IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, November 2022
Impact Factor: 6.252

Human Activity Detection by using Deep Learning

Mohammed Rizwan[1], Mikkili Dinesh[2], Midathada Saikumar[3], Tarak Ramarao[4], Mulla Sameer Basha[5], Venkat Sai Ajit[6]
GMR Institute of Technology, Rajam, Andhra Pradesh, India
20341A05B9@gmrit.edu.in[1], 20341A05C0@gmrit.edu.in[2], 20341A05C1@gmrit.edu.in[3], 20341A05C2@gmrit.edu.in[4], 20341a 05c3@gmrit.edu.in[5], 20341a05c4@gmrit.edu.in[6]

Abstract: *The subject of human activity recognition is considered an important goal in the domain of computer vision from the beginning of its development and has reached new levels. It is also thought of as a simple procedure. Problems arise in fast-moving and advanced scenes, and the numerical analysis of artificial intelligence (AI) through activity prediction mistreatment increased the attention of researchers to study. Having decent methodological and content related variations, several datasets were created to address the evaluation of these ways. Human activities play an important role but with challenging characteristic in various fields. Many applications exist in this field, such as smart home, helpful AI, HCI (Human-Computer Interaction), advancements in protection in applications such as transportation, education, security, and medication management, including falling or helping elderly in medical drug consumption. The positive impact of deep learning techniques on many vision applications leads to deploying these ways in video processing. Analysis of human behaviour activities involves major challenges when human presence is concerned. One individual can be represented in multiple video sequences through skeleton, motion and/or abstract characteristics. This work aims to address human presence by combining many options and utilizing a new RNN structure for activities. The paper focuses on recent advances in machine learning assisted action recognition. existing modern techniques for the recognition of actions and prediction similarly because the future scope for the analysis is to be developed in a high scale.*

Keywords: Artificial Intelligence, Machine Learning, Deep learning, Convolutional Neural Networks, Computer Vision

I. INTRODUCTION

Human activity recognition (HAR) is a widely studied computer vision problem. Applications of HAR include video surveillance, health care, and human- computer interaction. As the imaging technique advances and the camera device upgrades, novel approaches for HAR constantly emerge. This review aims to provide a comprehensive introduction to the video-based human activity recognition, giving an overview of various approaches as well as their evolutions by covering both the representative classical literatures and the state-of the-art approaches. Human activities have an inherent hierarchical structure that indicates the different levels of it, which can be considered as a three-level categorization. First, for the bottom level, there is an atomic element and these action primitives constitute more complex human activities. After the action primitive level, the action/activity comes as the second level. Finally, the complex interactions form the top level, which refers to the human activities that involve more than two persons and objects. In this paper, we follow this three-level categorization namely action primitives, actions/activities, and interactions. This three-level categorization varies a little from previous surveys and maintains a consistent theme. Action primitives are those atomic actions at the limb level, such as “stretching the left arm,” and “raising the right leg.” Atomic actions are performed by a specific part of the human body, such as the hands, arms, or upper body part. Actions and activities are used interchangeably in this review, referring to the whole-body movements composed of several action primitives in temporal sequential order and performed by a single person with no more person or additional objects. Specifically, we refer the terminology human activities as all movements of the three layers and the activities/actions as the middle level of human activities. Human activities like walking, running, and waving hands are categorized in the actions/activities level. Finally, similar to Aggarwal et al.’s review, interactions are human activities that involve two or more persons and objects. The additional person or object is an important characteristic of interaction.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, November 2022
Impact Factor: 6.252

Typical examples of interactions are cooking which involves one person and various pots and pans and kissing that is performed by two persons. This review highlights the advances of image representation approaches and classification methods in vision- based activity recognition. Generally, for representation approaches, related literatures follow a research trajectory of global representations, local representations, and recent depth-based representations. Earlier studies attempted to model the whole images or silhouettes and represent human activities in a global manner. The approach in is an example of global representation in which space-time shapes are generated as the image descriptors. Then, the emergency of space-time interest points proposed in triggered significant attention to a new local representation view that focuses on the informative interest points. Meanwhile, local descriptors such as histogram of oriented gradients and histogram of optical flow oriented from object recognition are widely used or extended to 3D in HAR area. With the upgrades of camera devices, especially the launch of RGBD cameras in the year 2022, depth image-based representations have been a new research topic and have drawn growing concern in recent years. On the other hand, classification techniques keep developing in step with machine learning methods. In fact, lots of classification methods were not originally designed for HAR. For instance, dynamic time warping and hidden Markov model (HMM) were first used in speech recognition, while the recent deep learning method is first developed for large amount image classification. To measure these approaches with same criterion, lots of activity datasets are collected, forming public and transparent benchmarks for comparing different approaches. In addition to the activity classification approaches, another critical research area within the HAR scope, the human tracking approach, is also reviewed briefly in a separate section. It is widely concerned especially in video surveillance systems for suspicious behavior detection. The writing of rest parts conforms to general HAR process flow. First, research emphases and challenges of this domain are briefly illustrated. Then, effective features need to be designed for the representation of activity images or videos. Thus, respectively, review the global and local representations in conventional RGB videos. Depth image- based representations are discussed as a separate part. Next, Section 6 describes the classification approaches. To measure and compare different approaches, benchmark datasets act an important role on which various approaches are evaluated. Section 7 collects recent human tracking methods of two dominant categories.

In representative datasets in different levels. Before we conclude this review and the future of HAR, we classify existing literatures with a detailed taxonomy including representation and classification methods, as well as the used datasets aiming at a comprehensive and convenient overview for HAR researchers.

II. LITERATURE SURVEY

A. A Review of Human Activity Recognition Methods

Pareek, P., & Thakkar, A. (2021). A survey on video-based human action recognition: recent updates, datasets, challenges, and applications. *Artificial Intelligence Review*, *54*(3), 2259-2322.

Recognizing human activities from video sequences or still images is a challenging task due to problems, such as background clutter, partial occlusion, changes in scale, viewpoint, lighting, and appearance. Many applications, including video surveillance systems, human-computer interaction, and robotics for human behavior characterization, require a multiple activity recognition system. In this work, we provide a detailed review of recent and state-of-the-art research advances in the field of human activity classification.

