Journal of Advances in Science and Technology
Vol. 17, Issue No. 1, March-2020, ISSN 2230-9659

A Study on Handcrafted Features Based Models in Human Actions Recognitions

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 - The vision-based comprehension in video sequences entices several real-life applications such as gaming, robots, patients monitoring, content-based retrieval, video surveillance, and security. One of the ultimate ambitions of artificial intelligence society is to produce an autonomous system that can be identified and interpret human behavior and activities in video sequences properly. Over the decade, numerous efforts are made to detect the human activity in films but nevertheless, it is a tough work owing to intra-class action similarities, occlusions, view variations and ambient factors. These methods are divided into handwritten features based descriptors and automatically learned feature based on deep architectures. The suggested action recognition framework is separated into handmade and deep learning-based architectures which are then employed throughout this study by incorporating the novel algorithms for activity detection.

Keywords - Handcrafted Features, Models, Human Actions, Recognitions, etc

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

INTRODUCTION

In computer vision and image processing, the recognition and understanding of human actions has become a hot topic. Applications include surveillance, human-machine interaction, and video indexing, to name just a few. For example, running, walking, leaping, and so on are all examples of ordinary activities that can be identified using a HAR system. A simple motion model may be used to depict each of these acts, which are carried out by a single individual over a predetermined length of time. Spatial-temporal and sequential modeling methods can both be used to model actions. (1) Human actions are modeled as a 3D volume in a spatio-temporal dimension or as a set of attributes extracted from the 3D volume in spatio-temporal approaches to find out how similar two images are, the concatenated volumes along the time axis are compared. Sequential approaches, on the other hand, view an activity as a series of specific observations. To be more exact, they characterize a human movement as a series of feature vectors derived from the photos and proceed to reduce those that are close to each other in terms of a particular distance. (2)

A video-based HAR system, on the other hand, consists of two simple steps: In the first step, the video input frames are analyzed to extract features of interest. The second stage is classification, which entails sorting video sequences into categories based on the primitive vectors that were retrieved in the first step. To understand more complex structures, video-based HAR systems employing deep learning (DL) are attracting considerable attention. Non-linear data processing is used to create hierarchically ordered layers in the DL approach, which is an extension of artificial neural networks. Many standard approaches to image processing and computer vision are outperformed by techniques based on DL. In two ways, DL approaches offer great promise for meeting the needs of HARs: First and foremost, it's possible to unearth details on how the human body moves. Using this method allows for the manipulation of complicated human actions in order to recognize them. It's also possible to go above and beyond what's possible with typical methods of recognition. (3)

Human Action Recognition

Human activity identification may be broken down into three representation levels: the basic technology used to extract data from sensors, the recognition of a wide range of activities, and the applications that make use of this data. Low-level approaches are the backbone of any conventional HAR system, allowing for human activity identification from video sequences. Pre-processing of input frames, feature extraction, and action categorization are only a few of the fundamental phases. (4) But approaches based on deep learning of features may automatically categories the action based on the characteristics retrieved from the pixels. The main technology follows these fundamental processes from data collecting through

the depiction and categorization of human activity: (5)

- **Pre-processing:** its purpose is to improve the quality of input video sequences so that more robust characteristics may be extracted from them. The process incorporated a number of methods, such as background segmentation, silhouette extraction, histogram equalization, optical flow estimation, etc.
- **Feature extraction and Representation:** Sensors' raw video sequences provide unnecessary details. These superfluous details were eliminated during the feature extraction process, allowing for a clearer look at the Spatio-temporal connection between recognizable human activity and the original movies. In addition, the feature extraction methods filter out any background noise that may have been present when capturing sensor readings. (6) It will cut down on the amount of storage space needed and speed up the categorization process. The robust features for activity representation may be extracted using a number of well-known and established hand-crafted feature extraction methods, such as MHI, MEI, STIP, SIFT, Optical flow, BoWs, HOG, HOF, dense trajectories, etc.
- **Activity Classification:** There is no activity recognition system without this last component. Feature extraction from input data is crucial to achieving high classification accuracy. Machine learning techniques were used to classify and identify activities; however this practice had previously relied on human labour. The list includes methods like the linear multiclass SVM, HMM, K-NN, Random forest, Bootstrap, k-means, and so on. After the fully linked layers, automatic deep learning techniques employed the Softmax classifier to categories the action class. (7)

