REAL-TIME HUMAN MOTION DETECTION AND ACTIVITY RECOGNITION

SANDAR WIN

UNIVERSITY OF COMPUTER STUDIES, YANGON

JUNE, 2024

Real-Time Human Motion Detection and Activity Recognition

Sandar Win

University of Computer Studies, Yangon

A thesis submitted to the University of Computer Studies, Yangon in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

June, 2024

Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research and has not been submitted for a higher degree to any other University or Institution.

…..…… ……..……

Date Sandar Win

ACKNOWLEDGEMENTS

First and beginning, I would like to express thanks my gratitude to His Excellency, the Minister of the Ministry of Science and Technology, for providing comprehensive facility support during my doctoral studies at the University of Computer Studies, Yangon.

I would especially want to thank Dr. Me Me Khin, the Rector of the University of Computer Studies, Yangon, for providing me with general supervision during my study period and for allowing me to work on this thesis.

I also express a lot of thanks to Dr. Si Si Mar Win, Professor, and Course-Coordinator, the University of Computer Studies, Yangon, for her encouragement, suggestions, valuable guidelines, motivation, and recommendations.

Additionally, I would like to express my gratitude and special appreciation to Dr. Tin Thein Thwel, a professor at the University of Computer Studies, Yangon, for her insightful remarks, suggestions, and support, all of which have been very helpful to me.

My supervisor, Dr. Thin Lai Lai Thein, Professor, Data Analytics Lab, the University of Computer Studies, Yangon, has given me great advice, compassion, and patience. She has also given me great ideas for my research suggestions for which I am incredibly grateful.

I would like to sincerely thank Daw Aye Aye Khine, Professor, and head of the English Department, for her invaluable assistance with language and for pointing out the proper usage of certain words in my dissertation.

I also thank my friends from Ph.D. 11th Batch for providing the support and friendship that I needed.

I am very much thankful to my parents and my younger sister for always believing in me and for their endless love and support. They have always provided and encouraged me during the years of my Ph.D. Thesis.

ABSTRACT

One of the interest areas of computer vision is real-time human motion detection, tracking, and activity recognition. It has many applications in a variety of fields, including video processing, abnormally detection, behavior prediction, human-computer interaction, video surveillance, and content-based image retrieval systems. This technology is essential in the fight against crime, terrorism, and threats to public safety. Due to variations in human appearance, changes in illumination, and the volume of data generated, video-based real-time human activity recognition is a difficult and demanding task. Supporting a safe and secure environment for real-time motion detection, tracking, and activity recognition is the aim of this research. The system detects human body parts with skeleton and to define activity based on joint sequence movement and to extract more reliable manner for overlapping area and to solve similar pose with different activities.

The goal of this proposed system is to enhance an automated video surveillance system that can identify and track people in both indoor and outdoor settings. The main step of the system involves motion detection, tracking and activity recognition through several steps: First, the system is designed to capture input video and extract region of interest for each frame. And generate features to estimate human and to detect 2D joint projected positions. Then, human detection is applied by using OpenPose detector and categorizes 2D joint sequence of body parts. The system recreates a human skeleton joint in three dimensions using spatial-temporal integration of human body parts. Finally, recognizes the activities such as standing, walking, sitting and running according to joint collection distance and displacement of skeleton joint position.

With a deep learning framework, the proposed method operates a robust human skeleton model that is unaffected by changes in the environment or various circumstances. Using joint estimation and position recognition, the system builds a skeleton model from the data perception. The objective of this research is more robust and efficient approach in human detection and activity recognition system from training and testing of multiple data generation by using deep learning approach to recognize different human activities changes in real life environment. The system's total accuracy is 94%, and the proposed approach performs better than expected when it comes to 3D skeleton model-based human detection and activity recognition.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST Of EQUATIONS

CHAPTER 1
INTRODUCTION

Real-Time human motion detection and activity recognition have a lot of advantages to wide application area in computer vision. Computer vision used image and video to know and understand real-world scene and has applied in many application areas such as safety workplace, robotics, healthcare systems, interactive system and behavior analysis, etc. To recognize human activity, human skeleton has stimulated to interest the system development that comprises valuable information and usable on real-time application of computer vision resources.

Human body recognition can interpret visual information of the surrounding environment and can record evidence to solve real-world problem. There are many research efforts and encouraging advances in the past decade, but accurate detection, tracking and activity recognition system have been required on different human body with different background. Human skeletal model can provide complex activity and can extract real-life information from the human movement through the order of joint sequence.

Several improvements have been occurred by using human skeleton that is directly taken from sensors through active sensing or passive sensing. Many researchers used depth maps that is captured from RGB-D camera. This means each pixel of image contains the distance of a point in the scene from the camera and applies information of pixel point to recognize human activity in their system. But that is cost effective and it is fine in limited range for outdoor applications.