We propose a categorization of human activity methodologies and discuss their advantages and limitations. In particular, we divide human activity classification methods into two large categories according to whether they use data from different modalities or not. Then, each of these categories is further analyzed into sub-categories, which reflect how they model human activities and what type of activities they are interested in. Moreover, we provide a comprehensive analysis of the existing, publicly available human activity classification datasets and examine the requirements for an ideal human activity recognition dataset. Finally, we report the characteristics of future research directions and present some open issues on human activity recognition.

B. Deep-Learning-Enhanced Human Activity Recognition for Internet of Things

Priyadarshini, I., Sharma, R., Bhatt, D., & Al-Noman, M. (2022). Human activity recognition in cyber-physical systems using optimized machine learning techniques. Cluster Computing, 1-17.

HAR can be viewed as one kind of artificial intelligent technology which analyzes and recognizes human activities and

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, November 2022
Impact Factor: 6.252

behavior patterns automatically through a series of observations of wearable device data. Generally, Tirage et al. divided HAR into three levels as:

1. The movement recognition;
2. Action recognition;
3. Activity recognition, which were referred to as the low-level vision middle level vision and high-level vision, respectively Different machine-learning and deep- learning-based schemes were explored to handle issues in HAR and achieved effective performance Recently, studies focusing on HAR could be classified into two important categories as ambient sensor-based and wearable sensor-based approaches. Ambient sensor-based approaches usually deployed surveillance camera, sound, temperature, and other indoor sensors to capture environment related context signals and recognize people's daily activities in the fixed space. location information and identify their sitting, standing, and walking behaviors. Mantyjarvi et al. utilized the independent component analysis and principal component analysis schemes to recognize one person's walking posture based on the acceleration data collected from buttocks. These results demonstrated that activity recognition techniques based on wearable sensors could work effectively for the low-level vision, but failed to handle high-level recognition tasks for complex activity recognition

C. Human Activity Recognition via Hybrid Deep Learning Based Model

Khan, I. U., Afzal, S., & Lee, J. W. (2022). Human activity recognition via hybrid deep learning based model. *Sensors*, *22*(1), 323.

Human activity recognition (HAR) has multifaceted applications due to its worldly usage of acquisition devices such as smartphones, video cameras, and its ability to capture human activity data. While electronic devices and their applications are steadily growing, the advances in Artificial intelligence (AI) have revolutionized the ability to extract deep hidden information for accurate detection and its interpretation. This yields a better understanding of rapidly growing acquisition devices, AI, and applications, the three pillars of HAR under one roof. There are many review articles published on the general characteristics of HAR, a few have compared all the HAR devices at the same time, and few have explored the impact of evolving AI architecture. In our proposed review, a detailed narration on the three pillars of HAR is presented covering the period from 2011 to 2022. Further, the review presents the recommendations for an improved HAR design, its reliability, and stability. Five major findings were:

1. HAR constitutes three major pillars such as devices, AI and applications;
2. HAR has dominated the healthcare industry;
3. Hybrid AI models are in their infancy stage and needs considerable work for providing the stable and reliable design. Further, these trained models need solid prediction, high accuracy, generalization, and finally, meeting the objectives of the applications without bias;
4. little work was observed in abnormality detection during actions; and
5. almost no work has been done in forecasting actions. We conclude that: (a) HAR industry will evolve in terms of the three pillars of electronic devices, applications and the type of AI. (b) AI will provide a powerful impetus to the HAR industry in future.

D. Deep Learning-Based Human Action Recognition with Key-Frames Sampling Using Ranking Methods

Miranda, L., Viterbo, J., & Bernardini, F. (2022). A survey on the use of machine learning methods in context-aware middlewares for human activity recognition. Artificial Intelligence Review, 55(4), 3369-3400.

Video surveillance has become a vital need in the smart city era to enhance the quality of life and develop the area as a safe zone. Surveillance cameras are usually installed at a certain distance for the proper coverage of an area. Therefore, better analysis and more in- depth understanding of videos are highly required, profoundly impacting the security system. A video data driven system also helps healthcare, transportation, factory, schools, malls, marts, etc.

The objective of every camera feed is to know the specific incidence, such as identifying suspicious activities at the airport, bus stop, railway station, unusual activities at public gathering events (S. Wang et al. 2021. Human activity recognition's primary objective is to accurately describe human actions and their interactions from a previously unseen data sequence. It is often challenging to accurately recognize humans' activities from video data due to several

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, November 2022
Impact Factor: 6.252

problems like dynamic background and low-quality videos. In particular, two main questions arise among various human activity recognition techniques is the localization task. The sequences of images are referred to as frames. Thus, the primary objective of an action recognition task is to process the input video clips to recognize the subsequent human actions. Human activity mimics their habits; therefore, all human activities are unique, which turns into a challenging task to recognize.

Moreover, developing such a deep learning-based model to predict human action within adequate benchmark datasets for evaluation is another challenging task. With the ImageNet (Jia Deng et al. 2021) dataset's immense success for image processing, several benchmark action recognition datasets (Kay et al. 2020; Soomro, Zamir, and Shah 2022) have also been released to pursue research in this area. Similarly, suppose we compare video data processing with image processing; it requires enormous computation power and a large number of input parameters to train the deep learning model.

E. Depth Images-based Human Detection, Tracking and Activity Recognition Using Spatiotemporal Features and Modified HMM

Jindal, S., Sachdeva, M., & Kushwaha, A. K. S. (2022). A Systematic Analysis of the Human Activity Recognition Systems for Video Surveillance. In IoT and Analytics for Sensor Networks (pp. 345-354). Springer, Singapore.

Human activity recognition using depth information is an emerging and challenging technology in computer vision due to its considerable attention by many practical applications such as smart home/office system, personal health care and 3D video games. This paper presents a novel framework of 3D human body detection, tracking and recognition from depth video sequences using spatiotemporal features and modified HMM.

To detect human silhouette, raw depth data is examined to extract human silhouette by considering spatial continuity and constraints of human motion information. While, frame differentiation is used to track human movements. Features extraction mechanism consists of spatial depth shape features and temporal joints features are used to improve classification performance. Both of these features are fused together to recognize different activities using the modified hidden Markov model (M-HMM). The proposed approach is evaluated on two challenging depth video datasets. Moreover, system has significant abilities to handle subject's body parts rotation and body parts missing which provide major contributions in human activity recognition.

