HUMAN DETECTION AND ACTIVITY RECOGNITION USING DEEP FUSION NETWORK (DFNeT)

H BAGUM FATHIMA
Research Scholar, Dept. of Computer Science And Engg, United Arab Emirates University, UAE

Abstract - Human activity recognition and identification in entire image sequences have made significant progress in recent years. Detecting and anticipating human activity early in a real-time video, on the other hand, remains a difficult task. Further-more, dynamical events like as lighting changes, camera jitter, and object size variations make them extremely vulnerable. The proposed feature learning algorithms, on the other hand, are less expensive and easier to implement since highly abstract and discriminative features may be generated automatically without the requirement for expert knowledge. In this paper, I propose human detection based on HOG features and a human activity detection system based on Deep Fusion Network (DFNet). The human descriptors for training images are obtained using a histogram of square blocks of fixed size, while the Deep Fusion Network is made up of an InceptionV3 base network and the Multi-Scale Attention Guided Module is used to classify activities. Finally, we put our method to the test on the publicly available KTH dataset, demonstrating its superior performance and accuracy. The suggested model's efficiency in activity recognition on a typical benchmark collective activity dataset demonstrated by experimental findings.

Key Words: DFNet; InceptionV3; HOG.

I.INTRODUCTION

Recognizing human activity and detecting person from a video series or image series is now a subject of research, with the goal of developing a robust system that identifies human action quickly and effectively, which includes video surveil-lance [1], traffic surveillance [2], human computer interaction [3], automotive safety [4], real-time tracking [5], and hu-mans' collective behavior analysis [6], pedestrian detection [7] etc. Standing up, sitting down, walking, and climbing stairs are just a few of an individual's daily responsibilities. In hospital, elderly care, and home automation are all areas where automatic recognition of human activities could be beneficial. Due to their varied appearance and the broad range of stances that they can take, detecting individuals in photos as well as in video frames is a difficult undertaking. The first necessity is a strong feature set which enables the human shape to be readily identified even in crowded contexts with bad illumination. The initial phase in the complete process of these applications is human detection. As a result, in this research, we use HOG features from films to first recognize persons. It records that the edge or gradient structure is so characteristic of the local form with a degree of in variance to locally

adaptable geometric and photo metric amendments. Following the detection of humans, a Deep Fusion

Network (DFNet) was utilized to extract features and characterize human activities. In short, this paper contributions are:

•For human detection, suggested a normalized Histogram of Gradient Orientations feature.
•A Deep Fusion Network based on the InceptionV3 Base Network and the Multi-scale Attention Guided (MAG)prediction network is proposed.
•The MAG Module categorizes activities. Not only is this module used to quickly remove multi scale characteristics, but it may also prioritize more discriminatory maps of functions and dampen maps that do not match the size of the prominent item within the input picture.
•Our methodology is state-of-the-art with KTH challenging data sets

The rest of the paper is organized as follows: We offer 1st overview of the essential approaches for the detection and identification of human activity in Section II. In Section III, the structure of our suggested model is described in-depth. In Section IV, the experimental setup is illustrated, and the findings and analysis are explained, demonstrating the superiority of our suggested model. The final section, Section VI, contains the conclusions.

