A FRAMEWORK FOR HUMAN ACTIVITY RECOGNITION USING DEPTH AND SKELETON EVIDENCE

A DISSERTATION

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE AWARD OF THE DEGREE OF

MASTER OF TECHNOLOGY

IN

SIGNAL PROCESSING AND DIGITAL DESIGN

SUBMITTED BY

NEHA MEENA

(2K16/SPD/11)

Under the guidance of

Dr. Dinesh Kumar Vishwakarma

Associate Professor

DEPARTMENT OF ELECTRONICS AND COMMUNICATION

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

DELHI- 110042 (INDIA)

MAY 2018

DECLARATION

I, (Neha Meena), 2K16/SPD/11 of M.Tech. (Signal Processing &Digital Design), hereby declare that the thesis titled "**A Framework for human activity recognition using depth and skeleton evidence**" which is submitted by me to the department of Electronics & Communication, Delhi Technological University, Delhi in partial fulfilment of the requirement for the award of the degree of Master of Technology, is original & not copied from any source without paper citation.

I further declare that the matter embodied in this report has not been submitted for the award of any other degree or diploma.

Place: Delhi NEHA MEENA

Date: (2K16/SPD/11)

CERTIFICATE

This is to certify that the thesis entitled **" A Framework for human activity recognition using depth and skeleton evidence "** being submitted by NEHA MEENA (Roll.No.: 2K16/spd/11) for the award of degree of Master of Technology to the Delhi Technological University is based on the original research work carried out by her. She has worked under my supervision and has fulfilled the requirements which to our knowledge have reached the requisite standard for the submission of this thesis.

Dr. D. K. Vishwakarma

(Supervisor)
Associate Professor, Department of Information and Technology
Delhi Technological University

ACKNOWLEDGMENT

I owe my debt and thereby would like to articulate my profound feelings of thankfulness to accomplish the research program with the support and direction of several persons. This challenging and rewarding experience definitely helped me to grow in character as well as academically. It is a pleasant aspect that I have now the opportunity to express my gratitude for all of them.

First and fore-most I would like to show gratitude to my supervisor, for his continuous help and belief all through, while I was studying. His considerable and methodical way of conducting research, along with his indisputable attention in the research subject, helped me learn a lot. Words cannot express my gratitude to him for his patience.

Last but not the least I want to thank my parents for always supporting and motivating me.

NEHA MEENA

(2K16/SPD/11)

ABSTRACT

In today's world of advancement, computer vision is gaining popularity in different spheres of life. Many areas like imaging, prediction of events, identification, encryption of different languages, etc are covered under it. Activity recognition is one of its area of application. Human activity recognition is helpful in prognosticating the actions of a person or a club of people of different views. The particular topic is inspired from real-world submissions, for example, visual observation, video understanding etc. The main building blocks of activity recognition consist of pre-processing, features extraction and representation, and classification.

Considering the various applications of activity recognition, this thesis investigates human activity recognition approach based on human skeleton and its associated depth images. The joints extracted from skeleton representation used in feature extraction and mapped on associated depth images. To achieve higher recognition accuracy of human activities, a three- step methodology is devised:

First step is obtaining the skeleton representation to estimate joint location, second step is extraction and representation of features, which is done using two major approaches: first is the histogram based and the second is the vector quantization using linear discriminate analysis and finally is the classification of human activities performed by the neural network.

This research is mainly focused on proposing neural network for extraction and representation of features in activity recognition methodology based on skeleton and depth information. The projected methodology is based on two public databases. It showed great accuracy on public dataset.

TABLE OF CONTENTS

List of Figures

List of tables

Chapter 1

INTRODUCTION

This chapter covers brief of human activity recognition, which includes the definition of basic terms used for human activity recognition, its need in today's life, applications, and difficulties involved therein. After that, we have described a flow diagram of human activity recognition system that includes extraction of features and their representation. And finally classifying them to their respective label. The various challenges that were encountered have also been discussed [1].

1.1 Definition of basic terms

Before starting defining the human activity recognition we will define the basic terms used in our approach-

1.1.1 Action and Activity

It is very difficult to define what constitutes an action. Even though there is a great demand for a specific and established action/activity hierarchy, there is not any recognized action hierarchy in computer vision till now.

- Action is an uncomplicated, infinitesimal progress performed by an individual.
- Commotion connotes towards a complicated outlook which involves a group of individuals even if such actions can be conducted by an individual only.

There is a well- known action hierarchy which can be defined as:

- Action primitives,
- Actions, and
- Activities

-though deeds, performances, uncomplicated deeds, complicated deeds, conducts, movements, etc. are most of the time used as synonyms by various people who conduct researches.

1.1.2 Modality

Modality is the method through which we gather and store data in the dataset. It can be considered as the capture to processing the input for getting the desired output. Its examples are video camera, depth camera, sensor, accelerometer, and many other sensing devices.

1.1.3 View

View can be defined as the observation recorded by a suitable apparatus. If the apparatus is able to capture from single direction, then it is termed as single view and if multiple angles can be recorded then it will be taken as multiple view.

1.1.4 RGB

RGB is the colour of object or we can say that it is the hue and saturation of the scene. It is a combination of red, green and blue components.

1.1.5 Depth Evidence

Depth evidence refers to the depth data which is used for pre-processing step. Depth data is contained in depth image. Depth is basically the difference between the object and the depth camera i.e. Kinect Camera. Kinect gives two images one is RGB and other is the depth image. For the depth image, the closer is the object the brighter it will look. Depth image contains more information than the RGB image. Depth image contains depth and angular information. The problem of illumination is reduced to some extent via Kinect Camera.

1.1.6 Skeleton Evidence

Skeleton evidence is the geometrical description that describes a shape using lesser number of pixels than original. A skeleton can be obtained using a technique called maximal disk. A disk is maximal if it fills the object inside which it is placed fully and there is no other disk that can be placed inside the object that can cover it at that point. The centre of such a maximal disk forms a skeleton point. The idea is to place all possible maximal disks inside the object and finding their centres. The skeleton of the object is a set of all such centre points of maximal disks.

1.1.7 Feature

Any property of an object that helps in making identification of one object from other object is called an image feature. Any feature is a function of one or more qualified measurements. The object can be characterised by one of the following:

Natural features- These are visual appearances of the image that are natural to the object, such as brightness, contrast and texture.

Artificial Feature- These are derived features that we get using image manipulations. Amplitude histograms and frequency spectrums are example of this category. Features are required for image recognition. The interaction between feature extraction and object classification is shown in below diagram-

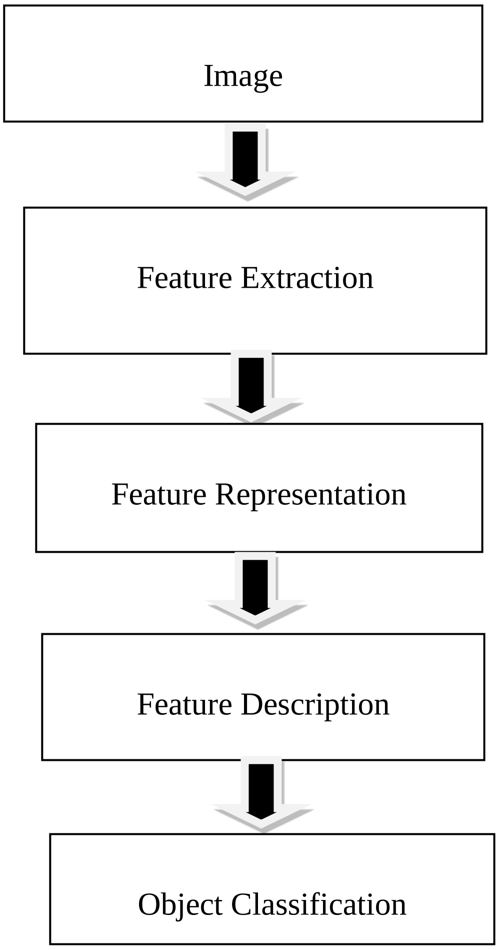


Figure 1 Interaction of feature extraction with the object classification process

1.2 Background of Human Activity Recognition

In the recent decade, human activity recognition is gaining popularity due to varied advantages. It has many real-time applications, such as video monitoring, recognizing activities and actions, and analysis of scene or view. However, it poses some problems also like cluttered backgrounds, illumination changes, camera motion, partial occlusions etc.

Most of the research work conducted in human activity recognition field had adopted a common assumption that the scene is noisy, where the actor is allowed to perform any activity. Each researcher has worked to develop a fully automatic activity recognition system which can identify the actions with greater accuracy, low error and low cost.

There is an inherent hierarchical structure associated with human activities which indicates various levels of it, which can be called as a three-level categorization . First category- bottom level, contains an atomic element and these atomic action primitives constitute more complex human activities. After the action primitive level, the action/activity comes at the second level. Finally, the complex interactions form the top level, which refers to the most complex level as it contains actions conducted by individuals that comprises of more than a couple of people and object. In the thesis, we follow this three-level categorization namely action primitives, actions/activities, and interaction. The work is based on only two levels i.e. action primitives and their respective labels.

