome the above-mentioned problems is acquiring only motion and shape cues form depth action video sequences. With this objective, in this paper, a new action representation approach is proposed based on Depth Motion Map (DMM), called as Difference Depth Motion Map (D²MM). Next, a well-designed CNN is trained especially to extract the features from two actions with a similar structure. The CNN model introduced in this paper involves five convolutional layers, three pooling layers, and one fully connected layer. The experimental results of the proposed method are compared with conventional methods on the publicly available dataset, MSR Action 3D. The comparative analysis proves that the proposed approach outperforms the state-of-art techniques.

Index Terms : *Human Action Recognition, Depth Motion Map, Deep learning, Convolutional Neural Networks, MAR Action 3D dataset, Accuracy.*

I. INTRODUCTION

In recent yeas, with the rapid growth in the technology (e.g. machines and computers) and its utilization in our activities of day-to-day life, Human-Computer Interaction (HCI) facilitation has developed as one of the most prominent research aspects. Henceforth there has been an increased focus by several research studies to construct or develop a new interaction model to meet this objective. The final and ultimate objective of this mechanism is to fill up the HCI gap such that the interaction should be like Human to Human. Among several potential applications of HCI, Human Action Recognition (HAR) based on computer vision has gained a more significance. The main aim of HAR is to understand the enduring human actions which have certain contextual importance. These actions may comprise the whole body movements or only specific portions of the body like Hands, Legs, head or arms, etc., contingent to the application requirement. Computer Vision centered HAR is more powerful, friendly, natural, and less harm to interactive, especially in the environments like where there is no possibility of physical contact and/or speech. The HAR has widespread applications starting from surveillance events understanding for public safety [1], entertainment [2], Robotics [3] and many other applications [4]. Hence, HAR has gained much consideration from both academic and industrial applications.

However, recognizing human actions in color images is a challenging issue due to various problems such as illumination variations, complex backgrounds, and clothing color, which makes the segmentation of human body much difficult in every scene. Especially, the color images won't have depth clues about the motion of a body which has a significant effect on action recognition. Due to the development of Microsoft Kinetic Sensors, the recent HAR studies are able to acquire the RGB-D image with uniform depth and color information and also illumination invariant that makes the recognition system robust to all human actions.

Recently HAR has been directed towards the utilization of depth maps due to their expressive and rich features provision [5]. The key to successful action recognition lies in the action representation such that they can provide discriminative features of each action for classification. Depth cues can provide a greater structural features of an action scene which are more sensitive to skin color, clothes and illumination variations. Even though extensive research has been carried out in this context [6], still there exists some ambiguity for some actions which leads to wrong classification, because, some depth frames include external effects like noises, cluttered scene backgrounds, occlusions, shadow backgrounds and small body shaking movements, etc. A better approach for depth sequence-based action recognition should provide a more clear depth map which is resilient to all these problems.

Regardless of data used for action representation, feature extraction and classification both play a vital role in the HAR. Basically, the handcrafted feature extraction approaches [7, 8] applied Support Vector Machine (SVM) classifier. However, in recent years, Convolutional Neural Networks (CNNs) have gained more popularity, have gained a huge success rate in the classification of images [9]. CNN is a powerful method and it is more effective at feature extraction as well as at classification, it can learn the discriminate features automatically from training data.

In this paper, a new HAR model is developed from depth maps and CNNs. An enhanced version of Depth Motion Map (DMM), called Difference Depth Motion Map (D^{2}MM) is developed in this paper which makes the action representation more resilient to external effects like noises, ghost shadow backgrounds and small body shaking movements. Next, a well-designed CNN is trained especially to extract the features from two actions with a similar structure. The action recognition introduced in this paper involves five convolutional layers, three pooling layers, and one fully connected layer. The experimental results of the proposed recognition model are compared with conventional methods on the publicly available dataset, "MSR Action 3D". The comparative analysis proves that the developed framework outperforms the earlier developed techniques.

Remaining paper is organized as follows; Section II summaries the related work details. After that, the complete details of the proposed HAR model is explained in section III. Further section IV explains the details of experimental results. Finally, the concluding remarks and future directions are represented in section V.

1. Depth based Action recognition

With the advancement in the development of low-cost depth sensors and due to the capability of fetching the depth data, the researches have gained an increased research interest over depth images and diverted from RGB images. Several action recognition models are constructed by considering the depth data from action videos. This section outlines several models based on two aspects: handcrafted and deep learning approaches. A detailed and summarized surey over these methods is described in [10-12].

A. Handcrafted methods:

In earlier, several methods are proposed by considering the depth action sequences as an action representators. Initially, Yang et al. [13] projected the depth maps onto three orthogonal planes and accumulate the global activity through the entire video sequence to obtain the DMM. Then "Histogram of Oriented Gradients (HOGs)" are generated through these DMMs for action representation. Next Wang et al. [14] focused on the characterization of human motion and Human to object interactions and extracts a novel feature set that is suitable to depth data which are robust to temporal and translational misalignments, and noise variations. Further, to capture the intra-class variability, an action let ensemble models learned. However, these approaches are failed to acquire complex motion sahpes of joints at pixel-level.

To overcome this problem, O Omar et al. [7] developed a new descriptor through the histograms which can capture the features like spatial coordinates, depth, and time in 4D space, distributed over the orientations of different surface normals. To obtain this histogram, the 4D projectors are created at first for quantization of 4D space and then focused over the representation of possible 4D normal directions. Next, a "Super Normal Vector (SNV)" [8] is constructed through the low-level polynomials, obtained through the grouping of huper surface normal in depth sequence. Finally these SNVs are fed to the action recognition model. A. W. Vieira et al. [15] proposed a new descriptor called "Space-Time Occupancy Pattern (STOP)". In STOP, the depth action sequences are described through 4D grids and it is achieved through the segmentation of space-time axis. This approach preserves both the temporal and spatial contextual data which is being flexible to accommodate the intra-class variations. To increase the robustness, another 4D descriptor, called Random Occupancy Pattern (ROP) proposed by J. Wang et al. [16] which deals with occlusions and noise combined with sparse coding methods.

With an inspiration of the success of Spatio-Temporal Interest Points (STIPs) in RGB videos based on HAR, Xia L and Agarwal J [17] proposed a new filtering method called DSTIP to extract the STIPs through which the noise is suppressed. Further, a novel feature called, "Depth Cuboid Similarity Feature (DCSF)" to represent the depth of the local 3D cuboid located around the DSTIPs with an appropriate size. Further, Lu et al. [18] modeled a new features set based on the binary range samples that can remove the complex occlusions and cluttered background to acquire the motion and body shape in the depth image sequences. Next, Chen at al. [19] used "Local Binary Patterns (LBPs)" to acquire a compact feature representation from Depth Motion Maps at multiple views such as front, side and upper. Once the features are extracted all three LPBs are merged and formed into a composite feature vector and then fed to a classifier. At the classification level, a soft-decision assisted fusion rule is applied to merge the outputs of classifiers.