II. RELATED WORKS

Hou [8] has proposed a quick approach for detecting and classifying humans using HOG and SVM.The suggested technique included three main stages: (a) the video series of moving areas; (b) the extraction of HOG features from moving areas; and (c) the classification of moving regions using SVM. [9] developed the partly occluded human sensing random sub spacious (RSM) method based on HOG and HOH-LBP. The characteristics extracted were subsequently categorized by SVM.Chung [10] suggested a video sequence hybrid method for classifying moving objects. Segmentation has been done via a histogram-based prominence approach and shadow reduction technology has also been applied to boost human categorization efficiency. In order to get the vector of fundamental values, HOG and wavelets were fused with local shape characters and classified eventually with the SVM classification. D Kim [11] presented two new characteristic descriptors: (a) binary HOG and Local Gradient patterns, and (2) local transform characteristics fused in the Ada boost selection method using a mixture of local characteristics, including LBP, LGP and HOG. The feature selection technique presented substantially enhanced human detection performance by playing a key role in the determination of extracted features. Lee [12] proposed a novel method to detect human head and shoulders by the extraction of information about their borders and geometrical characteristics. Statistical techniques to learning have often been used to address difficulties of activity recognition [13]. In order to identify seven distinct motions, including walking, running and leaping, the Gupta P and Dallas T [14] used a KNN/BayeNave classification. Nevertheless, they depended on hand-crafted characteristics and were unable to discover discriminate features that properly differentiated between distinct activities. Symbolic representation [15], raw data statistics [16], and transform coding [17] are popular feature extraction approaches in human activity identification, however they are heuristic and need expert knowledge to create features [18]. In order to extract features from potential picture areas in different resolutions for local and global depictions from outstanding objects some early works have been able to draw on deep learning models to extract high-level features [19]. These methods were not particularly efficient, despite their success because of the utilization of thick layers. Introduced by Mehrdad et al. [20], Multi-level Integrator Module and Multi-scale Attention Guided were utilized with Channel Attention Block and the experimental outcome indicated higher precision and faster operation at real time. The CA block resembles the previously implemented networks of squeeze and excitation (SE) [21].

III. PROPOSED METHOD

We discuss in this part our suggested approach for the detection of human activity. The suggested algorithm’s.

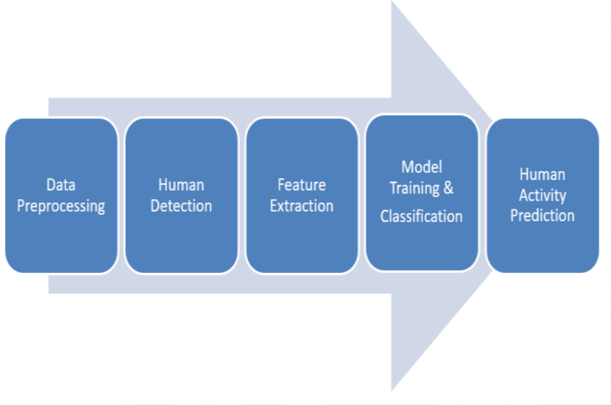


Fig.1: Overall Block diagram

Two key stages have been followed to validate the proposal algorithm: (a) Human detection and (b). Activity recognition, as seen in figure 4, is done in the first phase. Preprocessing is an ordinary name for actions using the lowest observation level input video frame. By the video sequence, the input frame is recorded. Improved frame data that improves certain primary characteristics for further processing is the main objective of preprocessing. HoG is then given for human detection from the preprocessed frames in this study. In terms of overhead processing, the suggested method worked well than various traditional strategies by reducing the number of scanned areas. We have launched in the second phase the two-part Deep Fusion Network; I the Feature Base Network and (ii) the Predictive Network. Figure 4 showed the overall schematic diagram. For the base network, the initial V3 structure is utilized. After the training for each base net is complete, coevolutionary kernels are preserved, except for those forming completely linked layers. We get several dynamically weighted characteristics out of a pre-training network on various abstraction levels by using the MAG modules in our Feature Prediction

Network. Selected activities such as boxing, walking, and hand waving were picked from KTH data sets,[22] consisting of 599 video sequences of six classes of human movement, each of which is carried out in Figure 2 by 25 people.