Recognizing human activities from video has made greater attention by researchers and become fundamental topic in pattern recognition research areas, including human machine interaction. In addition, several researchers faced other problems in the form of light sensitivity and motion ambiguities due to conventional cameras. To access high quality images and overcome the mentioned problems, depth cameras started new era for a variety of image recognition tasks including human activity recognition (HAR). These cameras facilitate to behave insensitive to lighting conditions, offering spatial characteristics and reducing body-occlusion. To review depth-based HAR research, Xia and Aggarwal described an algorithm to extract interest points from

III. METHODOLOGY

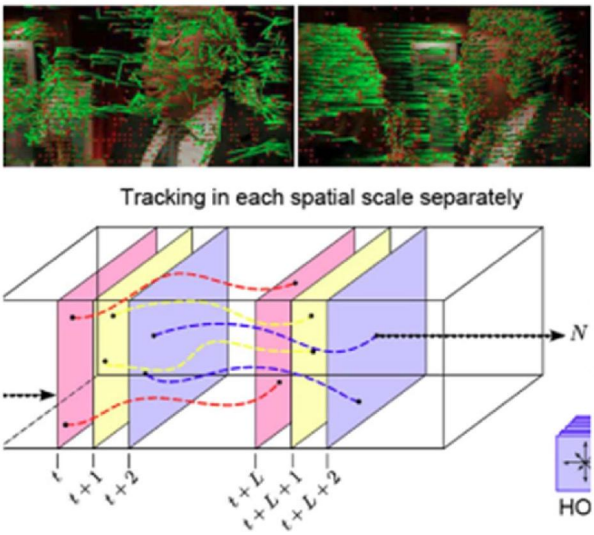
3.1 Methodology-1 [Space-Time Methods]

Space-time approaches focus on recognizing activities based on space-time features or on trajectory matching. They consider an activity in the 3D space-time volume, consisting of concatenation of 2D spaces in time. An activity is represented by a set of space-time features or trajectories extracted from a video sequence. Depicts an example of a space-time approach based on dense trajectories and motion descriptors

Visualization of human actions with dense trajectories (top row). Example of a typical human space-time method based on dense

Trajectories (Bottom Row). First, dense feature sampling is performed for capturing local motion. Then, features are tracked using dense optical flow, and feature descriptors are computed.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, November 2022
Impact Factor: 6.252



A. By Using Support Vector Machine[SVM]

The classification of a video sequence using local features in a spatiotemporal environment has also been given much focus. represented local events in a video using space-time features, while an SVM classifier was used to recognize an action.

considered actions as 3D space-time silhouettes of moving humans. They took advantage of the Poisson equation solution to efficiently describe an action by using spectral clustering between sequences of features and applying nearest neighbor classification to characterize an action.

Addressed the problem of action recognition by creating a codebook of space-time interest points. A hierarchical approach was followed by, where an input video was analyzed into several feature descriptors depending on their complexity. The final classification was performed by a multiclass SVM classifier. proposed spatiotemporal features based on cuboid descriptors. Instead of encoding human motion for action classification, proposed to incorporate information from human-to-objects interactions and combined several datasets to transfer information from one dataset to another.

B. By Using K-Nearest Neighbours[KNN]

Novelties proposed a novel representation of human activities using a combination of spatiotemporal features and a facet model, while they used a 3D Haar wavelet transform and higher order time derivatives to describe each interest point. A vocabulary was learned from these features and SVM was used for classification. used a mid-level feature representation of video sequences using optical flow features. These features were clustered using K-means to build a hierarchical template tree representation of each action.

A tree search algorithm was used to identify and localize the corresponding activity in test videos. also proposed a hierarchical representation of video sequences for recognizing atomic actions by building a codebook of spatiotemporal volumes. A probe video sequence was classified into its underlying activity according to its similarity with each representation in the codebook.

3.2 Methodology-2[Convolutional Neural Network (CNN) with Spatiotemporal Three-Dimensional (3D) Kernels]

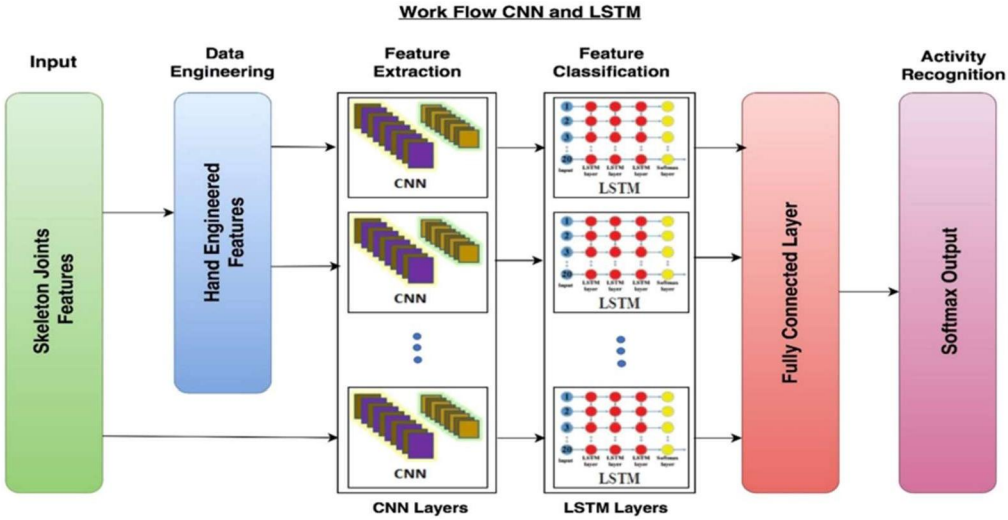
Skeleton-based Human Activity Recognition using ConvLSTM and Guided Feature Learning

The proposed methodology consists of detection of face, eyes, and mouth. The system architecture flow given below shows the complete process to be followed during detection.

