Chapter 6 COMBINING CNN STREAMS OF DYNAMIC IMAGE AND DEPTH DATA FOR ACTION RECOGNITION

6.1. Introduction

RGB-D sensors have been in great demand due to its capability of producing large amount of multimodal data like RGB images and depth maps, useful for better training of deep learning models. In this chapter, a deep learning model for recognizing human activities in a video sequence by combining multiple CNN streams has been proposed. The proposed work comprises the use of dynamic images generated from RGB images and depth motion maps for three different dimensions. The proposed model is trained using these four streams on VGG Net for action recognition purpose. Further, it is evaluated and compared with the other state-of-the-art methods available in the literature, on three challenging datasets namely MSR Daily Activity [290], UTD MHAD [291] and CAD-60 [292] datasets, in terms of accuracy, error, recall, specificity, precision and f-score. From obtained results, it has been observed that the proposed method outperforms other methods.

Human Activity Recognition in a real environment is a challenging task. In spite of lots of success in this field, computer vision researchers are still working to achieve newer techniques of activity recognition. The ease of availability of Microsoft Kinetic Sensors and with automatic learning capacity of a deep neural network such as Convolutional Neural Network (CNN) [337], there is a great peak in this field. Most of the previous works focus on Handcrafted feature extraction based approaches from RGB frames. However, by the introduction of Kinetic Sensors, different cues such as RGB, Depth Motion Maps and skeletal data are used for extraction of features from videos. Great

Success of Convolutional neural network over the images [338], [339] to videos and automatic learning capability has ignited the researchers to use deep neural networks.

In the recent past, many researchers have been exploring the field of Human Activity recognition, and still, experiments are going on. Most of the previously proposed methods for action recognition based on depth data uses some specific handcrafted feature descriptors. Some of the methods using depth modality has been presented in [290], [313], [316], [126]. Various advantages of training depth motion maps (DMMs) on dedicated CNN is discussed in [340], [341]. DMMs are formed by using raw depth frames consisting of motion representation of images, similar to MHI and optical flow images, which are constructed from RGB frames.

6.2. The Proposed Method

The main idea behind the proposed work is to fully utilize the data generated from RGB-D sensors in a productive way. In our work, we fused convolutional neural networks trained separately on dynamic image network and depth motion maps for three different views i.e. front, side and top. As we know RGB images are used previously in training of CNN's but due to more training time dynamic images came into scenario. Dynamic Images are in great demand as they contain the information of full video in single image. They contain all the actions and the motion of the video in single image. Dynamic images [77] are constructed using rank pooling methodology and act as the first stream for the proposed model. Also, Different views of depth motion maps [341] are trained for individual CNN to act as second, third and fourth stream for proposed model. The following subsections discuss the different streams in a detailed manner i.e. dynamic images and depth image in three different dimensions.

6.2.1. Construction of Dynamic Images

The proposed four-stream model works on the principle of utilizing most of the data generated by RGB-D sensors because of the cheaper availability of the sensors. Most of the previous works focus on using the different streams individually or in the group such as RGB+ Depth data, RGB+ Skeletal data or in single stream. Inspired by the success of two-stream models this four-stream model is proposed. As we know that RGB data contains the data in the form of three channels Red, Green and Blue making these images more in size. So, to overcome this problem the newer concept of dynamic images, proposed by Basura et al. [57] is used, which focuses on generation of single dynamic image per video containing all the temporal information over the time. These generated dynamic images are trained on pre-trained network resulting in preparation of first stream of proposed model. This subsection includes brief introduction about generation of dynamic images from the video sequence. In discussed approach, there is temporal video evolution over the full video and it is assumed that video consists of frames listed as $F_{1,}$…. F_T and feature vector $\psi(F_t) \in \mathbb{R}^d$ is extracted for each individual frame F_t of the video. Let us calculate time average of the extracted features by using equation