Development of human skeleton with 3D modelling systems have been proposed in recent years, but these are unsatisfied to get reliable manner. Currently human activity recognition with 2D to 3D skeletal model has limitations in various aspects and loss of information caused by projection error. Most of the system required for overlapping area and similar pose with different activities for better performance.

The principle sources of difficulties in moving object detection are changes in appearance, partial occlusions of the target by other objects, complexity of the background and environmental changes. Detection of human from the video has occurred the numerous variation of the human pose, shapes and illumination changes.

This needs a well-defined method to manage the different motions for different situations.

The proposed system has developed 2D to 3D skeletal model by using deep learning approach to get more accurate result for human activities. Deep learning algorithm can achieve high accuracy for human detection and recognize human actions from visual data. Human body recognition from 2D to 3D skeleton model is generally focused based on raw position, orientation, joint displacement, and integrated information.

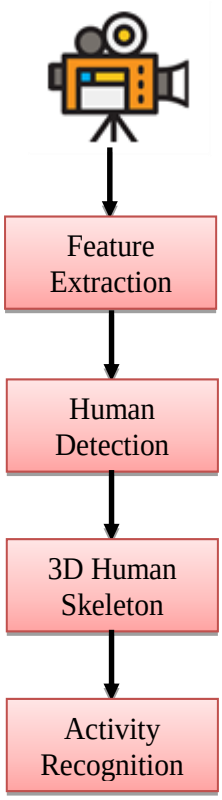


Figure 1.1 Overview of the Proposed System

The Figure 1.1 shows overview of the proposed system that includes feature extraction from video sequences, detection of human based on 2D joint sequences, reconstruction of 2D to 3D human skeleton and human activity recognition based on joint sequences activity. The proposed system is based on a human skeleton model, which may reveal hidden human body parts in three dimensions and demonstrate how humans interact with one another, overlap spaces, and alter their daily routines.

The underlying structure of motion states can be interpreted by the human pose component, which can then be analyzed to identify human activity. The twenty fundamental points that comprise the nose, right eye, left eye, right ear, left ear, neck, chest, waist, left shoulder, left elbow, left wrist, right shoulder, right elbow, right wrist, left hip, left knee, right hip, right knee, and right ankle are the definitions of the human body parts in this proposed system.

The suggested system is made to simulate the human skeleton robustly using a deep neural network that is resistant to changes in the environment and changing conditions. The system uses joint estimation and pose recognition to build a skeleton model from data perception. The goal is to use a deep learning methodology for training and testing with numerous data creation in order to build a more reliable and effective method for human detection and activity recognition.

1.1 The Scope of Computer Vision

A broad field of artificial intelligence known as "computer vision" uses photos and videos to identify, track, and categorize objects or events in order to comprehend scenes found in the real world. This facilitates all of the functions carried out by biological vision systems, such as "seeing" or seeing an image, comprehending what is observed, and deriving intricate information into a format that can be utilized by other systems.

A wide range of computer vision applications, such as those in e-commerce, gaming, automotive, manufacturing, and education, depend heavily on image recognition, clinical or biomechanical system, robotic, facial recognition, and identifying pedestrians and etc. Deep Learning becomes a major breakthrough in computer vision. To train metadata, the images must be assigned in the form of identifiers, captions, or keywords.

Current problems for human activity recognition system in computer vision that need to define complex actions recognizing and behaviors on realistic scenarios. When trained on complex computer vision tasks, such as autonomously summarizing the content of images and recognizing and localizing objects in video, they have shown to be highly effective. Real-time information is very helpful for decision-makers, but overlapping of joint sequences can lead to incorrect recognition of human activity.

1.2 Motivation of the System

Real-time video-based human activity detection is challenging due to variations in human appearance, lighting, and data generation. Misdetections have been occurred if multiple objects are close together, or if moving objects is away from the camera and appearance of moving object is similar to background and overlapping area.

There are many research efforts with sensor device and RGB depth camera, but that is costing, short range, limited people and does not recognize for lighting changes. Unreliable depth data under direct sunlight is unsuitable for multimedia video and other very large bit sequences. In a crowded location, it is more challenging to estimate the next positions and tracking sizes of moving kinematic patterns. that a person is obscured in part or entirely between camera frames by nearby people or a cluster of people standing close together.

Human motion has complicated structures in segmentation, modeling and occlusion handling. Therefore, moving object detection, tracking and activity recognition have been required to define robust method without affected by changes of the environmental features. The idea is to provide current trends and open problems for human detection, tracking and activity recognition for video sequences

1.3 Objectives of the System

The primary goals of this research are to provide a safe and secure environment for motion detection, tracking, and activity recognition in real time and to detect human body parts with skeleton and to define activity based on joint sequence movement and to extract more reliable manner for overlapping area.