Action primitives are those atomic actions at the limb level, such as “stretching the left arm,” and “raising the right leg” and many other. Atomic or we can say the basic actions are those which are performed by a particular part of the human body. Actions and activities are often used as same thing in this review, as it is referring to the whole-body movements composed of several action primitives in temporal sequential order and performed by a single person. We consider the terminology human activities as all movements of the three layers and the activities/actions as the middle level of human activities. Human activities such as walking, running, sitting, clapping and waving hands are categorized in the actions/activities level.

Human activity recognition involves database-which dataset are to be used for processing, feature representation-how the features to be passed through the system are to be represented, feature extraction-what features are to be extracted and finally the classification stage which deals with the classifier, and performance metrics.

1.2.1 Dataset Selection

Dataset selection is an important step in any research area. One has to be very clear which dataset is to be used and considered for processing. Here dataset refers to a set of images and videos to be taken into consideration.

Majorly there are two datasets used in human activity recognition field. One is RGB dataset [2] and other is the RGB(D) dataset. In recent generation RGB(D) dataset gaining popularity. With the advancement of technology, RGB dataset becomes obsolete in some way. RGB(D) dataset offers various advantages over traditional RGB dataset such as low- cost sensor, more information is being processed as it is recorded with different angles of view and most importantly it reduces the problem of illumination. With the help of this dataset now we are able to extract information from the images taken into the dark. There are many RGB(D) datasets available, some of them are mentioned below-

The Cornell Activity Dataset [2]-It holds 60 RGB(D) video shots of four subjects conducting 12 actions which are rinsing orifice, cleaning of teeth, using of contact lens instead of spectacles, conversing over the telephone, consuming water, opening medication box, food preparation, chopping, cooking-stirring, conversing on the sofa, lying down on the sofa, writing on white board and working on laptop in five diverse situations-office, kitchen, bedroom, bathroom and living room.

MSR Daily Activity 3D Data-set [3]-It has 16 activities which were being performed in front of Kinect Camera. Those activities are consuming any liquid, eating food, browsing a book, making a call over the mobile phone, scribing on a paper, playing any game, using a laptop, working on various cleaning appliances, gearing up, staying standstill, tossing paper, relaxing on a couch, strolling, striking guitar, reaching up and relaxing down. All of these work is conducted two times each, firstly in an upright position and secondly in the sitting position The samples are collected by ten subjects.

MSR Action 3D dataset- It has 20 actions repeated thrice by ten various individuals. The activities performed are waving high up, waving the arm in a horizontal manner, strike, hand catch, frontward thump, elevated fling, draw x, draw tick, draw circle, clapping, two hand wave, side boxing, twist, frontward punt, side kick, trotting, tennis sway, tennis serve up, golf swing, pick-up and throw.

UT Kinect Dataset- The UT Kinect dataset is composed of 10 different individuals comprising of 9 men and 1 woman who are asked to do each of the ten activities two times.. The subsequent actions are components of the dataset: walk, lie down, stand up, pick up, carry, throw, push, pull, wave, and clap hands.

1.2.2 Feature Representation and Description

We need a suitable medium to represent features for further processing. There are many methods proposed by researchers. Feature representation signifies how the features are to be represented.

Feature representation [4] involves storing the feature in a form that can be processed by image analysis program. It can be classified into data safeguarding and non-information safeguarding techniques. Non-information preserving methods include polygonal approximations where the irregular polygons are approximated to regular polygons before the representation. This process incurs information loss.

After the representation, the features should be described using descriptors. There are two types of descriptors-regional descriptors and boundary descriptors.

1.2.3 Regional Descriptors

The region descriptors include descriptors that characterize the object. These include approaches such as histogram, shape, topological, transform and texture features.

- Histogram features are known as amplitude or brightness features and are in terms of the luminance or intensity features. The measurement can be made at specific locations or over the entire neighbourhood.
- Shape features are one of the most important features of an object. It is difficult to describe a shape as most of the real-world objects do no exhibit standard geometrical shape.

- Topological features deal with the topology of an object. The main advantage of topological features is that they are not affected by the deformations and have been proven very useful for recognition process.
- Transform features are related with the Fourier transform of image contents. The transform coefficients specify the amplitude of the luminance patterns. If the basic pattern is the same as the feature pattern, then the feature detection can be done by monitoring the transform coefficients.
- Texture features are formed because of the variations in the intensity and colour. A texture is present in the form of repeated patterns.

1.2.4 Feature Selection Techniques

Feature selection is an optimizing set. Once the measurements are taken from the object, it is given to the classifier for recognition. Every feature adds a dimension to the computation. If there are two features, the features can be plotted as a 2D graph. Then a suitable decision boundary can be used to separate one object from another. But if there are three features then computation is a 3D plot involving plane calculation. As the number of dimensions increases, the amount of computation increases manifold. So, there is need for suitable optimization.

1.2.5 Recognition and Classification

A human being’s intelligence has a peculiar competence called classification. It is simply the aptitude to allocate a significant tag to an article. The procedure of conveying a significant tag to an unidentified article is named as acknowledgment. Further, the procedure of evaluating an unidentified article with the amass prototypes to distinguish the unidentified article is called categorization. Therefore, classification is the process of pertaining a tag or prototype class to an unidentified example. As defined above, feature extraction is a step for extracting the features relevant to the application. This consists of size, shape, location, and other visual features necessary for the application. The feature vector or pattern vector is a vector that contains the features that are extracted. The next stage is pattern recognition, where the feature vectors are learned and recognized.

A vital constituent in prototype identification is the capability of the structure to study from the information. Some of the important types of knowledge are administered and un-administered knowledge.

1.2.5.1 Supervised learning

This kind of learning involves a supervisor to make the system learn about the samples. Once the system becomes a learnt system, the supervisor supplies the test data, using which it tests the system.

1.2.5.2 Unsupervised learning

It is different from the supervised learning. In this type of learning, there is no explicit teacher or supervisor component. The learning system itself learns by trial and error. The instances themselves, based on similarity measures, form groups or clusters.

1.2.5.3 Classification

Classification is a supervised learning method. A classification model is supposed to 'learn' the complex relationships that exist flanked by the participation representation features using the training information. This resultant learning process is known as a concept or model. Post the time of learning, a classifier is termed as a 'learned system' plus it constructs one classification prototype. So, when an unknown instance is given as input, and if the classifier assigns the label correctly, it is known as a correct classification otherwise it is termed as incorrect classification and implies that more inputs are required to train the classifier.

Hence the classification process engrosses two training phases-training phase and testing phase. In the first phase, the classifier algorithm is fed with a huge set of identified information. This dataset is called training data. It is called so because the labels of known instances are given along with their attributes. The training data should generally be large and representative in nature. Once the training phase is over, a data driven classification model is created. The second phase is called the testing phase where the constructed model is tested and evaluated with unknown test data.

1.2.5.3.1 Classifier Design

There are many ways to design a classifier. Some of the popular design techniques are-

- Statistical Techniques- These are broadly divided into parametric and non- parametric. Parametric techniques are further classified into verdict theoretic practices and probabilistic practices.
- Non-statistical Techniques- These are divided into syntactic and structural practices.
- Cross breed practices

Statistical classifiers use statistical theories for obtaining models from the given training data with the help of statistical learning methodologies. Statistical classifiers are of two types-
Parametric classifier- Parametric classifiers are fed with a set of training data and a classification categorization is constructed. It is classified into the following two types based on the methods used for classification:

Decision-theoretic Technique-It is often called discriminant function examination. The thought is to design a decision function or a discriminant function that responds differently for each class, that is, its response for each class is unique. Examples are linear discriminant analysis, template matching, and Bayesian decision function-based techniques.

Probabilistic technique- Bayesian classifier is a popular probabilistic technique. It calculates the prior probability and conditional probability and uses these values to assign a label to the unknown instance.

Non-parametric classifier- When nothing is known about the densities of the data, no assumptions can be made. In such a case, non- parametric classifiers are useful. K-nearest neighbour and neural networks are popular non-parametric classifiers. Structural methods are of the following two types-

Syntactic- The objects of an image can be described using structural relations.

Structure-based techniques- In this technique, graphs are used instead of strings. An object can be modelled as a graph. The object is then matched with another object using graph matching techniques.

1.2.5.3.2 Evaluation of Classifier Algorithms

Accuracy is the ability of the classification models to correctly determine the class of an unknown test image. Classifiers are affected by noise and outliers present in the dataset. Some of the popular techniques for evaluation are-

Separate training/test sets- This is one of the simplest methods for testing the classifier. The dataset is separated into two sets. One of them is called training set. The other set, called the test dataset, is used for trying presentation of the classifier. k-fold cross validation- Another popular method that is used frequently is cross validation. K-fold cross validation is an improvement as compared to the previous methods. Leave-one-out cross validation- This is also called N-folding or jack-knifing technique. This is an extreme form where every instance is considered as dataset. The amount of computation involved is very large and therefore, this method is considered unsuitable for many real-world applications. Many objective metrics are available for quantifying the quality of the classifiers. Some of the useful parameters are as follows:

Accuracy, indicates the ability of the classifier to predict unknown instances.