$V_t = \frac{1}{t}\sum_{T=1}^{t} \boldsymbol{\psi}(F_T)$, here V_t represents the average of the above-stated features up to time t. After that ranking function associated to time t is used to calculate score as indicated below: -

$$S(t|d) = < d, V_t > \text{ where } d \in \mathbb{R}^d \quad (6.1)$$

$$q > t \Rightarrow S(q|d) > S(t|d) \quad (6.2)$$

$$d^* = \rho(F_1 \ldots., F_T; \psi) = argmin\, E(d) \quad (6.3)$$

Where E(d) is represented as below in equation 4

$$E(d) = \frac{\lambda}{2}\|d\|^2 + \frac{2}{T(T-1)} \times \sum_{q>t} \max\{0, 1 - S(q|d) + S(t|d)\} \quad (6.4)$$

$S(q|d) > S(t|d) + 1.$

Here, $\rho(F_1 \ldots, F_T; \psi)$ is written as optimizer to equation 6.1, which further maps a sequence of T video frames to a single vector $\mathbf{d}^*$. The constructed dynamic image by using above stated equations act as the first input for the proposed model.

6.2.2. Depth Motion Maps

Depth Motion Map [332] consists of 3D structure and motion information. From these Depth Motion maps Pichao et al.[342] introduced 3D Depth Motion maps. In this methodology, the depth data is projected on three different orthogonal planes after rotating them using 3D points cloud as discussed in [342]. The absolute difference from these orthogonal planes is used to make three different 2-dimensional depth motion maps namely Top DMM, Front DMM and Side DMM. The DMM's are generated using the following mathematical notation

$$DMM_v = \sum_{i=a}^{b} | map_v^i - map_v^{i-1} | \quad (6.5)$$

In above equation 6.5, variable i represents the index for a particular frame, map_v^i represents the projection corresponding to the i^{th} frame in view *v,* i.e. front, top or bottom and b is the limit from starting frame to last frame. Our main aim is to fully utilize as much of data generated from RGB-D sensor. Depth sequences are independent of lighting conditions and therefore no occlusions occur. In the proposed approach the generated depth motion maps for front, top and side view are trained individually on Pre-trained VGG model and acts as second, third and fourth modality.

6.2.3. Training the Model

Dynamic Images constructed using RGB frames and Motion maps of front, top and side views are used to train the pre-trained VGG-F model. VGG-F and VGG-16 [343], [344] are well-known pre-trained models available publically. As discussed in paper [343], VGG-F model consists of 5 different convolutional layers, 3 pooling layer and 3 fully

connected (FC) layers. SoftMax layer is embedded at the end. In our computation the SoftMax layer at end is interchanged with dropout layer and the third fully connected layer is removed. Dropout Layer consists of fully connected layer having 4096 neurons also ratio between last two fully connected layers is 0.7 to avoid overfitting. Dynamic Images, Side DMM, Front DMM and Top DMM are individually trained using VGG-F pre-trained model. Input to this model is matched by adjusting the size of generated dynamic images and three different DMM's to dimension of 224 x 224. Input images first get convolved through 64 kernels of first convolution layer having size of 11 x 11 x 3 with stride dimension of 4 pixels. Output filtered from first layer is convolved with second convolutional layer having 256 kernels of size 5 x 5 x 64 with stride dimension of 1 pixel. Similarly, it goes till five convolutional layers and output yielded from last convolutional layer gets feed to fully connected layer number 1 passing through ReLU activation function and pooling layer. 4096 neurons are contained by both First and second fully connected layers. Also, Dropout and Relu Layers are used to connect second fully connected and Third fully connected to their previous layers. Moreover, Pseudo coloring technique is used for increasing the number of training samples in case of dynamic and depth motion map images. The output from the proposed model is fused to get class score for particular activity.

The model combines the dynamic image and depth images as shown in Figure 6.1 which are individually trained on pre-trained VGG model and at last, the scores are fused to get the actual output.

Figure 6.1 Four-Stream Proposed Model for Recognition of Actions

6.2.4. Algorithm for Four-Stream Fusion Model

Input: Video Dataset with different activities.

