Journal of Advances in Science and Technology
Vol. 18, Issue No. 2, September–2021, ISSN 2230-9659

A Study of Deep Learning Models in Human Actions Recognitions (HAR) in Video Sequences

Shashikant Pathak[1]*, Dr. Girish Padhan[2]

[1] Research Scholar, Shri Krishna University, Chhatarpur M.P.

[2] Associate Professor, Shri Krishna University, Chhatarpur M.P.

Abstract - Identifying people in videos and figuring out how to fix the problem that comes with it. The most popular and significant state-of-the-art solutions are presented. The ConvNet topologies based on deep learning to address the shortcomings of existing hand-coded methods, These ConvNets frameworks use a pre-trained deep model for features ex-tractions to recognize human behaviors in videos, making them suitable for transfer learning. It is empirically shown that deep pre-trained model built on a big annotated dataset is ex-changeable to action recognition task using the smaller training dataset. a deeply linked ConvNet for human activity detection presented that utilize the RGB frames at the top layer with Bi-directional Long Short Term Memory (Bi-LSTM), and at the bottom layer, CNN model is trained using a single Dynamic Motion Image (DMI).

Keywords - Deep Learning, Models, Human Actions, Recognitions, Video Sequences, etc.

----------X----------

INTRODUCTION

Nowadays, we're inundated with a deluge of image and video data thanks to advances in camera technology. Smart cameras are utilized in public places, in schools, in banks, in shops, in the medical domain, in the air, and even underwater. Analysis of video footage has become increasingly important for maintaining public safety and security in light of the proliferation of recorded videos and the ease with which they may be accessed. When watching movies for extended periods of time, the human cognition system's efficiency starts to suffer. For this reason, automated systems must be developed to analyze and understand video content for long periods of time. (1) Video comprehension relies heavily on the ability to identify human behaviors in the footage. Deep learning and handcrafted feature extraction methods are used to identify human behaviors in videos in this study.

It comprises safety, surveillance, healthcare, robotics, animations, sports analysis, content-based video summary, and behavioral analysis, smart homes and many more. An autonomous system that can properly identify and interpret human behavior and actions is one of the ultimate goals of the artificial intelligence civilization. To better serve society, a robot assistant may, for instance, help a patient under observation at home by analysing the best technique to exercise and preventing further injuries. This kind of smart technology will be immensely useful to us since it eliminates the need for unnecessary doctor visits, lowers healthcare costs, and allows for constant remote monitoring of the patient. Many feature-based methods, both manually designed and automatically taught, have emerged in the past two decades for identifying human actions in video footage. (2) Traditional methods of human activity identification rely on meticulously designed characteristics that drill down to the most fundamental of motions. Later, deep models for video activity analysis using convolutional neural networks (CNNs) were developed because of their ability to automatically learn the characteristics and categorise from raw video alone. Spatial background removal, optical flow, dense trajectories, and hu- man posture changes provide the basis of the handmade feature extraction methods used for activity recognition. Human position estimation, object identification, segmentation, audio analysis, object tracking, and super-resolution are all areas where deep learning systems have made significant strides in recent years. (3) In visual recognition tasks, the deep learning model is also crucial. Instead of manually extracting the characteristics, deep learning-based methods provide a more efficient and time-saving alternative. Handcrafted features solutions were shown to be effective, however they over-relied on feature descriptors when attempting action categorization. This type of problem needed additional man hours and specialised knowledge to solve. However, the automated features extraction from raw movies and improved identification rate provided by deep learning-based systems have made them the de facto standard.

What is Action?

According to Oxford dictionary, “the fact or process of doing something, often to achieve an aim” is termed action and similarly, an activity is described as “a thing that a person or group does or has done”. There are different definitions of action presented by various authors in their writings. However, most properly stated by Herath et al., “Action is the most primitive human-surrounding interaction with a meaning”. Therefore, human activities may be divided based on contact with the surrounding into four broad groups as follows: (4)

- **Gestures:** basic or fundamental motions of the body, such as the waving of a hand, the tilting of the head, etc.
- **Actions:** have the broadest definition, including a wide range of physical activities from running and walking to jogging and acting. You may think of an action as a series of smaller motions put together.
- **Interactions:** are described as conversations between two or more persons who are not part of a larger group. Human-human interaction (HHI) and HIO might both play a role (HOI). Human-human interactions include things like shaking hands, exchanging pleasantries, arguing, and fighting; human-object interactions (HOI) include things like preparing tea and answering the phone. (5)
- **Group Activities:** are described as several individuals or groups participating in shared aims such as a group meeting, two groups battling with each other, etc.