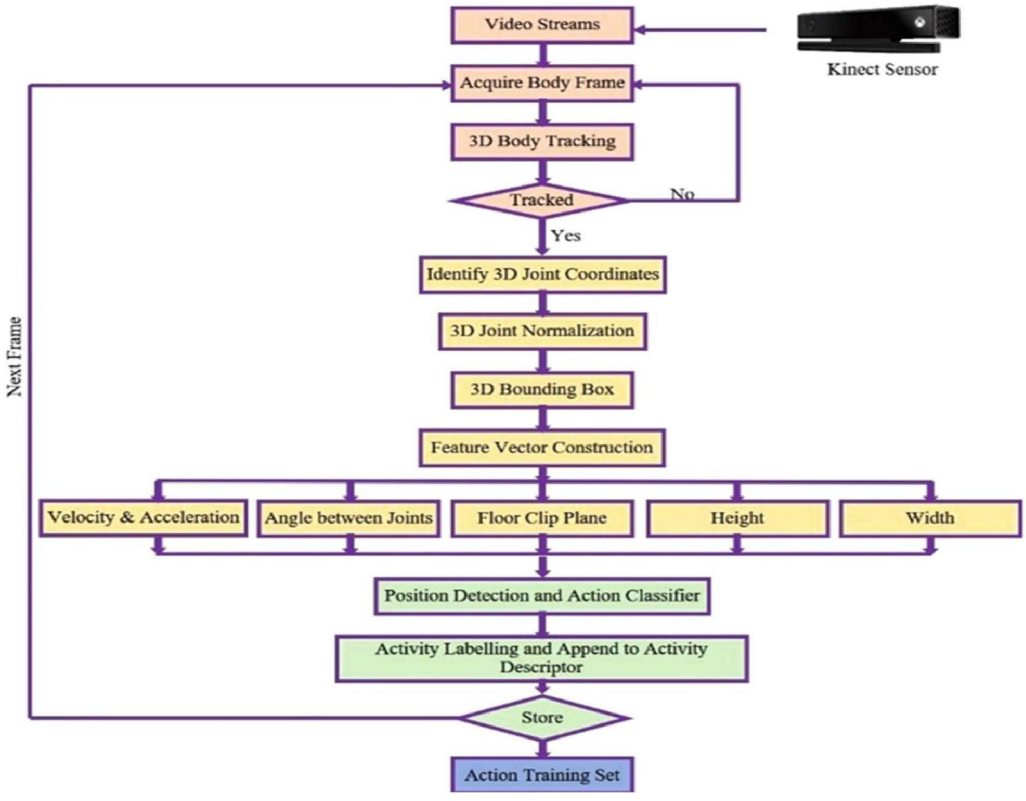
Convolutional Neural Network Algorithm: With the development of technologies and its integration with hardware and software has become the main adoption method for sundry projects predicated on CNN (Convolutional Neural

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, November 2022
Impact Factor: 6.252

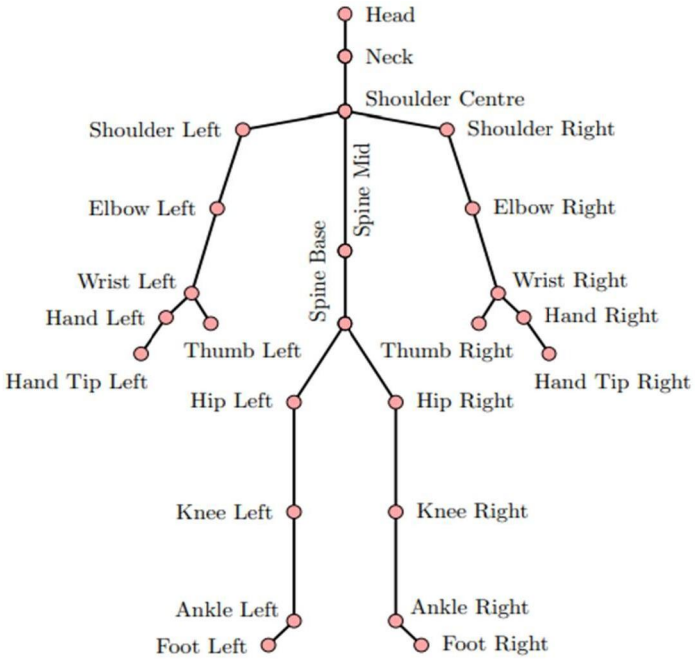
Network) specially for image relegation process. In this paper fundamental concepts of CNN are studied for the down sampling function in OpenCV to process the picture.



The proposed model of ConvLSTM. From the raw input videos 3D skeleton coordinates are extracted which are passed to calculate the geometrical and kinematic features. The extracted features along with raw skeleton joint coordinates are passed to computer then passed to LSTMs for extracting the temporal features. Finally, fully connected layers are applied to classify the activities and calculated the SoftMax scores

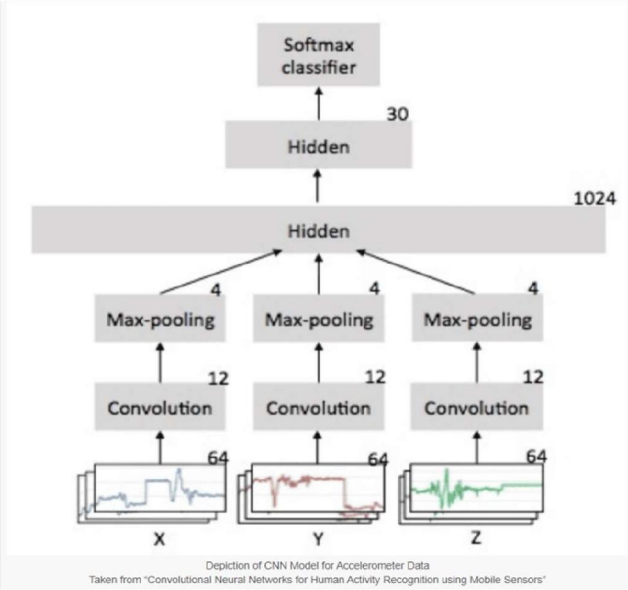


IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, November 2022
Impact Factor: 6.252



3.3 Methdology-3[Recurrent Neural Networks]

- Recurrent neural networks, or RNNs for short, are a type of neural network that was designed to learn from sequence data, such as sequences of observations over time, or a sequence of words in a sentence.
- A specific type of RNN called the long short-term memory network, or LSTM for short, is perhaps the most widely used RNN as its careful design overcomes the general difficulties in training a stable RNN on sequence data.
- LSTMs have proven effective on challenging sequence prediction problems when trained at scale for such tasks as handwriting recognition, language modeling, and machine translation.
- A layer in an LSTM model is comprised of special units that have gates that govern input, output, and recurrent connections, the weights of which are learned. Each LSTM unit also has internal memory or state that is accumulated as an input sequence is read and can be used by the network as a type of local variable or memory register.



