Activity Recognition using kinect v1 based CAD-60 Dataset

Rabia Asim

NUST201361608MPNEC45313F

Supervisor

Dr. Sajjad Haider Zaidi

DEPARTMENT OF ELECTRONICS AND POWER ENGINEERING
NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY
ISLAMABAD

AUGUST, 2017

Activity Recognition using CAD-60 dataset from kinect v1

Rabia Asim
NUST201361608MPNEC45313F

A thesis submitted in partial fulfillment of the requirements for the degree of
MS Electrical and power engineering

Thesis Supervisor
Dr. Sajjad Haider Zaidi

Thesis supervisor signature:________________________

DEPARTMENT OF ELECTRONICS AND POWER ENGINEERING
NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY
ISLAMABAD
AUGUST, 2017

I certify that this research work titled *Activity Recognition using CAD-60 dataset from kinect v1* is my own work. The work has not been presented elsewhere for assessment. The material that has been used from other sources, it has been properly acknowledged and referred.

Signature:__________
Rabia Asim
NUST201361608MPNEC45313F.

Contents

List of Figures

Abstract

The requirement for understanding human activities has gradually developed recently. It has a number of applications which include health-care services, particularly in senior citizen care establishments, disability rehabilitation, robotics, and assistive living and surveillance systems. A major number of assets can be saved by deploying these activity recognition systems. It can be used to detect abnormal activities in regard to surveillance and any irregular behavior for personal care as well. The recognition of a range of activities of daily living can expose important information about an individuals activity patterns. Many researchers have effectively recognized activities using wearable sensors like watch with high accuracy. Smartphones GPS data has also been used to identify activity patterns to identify anomalies in behavior over a long period of time. Smartphones have been acknowledged as a influential solution for activity recognition systems because of its increasing demand and user friendly aspects. 3-dimensional cameras are gaining much attention lately due to high accuracy we can achieve. It provides multiple sensors for image analysis in cahoots with trajectory data. These two aspects have been combined for recognition and classification. In our work, we have used only 3-dimensional joint position data for activity recognition. A system has been proposed to pre-process and extract features from CAD-60 dataset collected by Cornel University. We have normalized the data to make it position and person invariant. Each activity takes different time to complete, we have performed dynamic frame wrapping (DFW) to make all activities the same size without loss of relevant information. After DFW, we have performed Dynamic invariant joint removal. Almost all activities use some joints more than the other. For this reason, we have dynamically removed least variant joints from all activities. After invariant joint removal, we have applied histogram of oriented gradients on a single skeleton frame. Histograms are calculated with respect to a reference maximum gradient vector and bining proceess is done as usual my adding the magnitude

in one direction bin. Feature vector is then passed through Linear discriminant analysis (LDA). LDA is applied to maximize the ratio of across the classes spread and inside-class spread of data. The LDA algorithm searches for the vectors in the principal space to construct the best discrimination between different classes. Algorithm is tested using Support vector machine (SVM) for classification and has produced promising results. We used 5-fold cross validation and received precision over recall rate of 97.222%/97.222%. With new person metric

Chapter 1

Introduction

1.1 Motivation

Human activity analysis has been studied keenly since the 80s. It has a number of significance in fields of health-care systems, Surveillance, Robotics and smart homes and offices. Health-care related uses of activity recognition can be seen in nurses assistance systems and rehabilitation centers. The proportion of aged people in todays society is continuously growing. As a consequence, the problem of supporting older adults in loss of cognitive autonomy who wish to keep on living on their own in their home as opposed to being required to live in a hospital has been one factor in studying human activities to identify injurious behavior, fall detection systems and also to monitor diabetic patients. Smart settings have been developed in order to offer support to the elderly people with safety issues wishing to maintain an independent lifestyle. When it comes to surveillance, some time ago, airport and mall security was carried out by human beings scrutinizing human behavior. Now systems are being developed that can look for aggressive behavior in public areas to provide security. In robotics, human activities are analyzed and classified in order for robots to mimic behavior. These are just a few applications of human activity recognition systems that has motivated us to develop an improved system

1.2 Problem Statement

In the field of kinect based human activity recognition systems, numerous researchers have come up with novel algorithms and systems and have carried out experiments on public data sets bearing in mind accuracy of proposed system and computational complexity. Nevertheless, it is an emerging field which requires further improvement. The existing techniques on kinect based CAD-60 dataset have achieved accuracy up-to 94.5% leaving a margin for improvement. Previous work shows a fusion of image processing and stick figure modeling. We have proposed a single stick figure approach on spatio-temporal data, reducing the complexity and increasing the processing speed. The approach of Histogram of Oriented Gradient has been improved by using more discriminative features. The refined objective of the thesis is:

The primary objective of this thesis is to develop a 3-dimensional simple and accurate human activity recognition system that will increase efficiency as compared to previous similar systems. The proposed method will be tested using CAD-60 data-set.

1.3 History of activity analysis

Human activity recognition has been studied keenly since the 80s using a variety of sensors including 2D images and video analysis. In its most basic form, the aim of activity recognition system is to analyze and identify actions automatically and sometimes purpose of one or more subjects from a number of observations taken from a segmented video. Such a system can offer personalized support for many applications and it is used in a number of fields of research such as health-care, robotics, or sociology.

In health-care, activity recognition has been employed as a nurses assistance system to monitor patient conditions. Or in elderly home, it is used as a fall detection system with an alarm for quick response. Activity recognition is also necessary

for surveillance and other monitoring systems in public areas, for example shopping malls, hospital and airports. These systems are also used in intelligent homes and workplaces.
Aggarwal in [3] divided human activities based on their level of complexity into four distinct groups: gesture, action, interaction, and group activity. Gestures are basic movements performed by a human being, and are elemental motions. Waving of an arm and nodding of head are some common gesture types. Actions are single-person activities that can be an accumulation of more than one gesture movement performed for some time like, walking, jumping, and punching. Interactions are human activities that require two or more people or an object/person interaction. For example, badminton is an example of interaction in between individuals and a person carrying a mug is a human-object interaction which involves one human and one object. Finally, group activities are the activities done by groups comprising of a number of people and articles: A marching band or a business meeting are typical examples.

1.4 Human Activities

Humans perform various activities during a day. It is essential to label those activities into sets in order to develop a human activity recognition system. Aggarwal in [3] divided human activities based on their level of complexity into four different levels: gestures, actions, interactions, and group activities as shown in figure 1.1. Gestures are basic human body-part movement,and and for basic elements

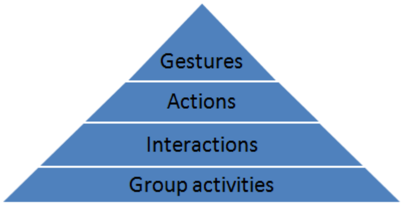


Figure 1.1: Human Activity Pyramid

which signify he meaningful motion of a person. Waving an arm and nodding of head are good cases of gestures. Actions describe single-person activities that can be an accumulation of more than one gesture movement performed for some time like, walking, jumping, and punching. Interactions describe human activities that require two or multiple individuals or an object/person interchange. For example, tennis is an interaction between two individuals and tennis-ball is a human-object interaction including one human and one object. Finally, group activities are the actions and motions done by groups comprising of many individuals and/or objects: A marching band or business meetings are typical examples.

Aggarwal divided activity recognition aproaches into two categories: single-layered methods and hierarchical methods. Single-layered methods are those that signify and classify motion straight forwardly based on concatenated images. Due to their characteristic, single-layered approaches are appropriate for the identification of gestures and actions with chronological properties. On the other hand, hierarchical methods characterize high-level human activities which stipulations of simpler activities, and are called sub-activities. For such activities, recognition systems are a collection of multiple layers enabling them for the analysis of complex activities.

These two layers are further divided into sub categories as shown in figure 1.2. Single-layered methods are further divided into two classes based on how they design human activities: that is, space-time and sequential approaches. Space-time approaches examine an input video as a 3-D (x,y,t) volume, which is divided into three more classes depending on the features they utilize from the 3-D space-time volumes: volume, trajectories, or local nterest point descriptors. Sequential methods on the other hand construe it as a sequence of observations. Space-time methods are classified by using exemplar-based methods or model-based methods. Hierarchical methods are classified based on the use of statistical approaches, syntactic approaches, and description-based approaches. Statistical approaches create state-based models. High level human activated is represented by concatenated hierarchy (e.g., layered hidden Markov models). Syntactic approaches use a gram-

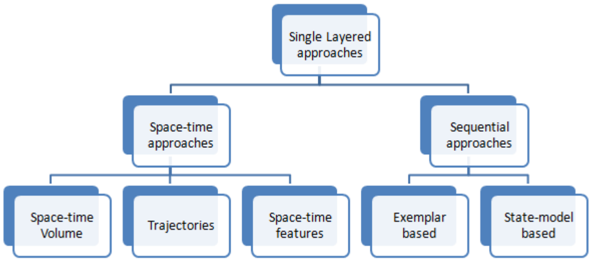


Figure 1.2: Human Activity Recognition methodologies: Single-layered approach

mar syntax model sequential activities done by modeling high-level activities as a sequence of atomic activities like gestures. Description-based approaches characterize human activities by defining sub-events of the activities and their time, space, and logical structures. Hierarchial methods are summarized in figure 1.3.

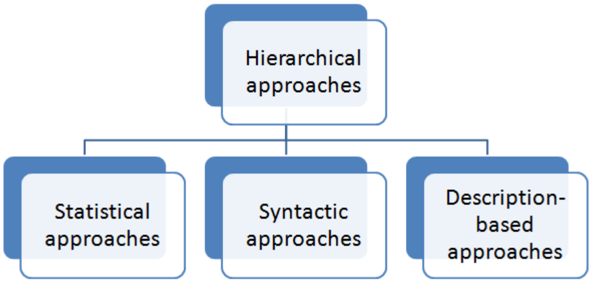


Figure 1.3: Human Activity Recognition methodologies: Hierarchial approach

1.5 Activity recognition sensors

The first step in the employment of an activity recognition system is the sensing of the activities. There are a number of sensors available commercially and the choice of the suitable sensor plays a critical role in the efficiency of the system. There is a compromise between the cost and effectiveness of sensor choice. To capture a broad variety, sensors can be of two types depending on its placement and how it interacts with the user: the ambient sensors and the wearable sensors [4]. Based on extraordinary advancements of 3D sensors, a lot of research focuses on the use of 3D cameras like kinect. Wearable sensors and smartphone have also become household item which has acquired a lot of attention from researchers in activity recognition systems.