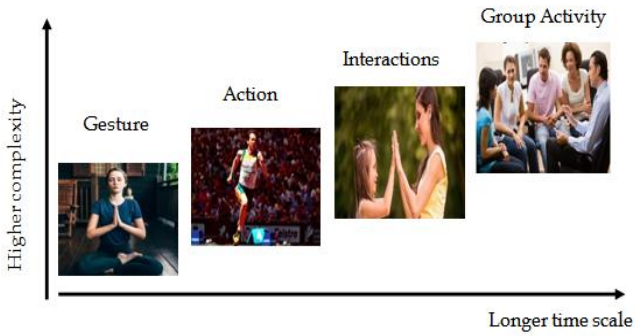

Figure 1: Levels of Human Activities

Figure 1 shows that moving from gesture recognition to group activity recognition increases the complexity of action representation.

Human Activity Recognition

The overall aim of human activity recognition is to analyze and understand human motion in a video sequence. Practicably, the goal is to categorize a video sequence or part of a video sequence as a particular class of human activity. In this context, the class of human activities can vary considerably from the simple to the much more complex. From simple to complex, these include: gestures or “atoms”, simple actions/activities, human to human interactions, human to object interactions and group activities. This literature review is largely focused on recognizing human activities, human interactions and group activities. (6) A single instance of a human activity typically lasts a few seconds in duration, although with periodic activities such as walking there may be no obvious end to the activity, and a video sequence may consist of the same simple activity repeated several times. An example of such a walking activity is shown in Figure 2. Further examples of human activities including a gesture activity, tennis server activity and a surveillance scenario are shown in Figure 3.

Figure 2: The KTH dataset has an example of a person walking

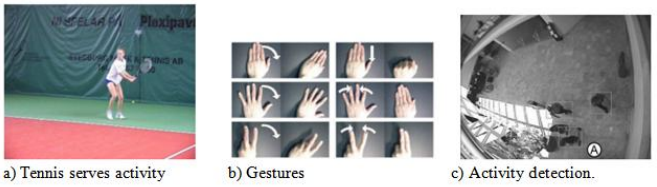

Figure 3: Recognizing human action through examples.

First, let's have a basic grasp of human acts and activities before diving into the specifics of how they might be identified. Gestures, actions, interactions, and group activities can all be categorized into four major categories: Raising a leg or stretching an arm is examples of gestures, which are the basic motion patterns of human body parts. Walking, punching, and waving are all examples of actions, (7) which are defined as single-person behaviors including a collection of various movements. Interactions often involve at least two persons or objects, such as a fistfight or a luggage theft. When a group of individuals or objects perform a group activity, it is called a "group fight" or "a meeting." This thesis focuses mostly on the recognition of human behaviors in video clips. Using a vision-based human action recognition system, action sequences are automatically analyzed to identify the activity. Pre-processing, feature extraction, and classification are three of the most important processes in the process of identifying human actions in movies. An identification system must be able to recognize basic human poses like standing, bending, walking, and sitting down, and so on in order to interpret complicated real-world behaviors and activities automatically. As a result, pre-processing of action sequences and feature extraction ensure that relevant spatio-temporal information of human positions, and hence the action done, can be recovered from the video footage. (8) The identification system makes a final conclusion based on the extracted features in the classification stage.