There are various methods for analyzing the intelligible motion of objects in a sequence to recognize human activity, but detection of overlapping area and recognition of similar pose for different activities have been required. The principal sources of difficulties in moving object detection are changes in appearance, partial occlusions of the target by other objects, complexity of the background and environmental changes.

Numerous varieties of human position, gesture, and activity have been identified as human activity in the film. The suggested approach created a 2D to 3D skeleton model based on the components of the human body in order to obtain a trustworthy method. A 3D skeleton model may highlight obscure human body parts in 2D photos, allowing people to engage and see overlapping groups and human activities.

This research's primary goal is to identify human behavior in a video or a series of photos and define skeleton model activities. These are the research's other goals:

- To support real time interpretation and robustness
- To develop intelligent visual surveillance system
- To improve performance of object detection and tracking
- To construct 3D human joint sequence model
- To recognize human activity changes overtime

1.4 Problem Statement

The principle sources of moving objects are different appearances, partial occlusions of body parts, complexities of the backgrounds and environmental changes. Detection of human from the video sequences has occurred numerous variation of the human pose and that need advanced methods to manage different motions and different situations. Occlusion of body part detection, tracking and activity recognition have remained a problem in computer vision. In order to have effective action recognition performance, the object's visual model may alter briefly if occlusion happens.

The displacement of joint movement is used to identify human motions; some joint points are noise-producing and irrelevant. The performance of recognizing human activities will be lowered by this issue. Consequently, many different human body poses may have very similar image projections. Each skeleton joint can vary from one activity to another.

In most actions, many joints contribute very little changes and not significant for action recognition. Images with similar human poses can appear different because of changing viewpoints, subjects, backgrounds, clothing, etc. Hence, they even bring noises which will affect the overall performance. The positions of joints are estimated

from low resolution score maps and this reduces the accuracy of joints location the current problem in action recognition with multiple viewpoints is still missing in the case of overlapping area.

The suggested effort is divided into two sections to address the issue of mixed human body parts: the first involves human detection using a deep learning algorithm, and the second one is activity recognition on 3D skeleton joint sequences. An action can be defined by spatial-temporal sequence of human body movements.

1.5 Contribution of the System

This research focuses on developing 2D to 3D skeletal model with deep learning method. These focused works include the following:

- Capture input video and extract region of interest for each frame.
- Generate feature to estimate human and to detect 2D joint projected positions.
- Detect human activities by using OpenPose detector and categorize 2D joint positions as sequence of body parts.
- Reconstruct 2D to 3D human skeleton using spatial-temporal integration of human body parts.
- Define activity recognition according to joint collection distance and displacement of skeleton joint position.

1.6 Organization of the System

Six chapters make up this research article, the first of which introduces the human activity recognition system, the motivation of this research work, problem statements with human detection and activity recognition system, goals, concentrates and contributions of the research work are described. The remainder of the research work is organized as follows:

Chapter (2) surveys the various approaches used in literature reviews that address dissertations.

Chapter (3) represents the theoretical context in research work. In this chapter, human detection from 2D to 3D model and activity recognition are carried out.

Chapter (4) describes the suggested system's design and suggested algorithms for human detection and idea are discussed.

Chapter (5) presents the planning and execution of the suggested approach and the evaluation of the experimental results by measuring with confusion metrics and processing time.

Finally, Chapter (6) concludes the research represent and illustrates potential areas for further study to carry out this work.

CHAPTER 2

LITERATURE REVIEW AND RELATED WORK

Because of its practical uses, human identification and activity recognition from video sequences has advanced significantly. However, there is still much work to be done in real-life. Various techniques and utilizations have been stated in research field and that developed more and more effective recognition results based on advanced artificial intelligence techniques described by a large number of literatures. There exist various methods and different technologies have been developed in the past decades, but perfect human motion detection and activity recognition is still a significant issue. Due to the many challenges that consist of changes in the human form, a cluttered background, camera settings, and lighting, view point changes, target detection, localization, subject appearance, and image occlusion.

Nowadays, Human detection and activity recognition is standing a fundamental role in real-world uses, including security cameras, robotic, automated security, health care system, virtual reality creating, safety support to IoT and human activity monitoring in private and public areas. The primary objective is to understand human or group behavior from video sequences, including the context in which the behavior occurs. This can be described as an instinctive interpretation of the actions that take place in a video sequence that feature human activity. Different activity recognition systems have been improved from year to year by many researchers.