Action recognition from multiple views has gained an appreciable performance when compared to the action recognition based on a single view. In the method proposed by D. Kim et al. [20], initially the front view depth action image is processed for side view generation. Next, the both side and front views of action iamge are into two descriptors, namely "Depth Motion Appearance (DMA)" and "Depth Motion History (DMH)". SVM classifier is accomplished here for action classification. Another approach based on multi-view projection is proposed by Chen at al. [21] in which the initial projection of the original depth actionimage is accomplished over three Cartesian planes. Next, the DMM is obtained by the accumulation of absolute difference between two successive depth maps. Depth maps can also be generating at the segment level in which the depth action sequnces are fragmented into a set of several segments with overlapping. Based on this concept, a new method is developed by Chen et al. [22] in which the depth action sequence is segmented into temporally overlapping segments and then they are used to generate DMMs. Next, the DMMs are LBPs are exploited over the DMMs after the partition into dense patched to characterize the local texture information which has rotation invariance. Finally, the patch descriptors are encoded through fisher kernel and then fed to a kernel-based "Extreme Learning Machine (ELM)" algorithm for classification.

B. Deep Learning Methods

CNN [23] is the most prevailing mechanism for extracting the features and also for classification. Recent HAR methods focused on CNN for classifying action rather than the machine learning algorithms like SVM, Decision tree etc. A summarized advance of CNN in terms of weight initializations, loss functions, activation functions, optimization, and regularization is described in [24]. CNN's use of trainable filters and neighborhood polling process to attain a feature hierarchy of complex actions. CNN can extract the features those are invariant to lighting, pose, and neighboring clutter [25]. Moreover, the CNN's has an excellent performance in the recognition if they are trained through a perfect regularization [26, 27]. Several authors are focused over the development of deep learning assisited HAR system [28].

So many HAR [29] approaches are proposed based on CNN, some applied CNN to extract the features and some applied for classification and some more approaches utilized CNN for both. L Cai et al. [30] developed a new method for HAR based on the DMM and improved CNN. In this approach, initially the depth motion maps are projected into the front, side and top view and then features are extracted. Based on this aspect, an enhanced CNN model is developed to realize the HAR. Next, Xu Ran Zhao et al. [31] introduced a novel "Spatio-Temporal Conditional Random Field (STCRF)" [32] into Deep CNN (DCNN) to apply the temporal uniformity among the depth maps of successive frames of action video. A temporal consistent superpixel (TSP) is initialized to establish a correspondence target in successive frames. Then the DCNN is applied to relapse the depth value of each TSP monitored by STCRF to construct the relationship of estimated depths.

III. PROPOSED APPROACH

1. Overview

The framework of the proposed Human action recognition method is depicted in figure 1. In this method, the depth maps are used for human action representation and CNNS are used for feature extraction and classification. An enhanced version of Depth Motion Map (DMM), called as Difference Depth Motion Map (D^2MM) is developed in this work to make the recognition system resilient to extra side effects like noises, background occlusions and shaking body movements. After representing the human action with D^2MM, CNN is applied for feature extraction followed by classification. Due to the extraordinary benefits of CNN in extracting the features from an image, no other mechanism is deployed for feature extraction. CNN extracts more significant and discriminative features form an image, hence this work not focused on feature extraction phase separately. The details of action representation through D^2MM and CNN are explored in the subsequent subsections.

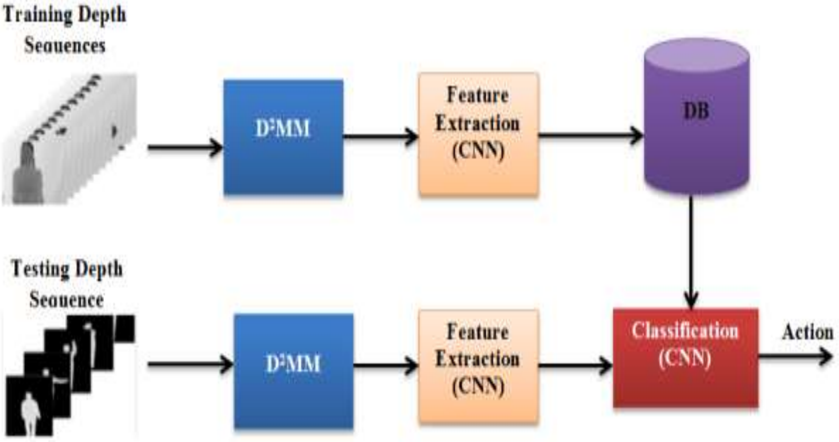


Figure 1 : Block diagram of the proposed recognition model

2. D^2MM

DMM is initially introduced by Yang et al. [13], which exploits the motion information from a depth sequence. DMM gives a visual perception of human activity and it is generated by the accumulation of motion energy throughout the entire depth sequence. Depth maps contain additional depth coordinates along with Cartesian coordinates which are generally provided by color images. Due to the presence of additional depth coordinates, depth maps are more informative than the normal color images. DMM based action representation transforms the action recognition issue from 3D to 2D and then applies for HAR. Particularly, the DMMs are constructed by the accumulation of energy after projecting the depth frames over an orthogonal Cartesian plane. Basically, the main intention of DMM is to signify the shape and motion of an action. The DMMs obtained at various levels are represented in figure 2.