Human Action Identification

The term "human action recognition" (HAR) is used to describe the automatic detection of different actions, activities, poses, and gestures performed by humans by means of a computer or machine. Human action detection in videos has garnered a lot of interest in computer vision. It's also due to the fact that it has such enormous potential in a variety of fields. Human-computer interfaces, sports, event analysis, robotics, intrusion detection systems, content-based video analysis, multimedia semantic annotation and indexing, etc. Ambient Assistive Living, healthcare for the elderly, Intelligent Video surveillance systems, HCI, etc. Real-time movies, on the other hand, show complicated movements with varying degrees of inter- and intra-class similarity as well as lighting and perspective fluctuations, partial or whole occlusion, a cluttered background, and a constantly shifting camera. (8)

Let's have a handle on what human acts and activities are before we go into the specifics of classifying them. Human activities fall into four major categories: gestures, acts, interactions, and group activity. Raise a limb, extend out a limb; these are examples of gestures that include fundamental patterns of human body movement. An action is a series of motions performed by a single individual, such as walking, punching, or waving. All interactions include at least three actors and/or a shared environment, such as a fight between two individuals or the theft of a luggage. (9) Group activities, such as group fighting or a meeting, are carried out by a collection of individuals or things working together. In this study, the author primarily concentrates on locating human activity in recorded videos. In order to detect human actions, vision-based systems automatically extract spatial and temporal information from action sequences. Human action identification in videos entails a number of basic processes, including pre-processing, feature extraction, and classification. For an identification system to be able to automatically comprehend the real world's complex actions and activities, it needs to have a firm grasp on even the most fundamental human poses, such as standing, bending, walking, sitting down, and so on. Pre-processing action sequences aids in identifying the person in action, and the feature extraction step guarantees that useful spatio-temporal information of human postures, and therefore the action done, is retrieved. Finally, the retrieved characteristics are used to produce a categorization judgment by the identification system. Therefore, the effectiveness of action recognition is profoundly impacted by the many identifier-related stages.

The R-Transform and Zernike Moments in Deep Videos: A Framework for Recognizing Abnormal Human Actions

Here, we'll go over an innovative approach to human action recognition that has proven especially helpful for the geriatric population. We examine the most prevalent abnormal behaviors, such as fainting, chest pain, headaches, and forward and backward falls, to help the elderly become less reliant on others. Calculating the R-transform and Zernike moments of Average Energy Silhouette Images is a built-in part of the system, and the result is a powerful feature vector (AESIs). The AESIs are the result of combining the segmented silhouette data collected from the Microsoft Kinect sensor v1. (10) Having features that are insensitive to scale, translation, and rotation makes the proposed feature descriptor more robust against noise and more efficient at eliminating superfluous information. It improves the accuracy of the classification process and the reliability of the proposed method. This novel aberrant human action (AbHA) dataset is utilized in conjunction with three existing public 3D datasets (UR fall detection dataset, Kinect Activity Recognition Dataset (KARD), and multi-view NUCLA dataset) to validate the efficacy of the proposed technique. The proposed framework improves on the current state of the art, as

measured by the Average Recognition Accuracy (ARA).