As a result, the performance of action recognition is substantially influenced by each level of the identification system. (9)

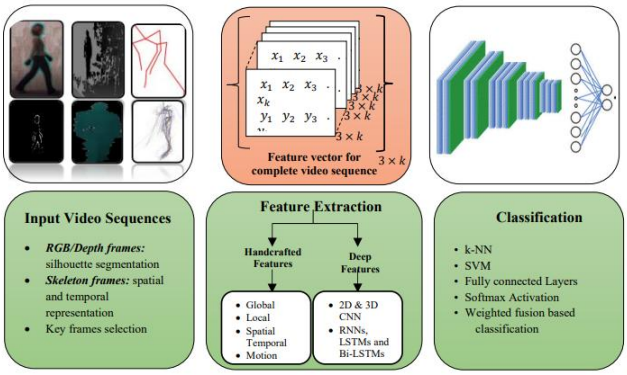

Figure 4: Overview of Human Action Identification (HAI) system

VIEW-INVARIANT DEEP HUMAN ACTION RECOGNITION MODELUSING MOTION AND SHAPE TEMPORAL DYNAMICS

It's challenging to identify human behavior in unfamiliar visual settings. Here, we provide a view-independent deep action detection paradigm by combining motion and shape temporal dynamics, two critical action signals (STD). RGB Dynamic Images (RGB-DIs) in the motion stream are used to record the action, which is subsequently processed using a modified version of the InceptionV3 model. (10) STD stream is capable of understanding the view-invariant shape dynamics of action over extended time periods by mining view-invariant properties from critical depth human pose frames using the structural similarity index matrix (SSIM). Predictions for the test sample's ultimate grade are made using late fusion techniques (maximum, average, and product) applied to scores from many streams. The value of the service provided is evaluated using cross-subject and cross-view validation techniques. Our method is demonstrated to be far more effective than the existing gold standard across a range of parameters, including accuracy, ROC curve, and Area under the Curve. (11)

Major Challenges

However much progress has been achieved in the field of human action detection in video sequences, it remains difficult to provide a discriminative action representation in naturalistic video. Obstacles in the environment include things like a busy background, shifting perspectives, varying lighting, shifting scales, a large degree of resemblance or dissimilarity amongst classes, and clothing of wildly contrasting colours and textures. A good descriptor that can deal with these issues is necessary for a successful action identification system.

The quality of an identification system based on visual cues is very sensitive to the conditions in which a video sequence was captured. When capturing video, the quality suffers because of the distortion of the collected picture data due to variations in the quantity of light or night vision. The result is that certain areas of a picture are bright while others are dark, making it difficult to make out details about the thing in question. For this reason, it is possible that the scene's foreground object and its backdrop are difficult to tell apart. It has an effect on the action identification process's pre-processing and feature extraction outcomes. Thus, in order to accurately identify the spatial information of the object in the frame, the influence of light must be decreased in the pre-processing step. To achieve this goal, histogram equalisation is employed to normalise the contrast of the pixels, however despite the increased system complexity, histogram equalisation based techniques were unable to produce steady performance across a wide range of light levels. Some studies used GMM, LBPs, etc. to process texture data in order to segment the object from the frame. (12) While GMM and LBP based object segmentation can be effective, they need a lot of time to complete. These days, you can get depth data about a scene in addition to the standard color data from a camera like the Microsoft Kinect. Since infrared (IR) light is not affected by external light, it may be used to provide accurate depth photographs. Thus, depth pictures are also independent of lighting conditions. Therefore, action recognition based on depth images may be the best option.

OBJECTIVES OF THE STUDY

- To study about the hybrid framework for human action recognition in RGB video sequences.

RESEARCH METHODOLOGY

View-invariant RGB-D action recognition framework with a deep neural network architecture is depicted in a block diagram format in Fig. 5. Over time, the object's motions and views are learned to inform the design of the building. The action's motion content is encrypted using dynamic images (DI), and then transfer learning is used with InceptionV3 to decode the action from RGB videos. We employ one Bi-LSTM and one LSTM layer, followed by dense, dropout, and softmax layers to train view-invariant Human Pose Model (HPM) features.

Depth-Human Pose Model (HPM) based action descriptor

When a human's attitude is represented in three dimensions, the information about the location of different portions of the body is maintained. The proposed technique employs view-invariant HPM features to express fine-grained information on human stance in depth from any perspective. To a large extent, the HPM model's architecture mirrors that of the AlexNet. Multi-viewpoint action data is used to train the system, with the data being

generated synthetically from 180 different viewpoints by employing human postural models fitted to the CMU motion capture data. This makes HPM completely immune to changes in perspective. To preserve chronological order over time, HPM features are learned on top of an LSTM sequential model.