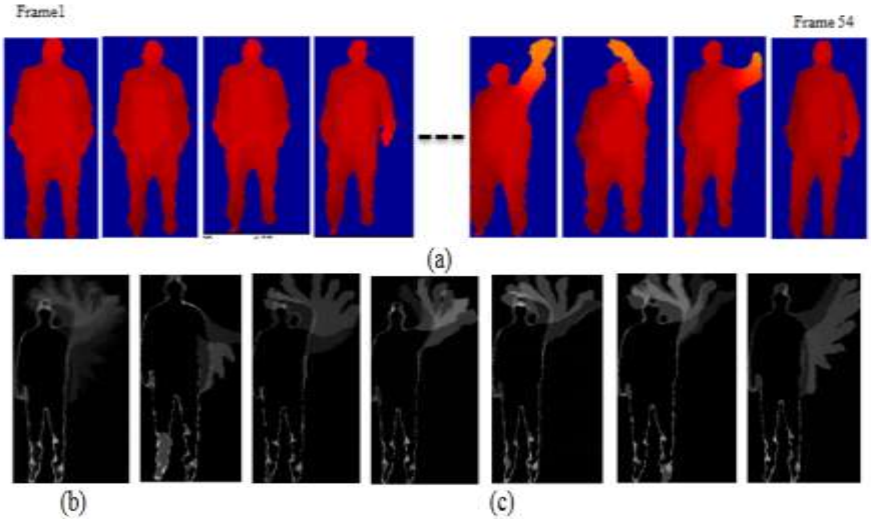


Figure 2 : DMMs representation (a) Depth action Sequence, High Wave, DMMs obtained through (b) 54 frames, (c) 1-9, 10-18, 19-27, 28-36, 37-45, and 46-54 subset of frames

Figure 2a shows a depth action sequence having high wave action, figure 2b shows the DMM obtained after considering the entire action sequence and figure 2c shows the different DMMs obtained after considering six different subsets of depth frames (e.g. frames 1-9, 10-18, 19-27, 28-36, 37-45, and 46-54 subset of frames) in the same action sequence. From figure 2c, we can observe that the defined action of hand waving is clearer than the DMM shown in figure 2b, i.e., the main shape and motion of action part is clearly visible in DMMs generated through subsets than the DMMs generated through the complete set of depth frames. Hence, the DMMs are generally generated through only the subsets of frames.

The concept of DMM considered by Yang et al. [13] is based on the threshold difference due to which there is a probability of information loss. Unlike the threshold-based DMM generation, Chen Liu et al. [21] evaluated the DMM based on the accumulation of motion energy of the absolute difference between consecutive frames. This method preserves the motion information more effectively and hence our method also considered it as a base for DMM evaluation. According to Chen Liu et al. [21], the DMM evaluation is formulated as follows;

$$DMM = \sum_{t=0}^{N-2} |D(i,j,t) - D(i,j,t-1)| \qquad (1)$$

Where D(i,j,t) is a pixel value at the position (i,j) of a depth frame at the instant of *t* and D(i,j,t-1) is a pixel value at the position (i,j) of a depth frame at the instant of *t-1*, where *t* varies from 0 to N-1.

DMM can acquire the shape and motion cues of a depth action image more effectively, results in a discriminative map that gives more discrimination between different actions. This discrimination is provided through the spatial distribution of energy. However, due to unstoppable reflection of depth cameras and low-resolution cameras, the depth data acquired in the depth sequence come with unnoticeable noises, resulting in a video with less quality and undefined energy regions in DMM.

Furthermore, there exist some regions which won't exist in some frames and exists in some other frames. Ghost shadows are better examples of this type of region which can appear in only some frames. These shadows result in undefined depth values. Along with these problems, the insignificant or small shaking movements of body occured in the successive frames may also introduce some narrow and unnecessary moving edges to the DMM. These narrow and unnecessary edges has a severe effect on the body size and there is no much use with these edges. Further, the narrow edges do not provide any significant information which helps in the provision of perfect discrimination between different actions. Several image processing techniques like mathematical morphology [33] and median filtering techniques are applied to overcome the above-specified constraints like noises, ghost shadows and small body shaking movements. However, these techniques result in a loss of original motion information contained in DMM. Hence in this model, we introduced a new version of DMM, called D^{2}MM based on weighted motion score (WMS) to remove the above problems. For this purpose, initially a binary motion image is constructed by comparing the pixels at same positions of different frames, as

$$B_M(i,j,t) = \begin{cases} 1, \; if\, D(i,j,t) \neq D(i,j,t-1) \\ 0, \qquad\qquad\qquad Otherwise \end{cases} \qquad (2)$$

Where B_M is a binary motion image. Next, based on the binary motion image and applying a sliding window over it, the Weighted Motion Score $W_M(i,j,t)$ is evaluated for every pixel, as

$$W_M(i,j,t) = \frac{1}{(w+1)^2} \sum_{i=x-(w/2)}^{x+(w/2)} \sum_{j=y-(w/2)}^{y+(w/2)} B_M(i,j,t) \qquad (3)$$

The size of the sliding window is considered as $(w+1) \times (w+1)$, where w is the height and width of sliding window. In the present paper, we set the height and width of the sliding window as 8, i.e., $w = 8$. Next, based on the obtained WMS, the difference map, $D_M(i,j,t)$ is constructed as follows;

$$D_M(i,j,t) = \begin{cases} |D(i,j,t) - D(i,j,t-1)|, & if\, W_M(i,j,t) > T_M \\ 0, & Otherwise \end{cases} \qquad (4)$$

Where T_M is the threshold of WMS and in our experiments, it is set as 0.6, i.e., $T_M = 0.6$. From difference map, we can notice that the pixels are considered as noisy when the WMS of the respective pixel is lower than the threshold T_M. Finally, based on the obtained difference map, the D^{2}MM is constructed as

$$D^2MM = \sum_{t=0}^{N-2} D_M(i,j,t) \qquad (5)$$

Figure 3 shows the example of a two depth maps of a single action generated through DMM and *D^{2}MM*. From these observations, we can notice that the depth map obtained through *D^{2}MM* has more visual-spatial energy distribution than the depth map obtained through DMM.

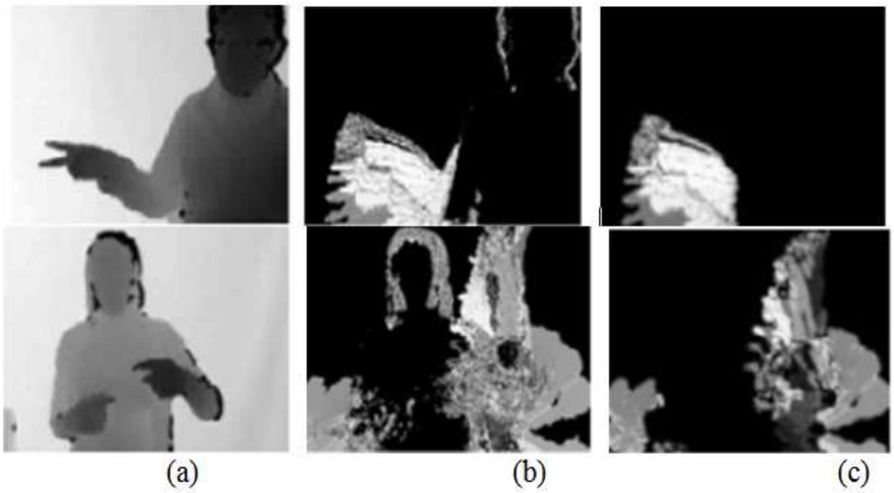


Figure 3 : (a) original depth action sequence, Depth Motion maps obtained through (b) DMM and (c) *D*^{2}*MM*

C. Convolutional Neural Networks Model

With an enormous growth in the development of powerful hardware like “Graphical Processing Units (GPU)” and Deep learning, “Deep Neural Networks (DNNs)”, especially CNNs [34] have gained more and more popularity in human action recognition. The main advantage of the CNNs is their capability to represt the image through rich mid-level features, which is the main drawback of conventional handcrafted representations. Hence this paper accomplished the CNN to extract the required features as well as to classify the action.