Challenges in HAR

At least one of the dataset's recorded videos has a problem, typically involving action repetition, a distracting background, inconsistent views or lighting, or occlusions. Under the past, human action datasets were less complex than modern ones because there were fewer action classes, fewer subjects, and the actions were captured in more ideal settings. These essential features of the datasets determine how well an algorithm performs. When compared to RGB-D (depth) datasets, RGB datasets provide greater difficulties, such as perspective changes, intra-class differences, crowded backgrounds, partial occlusions, and camera motion. (11)

Background and Environment Conditions

Trees, waves, rain, and water all play a role in the recognition process since they are all part of the natural world. Background subtraction methods and foreground detection have a direct impact on the performance of feature descriptors. The dynamic backdrop of the KTH Action dataset makes it more difficult than the static Weizmann dataset. In movies when other elements, such as objects or the backdrop, are constantly in motion, identifying human activity becomes an especially important challenge.

Intra and Inter-class Variations

It's clear that several people have carried out the same tasks in varying ways. A person might run slowly, quickly, or even jump and then run; all of these variations are considered "running." This suggests that there is room for variation within the broad category of human movement that we call "activity." As an added complication, there are individual differences in the amount of time it takes to carry out an action. Variations in attitude and look between classes result from all of these causes.

Occlusion

"Occlusion" when one thing blocks your view of another. Recognizing human actions in obscured movies is a difficult undertaking. Occlusion is a significant difficulty for many computer vision tasks, including human posture estimation, object tracking, video surveillance, 3D foreground reconstruction, and traffic monitoring. Caused by the constant and irregular shifting of obstructing objects, both static and in motion, as far as human posture estimation goes, occlusion comes in two flavors: self-occlusion and occlusion by another object. When various sections of the body occlude one other owing to perspective differences, this is called self-occlusion.

View-Variations

The human activity recognition algorithm relies heavily on the perspective of any action captured inside the video collection. Multiple-camera recordings, in contrast to those shot from only one perspective, include more detailed information. Having several perspectives in HAR, however, makes the system more complicated.

Lack of Labelled Data

Most HAR methods have been found to function exceptionally well, even on relatively little human activity datasets. Widespread implementation of these methods is a difficult challenge for real-time systems. The results of using deep network based designs on big size datasets are encouraging. (12) However, a lot of labelled training data was needed to train these deep models. However, there are millions of action videos in databases like YouTube-8M and Sports-1M. However, video annotations are typically made using retrieval approaches that aren't always reliable. Due to a lack of labelled training data, training on such datasets is difficult. As a result, poor label accuracy has a negative impact on the effectiveness of action descriptions.

OBJECTIVES OF THE STUDY

- To examine a new framework for detecting abnormal human behavior.

RESEARCH METHODOLOGY

The proposed approach combines the scale, translation, and rotation features of the R-Transform with the Zernike moment to establish a robust action descriptor. A series of depth pictures are used to characterize the motion, and these images are then encoded using the R-Transform and Zernike moments to create an AESI image. Figure 1 depicts the proposed technique.

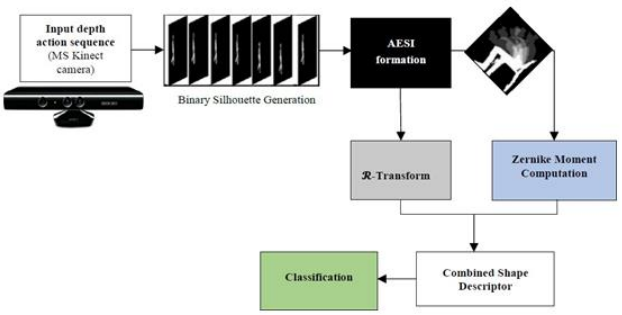

Figure 1: Flow diagram of proposed framework

Average Energy Silhouette Image (AESI) formation

When depth action sequences from the Microsoft Kinect camera v1 and skeleton joint locations per frame are combined, the fine binary silhouette extraction is greatly eased. The whole thing is condensed into one encrypted photo. An image of the mean energy silhouette, constructed from isolated binary shapes.

An R-Transform and Zernike Moments-based Descriptor of Geometry

The shape of a thing may tell you a lot about what it is doing. Recognizability relies on the accuracy of the form description. The activity that maintains the form and its alterations over time is represented compactly using AESI. To determine the shape descriptor, we employ a mixture of the R -transform and the Zernike moments.