Depiction of CNN Model for Accelerometer Data
Taken from "Convolutional Neural Networks for Human Activity Recognition using Mobile Sensors"

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, November 2022
Impact Factor: 6.252

Experiment 2: Shape Features are 300, Texture Features are 60, and Color Features are 9. In experiment 2, 2513 and 1628 images are taken from the Weizmann and KTH datasets respectively. The Weizmann dataset includes five categories. These five categories are bending, handshaking, jumping, running, and walking, while the KTH dataset includes 6 classes which are boxing, clapping, handshake, jogging, running, and walking. For experimental results, half of the images from each dataset are used for training and the remaining half are used for the purpose of testing. For assessment of the results, "10-fold" validation is used. The 10-fold validation is known as the evaluation method. For experiment 2, the maximum classification frequency is 99.5% for the Weizmann dataset on cubic-SVM, while for the KTH dataset, 99.9% is achieved in the subspace-KNN. The cubic-SVM applied to the model

Method	Classes	Experiment 1		Experiment 2		Experiment 3		Experiment 4		Experiment 5	
		Weizmann	KTH	Weizmann	KTH	Weizmann	KTH	Weizmann	KTH	Weizmann	KTH
Linear-SVM	C1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	C2	0.99	1.00	1.00	1.00	1.00	1.00	0.99	0.98	0.98	0.90
	C3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00
	C4	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.99	0.98	0.98
	C5	1.00	1.00	1.00	1.00	0.98	1.00	1.00	1.00	1.00	1.00
	C6		1.00		1.00		1.00		1.00		1.00
Cubic-SVM	C1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	C2	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.98	0.98	0.90
	C3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00
	C4	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.99	0.98	0.98
	C5	1.00	1.00	1.00	1.00	0.99	1.00	1.00	0.99	1.00	1.00
	C6		1.00		1.00		1.00		1.00		0.99
Complex tree	C1	0.98	1.00	0.99	1.00	0.98	1.00	0.98	0.99	0.98	0.99
	C2	0.87	0.98	0.89	0.96	0.85	0.98	0.88	0.99	0.87	0.99
	C3	0.94	1.00	0.93	1.00	0.94	1.00	0.94	0.99	0.94	0.99
	C4	0.90	0.99	0.88	0.98	0.90	0.99	0.88	0.99	0.87	0.96
	C5	0.97	0.98	0.95	0.96	0.97	0.98	0.96	1.00	0.96	1.00
	C6		1.00		1.00		1.00		1.00		1.00
Fine-KNN	C1	1.00	0.98	0.93	0.97	0.79	0.98	0.69	0.95	0.60	0.75
	C2	0.99	1.00	0.87	1.00	0.63	0.99	0.51	0.88	0.50	0.54
	C3	0.99	0.98	0.91	0.99	0.71	0.98	0.65	0.91	0.62	0.76
	C4	0.99	0.81	0.92	0.69	0.76	0.64	0.61	0.61	0.51	0.50
	C5	1.00	1.00	1.00	1.00	0.99	1.00	0.89	0.76	0.71	0.54
	C6		0.98		0.96		0.94		0.91		0.81
Subspace-KNN	C1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	C2	0.98	1.00	0.98	1.00	0.98	1.00	0.98	1.00	0.98	1.00
	C3	1.00	1.00	0.99	1.00	1.00	1.00	0.99	1.00	1.00	1.00
	C4	0.99	1.00	0.99	1.00	0.99	1.00	0.99	1.00	0.99	1.00
	C5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	C6		1.00		1.00		1.00		1.00		1.00

IV. RESULTS AND DISCUSSIONS

4.1 Result and Conclusion – 1

Performance Measures. Performance of the proposed algorithm is assessed on the basis of performance measures such as specificity, area under the curve, precision (PRE), sensitivity (SEN), and accuracy (ACU). Mathematically, it is represented by the following equations.

$$\mathrm{PRE} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}},$$

$$\mathrm{SEN} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}},$$

$$\mathrm{SPE} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}},$$

$$\mathrm{ACU} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{FP} + \mathrm{TP} + \mathrm{FN} + \mathrm{TN}},$$

$$\mathrm{AUC} = \int_{\infty}^{-\infty} \frac{\mathrm{TPR}(T)\mathrm{FPR}}{(T)dT}.$$

To get the experimental results, 50% of images are used for the purpose of training and the remaining 50% of them are used for testing. For assessment of the results, the "5-fold" validation is used. For, the maximum classification rate is 99.3% for the Weizmann dataset obtained with cubic-SVM. The linear-SVM and subspace-KNN obtained 99.8% accuracy simultaneously on the KTH dataset a. Cubic-SVM obtained a better sensitivity rate of 98.84, specificity of 99.81 and accuracy of 98.98 as compared to other classification methods using the Weizmann dataset. On the other hand, linear-SVM and subspace-KNN obtained a sensitivity rate of 99.86, specificity of 99.96, and precision of 98.74 which is better in comparison with other classification methods using the KTH dataset.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, November 2022
Impact Factor: 6.252

Experiment 2: Shape Features are 300, Texture Features are 60, and Color Features are 9. In experiment 2, 2513 and 1628 images are taken from the Weizmann and KTH datasets, respectively. The Weizmann dataset includes five categories. These five categories are bending, handshaking, jumping, running, and walking, while the KTH dataset includes 6 classes which are boxing, clapping, handshake, jogging, running, and walking. For experimental results, half of the images from each dataset are used for training and the remaining half are used for the purpose of testing. For assessment of the results, "10-fold" validation is used. The 10-fold validation is known as the evaluation method. For experiment 2, the maximum classification frequency is 99.5% for the Weizmann dataset on cubic- SVM, while for the KTH dataset, 99.9% is achieved in the subspace-KNN. The cubic-SVM applied to the MODEL