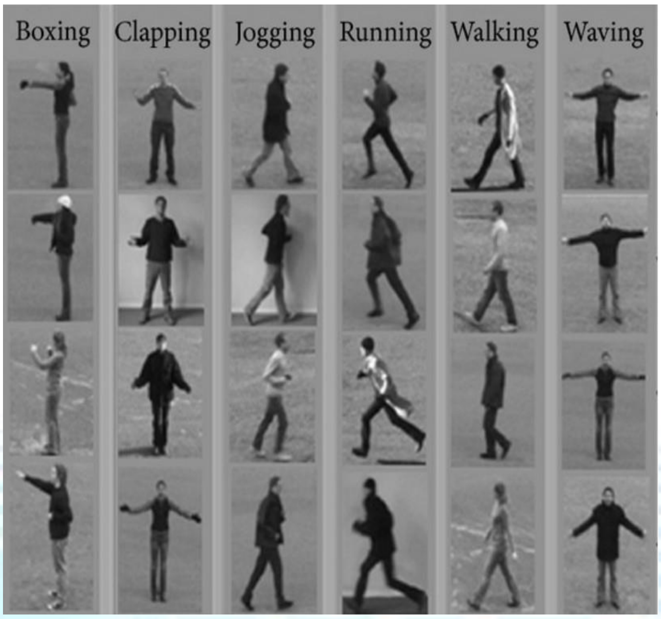


Fig.2: Selected Activity Images taken from KTH dataset

A. System of Human Detection

The categorization of the item as human or non-human is the detection of human beings. The picture of detection provides a minimum and maximum human height for all conceivable sub windows. Some of the sub-windows are filtered on the basis of the first pixel's percent. We then have to build a decent classification that can distinguish between people and non-individuals in a short period and with a low error rate. The project is essentially a combination of five basic substrates: first, frame standardization; second, gradient calculation blocks; third, the standardization block for contrast; the fourth, the HoG collection feature block; and fifth, sampling. The suggested approach is illustrated in Fig.3 in depth.

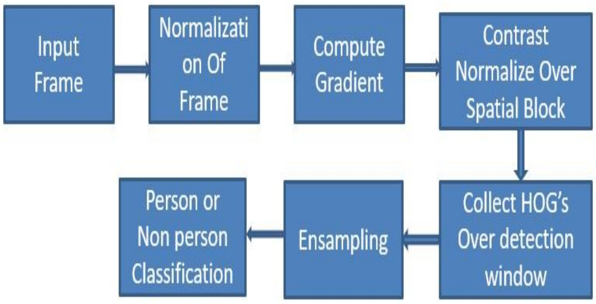

Input Frame
Normalization Of Frame
Compute Gradient
Contrast Normalize Over Spatial Block
Collect HOG's Over detection window
Ensampling
Person or Non person Classification

Fig. 3: Block diagram Proposed Human Detection

The technique is developed by the analysis of a dense grid of well-normalized local histograms of gradient orientations. The fundamental idea is that the distribution of local intensity gradients or borders may often describe the local object appearance and shape relatively efficiently, even without accurate knowledge of its gradient or edge positions. Itis also advantageous to standardize local reactions before using them to increase illumination, shadows, etc. The measurements of local histograms" energy" are then added toa certain spatial range (" blocks") and then the results are used to normalize all cells in the frame. This is achieved by using standardized descriptor blocks are called the descriptor histogram (HOG). The detection flap is utilized for our human detection chain with a dense grid of HOG descriptors and sampling.

B. Human activity recognition

We discuss our suggested human activity detecting technology in this part. First, we speak about the base network and the prediction network, our two DFNet sections. The fundamental characteristic of the Feature Base Network is to extract representational local and global features from the Feature Prediction Network at different levels.

This network has two key components: the V3 Feature Extraction and MAG Module. Figure. 4, shows the basic diagram of human activity where 80 Following detection, human segmentation will be performed. We categorized just four human activities in this detection such as hand waves, running, walking and boxing.