Confusion Matrix- Generally, the outcomes of the classifier are represented as table called confusion matrix. Confusion matrix indicates the performance of the classifier in classifying the instances. Some of the metrics are described as-

True positive rate (TP rate) - It indicates the sensitivity of the classifier and is described as the probability that it will produce a true positive rate. It can be calculated as TP/P , where $P = TP + FN$.

False Positive rate (FP rate)- It is also referred as specificity. It is the probability that a classifier produces erroneous results as positive results for negative instances. It can be calculated as FN/P , where $P = TP + FN$.

True negative rate (TN rate)- This is the probability of detecting the result as negative when the input fed was also negative. It can be calculated as TN/N , where $N = FP + TN$.

Positive predictive value (or precision)- This is the probability that an object is classified correctly as per the actual input. It is defined as $\frac{TP}{TP+FP}$.

Negative predictive value- This is the probability that an object is not classified properly as per the actual value. It is defined as $\frac{TN}{TN+FN}$.

Accuracy- This is also referred as recognition rate. The accuracy of the classifier can be shown as $\frac{TP+TN}{TP+TN+FP+FN}$. It indicates the ability of a classifier to classify instances rightly.

Error rate – This is also called as misclassification rate. The error rate indicates the proportion of instances that have n=been wrongly classified, and given as $\frac{FP+FN}{TP+TN+FP+FN}$.

1.3 Challenges of the Domain

1.3.1 Intra-class Variation and Interclass Similarity:

Different from speech recognition, there is no grammar and strict definition for human activities. This causes twofold confusions. On one hand, the same activity may vary from subject to subject, which leads to the intraclass variations. The performing speed and strength also increase the interclass gaps. On the other hand, different activities may express similar shapes (e.g., using a laptop and reading). This is termed as interclass similarity which is a common phenomenon in HAR. Accurate and distinctive features need to be designed and extracted from activity videos to deal with these problems.

1.3.2 Recognition under Real-World Settings

- Complex and Various Backgrounds**:** While applications like video surveillance and fall detection system use static cameras, more scenarios adopt dynamic recording devices. Sports event broadcast is a typical case of dynamic recording. In fact, with the popularity of smart devices such as smart glasses and smart phones, people tend to record videos with embedded cameras from wearable devices anytime. Most of these real-world videos have complex dynamic backgrounds. First, those videos, as well as the broadcasts, are noted in diverse and shifting conditions. Second, realistic videos abound with occlusions, lighting inconsistency, and perspective modifications, which make it harder to recognize activities in such complex and various conditions.

- Multisubject Interactions and Group Activities: Earlier research concentrated on low-level human activities such as jumping, running, and waving hands. One typical characteristic of these activities is having a single subject without any human-human or human-object interactions. However, in the real world, people tend to connect communicative activities with one or more persons and objects. An American football game is a good example of interaction and group activity where multiple players (i.e., human-human interaction) in a team protect the football (i.e., human-object interaction) jointly and compete with players in the other team. It is a challenging task to locate and track multiple subjects synchronously or recognize the whole human group activities as "playing football" instead of "running."

Chapter 2

Literature survey

1. Lu et al. [5], presented a fresh outlook for identifying the activities of humans via histograms of 3D joint locations (HOJ3D) as a compressed demonstration of positions. They hauled out the 3D skeletal joint locations from Kinect depth maps using Shotton et al.'s method. The HOJ3D calculated from the accomplishment depth progressions are re-projected using LDA and further huddled into k posture ocular words, that characterize the archetypal postures of performances. The secular progressions of those optical words are moulded by detached buried Markov models (HMMs). Along with the same, because of the design of the sphere-shaped synchronized scheme and the vigorous 3D skeleton inference from Kinect, their process displays noteworthy vision invariance on our 3D action dataset. Our dataset comprises of 200 3D sequences of 10 indoor actions conducted by 10 individuals in distinguishing views. Our method is real-time and attains finer outlooks on the demanding 3D action dataset. We also checked our algorithm on the MSR Action 3D dataset and our algorithm outperforms Li et al. [20] in the maximum stances..
2. Hossein Rahmani in [6]proposes methods for efficient modelling of depth videos with particular emphasis on designing algorithms that learn the complex structures of human actions without making prior assumptions about the camera viewpoint, visual appearance, noise, and action execution speed. In the beginning of the thesis, two view dependent action recognition techniques are proposed. The first one coalesces the discriminative data from depth descriptions and 3D human joint positions to achieve high action recognition accuracy. To avoid suppression of subtle discriminative information and handle local occlusions, a vector of many independent local spatiotemporal features is computed. Two random decision forests (RDF) are trained for feature selection and classification respectively. The second method employs the Locality-constrained Linear Coding (LLC) to arrest discriminative information of individual’s actions in spatiotemporal subsequence of colour or depth videos. Both representations outperform existing techniques in terms of speed and accuracy on benchmark datasets.
3. Xuelu Wang in [7]**,** presents a framework for automatic recognition of human actions in uncontrolled, realistic video data with fixed cameras, such as surveillance videos. In his

work, he divided human action recognition into three steps: description, representation, and classification of local spatio-temporal features. The bag-of-features model was used to build the classifier. Fisher Vectors were also studied. We focus on the potential of the methods, with the joint optimization of two constraints: the classification precision and its efficiency.

4. Heng et al. at [8] pioneers a visual depiction based on impenetrable trails and activity frontier descriptors. Trajectories confine the restricted signal information of the video. A detailed demonstration ensures a thorough reporting of forefront movement along with the adjoining perspective. A state-of-the-art visual surge algorithm facilitates a vigorous and proficient withdrawal of dense trajectories. As researchers we take out qualities associated with the trajectories to portray contour (point coordinates), look (histograms of oriented gradients) and movement (histograms of optical flow). Furthermore, they commenced a descriptor on the basis of motion boundary histograms (MBH) that are dependent upon differential visual stream. The MBH descriptor depicts to constantly do better than other state-of-the-art descriptors, specifically on real-world videos that comprise of an important quantity of camera motion. they assessed a video illustration in the structure of action categorization on nine datasets, namely KTH, YouTube, Hollywood2, UCF sports, IXMAS, UIUC, Olympic Sports, UCF50 and HMDB51.
5. Ionut et al. [9], has put forward a very competent methodology to imprison the movement information inside the video. Their ways and means is dependent upon an uncomplicated sequential and spatial cradle, which takes into account the alterations amongst the two uninterrupted surrounds. The planned descriptor, Histograms of Motion Gradients (HMG), is corroborating on the UCF50 human performance acknowledgment dataset. The HMG channel with quite a few supplementary speed-ups is capable to attain real-time video dispensation and outperforms various famous descriptors including descriptors based on the costly optical flow [5].
6. Abdul-Azim; Hemayed [10] in their research makes us aware about an enhancement of trajectory-based individual deed appreciation, advances to arrest discriminative chronological linkages. In their advancement, they extorted trajectories by tracing the identified spatio-temporal curiosity areas named “cuboid features” with matching its SIFT descriptors over the successive outlines. They anticipated some connecting furthermore investigating technique to acquire proficient trajectories toward changing illustration

within pragmatic situations. Later these numbers approximating the trajectories' spots are portrayed to reflect human responses based on the Bag-of-Words (BOW) model. Eventually, a sustain vector mechanism is noted to categorize individual's responses. The efficiency for the projected advance was assessed on three famous dataset 'KTH, Weizmann and UCF sports'. Investigational outcomes revealed, the projected methods capitulate significant recital development over the state-of-the-art methods.

7. Chun ; Zang [11], proposed an advancement to realise an individual's activities on the basis of motion history image (MHI) and deep convolution networks. Firstly, part of frames from the starting and end of the motion visual tapes is separated and the gray MHI are hauled out from the remaining. Post that it is colorized into 3 channels RGB by the rainbow encoding. At the end, we coach a bottomless convolutional neural network to categorize the RGB MHI by managing a pretrained model. The investigational outcomes show that the projected technique enhances the acknowledgment correctness rate by 13% remarkably at.
8. Jiali et al. discussed in his research [12] that**,** indvidual's communications and their acts categories protect brawny associations, and the recognition of the communication organization is of noteworthy significance to progress the deed gratitude consequence. On the other hand, communications are naturally approximated with the help of heuristics or looked as latent variables. The previous habitually fabricates wrong communication constitution whereas the latter pioneers demanding guidance predicament. Therefore, they planned a construction to together study connections and activities by formulating an impending role by utilising both qualities learned via deep neural networks and human interaction context. They suggested an iterative viewpoint to resolve the connected conjecture predicament ably and with accuracy. Investigational outcomes on genuine datasets exhibit that the anticipated approach dramatically outperforms baselines by an unbeatable margin, and is cut-throat matched up to with the state-of-the-arts.

2.1 Problem statement

The approach used so far in the field of individual's acts acknowledgment uses the histograms of 3D joint locations (HOJ3D) [5] as a dense depiction of poses. We used to get this from the action depth series, which are re-predicted using Linear Discriminant Analysis method and then grouped together into k posture visual words, which represent the prototypical postures of actions.

These k posture visual words are then modelled by disconnected veiled Markov models (HMMs). The choice of HMM as a classifier was not proved to be quite effective as it was very complex from evaluation's point of view.