Output: Recognized class of activities

1. Start
2. Prepare the input.

2.1. Take each video from the dataset and prepare dynamic images corresponding to each action sequence by using equation. 6.4.

2.2. Take each video from the dataset and prepare three different 2-dimensional depth motion maps namely Top DMM, Front DMM and Side DMM by using equation 6.5.

3. Train the pre-trained VGG Model for individual input streams (Dynamic images, DMM-Front, DMM-Side, DMM-Top) with network architecture discussed in section 6.2.3.
4. For each trained CNN models D-Score, DF-Score, DS-Score and DT-score is obtained for individual input streams.
5. Weighted Product Model (WPM) discussed in [345], is used for fusion of posterior probability scores of all stream in step 4 to effectively decide the class of activities.
6. END

6.3. Experimental Results

This section presents results and discussion of the proposed four-stream model on various state-of-the-art publically available datasets. We have presented the above-stated approach and their effectiveness on three publically available datasets, namely MSR Daily Activity dataset [290], UTD MHAD dataset [291] and CAD-60 dataset [292]. The proposed network is trained using MatConvNet toolkit in Nvidia P5000 GPU having Intel® Xeon® Silver 4110 Processor.

For efficiency analysis, the trained model has been tested on three datasets namely MSR Daily activity 3D Dataset [290], UTD MHAD Dataset [291] and CAD 60 Dataset [292]. As described in section 6.2, the first step is the construction of Dynamic images from RGB images and after that the depth images are used to generate three different views for top, bottom and side. These four streams are trained individually on pre-trained VGG-F model. Lastly, SoftMax classifier is used for classification purpose. For qualitative analysis purpose, proposed method is compared with different well-known methods proposed in literature along with their Accuracy, Recall, Error, F-score, Specificity and Precision. Also, Heatmap based on the proposed four-stream model for three different datasets, namely MSR Daily Activity[290], UTD MHAD [291] and CAD-60 [292] is presented.

For the quantitative analysis of the proposed models, various performance measures, as discussed in section 2.6, have been used. The proposed framework has been compared with different state-of-the-art frameworks in terms of accuracy, error, recall, specificity, precision and f-score. Three state-of-the-art datasets are evaluated using the proposed neural network based model for action recognition namely is evaluated namely MSR Daily Activity 3D [290], UTD-MHAD [291] and CAD-60 [292] datasets.

6.3.1. Experiment for MSR Daily Activity 3D Dataset

In this section, the robustness of the above stated proposed method is validated on MSR Daily Activity 3D dataset which is created by Zicheng Liu [290], during his research period in Microsoft Redmond Lab. Videos for 16 different activities such as call on the phone, drinking, eating, reading book, using laptop, sit down, stand up, playing guitar etc. are captured using Microsoft Kinetic Sensor. Ten different subjects perform actions one in standing position and one in sitting on Sofa. Total of 320 files for each individual channel is recorded resulting in 960 for all three channels. This dataset includes RGB, Depth as well as Skeletal Data but in our computation, we have used RGB and depth data only. The proposed method is evaluated on different datasets for qualitative analysis purpose. Heatmap for all 16 activities of MSR daily activity is generated (Figure 6.2). From the Heatmap it is observed that most of the actions are recognized correctly. Actions such as play the game and play guitar activities are recognized with less score.