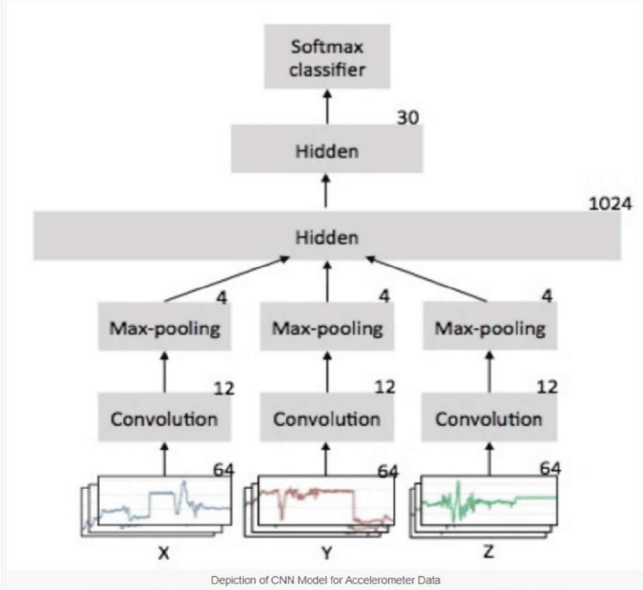
Exp no.	No. of classes KTH	Weizmann	Shape	Texture	Color	Folds
1	6	5	100	60	9	5
2	6	5	300	60	9	10
3	6	5	500	60	9	8
4	6	5	800	58	9	5
5	6	5	1100	55	9	7

Method	Classes	Experiment 1		Experiment 2		Experiment 3		Experiment 4		Experiment 5	
		Weizmann	KTH	Weizmann	KTH	Weizmann	KTH	Weizmann	KTH	Weizmann	KTH
Linear-SVM	C1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	C2	1.00	1.00	1.00	1.00	1.00	1.00	0.98	0.98	0.98	0.90
	C3	0.99	1.00	1.00	1.00	1.00	1.00	0.99	1.00	0.98	1.00
	C4	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.99	0.98
	C5	1.00	1.00	1.00	1.00	1.00	0.98	0.99	1.00	0.98	1.00
	C6		1.00		1.00		1.00		1.00	1.00	1.00
Cubic-SVM	C1		1.00		1.00		1.00		1.00		1.00
	C2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98	1.00	0.90
	C3	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00	0.98	1.00
	C4	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.99	0.98
	C5	1.00	1.00	1.00	1.00	1.00	0.99	0.99	0.99	0.98	1.00
	C6		1.00		1.00		1.00		1.00	1.00	0.99
Complex tree	C1	0.98	1.00	0.99	1.00	0.98	1.00	0.98	0.99	0.98	0.99
	C2	0.87	0.98	0.89	0.96	0.85	0.98	0.88	0.99	0.87	0.99
	C3	1.00	1.00	1.00	1.00	1.00	1.00	0.94	0.99	0.94	0.99
	C4	0.94	0.99	0.93	0.98	0.94	0.99	0.99	0.99	0.94	0.96
	C5	0.90	0.98	0.88	0.96	0.90	0.98	0.88	1.00	0.87	1.00
	C6	0.97	1.00	0.95	1.00	0.97	1.00	0.96	1.00	0.96	1.00
Fine-KNN	C1	1.00	0.98	0.93	0.97	0.79	0.98	0.69	0.95	0.60	0.75
	C2	1.00	1.00	0.87	1.00	0.63	0.99	0.51	0.88	0.50	0.54
	C3	0.99	0.98	0.91	0.99	0.71	0.98	0.65	0.91	0.62	0.76
	C4	0.99	0.81	0.92	0.69	0.76	0.64	0.61	0.61	0.51	0.50
	C5	1.00	1.00	1.00	1.00	0.99	1.00	0.89	0.76	0.71	0.54
	C6		0.98		0.96		0.94		0.91		0.81
Subspace-KNN	C1		1.00		1.00		1.00		1.00		1.00
	C2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	C3	0.98	1.00	0.98	1.00	0.98	1.00	0.98	1.00	0.98	1.00
	C4	1.00	1.00	0.99	1.00	1.00	1.00	0.99	1.00	1.00	1.00
	C5	0.99	1.00	0.99	1.00	0.99	1.00	0.99	1.00	0.99	1.00
	C6	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

3.3 Methdology-3[Recurrent Neural Networks]

Recurrent neural networks, or RNNs for short, are a type of neural network that was designed to learn from sequence data, such as sequences of observations over time, or a sequence of words in a sentence

A specific type of RNN called the long short-term memory network, or LSTM for short, is perhaps the most widely used RNN as its careful design overcomes the general difficulties in training a stable RNN on sequence data.



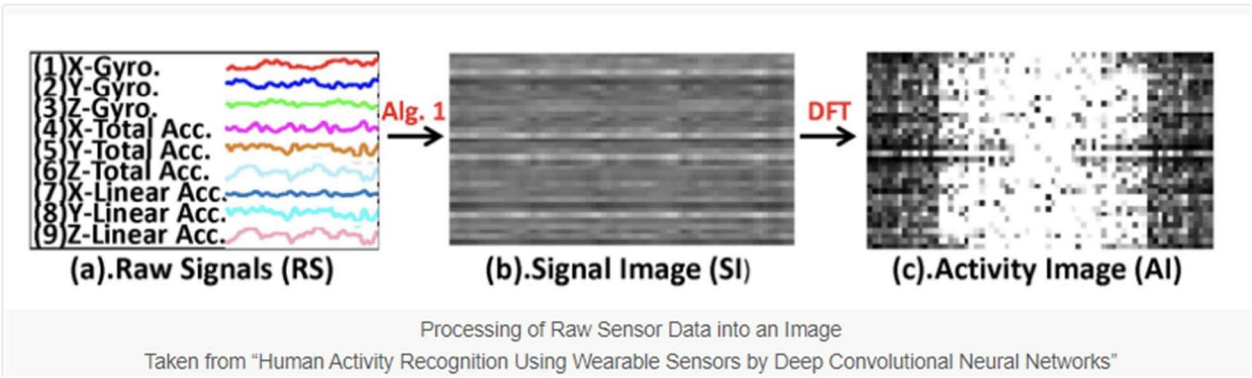
Depiction of CNN Model for Accelerometer Data
Taken from "Convolutional Neural Networks for Human Activity Recognition using Mobile Sensors"

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, November 2022
Impact Factor: 6.252

LSTMs have proven effective on challenging sequence prediction problems when trained at scale for such tasks as handwriting recognition, language modeling, and machine translation.

- A layer in an LSTM model is comprised of special units that have gates that govern input, output, and recurrent connections, the weights of which are learned. Each LSTM unit also has internal memory or state that is accumulated as an input sequence is read and can be used by the network as a type of local variable or memory register.
- Instead of using 1D CNNs on the signal data, they instead combine the signal data together to create "*images*" which are then fed to a 2D CNN and processed as image data with convolutions along the time axis of signals and across signal variables, specifically accelerometer and gyroscope data.
- Experiments show that convnets indeed derive relevant and more complex features with every additional layer, although difference of feature complexity level decreases with every additional layer. A wider time span of temporal local correlation can be exploited (1×9 – 1×14) and a low pooling size (1×2 – 1×3) is shown to be beneficial.