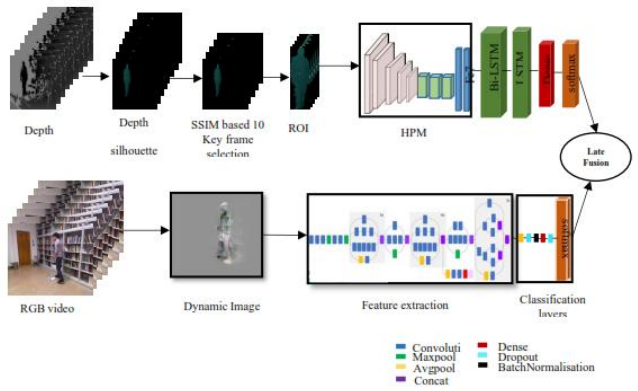

Figure 5: Block Diagrammatic Representation of the Proposed Method

Key Frame Extraction Using the Self-Similarity Index Matrix

To begin, a depth movie with n depth frames is created $\{f_1, f_2 \cdots \cdots, f_n\}$, the noise in these depth maps of human silhouettes was reduced using morphological pre-processing methods. With the help of the Structural Similarity Index Matrix, we can pick out the most crucial stance frames and cut out the unnecessary parts of the film (SSIM). It takes two successive depth frames and calculates their global structural similarity index value as well as their local SSIM map.

Model architecture and learning

To capture the long-term changes to an action's form, we suggest a deep convolutional neural network structure called shape temporal dynamics (STD) streams, which is identical to, except that we have coupled the last fc7 layer with a mix of Bidirectional LSTM and LSTM layers.

RGB-Dynamic Image (DI) based action descriptor

The inceptionV3 architecture is trained on the action sequence's motion using dynamic images (DIs) that accurately depict the video's visuals and motion. DIs is able to focus on the moving noticeable object while preserving the action dynamics over time by averaging out the pixels and motion patterns in the backdrop.

Experimental Work and Results

The final DI keeps the original film's proportions. To determine view-invariant motion properties of the action by being inserted into the motion stream, a combination of convolution layers from the InceptionV3 architecture with a variety of categorization layers. After the input image is processed by a pre-trained InceptionV3 model, the output is a high-dimensional representation of the image in the form of a vector of 8 8 2048 convolution features. To prevent excessive loss of accuracy after the dens (512, 'Relu') layer's 'ReLU' activations, a batch normalisation layer is used. We discovered that using the dropout layer in conjunction with batch normalisation was more successful than using either method alone, which would have led to a greater loss of weights as a result of the overfitting phenomena. The motion stream layers are learned from scratch for multiview datasets to fine-tune the weights of the InceptionV3 convolution layers. Maximum validation accuracy may be achieved by testing the sample with the best trained model weights. This will ensure a high identification rate is maintained despite changes in the viewpoint.

NUCLA multi-view action 3D Dataset

The Northern-UCLA multi-view RGB-D dataset was recorded in real time by Jiang Wang and Xiaohan Nie using Kinect v1 from three different locations throughout UCLA. Ten individuals' activities are included in the dataset, and they range from (1) one-handed picking up to (2) two-handed picking up, (3) dropping rubbish, (4) strolling, (5) sitting, (6) standing, (7) donning, (8) removing, (9) tossing, and (10) carrying. An issue with this dataset is that many actions have the same "walking" pattern both before and after the activity is finished. With the help of a support vector machine, 10 depth key frames are selected and transformed into 3D-HPM shape characteristics.

UWA3D Multi view Activity-II Dataset

The UWA3D multi-view activity-II dataset was created by recording 30 human actions from 10 persons from 4 different perspectives and at various times using the Kinect v1 sensor. There are 30 different types of body language, including: waving one hand, punching with one hand, punching with two hands, sitting, standing, vibrating, falling down, holding chest, holding head, sneezing, crouching, coughing, and Depending on your perspective, you may be looking at things from (i)the front, (ii)the left, (iii)the right, or (iv)the top.