1.5.1 Wearable sensors

Wearable sensors are becoming increasingly favoured in many areas such as healthcare, leisure, security and on commercial side aswell. They are practical in providing precise and consistent information on peoples activities and behavioral patterns. This ensures a safe and sound living environment. Wearable sensors in the shape of panic buttons in time of crisis have been used since a while now and are a commercial favourite [5].

The fast expansion of microelectronics and other associated technologies has facilitated the progress of a variety of of smart sensors measure data fast and efficiently, with lesser energy utilization and less processing power. Body temperature measurement is one of the general physiological measurement calculated by wearable sensors for human activity monitoring. It can detect signs of stress based on variation in skin temperature which can cause a number of health conditions like heart cardiac arrests, stroke and shock.

Other wearable sensors include accelerometers. They are frequently used in monitoring of human activity and measure acceleration along a responsive axis and over a particular range of frequencies. They are being employed in many fields like fall monitoring and detection [6] [7], analysis of motion [8] [7] or a subjects posture based orientation [9] [10]. Smartwatches with embedded accelerometers been in market for some time now and are also available for children for monitoring purposes

Wearable ElectroCardiogram (ECG) sensors are also common for a short time-period evaluation of cardiovascular diseases, particularly for people with chronic heart problems. The ECG signal provides helpful information about the rate and regularity of heart beats, which are used in diagnosis of cardiac diseases [5]. Smartphones have been accepted by researchers as a powerful solution for sensing applications since it includes a number of advantages that can defeat a number of current problems in the field [4] [11]. Smart-phone based sensing systems have been developed in the area of health monitoring, environmental monitoring, traffic monitoring, human behavioral monitoring and social networking [12]. The remarkable number of smart-phone users, the homogeneous worldwide architecture, the wireless network, the size of the device and lastly the assortment of the embedded sensors [11] makes the use of smartphone a priceless sensing device for activity recognition. Wearable activity recognition systems have a number of known issues in relation to their general applicability, most notably these include: (a)sensitivity to sensor placement and (b) the need for annotated datasets for supervised learning approaches.

1.5.2 Ambient Sensors

External sensors are able to sense several physical phenomena in various environments [13]. External sensors are generally used in the framework of smart spaces, homes and buildings. As per requirement of desired outcome, there is a broad variety of sensors which can be used such as video cameras, depth sensors, microphones, presence sensors, radio frequency identification (RFID) tags and thermometers [4]. There are a number of ambient sensors in the market available for commercial use that can provide us with relevant information for guessing human activity, such as accelerometers, video cameras, etc., but in recent years, the easily accessible revolutionary sensor developed particularly for tracking humans, as is the Microsoft Kinect has unfolded new research areas. In 2010, Microsoft uncovered a new Xbox 360 accessory they claimed would reinvent the way gamers played. Kinect was one of a kind motion sensor which excluded the need for a contemporary controller, instead permitting players to control the system with their body movements and hand gestures change the game. Microsoft Kinect v1 comes with three sensors in one device. A microphone, an RGB camera and an

Infra-red camera for depth perception. This has made it very popular in Human activity recognition research area.

1.6 Thesis Outline

The main objective of this thesis report is to develop a system that increases efficiency on CAD-60 data-set. Investigate performance of techniques used we developed to extract a feature vector for an action for classification using support vector machines(SVM). This technique has shown great promise and can be used to develop a recognition system for human activity recognition in changing environments. Our system is verified by using a standardized dataset CAD-60, developed by Cornel University. Dataset consists of eighteen activities. Each activity is performed by four different subjects in five different environments. Section 2 gives a thorough literature review of Action Recognition systems using variety of feature extraction methodologies and a number of different classifiers. Different sensors, from wearable sensors to smartphones to ambient sensors have been reviewed in context of activity recognition and their efficiency has been discussed. CAD-60 dataset has been explained in detail. its specification and results driven from it are also added. Next we explain the feature extraction methodology. We have used a trajectory-base method for activity recognition. Our activity recognition system comprises of several stages, which include 1) data acquisition 2) preprocessing 3) feature extraction and 4) classification. Data is acquired by Cornel university and is available publicly for research. Data is acquired using kinect version 1 sensor. Pre-processing involves a number of stages. It makes the data less redundant and fulfills the basic machine learning requirement of taking care of 'garbage in, garbage out' analogy. First we normalize the data to make location and skeleton size invariant. Next we perform dynamic frame wrapping. Each activity performed takes different time depending on the person and activity itself. In order to make all activities same size, great care must be involved in frame wrapping technique as to not loose any pertinent information. Last step in pre-processing in Dynamic invariant joint removal. Any activity performed does not require all joints to participate, for example, in brushing teeth activity, knee joints can be removed before feature extraction as they show minimal movement and can negatively effect the feature vector by having redundant zero information. Before last stage of classification. We have performed linear discriminant analysis (LDA) on our feature vector. LDA is applied to maximize the ratio of across the classes spread and inside-class spread of data. The LDA algorithm searches for the vectors in the principal space to construct the best discrimination between different classes. The LDA algorithm searches for the vectors in the principal space to formulate the ideal discrimination between multiple classes. by which we can achieve a more vigorous feature space that divides the feature vectors of all classes. SVM classifier is used for activity classification. Our results and previous results are compared for analysis

Chapter 2

Literature Review

Identification and recognition of human activities from a series of action is the goal of every human activity recognition system. A number of researchers have studies and published papers supporting the emerging need of activity recognition systems, using numerous approaches based on vision sensors, inertial sensors, smartphone sensors and a combination of the aforementioned. Activity recognition is currently being applied in a number of domains such as health-care, surveillance, human-computer interaction and rehabilitation [14] [15]. A brief review for comparison purposes is given here of activity recognition approaches and systems using external sensors.

B. Krausz and C. Bauckhage in [16] used an off-line procedure of nonnegative tensor factorization to extract basis images that correspond to body parts. The weighting coefficients use a group of regular image sequences to filter a frame. Filtering is done proficiently as basis images are acquired from nonnegative tensor factorization.

M. S. Cheema and A. Eweiwi and C. Bauckhage in [17] deal with the problem of concurrently recognizing actions and the underlying styles (actors) in videos. They proposed a hierarchical method based on straightforward action recognition and asymmetric bilinear modeling. Their method is exclusively based on dynamics of the underlying activity. Results on the multi-actor multi-action data set

IX-MAS show a high recognition rate. F. Liu and X. Xu and S. Qiu and C. Qing and D. Tao in [18] presented a simple to complex action transfer learning model (SCA-TLM) for complex human action recognition. Recognizing complex human actions is a demanding task as training a vigorous learning model needs a hefty labeled data, which is difficult to acquire. SCA-TLM enhances the performance of complex action recognition by leveraging the copious labeled simple actions. SCA-TLM is validated by conducting extensive experiments on two well-known action data sets: 1) Olympic Sports data set and 2) UCF50 data set.
C. Sun and I. N. Junejo and M. Tappen and H. Foroosh in [19] proposed that the an action in video file comprises of a sporadic self-similar manifold in the space-time volume, which is completely described by linear rank decomposition. Initially originated by the recurrence plot theory, they introduced the notion of Joint Self-Similarity Volume (Joint-SSV) for modeling sporadic action data, and therefore proposed an improved rank-1 tensor estimation of the Joint-SSV to collect a reduced-dimensional feature vector that efficiently signalize an action in a video sequence. R. Bhardwaj and P. K. Singh in [20] wrote a review paper on human activity recognition using video analysis with different actions and a number of of activities performed by human in video. To accomplish activity recognition, author's employed a distinct technique of object segmentation and feature extraction, Hidden markov model, bag of word approach. They included fundamental concepts of machine learning methodologies like supervised learning, LDA, clustering, K-Nearest Neighbour.

2.1 Activity Recognition using Wearable sensors and Smartphones

Accidents causing a fall correspond to one of the most widespread cause of injury-related morbidity and death in later life. P. Melillo, R. Castaldo and G. Sannino in [21] conducted three trials for assessing the effectivness of ECG monitoring with wearable devices for: risk estimation of falling in the next few weeks; deter-

rence of impending falls due to standing hypotension; and fall detection. Statistical and data-mining methods are adopted to expand classification and regression models, validated with the cross-validation approach. The three studies made it clear that ECG monitoring could accomplish suitable performances compared to other system for risk assessment, fall prevention and detection.
Sensing technologies, such as human sweat sending and wearable digital tracking technology for monitoring an individuals health condition have become easily available to the public in recent years. The creation of such technology has made it much easier to gather biological and physiological sensor data like blood pressure/oxygen level, electrocardiogram (ECG), electroencephalogram (EEG), heart rate (HR), body temperature, accelerometers, etc. By collecting and studying this data, it can assist us in better understanding of a persons health condition. Consequently, using wearable sensory data for health-care has been attracting notable research from researchers and industry both.
X. Liu and L. Liu and S. J. Simske and J. Liu in [22] introduced an activity recognition system, which assimilates a nonlinear SVM algorithm to recognize twenty distinct human activities from accelerometer and RGB-D camera data. Experimental results have shown promising and effective results
Sun, Lin and Zhang, Daqing and Li, Bin and Guo, Bin and Li, Shijian in [23] used accelerometer-embedded cell phones to record ones everyday physical motion for the sole purpose of altering an individuals inactive lifestyle. As opposed to the past putting it in a pre-defined position or specified orientation, this paper aspires to characterize the physical agility in the normal setting where the cell phones orientation and position is continuously changing, based on material, size and hosting position. By taking into account siz pocket placements, this paper develops a SVM based classifier to recognize seven basic physical activities. Based on ten fold cross validation result on a 48.2 hour data set collected from seven individuals, out system has shown better results over Yangs solution and SHPF solution by five to six percent. By using an orientation insensitive sensor data, they have increased the inclusive F-score from 91.5 percent to 93.1 percent. F-

score has increased to 94.8 percent by using pocket position.
N. elenli and K. N. Sevi and M. F. Esgin and K. Altunda and U. Uluda in [24] used acclerometer and gyroscopes from smart phones for increasing human activity recognition performance rates. Recognition results were acquired utilizing features like extrema, zero crossing rates extracted from time-windows. K-Star classifier led to the best performance among 6 classifiers tested, exceeding 98 percent recognition accuracy.
Sun, Lin and Zhang, Daqing and Li, Bin and Guo, Bin and Li, Shijian in [25] have proposed Spatial-Temporal Activity Inference Model (STAIM) to deduce user activities from data with those three features 1) geographical feature, representing where a user stands; 2) temporal feature, giving us a time vector of activities; and 3) semantic feature, representing the semantic idea of a place from location-based social. System investigational results show that STAIM is capable to effectively deduce user activities, achieving 75 percent accuracy on average. Moreover, STAIM could infer user activities even when there is no training data (with some performance loss). Moreover, sensitive analysis of parameters is also conducted to select the most optimal parameter.