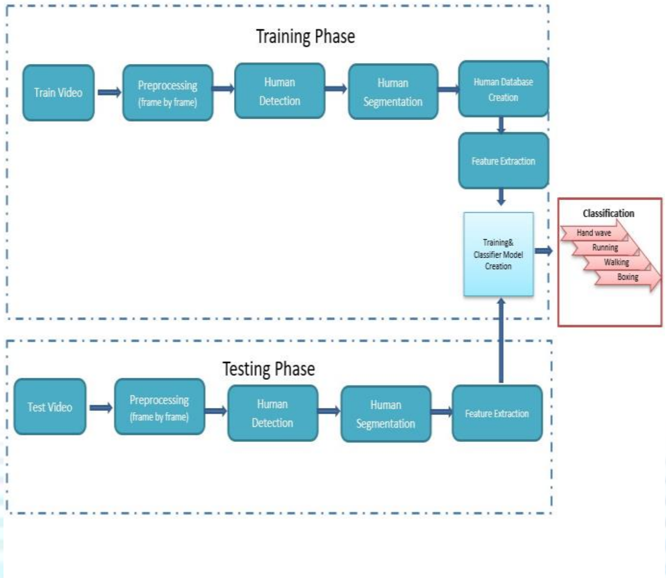


Fig. 4: General Schematic Diagram of Human Activity

1) Feature Base Network Using Inception V3: The key functions of the FAB-Network are to derive local and global relevant features to use the FAB-Network at various levels. A KTH data set frequently forms the backbone of activity detection to extract a network of characteristics that at many abstract levels are becoming increasingly challenging. One of our benefits is that it is highly adaptable and may be utilized on any core with no impact on the other model's design. In this view, the backbones of DFNet are VGG-16, MobileNets and V3 from DFNet, DFNet-M and DFNet-V3. The backbones are pre- trained for grading the picture. Thus adapted these models to satisfy the requirements of activity detection.

The model includes structural components such as symmetrical and asymmetric convolutions, the average pooling, max batches, concats, fully-connected layers and soft maxshown in Fig 5.

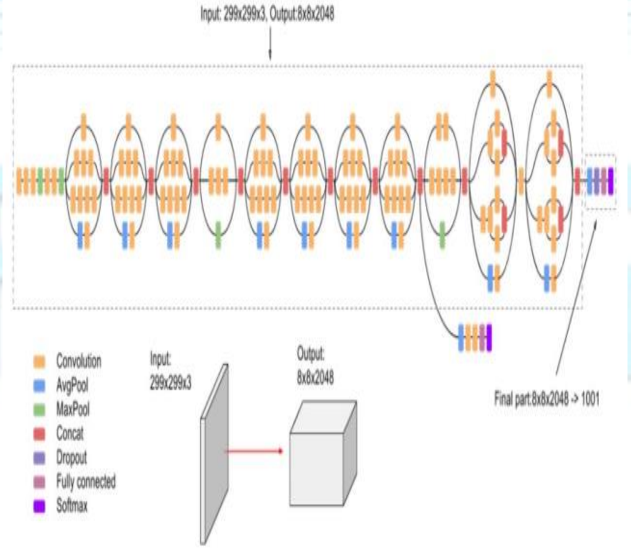
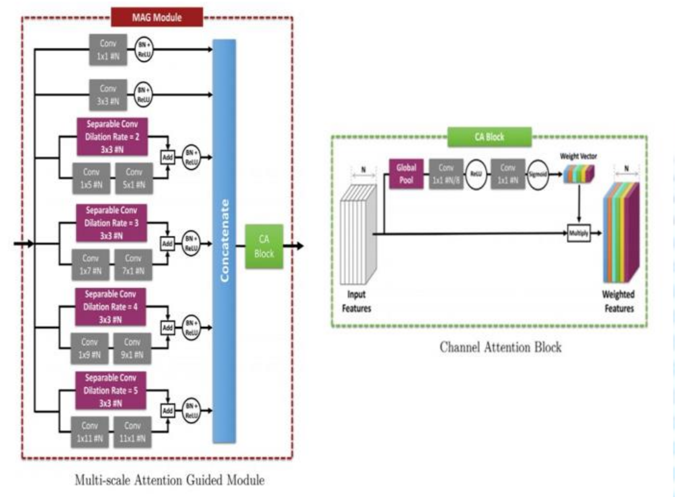


Fig. 5: High level model of Inception V3

If 3 × 3 parameters were changed by 2 x 2 as a secondary classifier between layers during training, which adds to the main network loss, the number of parameters would be somewhat larger. This would yield somewhat higher than the asymmetric chaos. An auxiliary classifier act as the regulator in the Inception v3 which usually reduces the sizes of the grid through the combination. In order to combat bottlenecks in computing costs, however, a more efficient technique is given.