Once Human action representation is completed through *D*^{2}*MM*, the next step is feature extraction and using them for recognizing human action. In this model, after representing the human action with *D*^{2}*MM*, then it is resized to 112X112 and processed as input to the proposed new CNN architecture. The CNN architecture is comprised of convolutional layers and pooling layers. Generally, the convolutional layers are accomplished for feature extraction and pooling layers are accomplished for dimensionality reduction. This model is composed of totally five convolutional layers and three max-pooling layers. The architecture of the proposed CNN model is depicted in the following figure 4.

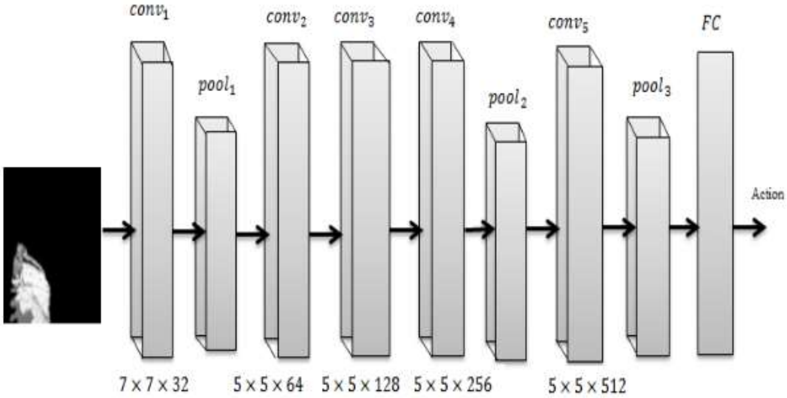
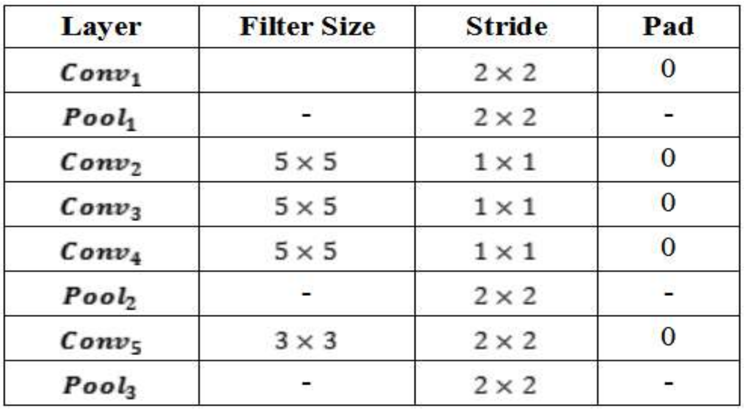


Figure 4 : proposed CNN model for HAR

In the proposed model, the first convolutional layer (*Conv$_1$*) used 32 convolutional filters and the size of each filter is 7X7. The next three convolutional layers such as (*Conv$_2$*), (*Conv$_3$*), and (*Conv$_4$*) used 64, 128, and 256 convolutional filters and the size of each convolutional filter is 5X5. The last convolutional layer (*Conv$_5$*) used 512 convolutional filters and the size of each filter is 3X3. A Rectifier Linear Unit (ReLu) follows every convolutional layer and it is an activation function to increase the non-linearity. Next, the proposed model uses three max-pooling layers. At every max-pooling layer, the stride is considered as 2. Finally, a Fully Connected Layer (FCL) with the size equal to the total number of actions is considered and it is the final stage of feature extraction. During the training process, a "Multi-Nominal Logistic Loss Function" is accomplished along with a "Gradient Descent" method to update the weights. The following table.1 shows the details of parameters and their values considered at convolutional layers and pooling layers.



Layer	Filter Size	Stride	Pad
$Conv_1$		2×2	0
$Pool_1$	-	2×2	-
$Conv_2$	5×5	1×1	0
$Conv_3$	5×5	1×1	0
$Conv_4$	5×5	1×1	0
$Pool_2$	-	2×2	-
$Conv_5$	3×3	2×2	0
$Pool_3$	-	2×2	-

Table 1 : CNN structures

As explained above, this model considered the filter size as 7X7 at convolutional layer 1 and it was gradually decreased to 3X3. The textures extracted through depth maps of two different action images make it difficult to obtain more discriminative and distinct features when a filter with a small size is applied over the images. Let's consider two different action images. If we apply a filter of size 3X3 over them, the involved outputs may not differ much from each other. Because at the very beginning the two images may have similar features, which is the main reason behind the consideration of a 7X7 sized filter at the first convolutional layer, *Conv$_1$*. Generally, the last layer of CNN, i.e., fully connected layer ends with 1 or 2 fully connected layers before the last phase of the classification layer. However, based on the experiments conducted at training phase, we discovered that one FCL is enough. The final FCL is applied over the output of the third max-pooling layer and it can preserve all significant features and also can generate better classification results. Further, at the testing, the soft-max regression is accomplished to obtain a score for every action class depends on the trained weights in the training phase. Further, the class label is determined based on the higher score obtained at the soft-max regression layer.

IV. SIMULATION EXPERIMENTS

In this section, we comprehensively evaluated our developed action recognition system on a publicly available benchmark data set, MSR Action 3D dataset [35]. We employ MATLAB software for the simulation of the proposed approach. This data set provides depth map data that are suitable to construct the difference depth motion maps. In this section initially, the details of the “MSR Action 3D dataset” are explored. Next, the obtained simulation results are described. Finally, a comparative analysis is described between the proposed and conventional approaches through the obtained results.

A. MSR Action 3D dataset

The “Microsoft Research (MSR) action 3D dataset” is an action dataset constructed through 20 different actions such as forward kick, side kick, jogging, throw, pick up, tennis serve, golf-swing, high arm wave, hammer, draw X, forward punch, draw circle, two hand wave, draw tick, side boxing, bend, hand clap, high throw, hand catch, horizontal arm wave. Every action is performed by ten actors and totally three times. This is a very challenging dataset which has a lot of speed variations and also the actions have more similarities. A single viewing point is only used at which the actions are faced with a frontal view with the camera while capturing. Similar to [14], the subjects with odd numbers such as 1, 3, 5, 7, and 9 are used for training and the subjects with even numbers such as 2, 4, 6, 8, and 10 are used for testing. Some depth action sequences of this dataset are shown in figure 5.