2.2 Activity Recognition using Microsoft Kinect sensor

Development in depth imaging technologies have made human activity recognition reliable without attaching optical markers or any other motion sensors to human body parts. L. Piyathilaka and S. Kodagoda in [26] presented human activity detection model that uses only 3-D skeleton features generated from an RGB-D sensor (Microsoft Kinect TM). To infer the human activities, they implemented Gaussian Mixture Modal (GMM) based Hidden Markov Model(HMM). GM outputs of the HMM were effectively able to capture multimodel nature of 3D positions of each skeleton joint.
K. Adhikari, H. Bouchachia and H. Nait-Charif in [27] used Convolutional Neural

Networks (CNN) to analyzie and classify various poses by using Kinect. Depth data in fusion with RGB images are used to construct their own dataset to record activites by a number of individual in indoor seting. Their result suggestes that a fusin of RGB and depth data with CNN gives the most optimal solution for monitoring indoor fall detection.

W. Zhao, R. Lun, A. B. M. Fofana and D. D. Espy in [28] reported that loss of efficiency caused by injured lower back in offices can cost billions of dollars in a year. A momentous portion of these office injuries are due to negligence of employees by not adhering to safety policies. They presented a new computer vision based system that intends to augment the employees observance of best practices. It consists of reasonably priced depth sensors, wearable devices, and smart phones. The system uses depth sensors to analyze and track the activities of consenting individuals, make them cautious inconspicuously on detection of noncompliant behaviour, and construct cumulative information on their performance.

M. Li and H. Leung in [29] uses multi-view data taken from depth sensor for analysis of human skeletal interaction. It provides relevant indications for human behavior in groups. Their focus is on modeling human on human skeletal interactions for human activity recognition. Each interaction is attributed by a graph, which is intended to conserve the complex spatial structure between skeletal joints pertaining to their activity levels as well as the spatio-temporal joint features evaluating the methodology on the M2I dataset and the SBU Kinect interaction dataset. Hao Xu and Yongcheol Lee and Chilwoo Lee in [30] presented an approach for activity recognition by using 3D skeleton data obtained with a Kinect sensor using simplified dynamic time wrapping (DTW) and calculated Euclidean geometry distance to obtain the probable activities from the trained data. A. Jalal and S. Kamal and D. Kim in [31] tracked and classified human silhouettes using a sequence of RGB-D images by extracting the shape and motion features to identify richer motion information and then these features are clustered and fed into Hidden Markov Model (HMM) to train model and recognize human activities.

E. Cippitelli and E. Gambi and S. Spinsante and F. Florez-Revuelta in [32] made use of low cost RGB-D sensors for Human Action Recognition. They evaluated the performance of a skeleton-based algorithm for Human Action Recognition on a large-scale dataset. The algorithm exploits the bag of key poses method, where a sequence of skeleton features is represented as a set of key poses. A temporal pyramid is adopted to model the temporal structure of the key poses, represented using histograms. Finally, a multi-class SVM performs the classification task, obtaining promising results on the large-scale NTU RGBD dataset.

Histograms have been used extensively using different x-axis bins for binning process to form a feature vector. N. Dalal and B. Triggs in [33] experimented with grids of histograms of oriented gradient (HOG) descriptors and it significantly outperformed existing feature sets for human detection. Other researchers have been developing similar histogram based binning procedures using orientation and direction of a values like pixel intensities and motion direction to form histograms based on HoG. Alexander Klaser, M. Marszaek and C. Schmid in [34] used a Spatio-temporal localized descriptor describing a local fragment in an image or a video. A histogram is computed for any random scale along x,y,t. The support area about POI is separated into a grid of gradient orientation histograms. Each histogram is calculated over a grid of mean gradients. Gradient orientation is quantized using regular polyhedrons and each mean gradient is computed using integral videos. Oreifej and Z. Liu in [35] created 4D projectors, which quantize the 4D space and represent the possible directions for the 4D normal. Projectors are initialized using the vertices of a regular polychoron. Scovanner, Ali and M. Shah in [36] calculated a 2D gradient magnitude and orientation for each pixel. To create sub-histograms, sub-regions adjacent to the interest point are used in experimentation, where each pixel contains a single magnitude value and two orientation values and . For each 3D sub-region orientations are added into a histogram. The final descriptor is a vectorized sub-histograms. Jie Liang1, J. Zhou, Y. Gao in [37] proposed a 3D high-order texture pattern descriptor for hyperspectral face recognition, which efficiently takes advantage of both spatial and spectral features in hyperspectral images. Based on the local derivative pattern, the hyperspectral faces with multi-directional derivatives and binarization is encoded in spatial-spectral space.

Table 2.1 gives an overview of some notable work done in activity recognition using kinect sensors.

Reference	Methodology	Dataset	Accuracy
[26]	Gaussian mixture based HMM	CAD-60	84
[27]	Convolutional Neural networks	Own	74
[30]	Eigen-joints based on HMM	MSRA	80
[38]	3-D Posture Data	CAD-60	77.3
[39]	Histograms of 3D Joints	MSRA	95
[29]	Active Joint Interaction Graph	M2I and SBU dataset	94.12 and 88.57
[40]	Simultaneous Feature and Body-Part Learning	MSRA and CAD-60	91.6 and 83.93

Table 2.1: Table to compare few kinect based AR systems

2.3 Kinect Working

Kinects software is patented and how it works is mostly based on assumptions. In 2010, Microsoft uncovered a new Xbox 360 accessory they claimed would change the way gamers played. The Kinect was the first motion controlled gaming system that didn't involve a controller, instead permitting players to use their whole body to move the game.

It uses an intricate system of sensors, lasers and cameras to reflect a player's motion and actions on screen. Kinect has three lenses. First there is a regular RGB camera. The one used for this project is a 640 x 480 resolution camera. It works like a generic webcam, and records the room. The Kinect uses this camera any time it displays your image in the game, or for other functions like video chat. It captures images approximately at normal video speeds of 32 frames per second. Better resolution cameras are now available for more efficient processing. The more complicated camera relies on infrared light to work. Second lens on the Kinect is an IR emitter, which immerses the play space in light that the camera can pick up. The camera sees these waves as they recoil off people and objects in the room; the brighter the light, the closer the object is. Objects too close to the camera become too bright and hard to distinguish, which is why Kinect has a minimum and maximum range that players have to stand in, in order to be detected by Kinect sensor bar. The camera in the Kinect encodes information in that light as it goes out, then measures the degradation over time with the sensor in the Kinect's Infrared camera.

The data fed in by the camera is immediately processed by the patented software of PrimeSense. It recognizes shapes as humans by heads and limbs. This software knows how a human body moves, it knows your head can't turn 360 degrees on your neck, and it captures movement through more than 48 points of articulation. The Kinect software has been programmed with more than 200 possible poses, giving it has an idea of where your body is probably going to go to.

The Kinect's other much advertised feature is its capability to recognize voice commands. It has four microphones in its sensor bar to pick up players' voices, all pointed down to pick up soundwaves more effectively [41].

2.4 Public Datasets

For activity recognition problem, it is necessary to collect a large amount of data for training the classification system. To meet this requirement, researchers have created datasets that agree with the processing approach. These approaches include vision-based methods i.e. 2D/3D video recording of activities, or body sensors i.e. accelerometers on straps and smart-phone sensors. Ziyun CaiJungong HanLi LiuLing Shao in [1] did a systematic survey of recognized depth datasets for various applications composing of object detection, scene recognition, hand gesture classification, parallel localization and mapping of 3D data, and pose estimation. They provided understanding of important attributes of each data-set, and gave a comparison of compared the marketability and the complexity of these data-sets. A data-set collected from Microsoft Kinect version 1 which includes a

number of activities in different has been made available publically by a number of researchers including the CAD-60 and CAD 120 dataset provided by Cornel university [2]. Chaquet et al. in [42] wrote a paper focusing on datasets dedicated to Vision-based human action and activity recognition. A different approach for collecting data describing a users activities is introduced by Sarkar [43]. A number of sensors are inserted into home appliances while the experience sampling tool (ESM) is given to the customer for getting self-reported activity label.
A comprehensive report on public datasets based on Kinect sensor is described below in figure 2.1

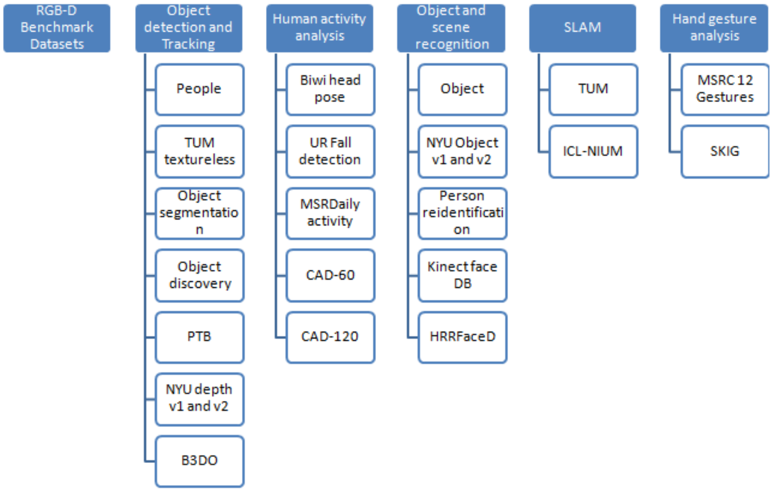


Figure 2.1: Kinect Datasets available [1]

2.5 Cornel CAD-60 Dataset

The CAD-60 data set includes RGB-D video sequences of humans performing activities which are recording using the Microsoft Kinect sensor version 1. CAD-60 dataset is generated using the Microsoft Kinect sensor version 1. Dataset is formed of video data and RGB image together with corresponding aligned depths indication at each pixel at a frame at the rate of 32 frames per second. Video data gives an image of 640x480 pixel resolution with depth range of 1.2m to 3.5m. The sensor is continently sized for it to be placed on any surface high or low or on the wall for better coverage. Data is collected considering five different environments: office, kitchen, bedroom, bathroom, and living room. Three to four common activities were identified for each location, giving a total of twelve unique activities [2]. Data is taken from four different people: two males and two females. Each activity is performed for approximately 45 seconds by each person. There is no occlusion of arms and body from sensor view. A sample image of how activities were performed is shown in figure 2.2