- $\mathcal{R}$ -Transform
- $\mathcal{R}$-Transform Properties
- Zernike Moment

Final Feature Vector Formation

Radon Transform generates 2D Projection of AESI [480x 640] along the angle (0^0, 179^0). In terms of RT features, it produces a matrix with dimensions of [803 180]. Using the integral sum of the squared values of the Radon transform (RT) and the radial basis function (RBF), the R-Transform creates a feature vector of size [1 x180] that is invariant with respect to. Thus, the RT feature matrix provided by FR is projected onto a 1-dimensional space through the R-Transform.

Experimental Work and Results

The effectiveness of the proposed method is tested experimentally using the AbHA dataset, as well as three publicly available datasets (UR fall detection dataset, KARD dataset, and multi-view NUCLA dataset). K-Nearest Neighbor (NN) and Support Vector Machine (SVM) classifiers are used to categories the tasks performed during the tests. It is not required that features have a single dimension of discreteness. So, a Radial Basis Function (RBF) kernel based support vector machine (SVM) is utilized for classification since it can deal with non-linearly separable action features. Each dataset's penalty C and gamma are optimized to maximize the non-linear SVM classifier's performance while minimizing over fitting. The efficacy of an algorithm may be evaluated by calculating its average recognition accuracy, which is described mathematically as Eq. specified below.

$$ARA = \frac{TP+TN}{TP+TN+FP+FN} \times 100\%$$

To clarify, TP indicates a correct diagnosis, TN a negative one, FP a positive one, and FN a negative one. Average Recognition Accuracy (ARA) is used to measure the success of the proposed work in comparison to other, similar state-of-the-arts. Since there aren't many articles covering both the UR Fall detection dataset and the KARD dataset, the available comparisons are limited.

UR Fall detection dataset

This dataset, created by Kwolek et al., combines RGB and 3D photos. Forty ADLs are demonstrated, and thirty falls are shown from the front and the top. Using just depth instances from the front view, AESI are developed in this study to assess 22 fall and 22 ADL actions. Illustrations of the RGB/Depth Fall and ADL processes, K-Nearest Neighbor (K-NN) classifiers trained with the Leave-One-Out Cross Validation (LOOCV) technique produce the greatest results when used to the UR fall dataset when 'K'=3.

Table 1 displays the data gathered during the experiments. In Table 2, we compare the ARA achieved using UR Fall dataset to that obtained using comparable state-of-the-art methods. As can be seen in Table 1, the recognition capability of the framework is much enhanced by the incorporation of Zernike moment and R -transform, which renders the framework view-insensitive. The suggested framework has lower recognition accuracy than various state-of-the-arts, as shown in Table 2. This is because the authors of these works used a cryptographic method that combined RGB-D and accelerometric data to conceal the users' movements. However, while just depth maps have been used in the study, only 90% accuracy has been attained.

Table 1: ARA of the proposed work for UR Fall Detection Dataset

Activities	Action descriptor	Fall (%)	ADL (%)	ARA (%)
SVM (%)	*R* -Transform	94.6	93.2	93.9
	ansform + Zernikemoments	95.5	95.5	95.5
K-NN (%)	*R* -Transform	94.7	96.28	95.89
	ansform + Zernikemoments	96	97	**96.5**

Table 2: Comparison of ARA with other state-of-the-arts for UR fall detection dataset

Method	Classifier	Input	ARA (%)
Riemannian manifold	SVM	RGB + v	96.77
V-DGP	SVM	Depth maps	90
	SVM	RGB-D + Accelerometer	94.22
	K-NN	RGBD + Accelerometer	95.71
HTP	SVM	RGB-D	87.76
EWMA	SVM	RGB + Accelerometer	96.77
Curvelet	SVM-HMM	RGB	96.88
OFFD	CNN	RGB	95
Proposed Method	**K-NN**	**Depth maps**	**96.5**