NTU RGB-D Human Activity Dataset

When it comes to cross-view RGB-D datasets for human activity analysis, the NTU RGB+D action recognition dataset is by far the largest and most complex one to date. The video was captured using three 450-degree, 0-degree, and 450-degree Microsoft Kinect cameras. There are a total of 56,880 motion capture samples in a wide variety of file types, including RGB movies, depth map sequences, 3D skeletal data, and infrared videos.

Table 1: **Comparison of ARA (percent) values from subject-to-subject cross-validation of the**

NUCLA, UWA3D II, and NTU RGB-D Activity Datasets

Dataset/ Method		NUCLA dataset	UWA3D II dataset	NTU RGB-D dataset
Motion stream		93	82.6	62
STD stream		76	73.5	68.3
Proposed Hybrid Approach	Max	83	81.8	71.6
	Avg	84.5	79.6	75.7
	Mul.	**87.3**	**85.2**	**79.4**

Table 2: Cross View validation results in terms of ARA for NUCLA Multi-View Action 3DDataset

Training/ Test View		[1,2]/3	[1,3] / 2	[2,3]/ 1	ARA
Motion stream		86.29	76.42	70.6	77.77
STD stream		58.88	73.67	63.83	65.46
Hybrid	Max	91.73	85.43	79.72	85.68
	Avg	90.46	80.65	74.50	81.87
	Mul.	**93.12**	**89.94**	**85.36**	**89.47**

Table 3: Cross View validation results in terms of ARA for UWA3D Multi View Activity-IIDataset (%)

TrainingView	[v1,v2]		[v1, v3]		[v1,v4]		[v2,v3]		[v2,v4]		[v3,v4]		
Test View	**v3**	**v4**	**v2**	**v4**	**v2**	**v3**	**v1**	**v4**	**v1**	**v3**	**v2**	**v4**	**mean**
Motion Stream	87.4	81.2	78.1	85.5	73.9	79.4	82.6	73.1	81.6	72.4	83.5	81.1	79.98
STD stream	62.1	73.5	69.6	79.6	65.4	75.9	64.3	69.5	66.3	69.8	78.6	68.8	70.2
Hybrid Approach (max, avg,mul)	86.6	**85.3**	81.8	86.5	78.3	82.8	85.1	83.6	85.1	81.2	85.3	82.3	83.65
	73.2	78.8	75.4	81.3	79.9	81.4	79.4	77.3	79.4	80.9	84.1	**84.2**	79.6
	88.2	84.3	**82.6**	**88.6**	**80.5**	**83.2**	**88.9**	**84.6**	**93.9**	**85.2**	**91.2**	83.0	**86.18**

Computation time: The proposed view-invariant deep model demonstrated state-of-the-art performance on the multi-view NUCLA, UWA3D II, and NTU RGB-D Activity Datasets by merging data from the motion stream with the view-invariant Shape Temporal Dynamics (STD) stream. In conclusion, the proposed two stream deep architecture achieves better results than the state-of-the-art view invaraint deep recognition models and saves significant amounts of time. Tests may be executed on a machine with 24GB of RAM and an NVIDIA Geforce RTX 2080 Ti GPU, because training and testing do not require extensive processing resources.

Table 4: Examining the NUCLA Multi-View Action 3D Dataset in comparison to other state-of-the-art methods

Train-Test View Methods	DataType	[1,2]/ 3	[1,3] /2	[2,3] /1	Mean
CVP	RGB	60.6	55.8	39.5	52
nCTE	RGB	68.6	68.3	52.1	63
NKTM	RGB	75.8	73.3	59.1	69.4
HOPC+STK	Depth	80	-	-	-
HPM_TM	Depth	92.2	78.5	68.5	79.7
HPM_TM+DT	RGBD	92.9	82.8	72.5	82.7
HPM	Depth	85.21	78.57	71.96	78.58
Motion Stream	RGB	86.29	79.7	70.6	77.77
STD stream	Depth	89.96	81.37	75.12	82.15
Proposed Hybrid Approach	RGBD	**93.12**	**89.94**	**85.36**	**89.47**

Table 5: Comparison of with other state-of-the-arts in terms of ARA (%) on NTU RGB-DActivity Dataset