The other disadvantage of HMM was that it performs categorization on the basis of trained dataset and the teaching is performed on mixed dataset as it mixes the data and then trained it.

Thus, the testing of dataset in HMM [13] is proven to be effective for mixed dataset but it is not capable to handle the exceptional cases. Therefore, there is a need to develop such a modus operandi for human action acknowledgment that is less complex, and should be capable enough to handle the exceptional cases.

2.2 Motivation for the work

The objective of human activity appreciation is to prognosticate the marker of the activity of a person or a group of people from a visual instance. This fascinating matter is enthused by numerous applications, such as visual inspection, video intelligence, etc. Thinking in a huge sphere, an online visible inspection for an accepting action of a group of individuals will be of immense value for public safety: an automated video accepting module will be very powerful to mark numerous online videos.

Though, in different scenarios (e.g. Vehicle collision and unlawful movement), intellectual arrangements do not hold the opulence of foreseeing for the complete video previous to having to respond to the activity/action enclosed in it. We can become competent to forecast a treacherous driving circumstance earlier than it happens. This job is called act prognostication wherein the methodology that can recognize increasingly experiential video segments, and diverse action credit methods.

The images or videos detected by simple camera poses different problems like the problem of illumination, close-far problem, other noises etc. With help of Kinect cameras, we can clarify this process as they provide more details of the subject from simple RGB cameras.

2.3 Objective of the work

To implement multi layered ANN model as decision making approach in human action recognition system.

2.4 Thesis Organization

In the First Chapter, an introduction for activity recognition has been covered along with its different challenges and applications. In the Second Chapter, Literature Review in the field of Activity Recognition has been carried out. With the literature review, we had defined problem statement, the motivation of the work proposed and the thesis flow. In the Third Chapter, the block diagrams, a flowchart of the related work, proposed work and its methodology is discussed in detail. In the Fourth Chapter, all the results that are obtained from the work are covered and a comparison with the previous work has been conducted. In the Fifth Chapter, analytical remark to overall achievement, limitation of proposed work and scope for further research work in this domain is presented.

Chapter 3

Methodology

The chapter introduces the proposed work for recognizing activities of human, describes the setup being used. It explains the methodology involved in the proposed work and the brief description of each step.

3.1 Proposed Work

The traditional human action recognition system firstly takes and input video and then frame extraction is done. After extracting the frame the classification was done by using Hidden Markov Method i.e. HMM.

But after having a review to the existing work, the HMM was found to be less effective as a decision deriving model. Thus, by considering the shortcomings of HMM, the proposed work implements the multi layered ANN (Artificial Neural Network) for the purpose of classification.

The reason behind using the multi layered ANN is that it is quite advantageous than traditional HMM. First advantage is that it is less complex, second is that it performs training on the basis of single layer thus the issue related to the training of mixed dataset did not occur in this. Another advantage is that it is capable to handle the unexpected cases sometimes on the basis of the trained dataset.

The flow diagram can be explained as first and foremost training samples are collected for which test images are taken for. In our work we took all the samples of datasets and some of them we kept at reserved for testing. From these collected samples each skeleton joint is estimated and extracted out. We had evaluated these joints through histogram representation for binning purpose.

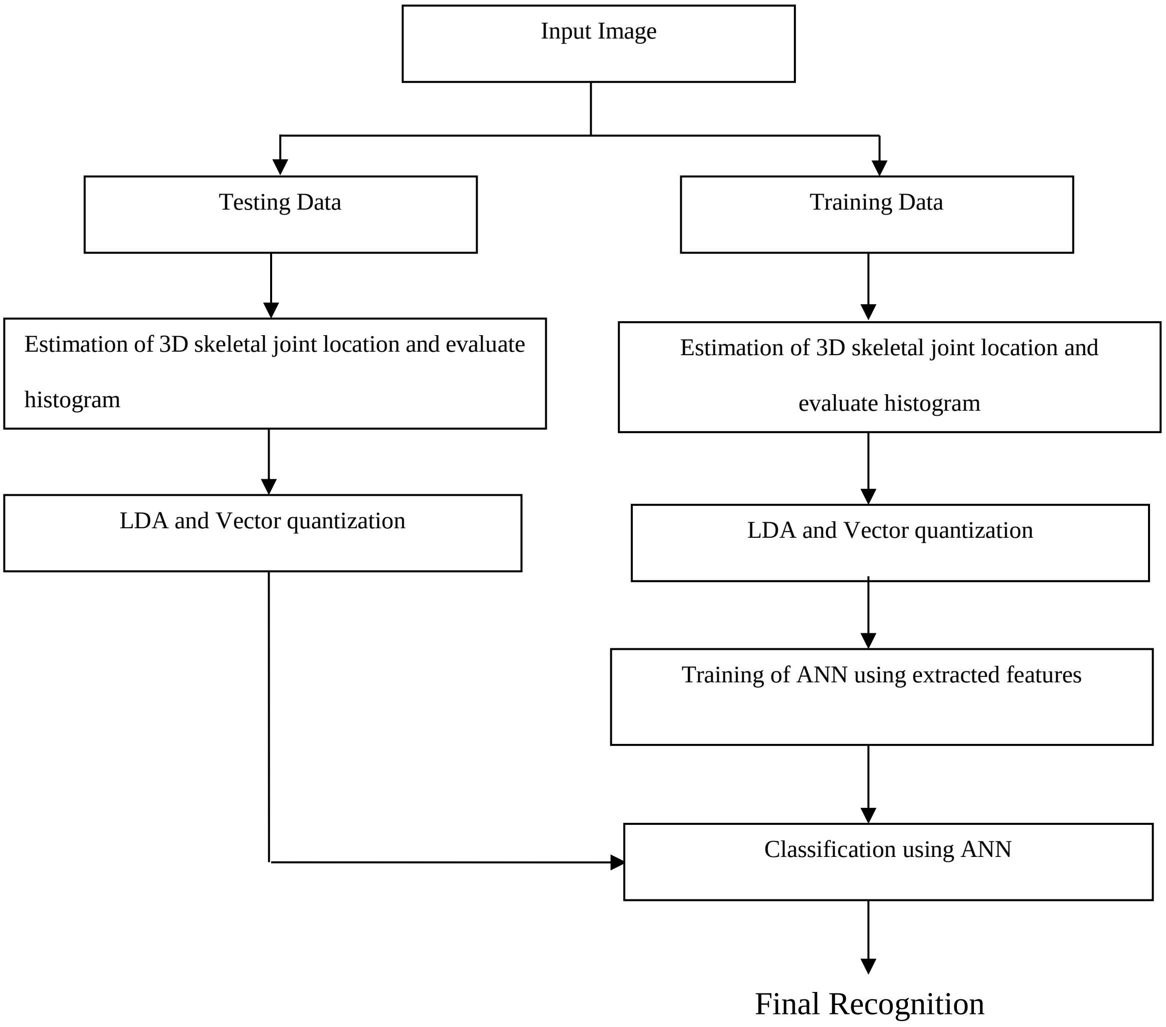

Input Image
Testing Data
Training Data
Estimation of 3D skeletal joint location and evaluate histogram
Estimation of 3D skeletal joint location and evaluate histogram
LDA and Vector quantization
LDA and Vector quantization
Training of ANN using extracted features
Classification using ANN
Final Recognition

Figure 6 Flow Diagram of proposed approach

As seen in the flow diagram these joints are grouped together via binning forms a feature for the methodology defined. These joints are grouped together to form feature vector and then appropriate quantization is performed. LDA is being employed to minimize the differences among similar classes. These feature vectors are being given as input to neural network that in turn gives the recognized results.