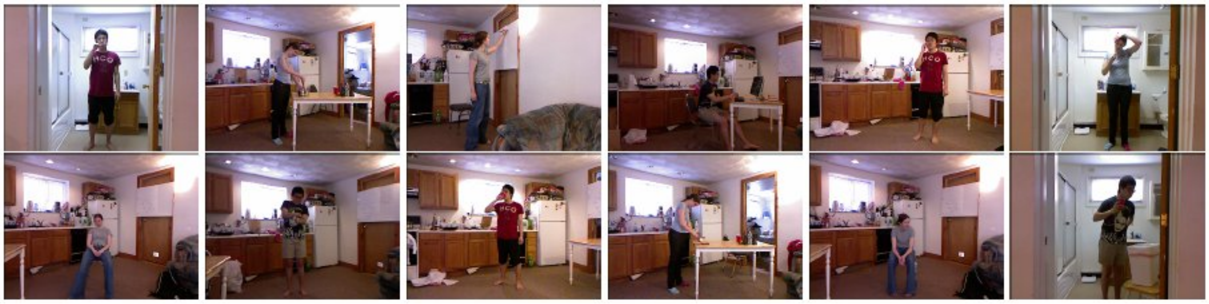
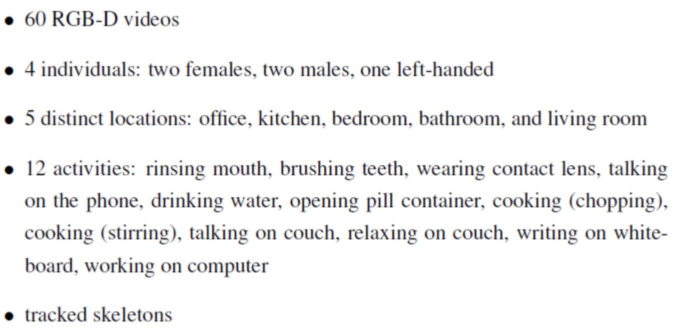


Figure 2.2: Samples from CAD-60 dataset [2]

2.5.1 CAD-60 dataset features

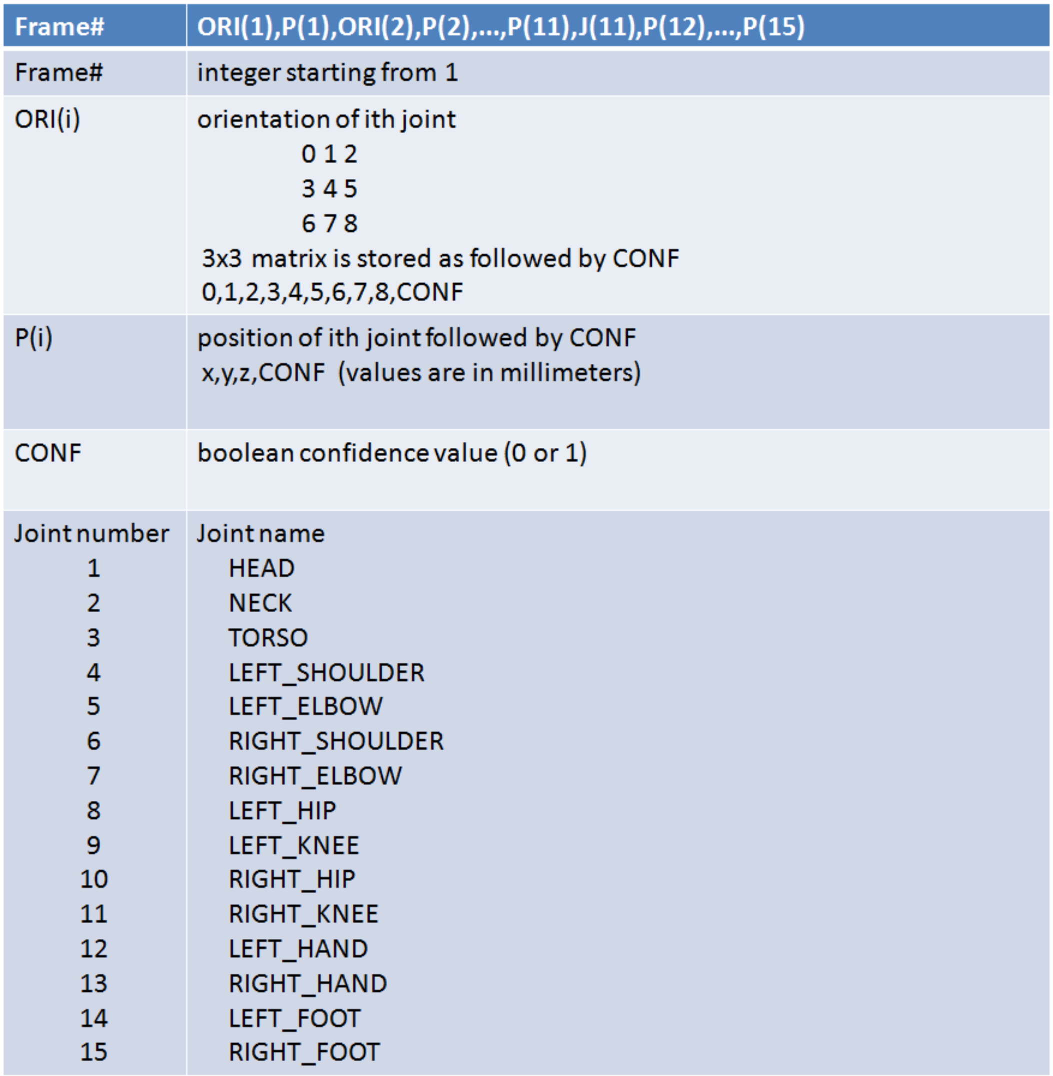


- 60 RGB-D videos
- 4 individuals: two females, two males, one left-handed
- 5 distinct locations: office, kitchen, bedroom, bathroom, and living room
- 12 activities: rinsing mouth, brushing teeth, wearing contact lens, talking on the phone, drinking water, opening pill container, cooking (chopping), cooking (stirring), talking on couch, relaxing on couch, writing on whiteboard, working on computer
- tracked skeletons

2.5.2 Skeleton Data Format

Skeleton data in the video sequence gives fifteen joints positions at each frame. 15 out of 11 joints have joint orientation and joint position vectors. 4 out of fifteen joints only have joint position. Each row follows the following format as shown in table 2.3.

Figure 2.4 shows frame 1 skeleton view of brushing teeth activity by subject 1



Frame#	ORI(1),P(1),ORI(2),P(2),...,P(11),J(11),P(12),...,P(15)
Frame#	integer starting from 1
ORI(i)	orientation of ith joint 0 1 2 3 4 5 6 7 8 3x3 matrix is stored as followed by CONF 0,1,2,3,4,5,6,7,8,CONF
P(i)	position of ith joint followed by CONF x,y,z,CONF (values are in millimeters)
CONF	boolean confidence value (0 or 1)
Joint number	Joint name
1	HEAD
2	NECK
3	TORSO
4	LEFT_SHOULDER
5	LEFT_ELBOW
6	RIGHT_SHOULDER
7	RIGHT_ELBOW
8	LEFT_HIP
9	LEFT_KNEE
10	RIGHT_HIP
11	RIGHT_KNEE
12	LEFT_HAND
13	RIGHT_HAND
14	LEFT_FOOT
15	RIGHT_FOOT

Figure 2.3: Skeleton data format

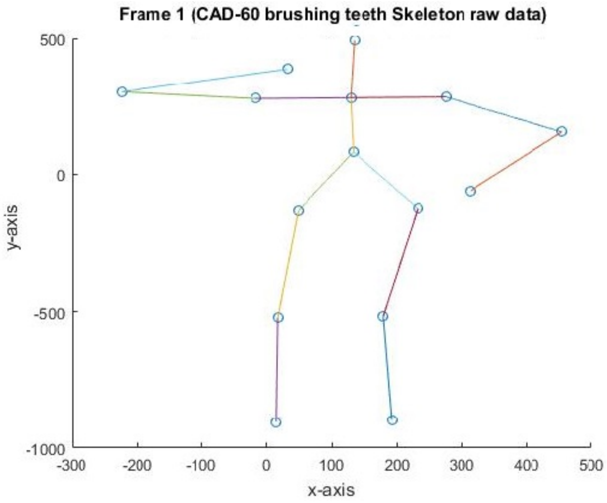


Figure 2.4: Skeleton data Frame 1

Chapter 3

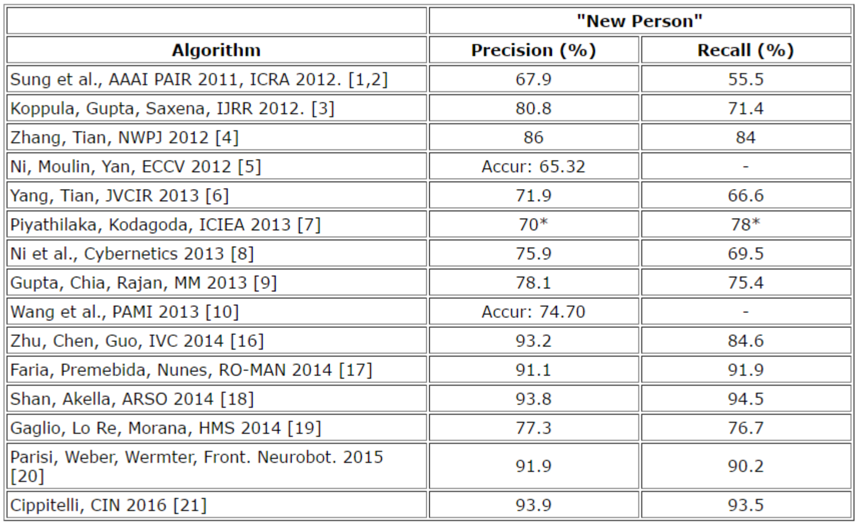
Activity Recognition

The promise of Kinect seems to be infinite when it comes to looking at the variety of applications that can be adapted from it. With a superior and one of a kind depth recognition ability and comprehensive motion detection system, people are coming up with all kinds of remarkable ideas that can take advantage of the technology-everything from quadrocopters that can evade impact to grocery carts that can tag along with you in the store, recording what you put in the cart. Activity recognition usefulness, in health-care to surveillance systems is undeniable. It is one of the leading research areas in Computer vision and Pattern recognition. We have employed computer vision and machine learning algorithms for activity recognition of CAD-60 dataset.