2) Feature Prediction Network Using MAG Module: This is the second stage of detection of human activities. The primary role of the Prediction Network is to categorize the creation model with a multi-scale care module (MAGModule). The functionality extracted is sent via the MAGmodule. Huge kernels are obviously ideal for capturing large things, while little kernels can grab small ones. It is not the ideal strategy to use basic, fixed-size kernels due to the size variety of objects. In the beginning, we thus employ kernels of different sizes as a fashion to simultaneously collect items of various scales. With kernels of several sizes this module executes convolutions. Then we utilize Channel Attention Block for weighting multi-scale characteristics after con-catenation. Computationally highly costly developments with huge kernel sizes such as 5 x 5 and greater. To minimize this issue, we take two solutions: i) We may factor n x n kernel to a combination of 1 × n and n × 1 kernels and (ii) An n x n to dilation rate r will have the same field receptive as the size kernel (n + (r 1) × 2) × (n + (r 1) × 2)). The combination of both techniques in our MAG Module is used to construct a nxn kernel. We utilize the CA block depicted in Figure 6 (b) to weight the multi-scale characteristics (c). Figure 6 (a) illustrates the installation of the MAG Module.



Multi-scale Attention Guided Module

Channel Attention Block

Fig. 6: a) Multi-scale Attention Guided Module b) Channel Attention Block

IV. RESULTS AND DISCUSSION

The performance of the several MIT and INRIA datasets detectors is presented in Fig. 7.

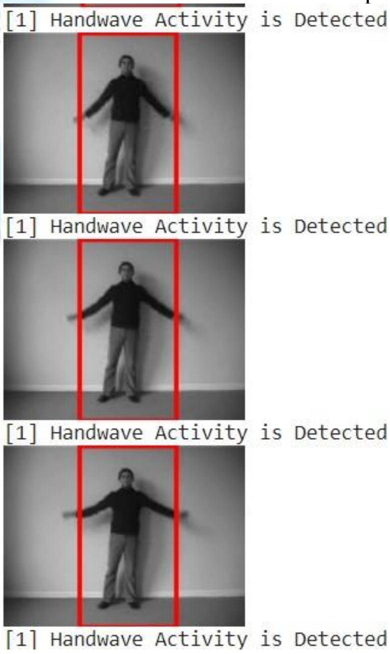

```
[1] Handwave Activity is Detected
[1] Handwave Activity is Detected
[1] Handwave Activity is Detected
[1] Handwave Activity is Detected
```

Fig. 7: Output image of human detection and Human Activity Recognition

The HOG detectors exceed the wavelet significantly. Figure 8 shows the output image of detected activities like boxing, hand wave, running and walking.

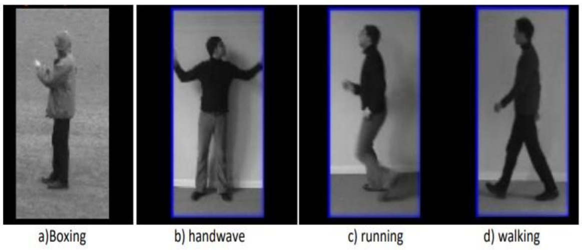


Fig. 8: Output image of detected activities a) boxing, (b) handwave, (c) running and (d) walking

Table 1 shows the output result of different parameters like accuracy, recall, F1-Score and Support for four different activities. Here Boxing, Handwave, Running and Walking were represented as Class 0, 1, 2 and 3 respectively.

TABLE I: Model Comparison-Training

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
MobileNetV2	.498	.56	.58	.56
Vgg-16	.9542	.80	.84	.81
InceptionV3	.9948	.87	.75	.74

	Accuracy	**Recall**	**F1-Score**	**Support**
Boxing	0.920	0.71	0.8	770
Handwave	0.5	0.51	0.5	390
Running	0.88	0.28	0.42	100
walking	0.43	0.98	0.6	182

In comparison to MobileNetV2 and Vgg-16, The V3 launch proved to be more effective in terms of number of network-generated parameters as well as the economic cost represented in Table 2. If a modification in the Inception Network is to be made, care must be taken to ensure that computer advice is no loss. Figure 6 illustrated the training and validation accuracy of MobileNetV2, Vgg-16and Inception V3.