Like the CNN that can read across an input sequence, the LSTM reads a sequence of input observations and develops its own internal representation of the input sequence. Unlike the CNN, the LSTM is trained in a way that pays specific attention to observations made and prediction errors made over the time steps in the input sequence, called backpropagation through time



Processing of Raw Sensor Data into an Image
Taken from "Human Activity Recognition Using Wearable Sensors by Deep Convolutional Neural Networks"

Testing Accuracy: 91.65252447128296%
Precision: 91.76286479743305%
Recall: 91.65252799457076%
f1_score: 91.6437546304815%
Confusion Matrix:

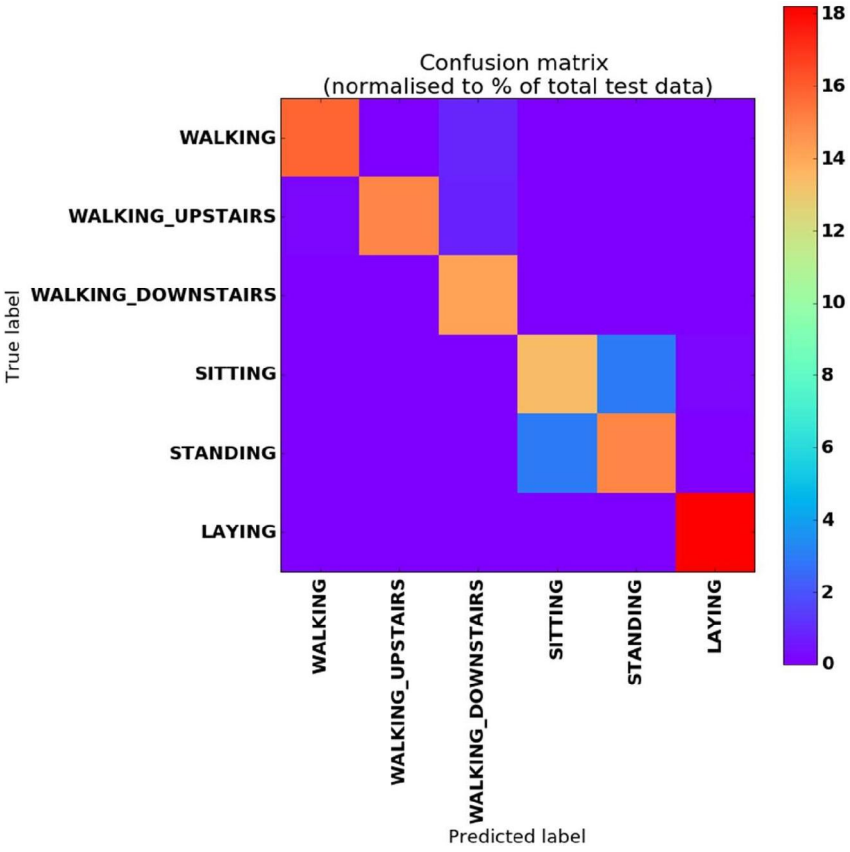
```
[[466   2  26   0   2   0]
 [  5 441  25   0   0   0]
 [  1   0 419   0   0   0]
 [  1   1   0 396  87   6]
 [  2   1   0  87 442   0]
 [  0   0   0   0   0 537]]
```

Confusion matrix (normalised to % of total test data):

```
[[ 15.81269073   0.06786563   0.88225317   0.           0.06786563   0.        ]
 [  0.16966406  14.96437073   0.84832031   0.           0.           0.        ]
 [  0.03393281   0.          14.21784878   0.           0.           0.        ]
 [  0.03393281   0.03393281   0.          13.43739319   2.95215464   0.20359688]
 [  0.06786563   0.03393281   0.           2.95215464  14.99830341   0.        ]
 [  0.           0.           0.           0.           0.          18.22192001]]
```

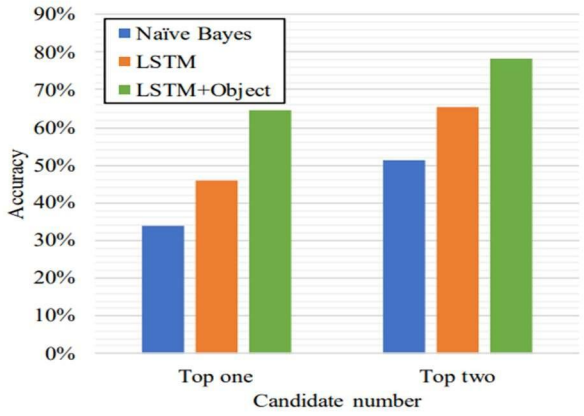
Note: training and testing data is not equally distributed amongst classes,
so, it is normal that more than a 6th of the data is correctly classifier in the last category.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, November 2022
Impact Factor: 6.252



4.2 Compartive Analysis

In the process of prediction, we first need to adjust and clean the original data. We treat the same activity that repeats in a short time as a pseudo record and merge them together. Furthermore, we delete some false records that do not meet the common sense. Then, we normalize the data to use it to train the LSTM model. Unlike the second stage, in the third stage, the initial 70% data are utilized for training and the remaining 30% data are for testing. In the experiment, we build a typical LSTM model on TensorFlow-GPU with Kera's as the high-level API. The training epoch is set to 10,000 to ensure the model is well trained. The LSTM model contains four layers: one input layer, two hidden layers and one output layer. The loss is set as categorical cross entropy, and the optimizer as Adam.



The timestep and neurons in the hidden layers are hyperparameters. We adjust the hyper parameters to ensure optimal performance of the model. We find that the test accuracy reaches its optimal value when timestep equals 3. This means that the LSTM model can utilize the past three activities to predict the next activity and achieve the highest accuracy,

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, November 2022
Impact Factor: 6.252

which accords with our assumption. Furthermore, the accuracy begins to decrease after that, meaning that the pattern of activities cannot be too large; otherwise, too much noise will be used. We also compare our model with the classical Naive Bayes method, as depicted in Figure 7. Because the Naive Bayes method only uses the current one activity to predict the next activity, we can see that our solution achieves much higher accuracy than that of Naive Bayes. The top two prediction accuracy reaches 65.2%. Moreover, when we apply the method to the process of prediction in the second stage, the accuracy will be as high is 78.3%.