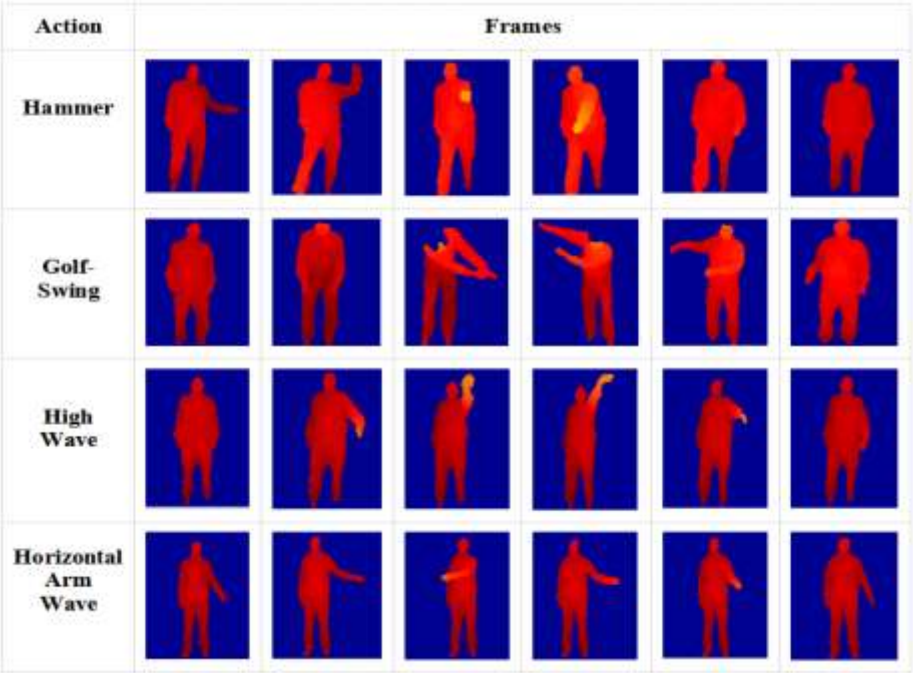


Figure 5 : some action samples of MSR Action 3D dataset

B. Results

In this paper, several performance metrics such as “True Positive Rate (TPR) Or recall, Precision or Positive Predictive Value (PPV), F-Score, False Negative Rate (FNR), False Discovery Rate (FDR)” and “Accuracy” are considered to evaluate the performance of proposed approach. After testing different actions, the obtained results are formulated into a confusion matrix. Depends on the obtained classification results, “True Positives (TPs), True Negatives (TNs), False Positives (FPs), and False Negatives (FNs)” are measured. Based on the obtained TP, TN, FP and FN values from the confusion matrix, accuracy is evaluated and the respective mathematical representation is given as;

$$TPR = \frac{TP}{TP+FN} \quad (6)$$

$$PPV = \frac{TP}{TP+FP} \quad (7)$$

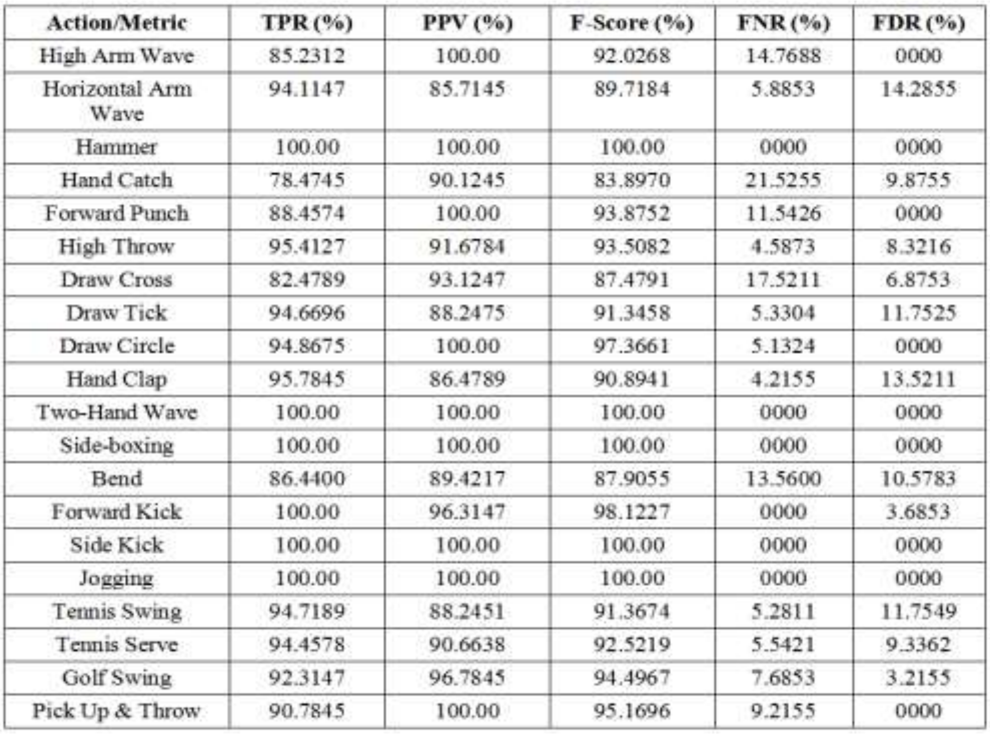
$$F - Score = \frac{2 \times TPR \times PPV}{TPR+PPV} \quad (8)$$

$$FNR = \frac{FN}{TP+FN} \quad (9)$$

$$FDR = \frac{FP}{TP+FP} \quad (10)$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (11)$$

Here totally 20 actions are considered for every actor. After the simulation of different action sequences through the proposed approach, the obtained results are represented as below.