TABLE II: Model Comparison-Training

V. CONCLUSION

By enhancing human detection and identification of ac-tions, we have provided a new technique. Two main stages are taken in the technique proposed. The initial stage is to recognize humans using HoG in the video sequences. We retrieved then the shapes and merged them according to their vector dimensions from certain sequences. Second, the classifications and action recognition utilizing the Deep Fusion Network are provided from selected characteristics. In comparison to other techniques, 99 approach was aesthetically as well as experimentally correct. In contrast to basic and state-of the-art techniques, there is consistently greater results and a well-performed generalization in the widely used public data set.

REFERENCES

[1] Y Xu et al., Detection of sudden pedestrian crossings for driving assistance systems. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 42(3), 729–739 (2012)

[2] W Fernando et al., in Information and Automation for Sustainability (ICIAfS), 2014 7th International Conference on. Object identification, enhancement and tracking under dynamic background conditions (IEEE, 2014)

[3] D Thombre, J Nirmal, D Lekha, in Intelligent Agent and Multi-Agent Systems, 2009. Human detection and tracking using image segmentation and Kalman filter (IAMA 2009. International Conference on, 2009) IEEE

[4] C Li, L Guo, Y Hu, in Image and Signal Processing (CISP), 20103rd International Congress on. A new method combining HOG and Kalman filter for video-based human detection and tracking (IEEE,2010)

[5] A Fakhar Ian, S Hosseini, T Gustafsson, in Control and Automation (MED), 2011 19th Mediterranean Conference on. Precise hybrid motion detection and tracking in dynamic background (IEEE, 2011)

[6] J Liu et al., in Image Processing (ICIP), 2013 20th IEEE International Conference on. Real-time human detection and tracking in complex environments using single RGBD camera (IEEE, 2013)

[7] W-C Cheng, D-M Jhan, A self-constructing cascade classifier withAdaboost and SVM for pedestrian detection. Eng. Appl. Artif. Intell.26(3), 1016–1028 (2013)

[8] H Beiping, Z Wen, Fast human detection using motion detection and histogram of oriented gradients. JCP 6(8), 1597–1604 (2011)

[9] J. Marin et al., Occlusion handling via random subspace classifiers for human detection. IEEE transactions on cybernetics 44(3), 342–354(2014)

[10] C-W Liang, C-F Juang, moving object classification using local shape and HOG features in wavelet transformed space with hierarchical SVMclassifiers. Appl. Soft Compute. 28, 483–497 (2015)

[11] D Kim, B Jun, in Theory and Applications of Smart Cameras. Accurate face and human detection using hybrid local transform features (Springer, 2016), pp. 157–185

[12] K-D Lee et al., Context and profile-based cascade classifier for efficient people detection and safety care system. Multimedia Tools and Applications 63(1), 27–44 (2013)

[13] Liu, W., Yang, J., Wang, L., Wu, C., Zhang, R. (2015). Movement Behavior Recognition Based on Statistical Mobility Sensing. AdhocSensor Wireless Networks, 25.

[14] Chavarriaga R, Sagha H, Calatroni A, et al. The Opportunity challenge: A benchmark database for on-body sensor-based activity recognition[J]. Pattern Recognition Letters, 2013, 34(15):2033-2042

[15] Ronao C A, Cho S B. Human activity recognition with smartphone sensors using deep learning neural networks[M]. Pergamon Press, Inc.2016.

[16] Ming Zeng, Le T Nguyen, Bo Yu, Ole J Mengshoel, Jiang Zhu, Pang Wu, and Juyong Zhang. Convolutional neural networks for human activity recognition using mobile sensors. In MobiCASE, pages197–205. IEEE, 2014.

[17] Lara O D, Labrador M A. A Survey on Human Activity Recognitionusing Wearable Sensors[J]. IEEE Communications Surveys Tutorials,2013, 15(3):1192-1209.

[18] Chen Z, Zhang L, Cao Z, et al. Distilling the Knowledge from Handcrafted Features for Human Activity Recognition[J]. IEEE Trans-actions on Industrial Informatics, 2018