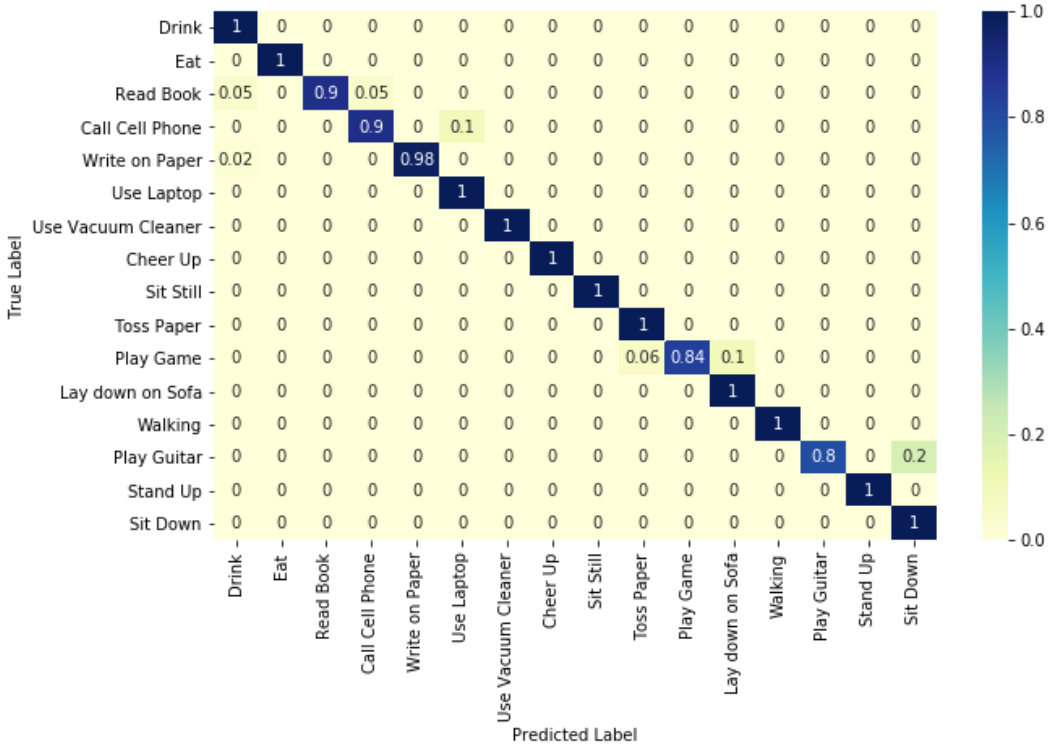

Figure 6.2 Heatmap on MSR daily Activity Dataset [290]

For quantitative analysis, the proposed model is compared with other methods in terms of different parameters such as accuracy, error, recall, specificity, precision and f-score calculated by using performance metrics discussed in Section 2.6 and presented in Table 6.1 listed below. It can be seen from the results that for proposed method values are better as compared to methods proposed by Salah et al.[346], Meng et al. [347], Pichao et al. [342] and Jiang et al. [348].

Table 6.1 Recognition results over MSR Daily Activity Dataset [290]

Method	Accuracy (%)	Error (%)	Recall (%)	Specificity (%)	Precision (%)	F-Score (%)
Salah et al.[346]	92.83	7.17	92.87	99.52	93.47	93.16
Meng et al. [347]	95.62	4.38	95.61	99.71	96.28	95.94
Pichao et al. [342]	81.88	18.12	81.88	98.79	83.96	82.90
Jiang et al. [348]	85.63	14.37	85.63	99.04	86.60	86.11
Proposed	96.38	3.63	96.38	99.76	96.73	96.55

6.3.2. Experiment for UTD MHAD Dataset

UTD MHAD Dataset [291] was captured by Researchers of the University of Texas at Dallas. UTD-Multimodal Human Action Dataset is captured by fusing the use of Microsoft Kinetic Sensor and Wearable Inertial Sensor. 27 different actions such as basketball shoot, two handclaps, lunges, squats, boxing, bowling, drawing triangle, circle etc. are captured in indoor environment by 8 different subjects consisting of 4 male and 4 females. All the different actions are performed twice by each subject resulting in total of 861 sequences. The data captured in UTD MHAD dataset is of four different types RGB, Depth, Skeletal and Inertial data. For analysis purpose, we have generated the Heatmap for all 27 activities of UTD MHAD activity dataset in Figure 6.3. From below Heatmap, it is observed that most of the actions are recognized correctly.