Action/Metric	TPR (%)	PPV (%)	F-Score (%)	FNR (%)	FDR (%)
High Arm Wave	85.2312	100.00	92.0268	14.7688	0000
Horizontal Arm Wave	94.1147	85.7145	89.7184	5.8853	14.2855
Hammer	100.00	100.00	100.00	0000	0000
Hand Catch	78.4745	90.1245	83.8970	21.5255	9.8755
Forward Punch	88.4574	100.00	93.8752	11.5426	0000
High Throw	95.4127	91.6784	93.5082	4.5873	8.3216
Draw Cross	82.4789	93.1247	87.4791	17.5211	6.8753
Draw Tick	94.6696	88.2475	91.3458	5.3304	11.7525
Draw Circle	94.8675	100.00	97.3661	5.1324	0000
Hand Clap	95.7845	86.4789	90.8941	4.2155	13.5211
Two-Hand Wave	100.00	100.00	100.00	0000	0000
Side-boxing	100.00	100.00	100.00	0000	0000
Bend	86.4400	89.4217	87.9055	13.5600	10.5783
Forward Kick	100.00	96.3147	98.1227	0000	3.6853
Side Kick	100.00	100.00	100.00	0000	0000
Jogging	100.00	100.00	100.00	0000	0000
Tennis Swing	94.7189	88.2451	91.3674	5.2811	11.7549
Tennis Serve	94.4578	90.6638	92.5219	5.5421	9.3362
Golf Swing	92.3147	96.7845	94.4967	7.6853	3.2155
Pick Up & Throw	90.7845	100.00	95.1696	9.2155	0000

Table 2 : Performance Metrics for different actions

Table 2 shows the details of performance metrics evaluated after testing all types of actions. The first measure, TPR measures the total number of true positives for a given total number of inputs. For example, let's consider the High Arm wave action, the TPR is measured as the ratio of total number of High Arm Wave action frames detected as walking to the total number of High Arm Wave action frames given as input for testing process. In this case, the TP is the total number of High Arm Wave action frames classified correctly and FN is the total number of High Arm Wave action frames classified incorrectly. This process is applied for all the remaining actions also and the respective TPRs are measured. From table 2 it can be noticed that the maximum TPR (100%) is obtained for a total of six actions and they are Hammer, jogging, side kick, forwardkick, side-bxing, and Two-hand Wave. The minimum TPR is obtained for Hand Catch (78.4745%).

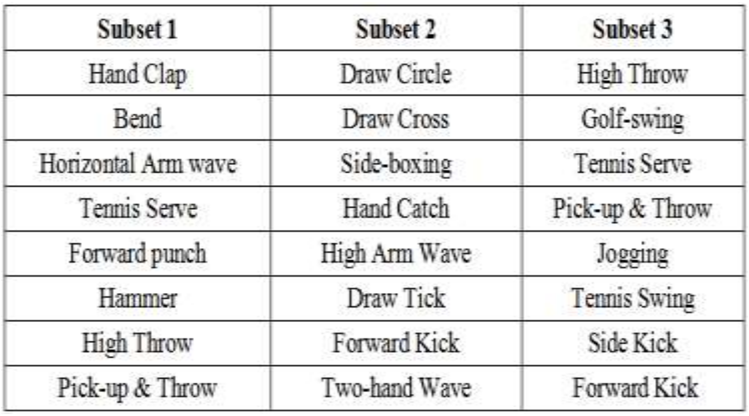
Next, the precision is measured as the ratio of TPs to the sum of TP and FP. In the above example of High Arm Wave action as input, the TP is the total High Arm wave action frames count those are classified correctly and FP is the total action frames count those are classified as High Arm wave when the input is not a High Arm wave action. In this case, the input is not required action but the output is required action. Next, the maximum PPV (100%) is obtained for a total of eight actions and they are two-hand wave, draw circle, forward punch, hammer, High Arm Wave, Side-boxing, Side Kick, and Pickup& Throw. The minimum PPV is obtained for Horizontal Arm Wave (85.7145%).

Next, the F-measure a simple harmonic mean of TPR and PPV. A higher value of F-score depicts the better performance and in this simulation, the maximum F-score (100%) is achieved for a total of five actions and they are Hammer, jogging, side kick, side boxing, and Two-Hand Wave. Further, the minimum F-score is obtained for Hand Catch (83.8970%).

The FNR is a measure which measures the total number of false negatives that are falsely classified. For example, if we had given a Hand catch action as an input and the system had shown it as Hand Clap, then it is counted as False Negative. The FNR is obtained by the accumulation of the total number of such False Negatives. A minimum value of FNR denotes better performance and during this simulation, we got minimum FNR (0%) for a total of six actions and they are Hammer, Jogging, side kick, forward kick, side boxing, and Two-hand Wave,. Further, the maximum FNR is obtained for an action Hand Catch (21.5255%).

Finally, the FDR is a measure which measures the total number of false positives that are falsely discovered. For a given input actual action, if the system had shown another action, then it is counted as False Positive, i.e., the total number of instances the action is falsely discovered. For example, if we had given a Hand catch action as an input and the system had shown it as Hand Clap, then it is counted as False Positives for Handclap action. The FDR is obtained by the accumulation of the total number of such False Positives. A minimum value of FDR denotes better performance and during this simulation, we got minimum FDR (0%) for totally eight actions and they are Two-hand wave, draw circle, forward punch, hammer, High Arm Wave, Side-boxing, Side Kick, and Pickup& Throw. Further, the maximum FDR is obtained for an action Horizontal Arm Wave (14.2855%).

Further, the simulation experiments are accomplished according to yang et al. [13] through three subsets. Under this simulation experiment, totally three types of experiments are conducted by dividing the entire action set into three different subsets. Similar to [14], the subjects with odd numbers such as 1, 3, 5, 7, and 9 are used for training and the subjects with even numbers such as 2, 4, 6, 8, and 10 are used for testing. The three action sets are shown in Table 3. Here the simulation study is accomplished by varying the frame length and pixel shift and the obtained accuracy is shown in figure 6 and figure 7.



Subset 1	Subset 2	Subset 3
Hand Clap	Draw Circle	High Throw
Bend	Draw Cross	Golf-swing
Horizontal Arm wave	Side-boxing	Tennis Serve
Tennis Serve	Hand Catch	Pick-up & Throw
Forward punch	High Arm Wave	Jogging
Hammer	Draw Tick	Tennis Swing
High Throw	Forward Kick	Side Kick
Pick-up & Throw	Two-hand Wave	Forward Kick