3.1 Activity Recognition with Space-Time Trajectories

Trajectory-based methods are recognition methods that infer an activity as a collection of space-time trajectories. An individual is commonly represented as a collection of 2-dimensional (x,y) or 3-dimensional (x,y,z) in this trajectory-based method corresponding to position in time-space [3]. We have used human body part inference method, commonly known as stick figure modeling, which has been used a lot lately to gather joint position/trajectory of a subject at predefined image frames. As a human performs an action, changes in his/her joint position are recorded as space-time trajectories, constructing 3D (x,y,t) or 4D(x,y,z,t)representations of the action. L. Piyathilaka and S. Kodagoda in [26] were effectively able to capture multi-model nature of 3D positions of each skeleton joint and achieve efficiency of 85 percent on CAD-60 dataset.

L. Xia and C. C. Chen and J. K. Aggarwal in [39] proposed a new approach of feature extraction from skeleton data for human action recognition with histograms of 3D joint locations (HOJ3D) as a concise descriptor of postures. They tested this new on the MSR Action3D dataset and showed 98.7 % efficient results. Table 3.1 gives an overview of results that have been derived by different researchers using CAD-60 dataset The new person metric was introduced by Jaeyong Sung, Colin Ponce, Bart Selman and Ashutosh Saxena in [2]. They used leave-one-out cross-validation to test each subjects data; i.e. the algorithm was trained on three people from whom data was collected and the fourth person data was used for



	"New Person"	
Algorithm	**Precision (%)**	**Recall (%)**
Sung et al., AAAI PAIR 2011, ICRA 2012. [1,2]	67.9	55.5
Koppula, Gupta, Saxena, IJRR 2012. [3]	80.8	71.4
Zhang, Tian, NWPJ 2012 [4]	86	84
Ni, Moulin, Yan, ECCV 2012 [5]	Accur: 65.32	-
Yang, Tian, JVCIR 2013 [6]	71.9	66.6
Piyathilaka, Kodagoda, ICIEA 2013 [7]	70*	78*
Ni et al., Cybernetics 2013 [8]	75.9	69.5
Gupta, Chia, Rajan, MM 2013 [9]	78.1	75.4
Wang et al., PAMI 2013 [10]	Accur: 74.70	-
Zhu, Chen, Guo, IVC 2014 [16]	93.2	84.6
Faria, Premebida, Nunes, RO-MAN 2014 [17]	91.1	91.9
Shan, Akella, ARSO 2014 [18]	93.8	94.5
Gaglio, Lo Re, Morana, HMS 2014 [19]	77.3	76.7
Parisi, Weber, Wermter, Front. Neurobot. 2015 [20]	91.9	90.2
Cippitelli, CIN 2016 [21]	93.9	93.5

Figure 3.1: Results of CAD-60 dataset

a new person metric. We have conducted the same leave-one-person data out to check our algorithms efficiency in case of a new person activity recognition.

3.2 System Framework

An activity recognition system comprises of several stages, which include 1) data acquisition 2) preprocessing 3) feature extraction and 4) classification, as it can be seen in Figure 3.2. Data is acquired by Cornel University via Microsoft Kinect version 1. Two types of datasets for each activity are available publicly, video data and skeleton data. Motion trajectories are acquired for all fifteen joints over the period of activity. This trajectory data goes through pre-processing and feature extraction before classification. The choice of the methodology followed in each processing stage plays a vital role in the concluding outcome of the recognition system. Each step is explained in detail in the next section

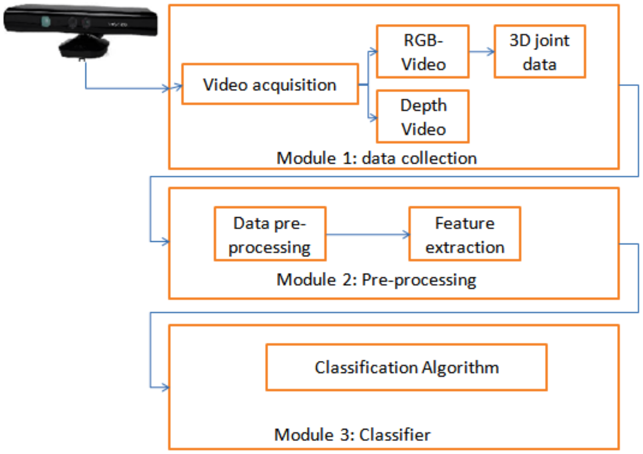


Figure 3.2: System Framework

3.3 Preprocessing data

CAD-60 dataset is available publicly online on Cornel University for research purposes. We have discussed earlier the specifications and formatting of Skeleton data recorded. Each activity duration and joint position number is also given in full detail in the folder provided. After data collection, comes preprocessing data. Machine learning algorithms train from data you use as input. It is significant that the data fed in is the right data for the problem you want fixed. Even if you have good data, you need to make sure that it is in a useful scale, format and consequential features are included. Preprocessing steps are show in figure 3.3 and explained below

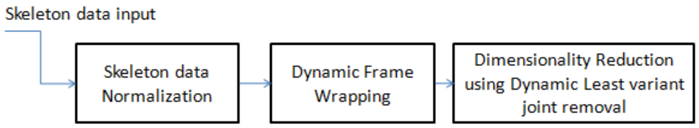

Skeleton data input
Skeleton data Normalization
Dynamic Frame Wrapping
Dimensionality Reduction using Dynamic Least variant joint removal

Figure 3.3: Preprocessing steps

3.3.1 Data Normalization

Data normalization is a preprocessing procedure that is generally employed to even out the range of values in our dataset.
If J_i where i =1,2,3,...15 represents a 3 dimensional joint position vector,
then P_x is a $[1 \times 45]$ where x = 1,2,3,...x
dimensional array representing all the skeleton joint positions in a single frame.
P_1 is first array representing joint positions of the first frame in any given activity. These frame are stored in a matrix **M** with dimensions $[x \times 45]$. There are a total of 72 **M** matrices. Each **M** uniquely represent a single activity. There are eighteen activities in total. Each activity is performed four times. x is different for each activity and varies from person to person depending on the amount of time it takes for a person to perform a given activity. In order to make joint position matrix invariant to the location of person performing any activity inside the coverage

area of Kinect; position invariance to the build and position of person is achieved by normalization of a single frame skeleton vector P_x by a scale factor d_x. d_x is calculated by formula given in eq 3.1

$$d_x = \sqrt{(J_3)^2 - (J_2)^2} \tag{3.1}$$

where
x=1,2,3,....x (number of frames in matrix M)
$J_3 =$ Torso Joint
$J_2 =$ Neck Joint

Data normalization for a single frame P_x in activity matrix M is performed by eq 3.2

$$P_x = \frac{P_x}{d_x} \tag{3.2}$$

Matrix representation is given in es 3.3

$$M_{(x,45)} = \begin{array}{c|ccccc|} & J_1 & J_2 & J_3 & \cdots & J_{15} \\ P_1 & (x,y,z) & (x,y,z) & (x,y,z) & (x,y,z) & \\ P_2 & (x,y,z) & (x,y,z) & (x,y,z) & (x,y,z) & \\ P_3 & (x,y,z) & (x,y,z) & (x,y,z) & (x,y,z) & \\ \vdots & \vdots & \ddots & \vdots & \vdots & \\ P_x & (x,y,z) & (x,y,z) & (x,y,z) & (x,y,z) & \end{array} \tag{3.3}$$

E. Cippitelli and E. Gambi and S. Spinsante and F. Florez-Revuelta in [32] performed data normalization by calculating center-of-mass first after taking average 3D position of main skeleton.

E. Cippitelli and E. Gambi and S. Spinsante in [44] calculated one d_x values for a complete matrix set.

Figure 3.4 shows first six frames of brushing teeth activity after data normalization

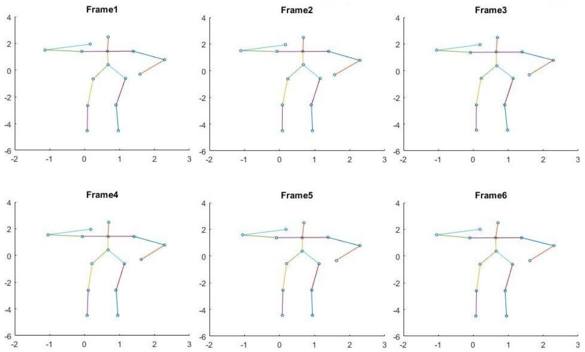


Figure 3.4: Data normalization result (brushing teeth frame 1 to frame 6)

3.3.2 Dynamic Frame wrapping

As stated in eq 3.3, matrix **M** row dimensions vary from activity to activity and person to person due to individual variation and condition environment. For this reason, number of frames for each activity matrix **M** has to be made equal with respect to temporal sequences for assessment. In order to evade loss of information in frames, error is reduced by employing a dynamic frame wrapping technique. Unlike Hao Xu and Yongcheol Lee and Chilwoo Lee in [30] who used simplified dynamic time wrapping by dividing each temporal sequence into several parts and taking average value in each part.

We have employed a different method where we can choose the resultant size of x of matrices **M**. All 72 matrices, where each matrix represent one activity are reduced to a fixed value of x_{new}. It can either be made equivalent to the smallest value of x or any value equal to one half of the longest duration activity.

For example, lets say we want x_{new} to be 200 for all 72 matrices **M**. We have value of $x_{new} = 200$. First step is to choose a value of the frame number x_{new} which you want for all activities. In our case we choose, $x_{new} = 200$. Now loop through all 72 matrices **M**. Take first Matrix **M**. Calculate the number of frames x it takes to perform activity 1. group value (GV) is calculated by eq 3.4

$$GV = \frac{x}{200} \tag{3.4}$$

End Value (EV) is calculated by eq 3.5

$$EV = 200 \times GV \tag{3.5}$$

In order to make new matrix **M** . Perform the calculation in eq 3.6

$$M_{(new,45)} = \sum_{x=i:i+(GV-1)}^{EV} \frac{M_{(n,45)}}{N} \tag{3.6}$$

where i = 1:GV:EV
N = Number of values from 1 to GV

3.3.3 Dynamic invariant joint removal

As a result of the step, dynamic frame wrapping, we get a new matrix **M** of the form given in eq 3.7 for all activities in CAD-60 dataset.

$$M_{(200,45)} = \begin{array}{cccccc} & J_1 & J_2 & J_3 & \cdots & J_{15} \\ P_1 & (x,y,z) & (x,y,z) & (x,y,z) & (x,y,z) & \\ P_2 & (x,y,z) & (x,y,z) & (x,y,z) & (x,y,z) & \\ P_3 & (x,y,z) & (x,y,z) & (x,y,z) & (x,y,z) & \\ \vdots & \vdots & \ddots & \vdots & \vdots & \\ P_{200} & (x,y,z) & (x,y,z) & (x,y,z) & (x,y,z) & \end{array} \tag{3.7}$$

When an activity is being performed by a person. There are certain joints in the body that exhibit minimal change in position. While pouring a glass of water or brushing your teeth and even working on a computer, midsection jonts like torso, knees and foot show minimal to zero movement. However these zero-to-minimum movement joints can vary from activity to activity. For this reason, these joints have to be checked for every activity matrix $M_{(200,45)}$ dynamically. Least variant joints are removed before feature extraction algorithm is applied to each activity.