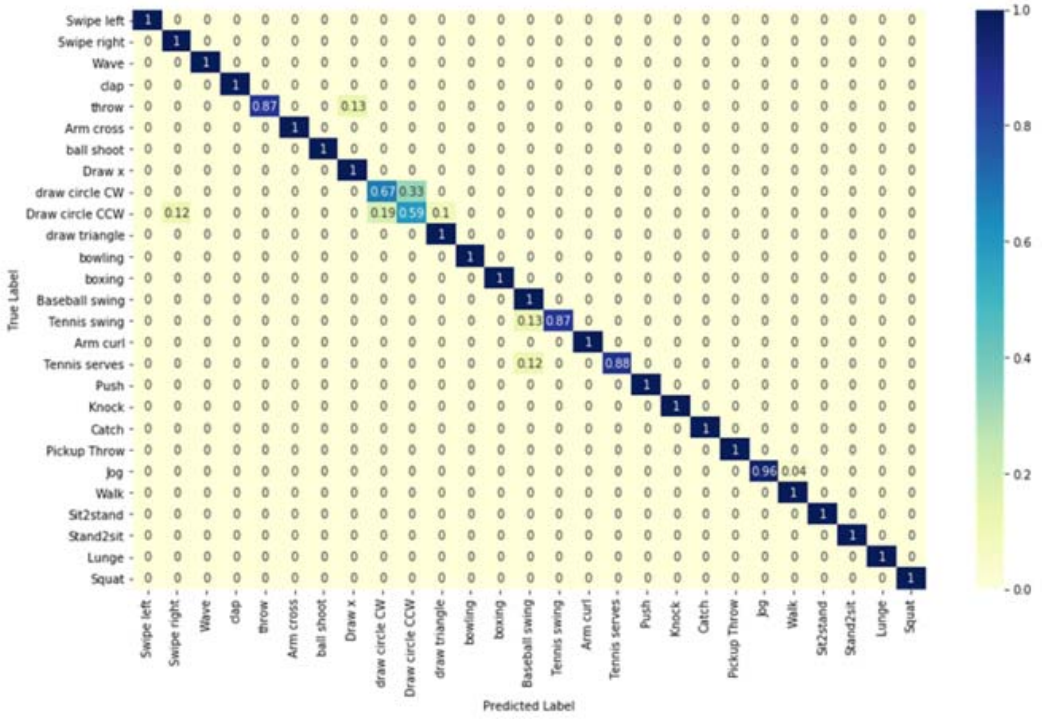

Figure 6.3 Heatmap on UTD MHAD Dataset [291]

For quantitative analysis, the proposed model is compared with other methods in terms different parameters such as accuracy, error, recall, specificity, precision and f-score calculated by using performance metrics discussed in Section 2.6 and presented in Table 6.2 listed below. As seen from below results that for proposed method values are better as compared to methods proposed by Pushpajit et al. [349], Suolan et al.[350], Pichao et al. [342] and Chuankun et al. [351].

Table 6.2 Recognition results over UTD MHAD Dataset [291]

Method	**Accuracy (%)**	**Error (%)**	**Recall (%)**	**Specificity (%)**	**Precision (%)**	**F-Score (%)**
Pushpajit et al. [349]	95.33	4.67	95.33	99.82	95.69	95.50
Suolan et al.[350]	93.47	6.53	94.74	99.75	95.31	95.02
Pichao et al. [342]	85.23	14.77	85.26	99.43	87.98	86.59
Chuankun et al. [351]	88.81	11.19	88.81	99.57	90.01	89.40
Proposed	95.70	4.30	95.70	99.83	95.81	95.75

6.3.3. Experiment for CAD 60 Dataset

CAD 60 dataset [292] is a dataset which consists of RGB-D videos performed by four different subjects. Among, which two are male subjects which are right-handed and two are female subjects in which one is left-handed and one is right-handed. Various daily indoor activities such as relaxing on the couch, talking on the couch, rinsing mouth, brushing teeth, wearing lens, talking on phone, working on computer, chopping vegetable, stirring etc. resulting in 60 videos are recorded. For analysis purpose, we have generated Heatmap for all 12 activities of MSR daily activity in figure 6.4. From below Heatmap, it is observed that most of the actions are recognized correctly. Actions such as talking on phone is recognized with less score.