KARD Dataset

In 2017, Gaglio et al. presented the Kinect Activity Recognition Dataset (KARD). There are eighteen distinct components that make up this diagram 2. There are three repetitions of each task, performed by ten different people. With this in mind, the dataset includes 540 (18x3x10) sequences shot in 640x480 at 30 frames per second. The 5-fold cross-validation method is used in conjunction with K-NN (K = 5) and SVM with a 1-vs-1 training set for activity recognition. Figure displays the results of the planned study on the KARD dataset. That's an increase of 1.74 percentage points from the most previous rating of 3.5. The better accuracy and generalizability of the suggested method, which compares the action recognition performance of the proposed work to that of other state-of-the-arts.

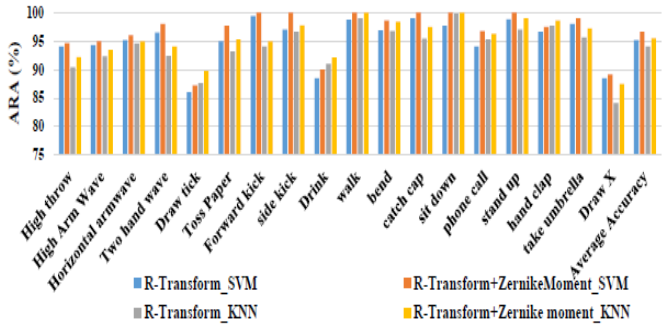

Figure 2: ARA for KARD dataset using SVM and K-NN

Abnormal Human Action (AbHA) Dataset

The Aberrant Human Action Dataset is an original compilation of often occurring abnormal behaviors in the daily lives of the elderly, such as "chest pain," "headache," "fainting," "falling backward," and "falling forward" (AbHA). Abnormal behavior describes situations in which a person needs assistance, such as when they are lying on the floor or sitting on a chair. Due to the lack of a publicly available dataset suitable for this sort of study, we were forced to compile our own, which we have named the AbHA dataset. To do this, we'll need eight employees to perform five separate tasks twice. This produces a total of 80 representative samples (8 x2 x5). The depth images captured by a Microsoft Kinect Depth sensor v1 are combined with the joint coordinates of the two skeletons to produce accurate binary silhouettes.

NUCLA Multi-View Action dataset

The NUCLA multi-view action dataset was created using the Microsoft Kinect v1 as RGB-D videos, and it contains ten actions performed by 10 people from three distinct perspectives: I left, (ii) front, and (iii) right. The dataset includes tasks such as one-handed and two-handed picking up, rubbish dropping, circling, sitting, standing, donning and doffing, tossing and carrying, and ten more. There are a lot of behaviours in the dataset that look quite similar to one another, making it hard to tell them apart. In the experiments, we only use half of the available views for training, and the other half for testing.

Table 3: Comparison of ARA with other state-of-the-arts for KARD dataset

Method	Classifier	Input Data	ARA (%)
Cippitelli et al.	SVM	Skeleton	94.9
Gaglio et al.	SVM	Skeleton	94.2
Madany et al.	ConvNet	Skeleton	98.5
Pham et al.	ResNet-44	Skeleton	99.97
Proposed method	**SVM**	**Depth**	**96.64**

Table 4: Comparison of ARA with other state-of-the-arts on multi-view NUCLA dataset

Method	Input	$V(3)$	$V(2)$	$V(1)$	ARA (%)
Depth-DVV	Depth	58.5	55.2	39.3	51.0
CV-CVP	Depth	60.6	55.8	39.5	52.0
NKTM	RGB	75.8	73.3	59.1	69.4
R-NKTM	RGB	78.1	-		78.1
HPM	RGB-D	91.7	73.0	79.0	81.3
Skepxels	Skeleton	91.5	85.5	79.2	85.4
Proposed method	$\mathcal{R}$-transfrom_KNN	89	82.4	75	82.13
	Hybrid vector_KNN	91.8	86.1	80.1	86
	$\mathcal{R}$-transform_SVM	90.6	85.6	79	85.06
	Hybrid vector_SVM	92.0	86.7	80.5	**86.4**

The proposed work's real-time performance has been tested on a single NVIDIA GeForce 940M graphics card, Intel core i5 processor, and 8GB of RAM. Table 3.6 lists the computations necessary to produce the feature vector and to test the action, providing concrete evidence of the proposed

framework's processing and testing time advantages in action recognition.