First step is to calculate variance of every xyz coordinate vector from 1 to 200 in all joint position as explained in eq 3.8

$M_{(200,45)} =$

$$\begin{array}{c} P_1 \\ P_2 \\ \vdots \\ P_{200} \\ \\ \end{array} \begin{pmatrix} x_1 & y_1 & z_1 & x_2 & y_2 & z_2 & \cdots & x_{14} & y_{14} & z_{14} \\ x_1 & y_1 & z_1 & x_2 & y_2 & z_2 & \cdots & x_{14} & y_{14} & z_{14} \\ \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_1 & y_1 & z_1 & x_2 & y_2 & z_2 & \cdots & x_{14} & y_{14} & z_{14} \\ \sigma(x_1) & \sigma(y_1) & \sigma(z_1) & \sigma(x_2) & \sigma(y_2) & \sigma(z_2) & \cdots & \sigma(x_{14}) & \sigma(y_{14}) & \sigma(z_{14}) \end{pmatrix} \tag{3.8}$$

Variance of x,y and z vectors is summed for a single joint vector J_i. Variance vector is a $[1 \times 15]$ dimensional vector. We sort this vector in descending order and remove joints with minimum variance values. Removing six out of fourteen joints gave best results. We have now most variant joints left for feature extraction. Fig 3.5 gives an example of joint removal that show minimum overall motion in an activity. Notice that we have used fourteen joints instead of fifteen as stated earlier. We remove the torso joint from this step as we require to use it as a reference joint is histogram of 3-dimensional directional derivative feature extraction process. We put torso joint back in matrix M for next step

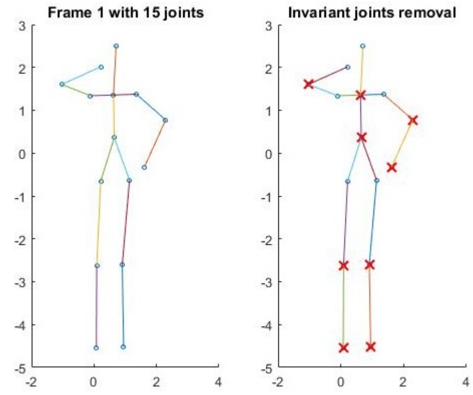


Figure 3.5: Frame before and after Dynamic invariant joint removal algorithm

3.4 Feature Extraction

Feature extraction plays a vital role in the any information analysis method. It principally conditions the information into a form to ensure success of any following statistics or machine learning algorithms. It is of importance to not exclude and relevant information during feature extraction module. Apart from refraining from excluding relevant information, it is also imperative that features are not redundant and each class has distinguishing properties for higher efficiency.
We have used the histogram of oriented gradient on 3-dimensional time-space trajectory data, x-axis of histogram has orientation angles for a single frame of an activity and magnitude is calculated and added to corresponding angle bin.We have applied it to each activity matrix after passing it through pre-processing stages. After histogram formation, linear discriminant Analysis is applied as a dimensionality reduction tool. System efficiency is checked my applying support vector machine (SVM)as a classifier and a result comparison is done with previous research and their results on the same dataset CAD-60.
Working of Histogram of 3 dimensional directional derivatives is given in the following section

3.4.1 Histogram of Oriented gradient formation

Directional Derivatives

In order to understand directional derivatives, we need to first recall what a partial derivative is. If we take a 2-dimensional example, we know that a partial derivative corresponds to the rate of change of a function $f(x, y)$ while changing x and holding y constant($\frac{\partial f}{\partial x}$) and by changing y and holding x constant($\frac{\partial f}{\partial y}$). In other words, partial derivatives ($\frac{\partial f}{\partial x}$)give slop in positive x-direction and partial derivatives ($\frac{\partial f}{\partial y}$) gives slope in positive y-direction. However, what if we want to determine the change in function f and allowing both x and y to vary at the same time. The question is how to determine where x and y vary as there can me a number of ways to change both x and y at the same timee. For example, x direction could be altering more rapidly than the y direction and then there is also the matter of whether or not each is escalating or declining. If we generalize partial derivatives to give slope in one given direction, result is a directional derivative. Step number one in taking a directional derivative, is to indicate the direction in which you want to calculate the slope. One way to identify a direction is with a vector $u = (u_1, u_2)$ that gives the position of the direction in which we want to calculate the slope. Lets suppose u is a unit vector. We write the directional derivative of $f(x, y)$ in the direction $\mathbf{u} = (\mathbf{u_1}, \mathbf{u_2})$ at any given point **a** as $\mathbf{D_u f(a)}$.

The theory of the directional derivative is straightforward. If a particle is standing at point a, $\mathbf{D_u f(a)}$ is the slope of f(x,y) facing the direction given by unit vector u.

Like a partial derivative, $\mathbf{D_u f(a)}$ is also a number and not a matrix. In nearly all cases, there is always one direction **u** where the directional derivative $\mathbf{D_u f(a)}$ is the largest, known as the uphill direction. If direction of this maximal slope is **m**, both the direction **m** and the maximal directional derivative $\mathbf{D_m f(a)}$ are calculated by the gradient of **f** and is represented by $\nabla \mathbf{f(a)}$. The gradient is a vector that points in the direction of **m** and whose magnitude is given by $\mathbf{D_m f(a)}$. Mathematical expression is given in eq 3.9 and eq 3.10.

$$\frac{\nabla f(a)}{\|\nabla f(a)\|} = m \tag{3.9}$$

$$\|\nabla f(a)\| = D_m f(a) \tag{3.10}$$

For a predetermined value of **a**, the maximum value of $\mathbf{D_u f(a)}$ take places when **u** and $\nabla \mathbf{f(a)}$ are pointing in the same direction (i.e., when $\theta = 0$ or $= 2\pi$), and the minimum value occurs when **u** and $\nabla \mathbf{f(a)}$ are pointing in opposite directions (i.e., when $\theta = \pi$). Hence range of values of $\mathbf{D_u f(a)}$ always lies between $-\nabla \mathbf{f(a)}$ and $\nabla \mathbf{f(a)}$. It so happens that the connection between the gradient and the directional derivative can be summarize by the eq 3.11

$$\begin{aligned}\mathbf{D_u f(a)} &= \nabla\mathbf{f(a).u} \\ &= \|\nabla f(a)\| \|\mathbf{u}\| cos\theta \\ &= \|\nabla f(a)\| cos\theta\end{aligned} \tag{3.11}$$

in the above equation θ is the angle between **u** and the gradient $\nabla\mathbf{f(a)}$. Also, **u** is a unit vector, hence $\|\mathbf{u}\|$ is equal to 1. We have applied this 2 dimensional theory to our 3 dimensional joint position data and formed a histogram of angle and directional derivatives as a feature vector.

After pre-processing steps, our activity matrix **M** is of the form, as given in eq 3.12

$$M_{(200,45)} = \begin{array}{c|ccccc|} & J_1 & J_2 & J_3 & \cdots & J_9 \\ P_1 & (x,y,z) & (x,y,z) & (x,y,z) & (x,y,z) & \\ P_2 & (x,y,z) & (x,y,z) & (x,y,z) & (x,y,z) & \\ P_3 & (x,y,z) & (x,y,z) & (x,y,z) & (x,y,z) & \\ \vdots & \vdots & \ddots & \vdots & \vdots & \\ P_{200} & (x,y,z) & (x,y,z) & (x,y,z) & (x,y,z) & \end{array} \tag{3.12}$$

Note that our number of frames is reduced to 200 from a varying range of frames between the range of 1960 to 147, and the number of joints left for feature extraction algorithm is reduced from fifteen joints to nine joints.

First step of histogram formation is to set a reference joint, joint torso $\mathbf{J_t}$. This reference joint position vector was removed from the matrix **M** before Redundant data removal using Dynamic Least variant joint removal step as there was a high likelihood of loosing this reference joint due to its minimal overall change in position. After least variant joint removal, torso joints are added back into the matrix to form resultant matrix **M** as given in eq3.12. In this case $\mathbf{J_t}$ = $\mathbf{J_1}$ $=$ in matrix **M**.

Next step is to consider each frame to calculate the gradient vector and directional derivatives with respect to the reference joint $\mathbf{J_t} = \mathbf{J_1}$.
Gradient vectors are calculated using eq.

$$\nabla J = \Big(\frac{\partial J_1}{\partial J_2}, \frac{\partial J_1}{\partial J_3}, ..\frac{\partial J_1}{\partial J_9}\Big) \tag{3.13}$$

$\mathbf{J_1}$ and $\mathbf{J_2}$ comprises of x,y and z coordinates. These values are gradient vectors. In order to find which direction gives maximum gradient magnitude, we calculate the distance from reference joint $\mathbf{J_1}$ to all the joints from $\mathbf{J_2}$ to $\mathbf{J_9}$ using eq 3.14 and save the direction of maximum gradient.

$$\nabla_{max} J = \sqrt{(x_1 - x_i)^2 + (y_1 - y_i)^2 + (z_1 - z_i)^2} \tag{3.14}$$

where i = 2,3,4...9 For directional derivatives, we needed two things, the direction in which we want to calculate the slope, joints from two to nine, and maximum gradient vector which we calculated from eq 3.14.
Directional derivatives is an array of scaler values calculated from dot product in eq 3.15

$$\mathbf{D_u f(a)} = \nabla_{\mathbf{max}} \mathbf{J}.\nabla \mathbf{J_i} \tag{3.15}$$

where i = 2,3,4...9 After calculating directional derivatives, we now calculate the angle between the maximum gradient vector $\nabla_{\mathbf{max}}\mathbf{J}$ and the direction vector $\nabla\mathbf{J}$ in which we calculated the slope.
We know from basic trigonometry, angle between two 3-dimensional vectors can be calculated in the same way as angle between two 2-dimensional vectors using eq 3.16 using direction cosines.

$$\theta = arccos\Big(\frac{\nabla_{\mathbf{max}}\mathbf{J}.\nabla\mathbf{J_i}}{|\nabla_{max} J|.|\nabla\mathbf{J_i}|}\Big) \tag{3.16}$$

where i = 2,3,4...9
Eq 3.16 gives us the shortest angle between two 3-dimensional vectors. In order to get theta from 0 to 360 degrees, we have used direction cosines to get any angle

from 0 to 360 degrees by taking an xy,xz or yx plane as a reference for counter clockwise measurement as shown in fig 3.6 Since z-axis shows minimal overall movement, we have chosen xy-plane to determine the counter clockwise angle between the gradient vectors. XY-plane in fig 3.6 is represented by no grid lines. If gradient vector between $\mathbf{J_1}$ and $\mathbf{J_2}$ in maximum gradient vector, θ is the angle between the maximum gradient vector formed by joints $\mathbf{J_1}$ and $\mathbf{J_2}$ and position vector formed by joints $\mathbf{J_1}$ and $\mathbf{J_3}$.