3.2 Methodology

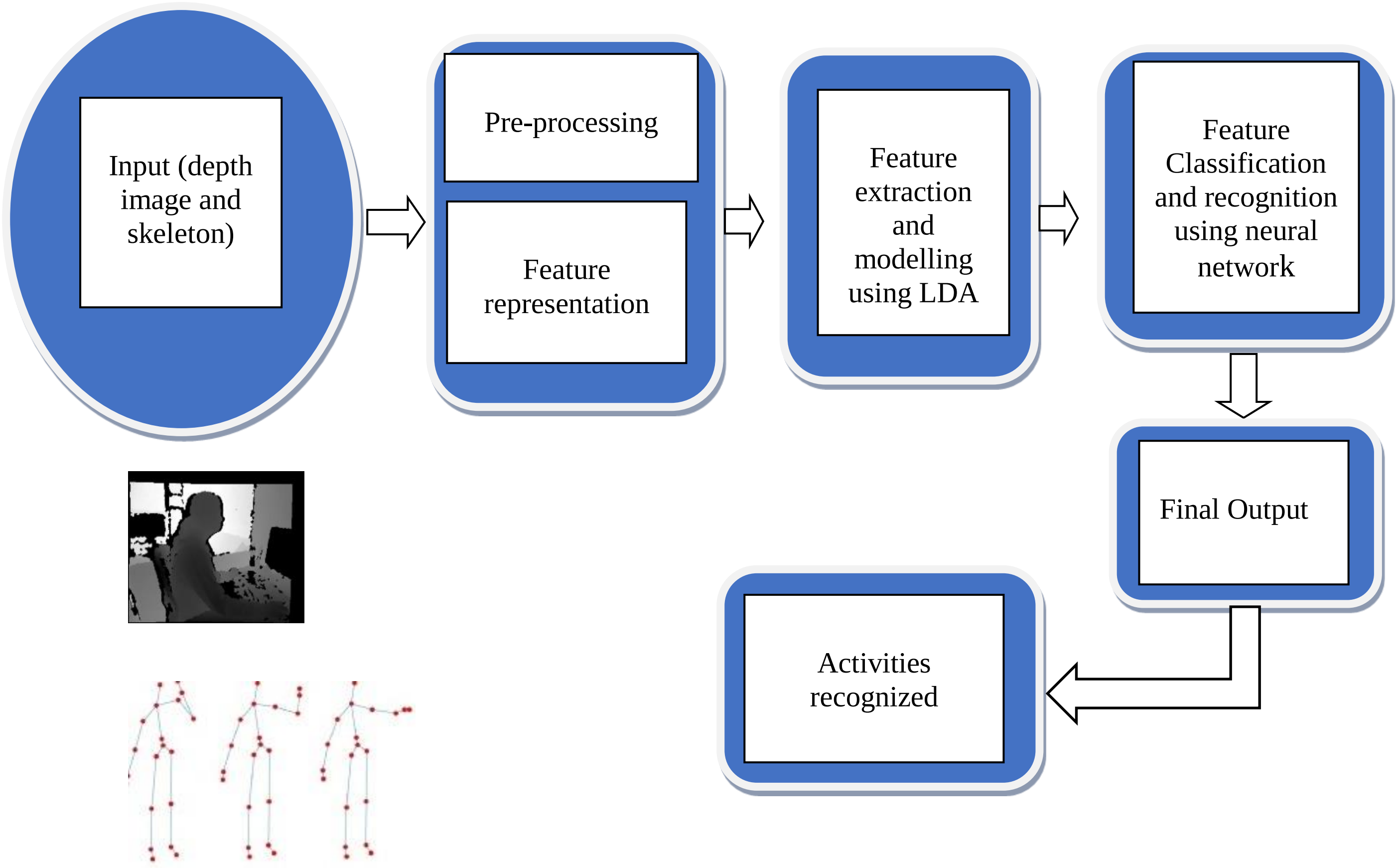

Input (depth image and skeleton)
Pre-processing
Feature representation
Feature extraction and modelling using LDA
Feature Classification and recognition using neural network
Final Output
Activities recognized

Figure 2 Block diagram of methodology

The block diagram of proposed work is described as-For the feature representation we have used skeleton and depth evidences. After the representation, appropriate features are extracted and modelled. LDA is used for modelling the features and at the last classification and recognition is done using artificial neural network.

3.2.1 Depth Silhouette Pre-processing (Depth image)

In order to get depth images of activities, we appropriated Kinect Camera [1]. Activity outlines are taken out from depth images and resized.

In the video based HAR, characteristically binary outlines [10] are taken for human activity depiction still they create imprecision in the middle of the same outline for diverse poses of diverse actions because of their restricted image value (i.e., “0 or 1”, therefore within the boundary of a posture, no information presents). The binary silhouettes produce restricted appreciation presentation because of some deficiency of information in the flat pixel concentration and hitches to categorize within the distant and proximate length of the individual body segments. For restraining the restraint of binary silhouette we take into account in depth based silhouette demonstration for human actions since depth silhouettes distinguish the body parts with the help of diverse concentration values. They reveal visible parts in accumulation to shape information.

3.2.2 Feature representation

3.2.2.1 Skeleton features

The step involves the identification of features from the interested images. As our purpose continues to probe what movement user is doing in an assigned time, we must tail actions of those body portions, that are commonly carried out while doing a specific activity.

There are many interacting systems built in the human body, out of which no system can operate in solitude. In specific, let’s commence from the musculoskeletal system [14], which is capable of supporting the human body and facilitating its actions in response to the incentive endow with the nervous system. To define one’s acts, we estimate tracking the human skeleton. The various portions regarding the human skeleton can be detailed as fragments correlated with each other over nodes, called joints, that restricts mobility of an individual’s body segment in the 3D term [2]. These joints are depicted below.

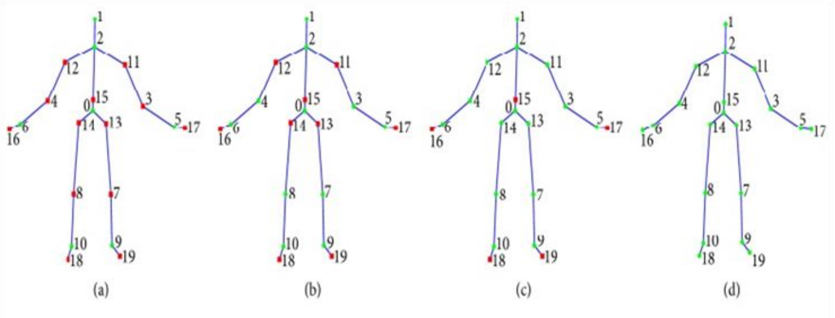


Figure 3 subset of joints

These joints are evaluated by simple mathematical model defined as-

Let J_i be the joints detected by means of the Kinect, the feature vector f is as

$$\mathrm{f} = \left[\mathrm{j}_1, \mathrm{j}_2, \mathrm{j}_3, \mathrm{j}_4, \mathrm{j}_{5,}\mathrm{j}_6, \mathrm{j}_7, \mathrm{j}_8, \mathrm{j}_{9,}\mathrm{j}_{10}, \mathrm{j}_{11}\right], \qquad (1)$$

where each $J_i$is the vector consisting of the 3-D normalized coordinates of the ith joint J_i captured by the Kinect.
So,

$$J_i = \frac{J_i}{S} + T \qquad \text{for all i ranging from 1-11} \qquad (2)$$

Being 's' the scale factor that normalizes the skeleton in accordance with the distance

$$s = \frac{\|j_4 - j_2\|}{h} \qquad (3)$$

h is the distance and T the translation matrix needed to set the origin of the coordinate system.

3.2.2.2 Posture Analysis

The plan of the system is to describe an action using a number of successions of some basic poses [15]. Each pose is the attribute vector f of Equation (i), i.e., the 3D coordinates of the joints constituting the skeleton. The central objective of posture choice stage is to choose prevailing and enlightening postures of various activities via clustering technique [12].

Consequently, sole intention is to arrange alike poses within definite clumps having comparable characteristics. Nonetheless, some movements may happen to be extra obscure compared to others, plus the fraction of clusters alter the aforementioned complexity. So, each system produces manifold prototypes concerning individual activity without defining the fraction of elementary postures that describe an activity. Accordingly, the input array is gathered in repetitious terms utilizing a diverse clusters (K) to cause various units factoring the same activity. The K-mean algorithm is employed various times to the input sequence, diversifying the aspired number of K clusters. In detail, given an activity composed of M posture features $[f1, f2, \dots, fM]$, the K-means delivers k clusters $[C1, C2, \dots, Ck]$, so as to minimize the intra cluster sum of squares

$$\text{argmin} \sum_{j=1}^{k} \sum_{fi \in Cj} \|fi - \mu j\|^2 \qquad (4)$$

where μj is a mean estimation of a cluster Cj.

At this instance, the collection of features which represents an action sequence is replaced with the centroid to which the posture feature fits.

Therefore, the centroid can be viewed as the principal pose of the activity. Besides, the posture-based design does not alter the potential concerning the model to discriminate amidst various actions including a difference in duration.

3.2.3 Feature Extraction and Modelling

Feature Extraction is the pith of the system, and it remains same for training as well as for classification phases. Its intention is to produce proper features which are applied in encoding an activity. The feature size must be scaled down, to increase the universality of the description including to lessen the complexity. It may happen several people perform the activities with varying velocities. There are many techniques used for processing feature generation and in this work, LDA is used.

3.2.3.1 Linear Discriminant Analysis (LDA)

LDA is a dimensionality reduction technique used for pattern recognition and data analysis in computer vision field. It presents differences among different classes of data by maximizing the separation between different classes and minimizing the scattering of data from the same class.

The approach of LDA is to adopt the resolution purposes to distinguish the features. For the two features $X1$ and $X2$, the decision boundary would be $d(X) = X1 - mX2 - c$. The basic idea is to model the plane such that it is capable enough to identify a particular feature. All the points at the decision boundary should satisfy the condition. This idea can also be extended for multiple features. Let represent the n-dimensional vectors. Let the number $d(X) < 0$ would identify a particular feature and $d(X) > 0$ would identify another feature. All the points at the decision boundary would satisfy the condition $d(X) = 0$. This idea can also be extended for multiple features. Let $x = (x1, x2, \dots \dots \dots ., xn)T$ represent the n-dimensional vectors. Let the number of classifiers be k. The problem is to assign an unknown instance x to any of the classes. This is done by designing k decision functions $d1(x), d2(x) \dots . dk(x)$.