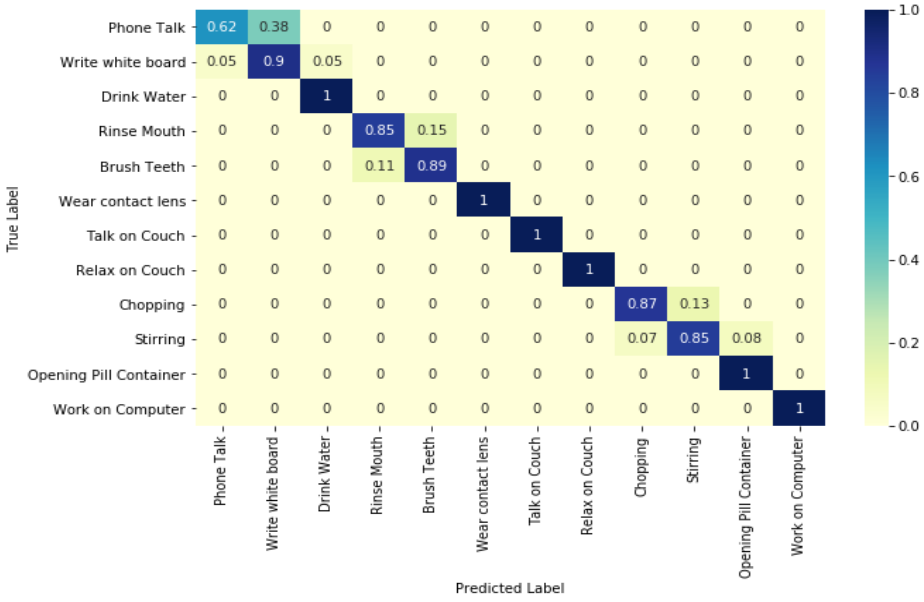

Figure 6.4 Heatmap on CAD 60 Dataset [292]

For quantitative analysis, proposed model is compared with other methods in terms of different parameters such as accuracy, error, recall, specificity, precision and f-score calculated by using performance metrics discussed in Section 2.6 and presented in Table 6.3 listed below. As seen from below results that for proposed method values are better as compared to methods proposed by Meng et al.[347], Jiang et al. [348], Salvatore et al. [352] and Jian et al. [353].

Table 6.3 Recognition results over CAD 60 Dataset [292]

Method	Accuracy (%)	Error (%)	Recall (%)	Specificity (%)	Precision (%)	F-Score (%)
Meng *et al.*[347]	81.48	18.52	83.68	98.58	87.30	85.45
Jiang *et al.* [348]	72.18	27.82	72.50	97.85	77.98	75.14
Salvatore *et al.* [352]	76.67	23.33	76.67	97.88	77.26	76.95
Jian *et al.* [353]	82.57	17.43	82.57	98.66	87.44	84.93
Proposed	94.80	5.20	94.80	98.70	94.82	94.81

6.4. Conclusion

Human Activity Recognition by a combination of different modalities from RGB-D sensor is presented in this article. In this work dynamic images trained on pre-trained VGG-F network and depth motion maps for different views such as top, side and front separately trained on pre-trained VGG-F networks are combined. We have tested the network on most promising datasets such as MSR Daily Activity, UTD MHAD and CAD 60 and achieved state-of-the-art results. Also, we have compared our results for different datasets and found that the proposed method outperforms most of the available methods. Results also indicated that method is able to recognize similar actions like “circle drawing clockwise” from “circle drawing counter-clockwise”,” pushing” from “punching” in a good way. Proposed approach is able to recognize the actions in the indoor environment having stable/still background making this approach to be deployed in indoor surveillance. In future we will combine some another modality to get better results. Further, proposed approach can be extended to Banks, ATMs and Airports, so that security at public places can be enhanced.