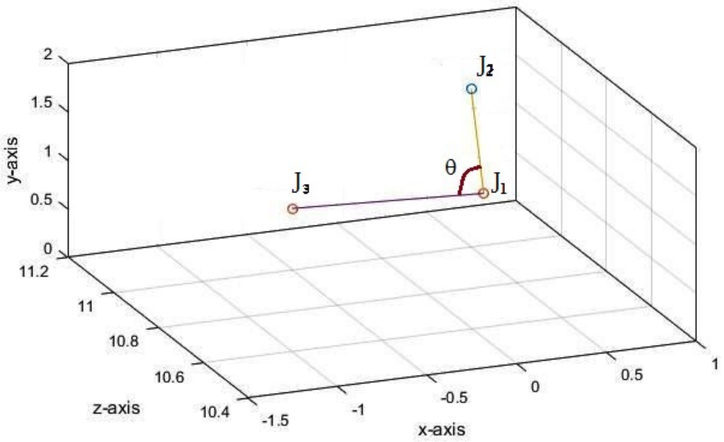


Figure 3.6: counter clockwise angle with xy-plane as normal

Histograms

Next is the bining process to form a histogram od directional derivatives. A histogram is kind of a bar graph that gives graphical presentation of information using bars of different heights. Data bining is over fixed intervals of 40 degrees from 0 to 360, making nine bins in one frame histogram. Histogram shows the magnitude of directional derivative as the height of the bar over each direction from 0 to 360. We have used signed gradients such that the orientations ranges from 0 to 360 degrees. The 200×27 activity matrix forms a $[1 \times 1800]$ size feature vector. Each frame of 27 joint position values is replaced by a histogram that gives joint direction and magnitude with respect to reference joint in a skeleton data. A single θ value between ∇J_i and $\nabla_{max} J_1$ is calculated as shown in Figure 3.7 Each frame contains 27 values and with a 9-bin histogram for each frame, for a total of 200 frames per activity, brings the final vector size to 200 frames per activity $\times$ 9-bins per histogram $= 1800$ values. Fig 3.8 shows how histogram for each frame is concatenated to form a $[1 \times 1800]$ feature vector for a single activity.

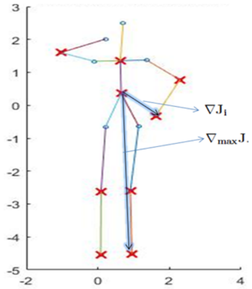


Figure 3.7: θ calculation for bining process

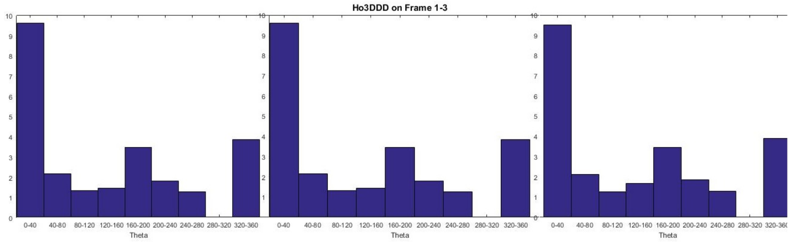


Figure 3.8: Histogram for frame 1 to 3 $[1 \times 27]$

3.4.2 Dimensionality Reduction

Linear Discriminant Analysis (LDA) is a classification method originally developed in 1936 by R. A. Fisher. It is straightforward, mathematically vigorous and often produces models with good accuracy. LDA depends on searching for a linear combination of predictors that best separates classes.

LDA in our system is used to extract the dominant features. It is used maximize

the ratio of the between - class spread and the within-class spread of data. The LDA algorithm searches for the vectors in the principal space to construct the best discrimination between different classes. In this way, a more vigorous feature space can be generated that divides the feature vectors of each class. In our experimentation, we decrease the dimension of the Histogram feature from $n = 1800$ dimensions to $No.of.Class - 1$ dimensions = 17 dimensions.
Common steps for performing a linear discriminant analysis are given below

- Calculate the n-dimensional mean vectors for 18 classes from the dataset where n = 1800.

- Work out the inside class and in-between class scatter matrices

- Compute the eigenvectors and equivalent eigenvalues

- Arrange the eigenvectors by declining eigenvalues and select k eigenvectors with the highest eigenvalues to form a $n \times k$ dimensional matrix W.

- Employ the $d \times k$ eigenvector array to convert the model into eigen-space. This can be done by recapitulating the matrix multiplication: $Y = X \times W$ (where X is an $d \times n$-dimensional matrix constituting the d classes, and y are the transformed $d \times k-$dimensional samples in the new subspace.

Result of Linear discriminant analysis is then used as input to classification algorithm, support vector machines(SVM).

3.5 Classification

Two foremost kinds of classifiers consist of unsupervised and supervised classifiers.
Unsupervised or unlabeled classification is where the results are dependent on the software analysis is absence of a user provided sample classes. The computer uses procedures to establish which input sample is correlated and groups them into

classes. The user can indicate which algorithm the software will employ and the preferred number of output classes but otherwise does not assist in the classification process. However, the user must have familiarity of the data being classified and grouping of data with frequent distinctive features formed by the computer have to be connected to actual features. Examples of unsupervised learners are clustering, neural networks, PCA etc.

Supervised classification is based on the idea that a user can select data samples that are representative of exact class and then direct the software to use these training samples in algorithms as references for the classification of all other incoming data. Training samples are chosen based on the information of the user. The user also sets the limits for how related the data must be to group them together. The user also assigns the number of classes that the data is classified into. We have used supervised machine learning algorithm, support vector machines (SVM) for activity recognition.

Support Vector Machines is supervised learning process known for classification. The biggest benefit of Support Vector Machines is that it can utilize a number of kernels for the purpose of transforming the data. It enables us to implement linear classification methodologies to non-linear data. These kernels are applied in such a way that a kernel equation forms a hyper-plane that splits data occurrences of one class from those of other within a multi-dimensional space.

The kernel equations could be any function that transforms data, originally linearly non-separable in one domain into another domain where the occurrences develop into linearly separable. Kernel equations examples include linear, quadratic, Gaussian, cubic or any kernel that accomplishes this exact purpose.

We have used a one-against-all multi svm to classify eighteen activities instead of a binary classifier.One significant thing to note down about Support Vector Machines is that the data to be classified has to be binary. If the data is not binary, Support Vector Machines takes it as though it is, and performs the classification through a series of binary estimation on the data.

Chapter 4

Results

The purpose of our system is to classify activities performed in CAD -60 dataset. For testing like, Sung et all in [2], we also validated our algorithm by producing results using 'new-person' settings. We have used 5-fold cross validation to test the data in the new person scenario, we used leave-one-out crossvalidation to test each persons data; i.e. the model was trained on three of the four people from whom data was collected, and tested on the fourth. Confusion matrix using linear SVM is shown in fig 4.1 Confusion matrix plot shows how SVM classifier performed in each class. It helps to recognize the areas where the classifier has performed poorly. Diagonal cells display the correctly classified observations in the trained network. It displays in percentage the true and predicted classes. Rest of the cells gives a percentage where the classifier makes a misclassification. The right most column shows the accuracy for each predicted class, while the bottom row plot gives accuracy for each true class. Over all accuracy is given at the bottom. Confusion matrix shows that we have attained excellent performance on all actions except for drinking water and opening pill container.

Histograms of Oriented gradients feature extraction method has been used extensively in the past in human detection systems and lately, different variations of histogram feature has been used in activity analysis and recognition. Our algorithm was successful in detecting and classifying with a precision/recall measure

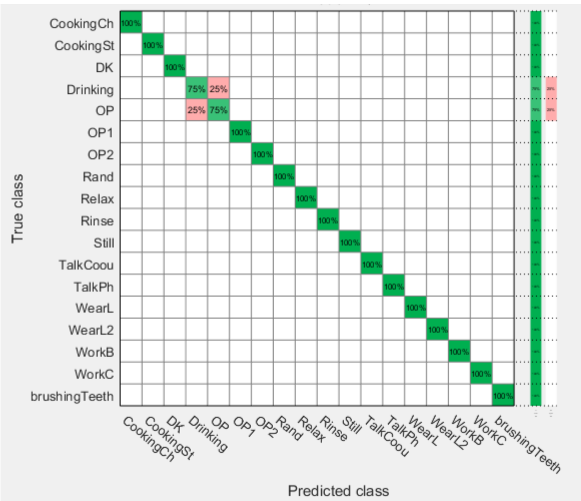


Figure 4.1: Confusion Matrix for SVM (Accuracy 0f 97.22)

of 97.2%/97.22% in ’ new person’ 5-fold cross-validation model. Our system has improved pre-processing steps that have reduced the number of invariant joints from feature vector by dynamically checking for joints in any given activity that show minimal movement. In our system, Histogram of oriented gradients is taken over a single frame with respect to mid section torso joint. Bining process is done by calculating directional derivatives from reference joint to rest of the joints. Signed angles forms the bins in the histogram. Achieved results in ’new person’ setting are compared with previous results using similar histogram based approaches, as shown in Table 4.1.