The instance is categorised as class i and not j if

$$\mathrm{d_i(x) > d_j(x); \quad i \neq j \text{ for } i, j = 1,2,3,,,,,,k} \tag{5}$$

Then the decision boundary is given as

$$\mathrm{d_i(x) - d_j(x) = 0} \quad (6)$$

A decision rule can be designed as follows:

Assign the instance to the class I if d_{ij}>0 and assign the instance to j if d_{ij}<0.There are many ways in which the decision functions can be designed. The simplest method is to design a decision boundary perpendicular to the line that connects the mean of the class.

Activities can be repetitive sometimes so to decrease the redundancy LDA is very much effective. Due to the repetitive nature of activities sometimes there may be scarcer feature prototypes than the ones with more variations amidst key poses. LDA will give a new dataset holding the action features, whose occurrences have a weight that is improved if the same actor appears in the sequence. By the completion of each processing step, the input activity is realized with numerous new activity features produced from various sets of elementary clusters. The resultant samples contain a set of features. Let’s use a simple illustration to understand this concept. Say, an activity comprises of three clusters, a possible compressed sequence will be

$$A = \ [\ \mathrm{C1, C2, C3, C3, C2, C3, C2, C3, C2}] \quad (7)$$

A is composed of many activity subsets which can be described as

$$\begin{aligned} A_1 &= [C1, C3, C2, C3, C2] \\ A_2 &= [C3, C2, C3, C2, C3] \\ A_3 &= [C2, C3, C2, C3, C2] \quad (8) \end{aligned}$$

From programming point of view LDA can be implemented by basic following steps

Step 1-Computation of mean vector

Mean vectors mi where $(i = 1,2,3 \ldots \ldots.)$ of different action classes is calculated.

Step 2- Scatter Matrices computation

Within class scatter matrix Sw

It is can be obtained from-

$$\mathrm{S_w} = \sum_{i=1}^{c} S_i \quad (9)$$

and

$$\boldsymbol{S_i} = \sum_{x\in D_i}^{n}(\boldsymbol{x} - \boldsymbol{m_i})(\boldsymbol{x} - \boldsymbol{m_i})^{\boldsymbol{T}} \tag{10}$$

Si=scatter matrix for ith class and mi is the mean vector

$$\mathrm{m_i} = \frac{1}{\mathrm{n_i}\left(\sum_{\mathrm{x}\in \mathrm{D_i}}^{\mathrm{n}} \mathrm{x_k}\right)} \tag{11}$$

a) Between class Scatter Matrix

This is computed by following equation

$$S_B = \sum_{i=1}^{c} N_i(m_i - m)(m_i - m)^T \tag{12}$$

Where m is an averaged mean and m_i and N_i are the sample mean and sample sizes of respective classes

Step 3- Solving the generalized Eigen value problem

To obtain linear discriminant we solve the generalized Eigen value problem $\mathrm{S_{w-1}}S_B$.

Step 4- Selection of linear discriminant for new feature sub-space

After solving the Eigen values we will select the discriminant that will make new feature sub-space

a) Sorting the Eigen vectors by decreasing Eigen values

We don’t want mere projection of data into a subspace with reduced class separability rather we want a subspace having reduced seperability clubbed with reduced dimensionality of feature space.

However, the Eigenvectors only gives the direction of the new axis of separation, as they all have the same unit length.

Hence to decide which Eigenvector(s) we want to consider and which not for our lower-dimensional subspace, we need to look at the corresponding Eigen values.

The Eigen vectors having lowest Eigen value holds least information about the distribution of data, and those are the ones we want to reject. The common approach is to arrange the Eigen vectors in decreasing order corresponding Eigen values and choose the one having highest value because that will be holding most of the information.

Step 5- Selecting k Eigenvectors with the largest Eigen value

After arranging in descending manner, it is now time to construct k*d dimensional Eigen vector matrix W and thereby reducing the dimensionality of feature sub-space.

Step 6- Transformation of instances onto the new sub-space

Samples or instances can be transformed onto the new sub space by applying the below mentioned equation-

$$Y = X \times \mathrm{W} \tag{13}$$

Where $X = n * d$ dimensional matrix representing n samples
Y=transformed $n * k$ dimensional samples in new subspace

3.2.4 Activity classification and recognition

After feature extraction and posture analysis the next work is to generate feature generation. These generated features are fed to activity recognition system which is the final step.

Activity classification and recognition will give the output with the label of recognized activity along with each actor. There are many techniques employed for recognizing actions like SVM [16], Fuzzy logic, Neural Network [11], SVM with HOG [17] and many other. Out of these Neural Network is the recent advancement in the field of computer advancement.

3.2.4.1 Neural Network

A neural network is a group of algorithms that attempt to recognize underlying relations in a set of data through a process that imitates the way the human brain amputates. Neural networks [11] can accustom to varying input hence the chain makes the genuine possible result without demanding to re-considering the output criteria. A neural system is mentioned as “connectionist” mathematical system. The computational systems are sequential in nature; a curriculum gets to commence by the opening line of code, then execution, and it progresses on to the subsequent, trailing up the directions in a linear fashion.

The neurons, which are very simple are considered as individual elements of the network.

Neural network has many kinds depending upon its usage and complexity of dataset to be analysed. For my work I chose Artificial Neural Network which outperforms the other techniques.

Artificial Neural Network

For time-varying data, the artificial neural network is an interesting and effective approach. ANN stands as a pretty prevailing arithmetical framework for prototyping posterior probabilities given an assortment of samples (the input data). The fundamental structure of an (artificial) neural network (ANN) is the neuron. A neuron is a processing part which has various (commonly more than one) inputs plus single output.

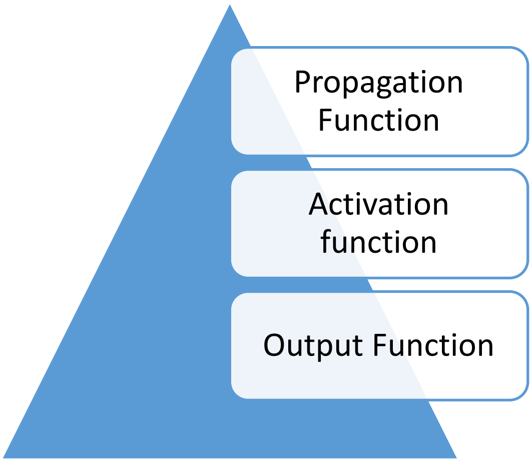


Figure 4 data processing steps

The above shows the flow of data processing via neural network

Data input of other neurons is fed to above structure which further processes as

Propagation Function-it is the weighted sum, it transforms the output of other neurons to net input.

Activation Function-it transforms the net input to new activation

Output Function-it transforms the activation function to output for other neurons.

First each input xi is weighted by a factor wi and the whole sum of inputs is calculated as

$$\sum_{for\ all\ inputs} w_i x_{i=} a. \qquad (14)$$

Then an activation function f is applied to the result a.

We take neuronal output as f (a). Generally, the artificial neural network is constructed by arranging the neurons in the arrangement of layers and make a nexus of the products of neurons of one layer to the inputs of the neurons of the subsequent layer.

Consequently, for computing the output, an activation function is applied to the weighted sum of inputs:

$$total\ inputs = a = \sum_{for\ all\ inputs} w_i\, x_i \qquad (15)$$

$$output = activation\ function\left(\sum_{for\ all\ inputs} w_i\, x_i\right) = \quad f(a) \qquad (16)$$

Back propagation network

The back-propagation network depicts one of the most classical example of an ANN, and it is also one of the most simple in terms of the overall design.

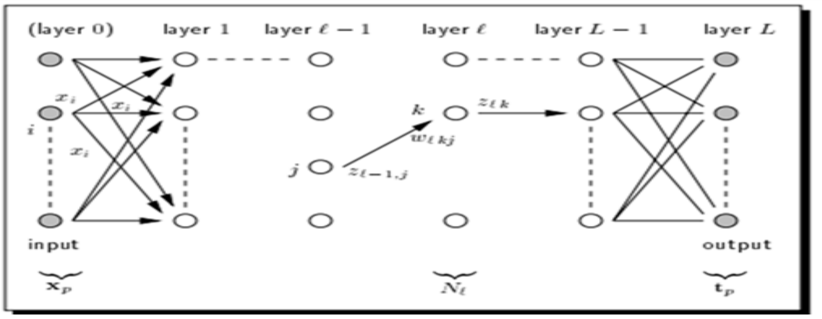


Figure 5 back propagation network

The Algorithm

The algorithm is based on discrete time approximation, i.e. time is $t = 0,1,2, \ldots$

A. The activation, error functions and the stop condition are supposed to be determined and fixed. Network running procedure:

1. The input layer is initialized, i.e. the output of input layer is made to be x;

$$z_o \equiv x \qquad (17)$$

for all layers l=1;L-starting with first hidden layer

$$\mathrm{z_l} = \mathrm{f(W_l z_{l-1})} \qquad (18)$$

2. The output layer is obtained from the output of the network, i.e.

$$y = z_l \tag{19}$$

B. Network Training mode:

1. Initialise all weights with small random values.

2. During every training set as long as stop condition is not sufficed:

Run the network to find activations on all neurons a_l and then the derivatives $f'(a_l)$.The network output will be required on next step.

$$Output - y_p \equiv zl(xp) = f(al) \tag{20}$$

- Using (yp, tp) ,calculate gradient

  Eg. for sum of squares it is computed by

  $$\nabla_{zl} E = z_l(x) - t \tag{21}$$

- Compute the error gradient.
- Update the W weights according to delta rule
- Check the stop condition and exit if they have been met.