Table 3 : Subsets of MSR Action 3D dataset

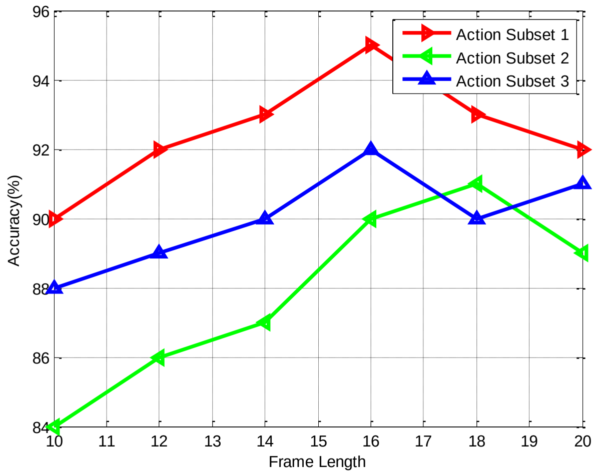


Figure 6 : Recognition Accuracy with varying frame length

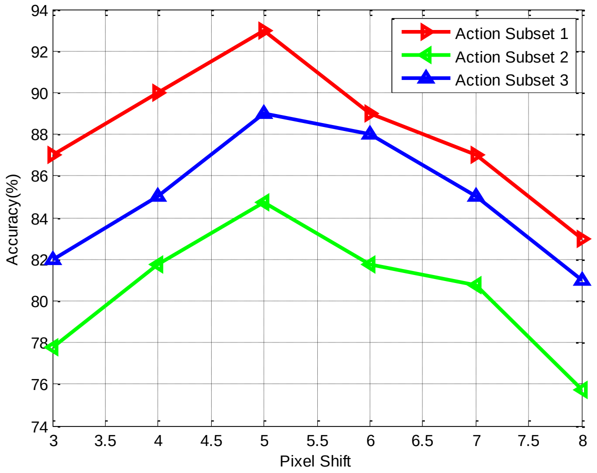


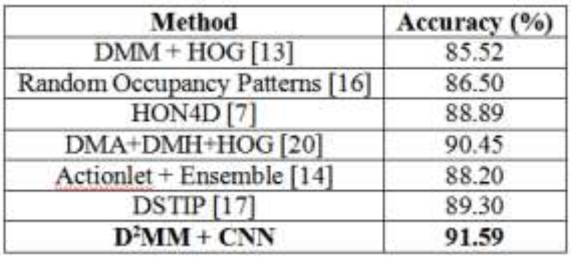
Figure 7 : Recognition accuracy with varying pixel shift

Figure 6 shows the details of recognition accuracies with varying frame length. From this figure, we can notice that as the frame length increases, the recognition accuracy also increases and this increment is only up to a particular level. As can be seen, the accuracy is gradually increased up to frame length 16 and then onwards it is decreased. A smaller frame length results in less depth information by which the detection system can't recognize the perfect action present in that frame. Similarly, a larger frame length also results a more confusion to the detection system due to the excessive depth information. For instance, even in a single video, some actions are carried out multiple times and an increased frame length results in more confused depth information. Moreover, the accuracy of the actions considered under action subset 1 has higher values and the actions considered under action subset 2 are having fewer values. The main reason behind this problem is three similar actions in subset 2, they are Draw Tick, Draw Cross, and Draw Circle. These three actions have similar characteristics at initial frames and hence the detection system has gained less accuracy.

Figure 7 shows the details of recognition accuracies with varying pixel shift. From this figure, we can notice that as the pixel shift increases, the recognition accuracy also increases and this increment is only up to a particular level. As can be seen, the accuracy is gradually increased up to a pixel shift of 5 and then onwards it is decreased. Here the pixel shift is accomplished through the extraction of overlapping frames. After 5, the recognition accuracy is decreased because for a greater pixel shift the spatial correlations between pixels become weak and hence the obtained Depth Motion Map won't have effective feature through which the system can discriminate between different actions. Moreover, the obtained accuracy is high for actions under subset 1 and the accuracy is less for actons under subset 2.

C. Comparative analysis

In this section, we discuss the comparison of recognition accuracies between the proposed and conventional approaches such as DMM + HOG [13], Random Occupancy Patterns [16], HON4D [7] and DMA+DMH+HOG [20], Action + Ensemble [14], and DSTIP [17].



Method	Accuracy (%)
DMM + HOG [13]	85.52
Random Occupancy Patterns [16]	86.50
HON4D [7]	88.89
DMA+DMH+HOG [20]	90.45
Actionlet + Ensemble [14]	88.20
DSTIP [17]	89.30
D^{2}MM + CNN	**91.59**

Table 4 : Comparison of recognition accuracy

Table 4 represents the accuracy comparison of proposed and conventional approaches. It can be seen from the table 4, the accuracy of the proposed system is higher compared to the conventional approaches. The most nearby method for the proposed model is DMA+DMH+HOG [20] and it achieved an accuracy of 90.45%. However, in this approach, the DMM evaluation didn't consider the external effects like noises, Ghost Shadows and shaking body movements and hence the obtained DMM is not effective to recognize the action more accurately. By generating a binary motion image before DMM evaluation, the proposed model has gained resilience to all these side effects. Furthermore, the one more conventional approach, i.e., DMM + HOG [13] is a basic approach for

DMM based action recognition and in this method, the DMM is evaluated based on the thresholding by which the obtained DMM won't have significant information which is more helpful in the action recognition. The main drawback of this approach is information loss due to thresholding. Though the DSTIP method suppresses the noises effectively, they can't capture the small body shaking movement sin depth action sequences.

V. CONCLUSION AND FUTURE SCOPE

In this paper, we have presented a Difference based Depth Motion Map which can nullify the external effects present in the Depth Action Video Sequences. The inclusion of difference map helps in the reduction of unnecessary constraints like shadows, small body movements and represents only action movements with high precision. This approach copes up with shape and motion cues for all types of actions. Further, the new CNN model helps in the extraction of more significant and distinct features of every action. The proposed model is extensively simulated over a standard benchmark dataset. The experimental results show that the developed method outperforms the conventional approaches after evaluating it over the standard "MSR Action 3D dataset". Moreover, the additional simulation experiments on the "MSR Action 3D dataset" confirm that the proposed HAR model can handle depth sequences that contain different frames and pixel shifts. With the implementation of five-layered CNN, our method can classify real-time actions also with more accuracy.

In the standard MSR Action 3D dataset, Skelton based action sequences are also available which also represents human actions. The main advantage with Skelton model is a reduced computational burden. Even with fewer features, the Skelton based actions can represent a human action more effectively which helps in the accurate recognition of human action. Hence, the further work of this paper is intended to apply the new CNN model with Skelton joint based motion action descriptors.

REFERENCES

[1] Dikmen, M., Ning, H, Lin, D. J, (2008, November). "Surveillance event detection". TRECVID.

[2] R. A. Seger, M. M. Wanderley, & A. L. Koerich, (2014). "Automatic detection of musicians ancillary gestures based on video analysis", Expert Systems with Applications, 41 (4), 2098–2106.

[3] Chi W, Wang J, and M. Q. H. Meng, "A gait recognition method for human following in service robots", IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2017.