Table 5: Comparison of computation time of the proposed framework on multi-view NUCLADataset

Method	ature vector formation/ action (per video)	ting time per sample (per video)
HPMRGB-D GAN Refined Model	49.1ms	0.68ms
Proposed framework	**0.95ms**	**0.33ms**

CONCLUSION

It is seen that the success of any human action identification system strongly depends on how the action characteristics are developed. The evaluation of Handcrafted features extraction techniques. When it comes to efficient action identification in films, however, the difficulty of developing handmade action descriptors for actions learned in harsh environments grows. Researchers are now looking to deep features to help them overcome the practical difficulties of action recognition in videos. The expected consequence of this research is that depth images of action sequences simplify foreground segmentation. The employment of a human posture descriptor based on the R-transform and Zernike moments offers translation, scale, and rotation invariant action description, which improves human action recognition performance.

REFERENCES

1. Chang Li, Qian Huang, Xing Li, Qianhan Wu (2017) “Human Action Recognition Based on Multi-scale Feature Maps from Depth Video Sequences”, 21, 7941

2. Khan, S.; Khan, M.A.; Alhaisoni, M.; Tariq, U.; Yong, H.-S.; Armghan, A.; Alenezi, F. Human Action Recognition: A Paradigm of Best Deep Learning Features Selection and Serial Based Extended Fusion. Sensors 2017, 21, 7941.

3. Mehrez Abdellaoui, Ali Douik (2020) “Human Action Recognition in Video Sequences Using Deep Belief Networks” Vol. 37, No. 1, February, 2020, pp. 37-44.

4. Pham, H.H., Khoudour, L., Crouzil, A., Zegers, P., Velastin, S.A. (2017). Video-based human action recognition using deep learning: a review, pp. 1-34.

5. Wang, L.; Xu, Y.; Cheng, J.; Xia, H.; Yin, J.; Wu, J. Human action recognition by learning spatio-temporal features with deep neural networks. IEEE Access 2018, 6, 17913–17922.

6. Gumaei, A.; Hassan, M.M.; Alelaiwi, A.; Alsalman, H. A hybrid deep learning model for human activity recognition using multimodal body sensing data. IEEE Access 2019, 7, 99152–99160.

7. Gao, Z.; Xuan, H.-Z.; Zhang, H.; Wan, S.; Choo, K.-K.R. Adaptive fusion and category-level dictionary learning model for multiview human action recognition. IEEE Internet Things J. 2019, 6, 9280–9293.

8. Khan, M.A.; Zhang, Y.-D.; Khan, S.A.; Attique, M.; Rehman, A.; Seo, S. A resource conscious human action recognition framework using 26-layered deep convolutional neural network. Multimed. Tools. Appl. 2020

9. Al Ghamdi, Manal. (2016). Human Action Recognition In video streams. 10.13140/RG.2.2.17622.65601.

10. Ullah, A., Muhammad, K., Haq, I.U., Baik, S.W. (2019). Action recognition using optimized deep autoencoder and CNN for surveillance data streams of non-stationary environments. Future Generation Computer Systems, 96: 386-397.

11. Perez, M., Liu, J., Kot, A.C. (2019). Interaction recognition through body parts relation reasoning. Springer Asian Conference on Pattern Recognition (ACPR), Auckland, pp. 1-12.

12. Majd, M., Safabakhsh, R. (2019). Correlational Convolutional LSTM for human action recognition. Neurocomputing.

Corresponding Author

Shashikant Pathak*

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