During most events, a satisfying performance is gained when training is done regularly with the whole training set. A shuffling from patterns is recommended, between repeats. These all are the principles involved behind the methodology of the proposed work. Each step has its significance and played an important role.

Chapter 4

Results

In this, we will detail about the outcomes that we get by processing above mentioned steps. We have tested our algorithm on two 3-D activity datasets: Kinect Activity Dataset, MSR activity dataset [18].

The algorithm performance is being measured by two publicly available datasets. In order to perform an objective comparison to earlier tasks, these indication trial modes have been accounted for every dataset. The presentation signs are assessed via four diverse subsets of joints [19]. Lastly, so as to appraise the presentation in practical situation, some acceptable actions linked to it are picked up from the datasets, and the identification approximacies are estimated.

Our result shows, framed outlook, delivers exactness above the existing one technique and approaches. It outperforms the other state of artwork.

4.1 Experimental setup

The Kinect camera is being used in getting the datasets and in its evaluation. We took two datasets and we executed our algorithm on these datasets and compare its performance with the existing approaches.

4.1.1 MSR-Activity Dataset

One of the most used datasets in human activity recognition is MSR Activity dataset [2]. 20 activities are included in this dataset and these activities are being performed by many individuals separately. Many sequences of depth (320 × 240) and skeleton frames are presented, but 10 of them have to be abandoned because the skeletons are either not present or influenced by numerous errors. High throw, standing, hand clapping, horizontal arm waving sitting, standing upright are some of the activities included in this dataset.

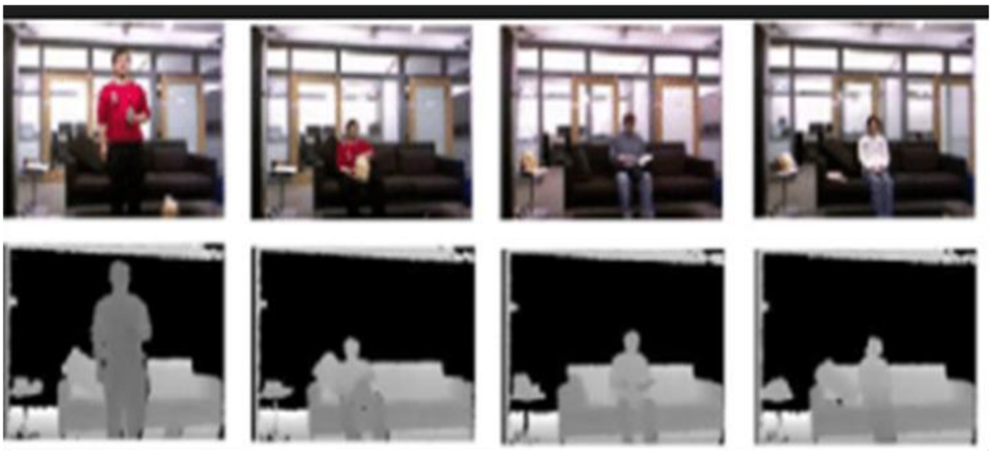


Figure 7MSR activity dataset

The samples of this Dataset are being processed by the proposed algorithm and the table showing the comparison of recognition accuracy from the existing approaches with the proposed one are discussed in the following section.

4.1.2 Kinect Dataset

There is another dataset which is commonly used in depth analysis of human activities. That dataset is UT-Kinect depth dataset [1]. The dataset holds 10 different individuals performing individual activities for a couple of time. The activities that the dataset holds are-clapping hands, waving, pushing, throwing, standing upright, picking up, walking, and lying down. The dataset provides 20 skeleton joints involved in performing activity.

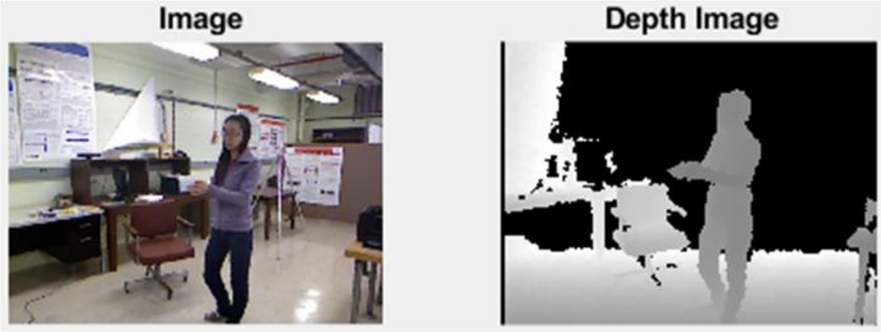


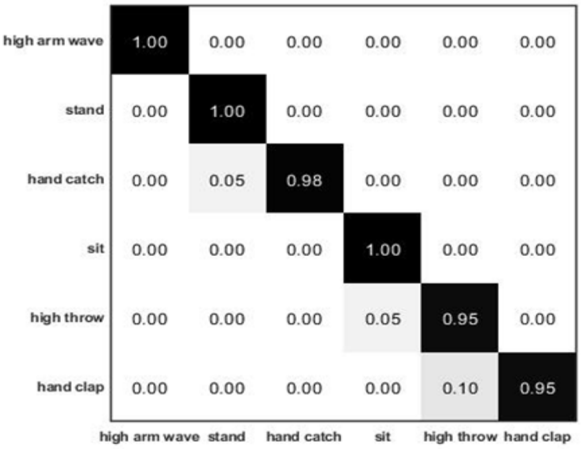
Figure 8: UT Kinect dataset images

4.2 Performance Evaluation

The performance of recognition is measured by confusion matrix. We have obtained confusion matrix for each dataset representing different activities.

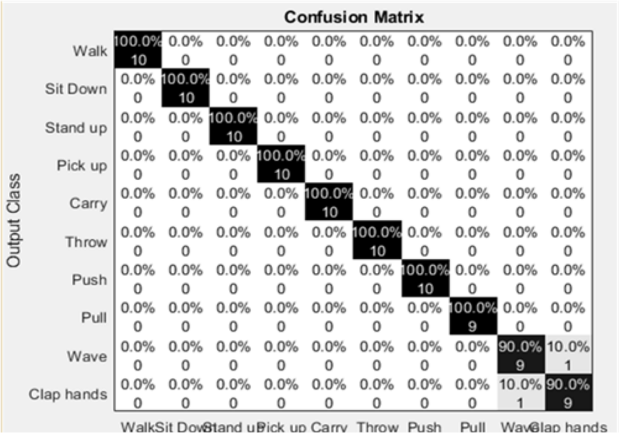
The accuracy of confusion matrix is computed by following parameter

$Accuracy = (\frac{true\ positive + true\ negative}{true\ positive + true\ negative + false\ positive + false\ negative}) * 100.$ There are many ways to compute the recognized results. Confusion matrix is one of that ways. Confusion matrix that we get for both the datasets are shown below-



	high arm wave	stand	hand catch	sit	high throw	hand clap
high arm wave	1.00	0.00	0.00	0.00	0.00	0.00
stand	0.00	1.00	0.00	0.00	0.00	0.00
hand catch	0.00	0.05	0.98	0.00	0.00	0.00
sit	0.00	0.00	0.00	1.00	0.00	0.00
high throw	0.00	0.00	0.00	0.05	0.95	0.00
hand clap	0.00	0.00	0.00	0.00	0.10	0.95

Figure 9 confusion matrix of MSR Activity Dataset



Confusion Matrix

Output Class

	Walk	Sit Dow	Stand u	Pick up	Carry	Throw	Push	Pull	Wave	Clap hands
Walk	100.0% 10	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0
Sit Down	0.0% 0	100.0% 10	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0
Stand up	0.0% 0	0.0% 0	100.0% 10	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0
Pick up	0.0% 0	0.0% 0	0.0% 0	100.0% 10	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0
Carry	0.0% 0	0.0% 0	0.0% 0	0.0% 0	100.0% 10	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0
Throw	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	100.0% 10	0.0% 0	0.0% 0	0.0% 0	0.0% 0
Push	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	100.0% 10	0.0% 0	0.0% 0	0.0% 0
Pull	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	100.0% 9	0.0% 0	0.0% 0
Wave	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	90.0% 9	10.0% 1
Clap hands	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	0.0% 0	10.0% 1	90.0% 9

Figure 10 Confusion matrix of UT Kinect dataset

By implementing our proposed methodology, we obtain the above depicted confusion matrices. These confusion matrices show that most of the samples that we fed to our designed model are correctly recognized. For MSR Activity dataset three activities-high throw, high catch and high clap are mis-detected with other activity giving an accuracy of 98%. On the other hand, for UT-Kinect dataset two activities-wave and clap hands are mis-detected with the wrong ones thereby giving 98% accuracy.