V. CONCLUSION AND SCOPE FOR FUTURE RESEARCH

Human activity recognition remains to be an important problem in computer vision. HAR is the basis for many applications such as video surveillance, health care, and human computer interaction. Methodologies and technologies have made tremendous development in the past decades and have kept developing up to date. However, challenges still exist when facing realistic sceneries, in addition to the inherent intraclass variation and interclass similarity problem. In this review, we divided human activities into three levels including action primitives, actions/activities, and interactions. We have summarized the classic and representative approaches to activity representation and classification, as well as some benchmark datasets in different levels. For representation approaches, we roughly sorted out the research trajectory from global representations to local representation and recent depth-based representations. The literatures were reviewed in this order. State-of-the-art approaches, especially those depth-based representations, were discussed, aiming to cover the recent development in HAR domain. As the next step, classification methods play important roles and prompt the advance of HAR.

For human tracking approaches, two categories are considered namely filter based and kernel-based human tracking. Finally, 7 datasets were introduced, covering different levels from primitive level to interaction level, ranging from classic datasets to recent benchmark for depth-based methods. Though recent HAR approaches have achieved great success up to now, applying current HAR approaches in real-world systems or applications is still nontrivial. Three future directions are recommended to be considered and further explored.

First, current well-performed approaches are mostly hard to be implemented in real time or applied to wearable devices, as they are subject to constrained computing power. It is difficult for computational constrained systems to achieve comparable performances of those offline approaches. Existing work utilized additional inertial sensors to assist in recognizing, or developed microchips, for embedded devices. Besides these hardware-oriented solutions, from a computer vision perspective, more efficient descriptor extracting methods and classification approaches are expected to train recognition models fast, even in real time.

Another possible way is to degrade quality of input image and strike a balance among input information, algorithm efficiency, and recognizing rate. For example, utilizing depth maps as inputs and abandoning colour information are ways of degrading quality. Second, many of the recognition tasks are solved case by case, for both the benchmark datasets and the recognition methods. The future direction of research is obviously encouraged to unite various datasets as a large, complex, and complete one. Though every dataset may act as benchmark in its specific domain, uniting all of them triggers more effective and general algorithms which are closer to real-world occasions. For example, recent deep learning is reported to perform better in four-dataset-combined larger datasets. Another promising direction is to explore an evaluation criterion which enables comparisons among wide variety of recognition methods. Specifically, several vital measuring indexes are defined and weighted according to specific task, evaluating methods by measuring indexes such as recognition rate, efficiency, robustness, number, and level of recognizable activities. Third, mainstream recognition system remains in relative low level comparing with those higher-level behaviours. Ideally, the system should be able to tell the behaviour “Having a meeting” rather than lots of people sitting and talking, or even more difficult, concluding that a person hurried to catch a bus rather than just recognizing “running.” Activities are analogous to the words consisting behaviour languages. Analysing logical and semantic relations between behaviours and activities is an important aspect, which can be learned by transferring from Natural language processing (NLP) techniques. Another conceivable direction is to derive additional features from contextual information. Though this direction has been largely exploited, current approaches usually introduce all the possible contextual variables without screening

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, November 2022
Impact Factor: 6.252

REFERENCES

[1]. R. Poppe, "A survey on vision-based human action recognition," Image and Vision Computing, vol. 28, pp. 976–990, 2020.
[2]. J. K. Aggarwal and M. S. Ryoo, "Human activity analysis: a review," ACM Computing Surveys, vol. 43, p. 16, 2021.
[3]. T. B. Moeslund, A. Hilton, and V. Krüger, "A survey of advances in vision- based human motion capture and analysis," Computer Vision and Image Understanding, vol. 104, pp. 90–126, 2020.
[4]. J. M. Chaquet, E. J. Carmona, and A. Fernández-Caballero, "A survey of video datasets for human action and activity recognition," Computer Vision and Image Understanding, vol. 117, pp. 633–659, 2022.
[5]. M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," in Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1, pp. 1395–1402, Beijing, China, 2020
[6]. Christopher Reining 1, *, Friedrich Niemann 1, Fernando Moya Rueda 2, Gernot A. Fink 2 and Michael ten Hompel in Tenth IEEE International Conference on Computer Vision (ICCV'05)
[7]. Neha Gupta1,3 • Suneet K. Gupta1 • Rajesh K. Pathak2 • Vanita Jain3 • Parisa Rashidi4 • Jasjit S. Suri5,6 Published online: 18 January 2022 © The Author(s), under exclusive licence to Springer Nature B.V. 2021
[8]. Sharma, V., Gupta, M., Pandey, A. K., Mishra, D., & Kumar, A. (2022). A Review of Deep Learning-based Human Activity Recognition on Benchmark Video Datasets. Applied Artificial Intelligence, 36(1), 2093705.
[9]. Kamal, S., Jalal, A., & Kim, D. (2020). Depth images-based human detection, tracking and activity recognition using spatiotemporal features and modified HMM. Journal of Electrical engineering and technology, 11(6), 1857-1862.
[10]. Kim, Y., & Ling, H. (2009). Human activity classification based on micro- Doppler signatures using a support vector machine. IEEE transactions on geoscience and remote sensing, 47(5), 1328-1337.
[11]. Chen, Y. P., Yang, J. Y., Liou, S. N., Lee, G. Y., & Wang, J. S. (2020).
[12]. Online classifier construction algorithm for human activity detection using a tri-axial accelerometer. Applied Mathematics and Computation, 205(2), 849- 860.
[13]. Jobanputra, C., Bavishi, J., & Doshi, N. (2019). Human activity recognition: A survey. Procedia Computer Science, 155, 698-703.
[14]. Xia, K., Huang, J., & Wang, H. (2020). LSTM-CNN architecture for human activity recognition. IEEE Access, 8, 56855-56866.
[15]. Franco, A., Magnani, A., & Maio, D. (2020). A multimodal approach for human activity recognition based on skeleton and RGB data. Pattern Recognition Letters, 131, 293-299.
[16]. Gumaei, A., Hassan, M. M., Alelaiwi, A., & Alsalman, H. (2019). A hybrid deep learning model for human activity recognition using multimodal body sensing data. IEEE Access, 7, 99152-99160.