Method	Data type	Cross subject validation	Cross view validation
Skepxel1oc+vel	Joints	**81.3**	**89.2**
STA-LSTM	Joints	73.4	81.2
ST-LSTM	Joints	69.2	77.7
HPM(RGB+D)_Traj	RGB-D	**80.9**	**86.1**
HPM_TM+DT	RGB-D	77.5	84.5
Re-TCN	Joints	74.3	83.1
dyadic	RGB-D	62.1	-
DeepResnet-56	Joints	78.2	85.6
HPM	Depth	65.8	70.9
Motion Stream	RGB	62	68.7
STD stream	Depth Maps	68.3	72.4
Proposed Hybrid Approach	RGB-D (max fusion)	71.6	79.8
	RGB-D (late fusion)	75.7	83
	RGB-D (product fusion)	**79.4**	**84.1**

Table 6: Average Computation Speed (frame per sec: fps)

Method	Training	Testing
NKTM	12fps	16fps
HOPC	0.04fps	0.5fps
HPM+TM	22fps	25fps
Ours	**26fps**	**30 fps**

CONCLUSION

Human action identification is broken down into four distinct methods using both conventional handmade characteristics and deep features. In order to identify

anomalous behaviour in the elderly, a technique based on human action identification is proposed for monitoring normal behaviour. The issues that significantly degrade the efficiency of automatic human action recognition in videos: (1) view variations in action sequences, and (2) high inter class similarities and intra class variability of the action. A number of interesting findings may be drawn from the experiment results. Although late fusion results in more comprehensive data, it requires lengthy deep model learning for each feature. The computational cost of feature extraction is drastically decreased when based on transfer learning.

REFERENCES

1. Huang, Y., Tian, K., Wu, A., Zhang, G. (2019). Feature fusion methods research based on deep belief networks for speech emotion recognition under noise condition. Journal of Ambient Intelligence and Humanized Computing, 10(5): 1787-1798.
2. Hinton, G.E., Osindero, S., Teh, Y.W. (2006). A fast learning algorithm for deep belief nets. Neural Computation, 18(7): 1527-1554.
3. Laptev, I., Marszałek, M., Schmid, C., Rozenfeld, B. (2008). Learning realistic human actions from movies. IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, Alaska, pp. 1-8.
4. Klaser, A., Marszalek M., Schmid, C. (2008). A spatiotemporal descriptor based on 3D-gradients. in M. Everingham, C. Needham and R. Fraile (eds.), BMVC 2008 - 19th British Machine Vision Conference, Leeds, United Kingdom, pp. 1-10.
5. Huang, Y., Yang, H., Huang, P. (2012). Action recognition using hog feature in different resolution video sequences. International Conference on Computer Distributed Control and Intelligent Environmental Monitoring, Zhangjiajie, Hunan, China, pp. 85-88.
6. Ali, K.H., Wang, T. (2014). Learning features for action recognition and identity with deep belief networks. International Conference on Audio, Language and Image Processing, Shanghai, China, pp. 129-132.
7. Zhang, H., Zhou, F., Zhang, W., Yuan, X., Chen, Z. (2014). Real-time action recognition based on a modified deep belief network model. IEEE International Conference on Information and Automation, Tianjin, China, pp. 225-228.
8. Geng, C., Song, J. (2016). Human action recognition based on convolutional neural networks with a convolutional auto-encoder. 5th International Conference on Computer Sciences and Automation Engineering, Sanya, China, pp. 933-938.
9. Ijjina, E.P., Chalavadi, K.M. (2016). Human action recognition using genetic algorithms and convolutional neural networks. Pattern Recognition, 59: 199-212.
10. Ji, S., Xu, W., Yang, M., Yu, K. (2013). 3D convolutional neural networks for human action recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(1): 221-231.
11. Ke, Q., Bennamoun, M., An, S., Sohel, F., Boussaid, F. (2017). A new representation of skeleton sequences for 3d action recognition. IEEE conference on computer vision and pattern recognition, Hawaï, pp. 3288-3297.
12. Uddin, M., Kim, J. (2017). A robust approach for human activity recognition using 3-D body joint motion features with deep belief network. KSII Transactions on Internet & Information Systems, 11(2): 1118-1133.

Corresponding Author

Shashikant Pathak*

Research Scholar, Shri Krishna University, Chhatarpur M.P.