Now we compare our proposed process flow with the existing approaches. The comparison is listed below through table defined as-

Table 1 Table of comparison of proposed approach with existing methods for MSR Activity Dataset

Method	**Features**	**Classifier**	**Accuracy**
HOJ3D [4]	Nodes of action graph	HMM	90.92%
Actionlet Ensemble [20]	Concatenated Fourier coefficients of 3D joints	SVM	82.22%
3D Posture [2]	Joints of skeleton evaluated by histogram representation	HMM	77.3%
HON4D [21]	Super normal vector by hyper plane separation	HOG	81.88%
Proposed method	Skeleton joints clustered by histogram representation	LDA with KNN	**98%**

Besides the confusion matrix, our methodology outperforms many existing approaches. We have demonstrated the comparison of proposed methodology for each dataset i.e. MSR activity dataset and UT-Kinect dataset. The comparison for MSR Activity dataset is defined in table 1 and for the other dataset we have used table 2. [4] in his approach proposed to describe bag of points (vocabulary of postures) as histogram of 3D joints to describe the important postures. But their scheme was view dependent and on the other side we have proposed a method which is view independent. [2]has overcome the problem of HOJ3D method and proposed a method which deals with time invariance if the person interacting object/person. This method used HMM which is based on probabilities and also giving less performance than our proposed approach. Some accuracy has been approved by [22] but due to limited capability of HOG, it could not recognize all actions with precision. There our approach proves to be better than all the algorithms.

Table 2 comparison of accuracy with exsting approach of UT Kinect dataset

Method	**Features**	**Classifier**	**Accuracy(%)**
DSTIP [1]	Spatio-temporal points using corner detection algorithms	SVM+Histogram intersection kernel	85.8 %
Random forest [18]	Spatio-tempral features	Random-forest classifier	87.90 %
SNV [21]	Polynormal vector using hypersurface-normals	SVM Libliner	88.89%
HOJ3D [5]	Skeleton joints using histogram	LDA with HMM	90.92%
Proposed method	Skeleton features	LDA with KNN	98 %

For the Kinect dataset we have compared the result of our methodology with the existing approaches [23]. In [21] ,he proposed a clustered technique called super surface normal. He made groups of hypersurface normal vectors in a depth sequence to form the polynomial used to distinguish the local motion and shape information. He had used the SVM Lib-linear for classification. [18]represented a novel approach using random forest classifier for detecting actions out of the depth instances. There are many plus points of random forest classifier like ineffectiveness to noise, efficiency for classification and the improvement in detecting the correct instances but it failed to deliver the desired performance. In [1] spatio-temporal points are extracted using corner detection algorithms. These features are clubbed together to derive depth cuboid similarity feature vector. Though the algorithm proposed is very useful as it is independent of skeleton information but it failed at identifying some instances. Our method outperforms all the methodologies depicted by many researchers.

From the comparison we infer that our methodology is counter performing the existing approaches [24]. Thus, it is said to be the optimal algorithm for detecting actions performed by different persons and it also shows good precision while detecting in the dark. This is due to the fact that we are using the Kinect camera which is a good at capturing things in the dark.

Chapter 5

Conclusion and future work

5.1 CONCLUSION

The above confusion matrices lead to us a result of 98 % which proves to be an efficient than the existing works. The standard datasets give more accuracy than expected. Neural network or any other classification methods gives more accurate results when the samples or classes are more as it gives them an opportunity to recognize the things with greater precision.

We had successfully managed to test a framework of human activity recognition using depth and skeleton evidences. Particularly, we took a scenario which include an environment equipped with many unwanted things such as door, table, pen, book etc which are present in the closed room. During the designing model, the Kinect is taken into consideration for capturing high-level data concerning what the user is executing.

This work exhibits a method to represent action/activities as a series of joints detected by Kinect Camera. In this, depth data is used instead of RGB dataset. The depth dataset gives more noticeable information and the problem of illumination is also removed. Human postures are characterized as histograms of 3D joint locations within a modified spherical coordinate system. The feature vector is generated using LDA and then this feature vector is served to a neural network. Our proposed algorithm makes use of back propagation two-layer network which successfully recognize actions performed by different actors.

We trained a neural network to analyze sequential postures into action types. The principal elements of our algorithm are real-time, which include the extraction of 3D skeletal joint locations, using HOJ3D and LDA techniques for feature vector generation, and classification by artificial neural network. The main advantage is the algorithm deals with time invariance representation.

Finally, the research work is concluded and future research direction as well as possible future applications are highlighted and discussed.

5.2 Future Scope

Kinect camera proves to be an advantageous factor than RGB camera as it removes the illumination factor, and gives more prominent information of targeted image or video. Although it is beneficial in the domain of computer vision, it holds some limitations like it captures the date only indoor which is the biggest disadvantage for recognition. The outdoor activities are difficult to detect as many occlusions occur like shadowing, partial occlusion by others and many environmental factors. A Kinect camera is unable to detect and classify the activities. The future work can be evolving the technique to detect and classify activities with better accuracy.

Usage of an artificial neural network is only good for recognizing or classifying simple level tasks but for complex tasks, we need a better version of a neural network. Convolution neural network and a recurrent neural network may be used for future work.

References

[1] J. Aggarwal and L. Xia, “Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera,” in *IEEE Conference on computer vision and pattern recognition,* 2013.

[2] S. Gaglio and M. Morana, “Human Activity Recognition Process Using 3-D,” in *IEEE TRANSACTIONS ON HUMAN-MACHINE SYSTEMS,* 2015.

[3] X. Yang, C. Zhang and Y. Tian, “Recognizing actions using depth motion maps-based histograms of oriented gradients,” in *ACM International Conference,* 2012.

[4] W. Li, Z. Zhang and Z. Liu, “Action recognition based on a bag of 3D,” in *ieee computer society on computer vision and pattern recognition,* 2010.

[5] X. Lu, C.-C. Chen and J. K. Aggarwal, “View Invariant Human Action Recognition Using Histograms of 3D Joints,” in *IEEE conference on computer vision and pattern recognition,* 2012.

[6] H. Rahmani and A. Mian, “3D Action Recognition from Novel Viewpoints,” in *IEEE,* 2016.

[7] X. Wang, “human action recognition from gradient boundary histogram,” *research gate,* 2017.

[8] H. Wang, A. Kläser, C. Schmid and C.-L. Liu, “Dense Trajectories and Motion Boundary Descriptors for Action Recognition,” *International Journal of Computer Vision,* 2013.

[9] I. C. Duta, J. R. Uijlings, T. A. Nguyen, A. G. Hauptmann, B. Ionescu, N. Sebe and K. Aizawa, “Histograms of Motion Gradients for Real-time Video Classification,” in *IEEE,* 2016.

[10] H. A. A. Azim and E. E. Hemayed, “Human action recognition using trajectory-based representation,” *Elseveir,* 2015.

[11] Q. Chun and E. Zhang, “Human action recognition based on improved motion history image and deep convolutional neural networks,” 2017.

[12] J. Jin, Z. Wang, S. Liu, J. Zhang, S. Chen and Q. Guan, “Joint label-interaction learning for human action recognition,” 2017.

[13] J. Yamato, J. Ohya and K. Ishii, “Recognizing human action in time-sequential images using hidden Markov model,” in *IEEE Comput. Soc. Conf. Comput. Vision Pattern Recog. Proc,* 1992.

[14] Y. Qin, L. Mo and B. Xie, “Feature fusion for human action recognition based on classical descriptors and 3D convolutional networks,” in *IEEE,* 2018.

[15] M. P. V. K. a. J. Heikkila, “Human activity recognition using sequences of postures,” in *Proc. IAPR Conf. Mach. Vision Appl.,* 2005.

[16] B. Schölkopf and Alexander, “Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond,” *MIT Press,* 2001.

[17] B. Lin and B. Fang, “Spatial-temporal histograms of gradients and HOD-VLAD encoding for human action recognition,” in *IEEE,* 2017.

[18] W. C. a. G. G. Y. Zhu, “Fusing Spatiotemporal Features and Joints for 3D Action Recognition,” in *IEEE Conference on Computer Vision and Pattern Recognition Workshops,* 2013.

[19] A. Oikonomopoulos, I. Patras and M. Pantic, “Spatiotemporal salient points for visual recognition of human actions,” in *IEEE Trans. Syst., Man, Cybern. B, Cybern., vol. 36, no. 3, pp. 710–719,* 2005.

[20] J. Wang, Z. Liu, Y. Wu and J. Yuan, “Learning actionlet ensemble for 3D human action recognition,” in *IEEE Trans. Pattern Anal. Mach. Intell.,,* 2014.

[21] Tian, X. Yang and YingLi, “Super Normal Vector for Activity Recognition Using Depth Sequences,” in *IEEE Conference on Computer vision and pattern recognition,* new york, 2014.

[22] J. Wang, Z. Liu, Y. Wu and J. Yuan, “Mining actionlet ensemble for action recognition with depth camera,” in *IEEE Conference on computer vision and pattern recognition,* chicago, 2012.

[23] Intille, L. Bao and S. S., “Activity recognition from user-annotated acceleration data,” in Pervasive Computing,” in *Lecture Notes in Computer Science, vol. 3001,* springer, 2004.

[24] G. Willems, T. Tuytelaars and L. Gool, “An efficient dense and scaleinvariant spatio-temporal interest point detector,” in *10th Eur. Conf. Comput. Vision,* 2008.