Reference	Methodology	Dataset	Accuracy
[26]	Gaussian mixture based HMM	CAD-60	84
[38]	3-D Posture Data	CAD-60	77.3
[2]	Unstructured Activity model	CAD-60	67.9
[45]	Depth and Image Fusion	CAD-60	75.9
[46]	Skeleton Data based	CAD-60	93.5
	Our Method	CAD-60	97.222

Table 4.1: Table to compare our results with previous results

Our algorithm has out-performed all histogram based systems in activity analysis. Calculating angles and directional derivatives between torso and rest of joints gave us a feature vector resulting in 97.222% accuracy. We achieved classification accuracy of 97.222% using Histogram of oriented gradients to extract features and support vector machine for classification. Results from our setup are given in Table 4.1 and show a significant improvement over previous work results.

Bibliography

[1] Z. Cai, J. Han, L. Liu, and L. Shao, “Rgb-d datasets using microsoft kinect or similar sensors: a survey,” *Multimedia Tools and Applications*, vol. 76, no. 3, pp. 4313–4355, 2017.

[2] B. S. Jaeyong Sung, Colin Ponce and A. Saxena, “Unstructured human activity detection from rgbd images,” *ICRA*.

[3] J. K. Aggarwal and M. S. Ryoo, “Human activity analysis: A review,”

[4] J. L. R. Ortiz, “Smartphone-based human activity recognition,” 2015.

[5] S. C. Mukhopadhyay, “Wearable sensors for human activity monitoring: A review,” *IEEE Sensors Journal*, vol. 15, pp. 1321–1330, March 2015.

[6] T. Shany, S. J. Redmond, M. R. Narayanan, and N. H. Lovell, “Sensors-based wearable systems for monitoring of human movement and falls,” *IEEE Sensors Journal*, vol. 12, pp. 658–670, March 2012.

[7] L. T. D’Angelo, J. Neuhaeuser, Y. Zhao, and T. C. Lueth, “Sensor set for wearable movement and interaction research,” *IEEE Sensors Journal*, vol. 14, pp. 1207–1215, April 2014.

[8] Y. C. Kan and C. K. Chen, “A wearable inertial sensor node for body motion analysis,” *IEEE Sensors Journal*, vol. 12, pp. 651–657, March 2012.

[9] Y. Chuo, M. Marzencki, B. Hung, C. Jaggernauth, K. Tavakolian, P. Lin, and B. Kaminska, “Mechanically flexible wireless multisensor platform for human physical activity and vitals monitoring,” *IEEE Transactions on Biomedical Circuits and Systems*, vol. 4, pp. 281–294, Oct 2010.

[10] E. S. Sazonov, G. Fulk, J. Hill, Y. Schutz, and R. Browning, “Monitoring of posture allocations and activities by a shoe-based wearable sensor,” *IEEE Transactions on Biomedical Engineering*, vol. 58, pp. 983–990, April 2011.

[11] C.-Y. Chen, Y.-H. Chen, C.-F. Lin, C.-J. Weng, and H.-C. Chien, “A review of ubiquitous mobile sensing based on smartphones,” *International Journal of Automation and Smart Technology*, vol. 4, no. 1, 2014.

[12] F. Xia, C.-H. Hsu, X. Liu, H. Liu, F. Ding, and W. Zhang, “The power of smartphones,” *Multimedia Systems*, vol. 21, no. 1, pp. 87–101, 2015.

[13] B. H. Gerritsen and I. Horvath, “The upcoming and proliferation of ubiquitous technologies in products and processes,” 2010.

[14] L. Chen, J. Hoey, C. D. Nugent, D. J. Cook, and Z. Yu, “Sensor-based activity recognition,” *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, pp. 790–808, Nov 2012.

[15] X. Su, H. Tong, and P. Ji, “Activity recognition with smartphone sensors,” *Tsinghua Science and Technology*, vol. 19, pp. 235–249, June 2014.

[16] B. Krausz and C. Bauckhage, “Action recognition in videos using nonnegative tensor factorization,” in *2010 20th International Conference on Pattern Recognition*, pp. 1763–1766, Aug 2010.

[17] M. S. Cheema, A. Eweiwi, and C. Bauckhage, “Who is doing what? simultaneous recognition of actions and actors,” in *2012 19th IEEE International Conference on Image Processing*, pp. 749–752, Sept 2012.

[18] F. Liu, X. Xu, S. Qiu, C. Qing, and D. Tao, "Simple to complex transfer learning for action recognition," *IEEE Transactions on Image Processing*, vol. 25, pp. 949–960, Feb 2016.

[19] C. Sun, I. N. Junejo, M. Tappen, and H. Foroosh, "Exploring sparseness and self-similarity for action recognition," *IEEE Transactions on Image Processing*, vol. 24, pp. 2488–2501, Aug 2015.

[20] R. Bhardwaj and P. K. Singh, "Analytical review on human activity recognition in video," in *2016 6th International Conference - Cloud System and Big Data Engineering (Confluence)*, pp. 531–536, Jan 2016.

[21] P. Melillo, R. Castaldo, G. Sannino, A. Orrico, G. de Pietro, and L. Pecchia, "Wearable technology and ecg processing for fall risk assessment, prevention and detection," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 7740–7743, Aug 2015.

[22] X. Liu, L. Liu, S. J. Simske, and J. Liu, "Human daily activity recognition for healthcare using wearable and visual sensing data," in *2016 IEEE International Conference on Healthcare Informatics (ICHI)*, pp. 24–31, Oct 2016.

[23] L. Sun, D. Zhang, B. Li, B. Guo, and S. Li, "Activity recognition on an accelerometer embedded mobile phone with varying positions and orientations," in *Proceedings of the 7th International Conference on Ubiquitous Intelligence and Computing*, UIC'10, (Berlin, Heidelberg), pp. 548–562, Springer-Verlag, 2010.

[24] N. elenli, K. N. Sevi, M. F. Esgin, K. Altunda, and U. Uluda, "An unconstrained activity recognition method using smart phones," in *2014 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pp. 1–7, Sept 2014.

[25] G. S. Njoo, X. W. Ruan, K. W. Hsu, and W. C. Peng, “A fusion-based approach for user activities recognition on smart phones,” in *2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 1–10, Oct 2015.

[26] L. Piyathilaka and S. Kodagoda, “Gaussian mixture based hmm for human daily activity recognition using 3d skeleton features,” in *2013 IEEE 8th Conference on Industrial Electronics and Applications (ICIEA)*, pp. 567–572, June 2013.

[27] K. Adhikari, H. Bouchachia, and H. Nait-Charif, “Activity recognition for indoor fall detection using convolutional neural network,” in *2017 Fifteenth IAPR International Conference on Machine Vision Applications (MVA)*, pp. 81–84, May 2017.

[28] W. Zhao, R. Lun, C. Gordon, A. B. M. Fofana, D. D. Espy, M. A. Reinthal, B. Ekelman, G. D. Goodman, J. E. Niederriter, and X. Luo, “A human-centered activity tracking system: Toward a healthier workplace,” *IEEE Transactions on Human-Machine Systems*, vol. 47, pp. 343–355, June 2017.

[29] M. Li and H. Leung, “Multiview skeletal interaction recognition using active joint interaction graph,” *IEEE Transactions on Multimedia*, vol. 18, pp. 2293–2302, Nov 2016.

[30] H. Xu, Y. Lee, and C. Lee, “Activity recognition using eigen-joints based on hmm,” in *2015 12th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI)*, pp. 300–305, Oct 2015.

[31] A. Jalal, S. Kamal, and D. Kim, “Shape and motion features approach for activity tracking and recognition from kinect video camera,” in *2015 IEEE 29th International Conference on Advanced Information Networking and Applications Workshops*, pp. 445–450, March 2015.

[32] E. Cippitelli, E. Gambi, S. Spinsante, and F. Florez-Revuelta, “Evaluation of a skeleton-based method for human activity recognition on a large-scale rgb-d dataset,” in *2nd IET International Conference on Technologies for Active and Assisted Living (TechAAL 2016)*, pp. 1–6, Oct 2016.

[33] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05)*, vol. 1, pp. 886–893 vol. 1, June 2005.

[34] A. Klser, M. Marszalek, and C. Schmid, “A spatio-temporal descriptor based on 3d-gradients.,” in *BMVC* (M. Everingham, C. J. Needham, and R. Fraile, eds.), British Machine Vision Association, 2008.

[35] O. Oreifej and Z. Liu, “Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences,” in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 716–723, June 2013.

[36] P. Scovanner, S. Ali, and M. Shah, “A 3-dimensional sift descriptor and its application to action recognition,” in *Proceedings of the 15th ACM International Conference on Multimedia*, MM ’07, (New York, NY, USA), pp. 357–360, ACM, 2007.

[37] J. Liang, J. Zhou, and Y. Gao, “3d local derivative pattern for hyperspectral face recognition,” in *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, vol. 1, pp. 1–6, May 2015.

[38] S. Gaglio, G. L. Re, and M. Morana, “Human activity recognition process using 3-d posture data,” *IEEE Transactions on Human-Machine Systems*, vol. 45, pp. 586–597, Oct 2015.

[39] L. Xia, C. C. Chen, and J. K. Aggarwal, “View invariant human action recognition using histograms of 3d joints,” in *2012 IEEE Computer Society Con-*

ference on Computer Vision and Pattern Recognition Workshops, pp. 20–27, June 2012.

[40] F. Han, X. Yang, C. Reardon, Y. Zhang, and H. Zhang, "Simultaneous feature and body-part learning for real-time robot awareness of human behaviors," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2621–2628, May 2017.

[41] "Kinect working principle." `http://www.microsoft.com`. Accessed December 25, 2014.

[42] J. M. Chaquet, E. J. Carmona, and A. Fernández-Caballero, "A survey of video datasets for human action and activity recognition," *Comput. Vis. Image Underst.*, vol. 117, pp. 633–659, June 2013.

[43] A. M. J. Sarkar and A. M. Khan, "An active process for sensor-based activity data collection," in *2011 24th Canadian Conference on Electrical and Computer Engineering(CCECE)*, pp. 001536–001539, May 2011.

[44] E. Cippitelli, E. Gambi, S. Spinsante, and F. Florez-Revuelta, "A human activity recognition system using skeleton data from rgbd sensors," in *2nd IET International Conference on Technologies for Active and Assisted Living (TechAAL 2016)*, pp. 1–6, Oct 2016.

[45] B. Ni, Y. Pei, P. Moulin, and S. Yan, "Multilevel depth and image fusion for human activity detection," *IEEE Transactions on Cybernetics*, vol. 43, pp. 1383–1394, Oct 2013.

[46] E. Cippitelli, S. Gasparrini, E. Gambi, and S. Spinsante, "A human activity recognition system using skeleton data from rgbd sensors," *Intell. Neuroscience*, vol. 2016, pp. 21–, Mar. 2016.