[4] A. I. Maqueda, C. R. del Blanco, F. Jaureguizar, & N. Garcia, (2015). "Human-computer interaction based on visual hand-gesture recognition using volumetric spectrograms of local binary patterns". Computer Vision and Image Understanding, 141, 126–137.

[5] A. S. Keceli, and A.B. Can, "Recognition of Basic Human Actions using Depth Information". International Journal on Pattern Recognit. Artificial Intelligence, 2014, 28, 1450004.

[6] S. Escalera, V. Athitsos, and I. Guyon, (2017). "Challenges in multi-modal gesture recognition", In Gesture recognition (pp. 1–60). Cham: Springer.

[7] Oreifej O, and Liu Z. "Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 716– 723.

[8] Yang X, and Tian Y, "Supernormal vector for activity recognition using depth sequences," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 804–811.

[9] Krizhevsky A, Sutskever I, and Hinton G. E., "Imagenet classification with deep convolutional neural networks," in Advances in neural information processing systems, 2012, pp. 1097–1105.

[10] L. Chen, H. Wei, and J. Ferryman, "A survey of human motion analysis using depth imagery". Pattern Recognit Letters, 34: 1995–2006, 2013.

[11] S.Herath, M. Harandi, and F. Porikli, "Going deeper into action recognition: a survey", Image Vision Computing, 60: 4–21, 2017.

[12] P. Wang, W. Li, P. Ogunbona, "RGB-D-based human motion recognition with deep learning: a survey". Computer Vision & Image Understanding, 171: 118–139, 2018.

[13]X. Yang, C. Zhang, and Y. Tian, "Recognizing actions using depth motion maps-based histograms of oriented gradients", In: Proc., of the 20th ACM international conference on Multimedia, Nara, Japan, pp. 1057–1060, 2012.

[14] J. Wang, Z. Liu, and Y. Wu, "Mining action let ensemble for action recognition with depth cameras", In Proc., of IEEE conference on Computer vision and pattern recognition (CVPR), Providence, Rhode Island, USA, pp.1290–1297, 2012.

[15] Vieira A, Nascimento E, Oliveira G, Liu Z, and Campos M, "Stop Space-time occupancy patterns for 3D action recognition from depth map sequences," Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications, pp. 252–259, 2012.

[16] Wang J, Liu Z, Chorowski J, Chen Z, and Wu Y, "Robust 3d action recognition with random occupancy patterns," in Computer vision–ECCV, Springer, pp. 872–885, 2012.

[17] L. Xia and J. Aggarwal, "Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera". In: Proc., of the IEEE conference on computer vision and pattern recognition, Portland, OR, USA, , pp. 2834–2841, 2013.

[18] C. Lu, J. Jia, and C. K. Tang, "Range-sample depth feature for action recognition". In: Proc.,s of the IEEE conference on computer vision and pattern recognition, Columbus, OH, USA, pp. 772–779, 2014.

[19] C. Chen, R. Jafari, and N. Kehtarnavaz, "Action recognition from depth sequences using depth motion maps-based local binary patterns", In: Proc., of IEEE winter conference on applications of computer vision (WACV), Waikoloa, HI, USA, pp. 1092–1099, 2015.

[20] Kim D, Yun W. H, Yoon H. S, and Jaehong H. S, "Action recognition with depth maps using hog descriptors of multi-view motion," in proc., of 8th International Conference on Mobile Ubiquitous Computing, Systems, Services, and Technologies, UBICOMM, pp. 2308–4278, 2014.

[21] C. Chen, K. Liu, and N. Kehtarnavaz, "Real-time human action recognition based on depth motion maps". Journal of Real-time Image Processing, 12 (1), 155–163, 2016.

[22] C. Chen, M. Liu, B. Zhang, J. Han, J. Jiang, and H. Liu, "3D action recognition using multi-temporal depth motion maps and fisher vector", IJCAI, pp. 3331–3337, 2016.

[23] Koushik J, "Understanding convolutional neural networks," arXiv preprint ar X iv:1605.09081, 2016.

[24] Gu J,Wang Z, Kuen J, Ma L, Shahroudy A, Shuai B, Liu T, Wang X, and Wang G, "Recent advances in convolutional neural networks," arXiv preprint /; ar X iv:1512.07108, 2015.

[25] S. Ji, W. Xu, M. Yang, K. Yu,"3D convolutional neural networks for human action recognition". IEEE Trans. Pattern Analysis Machine Intelligence, 35, 221–231, 2013.

[26] H. Wu, and X. Gu, "Max-pooling dropout for regularization of convolutional neural networks", In Proc., of the International Conference on Neural Information Processing, Istanbul, Turkey, 9–12, Springer, pp. 46–54, 2015.

[27] K. Yu, W. Xu, and Y. Gong, "Deep Learning with Kernel Regularization for Visual Recognition", In Advances in Neural Information Processing Systems; Neural Information Processing Systems Foundation, San Diego, CA, USA, pp. 1889–1896, 2009.

[28] Y. Han, P. Zhang, T. Zhuo, W. Huang, and Y. Zhang, "Going deeper with two-stream ConvNets for action recognition in video surveillance". Pattern Recognition Letters, 107, 83–90, 2018.

[29] W. Rawat, and Z. Wang, "Deep convolutional neural networks for Image Classification: A Comprehensive Review", Neural Computing, 29, 2352–2449, 2017.

[30] C. Linqin, L. Xiaolin, F. Chen, M. Xiang, "Robust Human Action recognition based on Depth Motion Maps and improved Convolutional Neural Networks", Journal of Electronic Imaging, 27(5), 2018.

[31] X. R. Zhao, X. Wang, and Q. C. Chen, "Temporally Consistent Depth Map Prediction Using Deep Convolutional Neural Network and Spatial-Temporal Conditional Random Field", JCST, 32(3): 443–456, 2017.

[32] P. M. A. Diaz, and R. Feitosa, "Spatio-temporal Conditional Random Fields for recognition of sub-tropical crop types from multi-temporal images", XVIII SBSR conference, SP, Brazil, 2017.

[33] Z. Zhang, S. Wei, Y.Song, &Y. Zhang, "Gesture recognition using enhanced depth motion map and static pose map". In Proc., of IEEE international conference on Automatic face & gesture recognition, pp. 238–244, 2017.

[34] Cun L, Bottou Y, Bengio L, & P. Haffner, "Gradient-based learning applied to document recognition". In Proceedings of the IEEE, 86 (11), 2278–2324, 1998.

[35] W. Li, Z. Zhang, Z. Liu, "Action recognition based on a bag of 3D points". In Proc., of IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), San Francisco, CA, USA, pp. 9–14, 2010.