Zarka et al. [69] described a real-time motion analysis, tracking, and person detecting system. Adaptive Background Model is used to get foreground pixels, and detect the silhouette shape. They applied human model based on the star skeleton and recognized the dynamic variation of human motion with standing, walking and running Haritaoglu et al. [15] proposed silhouette shape-based analysis for 2D body modeling. They establish a dynamic appearance model by utilizing the boundary's centroid, major axis, and contour. The system detected body parts (head, hands, feet, torso) of people and analyzed a person carrying an object or two people exchanging bags. An infrared camera is used to record videos in real time, and capable in detecting a single person, multiple people, and individuals carrying of various objects. However, this system does not utilize color cues for detection.

Banerjee et al. [4] used the Gaussian Mixture Model in conjunction with the Adaptive Background Model to build machine learning. They used features vectors based on Histogram of Oriented Gradients approach and classified human and non-human by using SVM. N. Dalal and B. Trigs [10] proposed human detection process using the HOG feature extraction. The basis of this identification method is the difference between the silhouette shape and the background. It also identifies human presence in areas that overlap. J. Grahn and H. Kjellstrom [14] pointed out issues for varying image pattern sizes in video sequences. Linear Spatial-Temporal difference filters and SVM were employed, but some problems with detecting persons stepping away from the camera's lens and false positives state between two humans have been remained.

Viola et al. [57] defined a collection of data regarding image intensity and motion appearance information Using a Cascade classifier, they trained and evaluated both static and dynamic human motion patterns. However, their method fails to detect human motion in windows smaller than 20x15 pixels. With advancements in sensor devices and visual technology, HAR based systems are more powerful and valuable in many real-world systems. Particularly, the production of small size sensor devices has enabled to recognize the human activities [68]. Depending on the design process and data collection strategy, these approaches fall into three categories: multi-modal, non-visual sensor-based, and visual sensor-based. How these sensors interpret the data is the main way that they differ from one another.

While conventional sensors only supply data as a one-dimensional signal, visual sensors can deliver data as 2D or 3D images or films [44]. In recent years, the wearable devices such as smart-watches, smart-phones, and health wristbands are popular and useful. These gadgets are available with available connectivity capabilities and have computing capacity. and that is suitable for HAR [52]. With the development of advanced technology, various improved methods with deep learning have been occurred and skeleton based human representation has been used extensively for activity recognition due to its ability to handle complex situations and dynamic environments. Along with several developing changes of human detection and recognition approaches, deep learning-based recognition result is a core component for researcher.

Luvizon et al. [37] presented a multitask framework for 2D joint that can recognize human motion from video sequences, estimate 3D position from still images.

In this system, generation of 3D is from 2D annotated data and action recognition is based on estimated pose and visual information. They trained multiple types of datasets to create 3D predictions and demonstrated a productive approach to action recognition using skeletal data. Next interesting as Vosoughi et al. [58] applied CNN based two parallel network; joint detection network and pose regression network. First, they analyze joint present vector or not with joint detection network and reconstruct full pose on regression network. The result is integrated as partial pose estimation and classified presence and absence of the body joints in image.

Using the Covariance of 3D joints matrix over the sub-sequences frame in a temporal hierarchy manner. Hussein et al. [20] suggested the human skeleton. Its length is set and is independent of the length of the sequence. A random joint probability distribution variable is used to generate corresponding feature maps in the area. The system used SVM for action classification and cooperative location deployment over time. Plagemann et al. [42] examined human shape by focusing on prominent spots on the human body, leading to the creation of further approaches. The method used a Bayesian network to learn the estimated positions of the body parts and directly estimated the 3D orientation vector from the body component in space.

The shadow silhouette-based skeleton extraction (SSSE) approach, which examined silhouette data, was presented by Hou et al. [17]. The process of skeleton synthesis and 3D joint estimate is based on the extraction of 2D joint positions from the ground's shadow area. Major joint position is determined and contrasted with RGB-D using the 3D skeleton configuration data. Especially deep learning is become superficial research as continue in applications, Wu and Shao [62] suggested employing deep neural networks with a hierarchical dynamic framework for the recognition of human actions. Using probability distribution models and 3D skeleton data, they successfully established an effective technique for action recognition based on skeleton joint information.

A dual-source network for 3D human posture estimation was created by Iqbal et al. [23]. They found the closest 3D pose after gathering a huge quantity of unrestricted data in both 2D and 3D stances. By reconstructing the closest 3D position, the system was able to estimate a single image and produce the desired recognition outcome. suggested the Double-feature Double-motion Network (DD-Net), which makes use of skeleton sequence properties and motion scale variances to create a lightweight network architecture. A translation scale-invariant technique that works

well with 2D skeleton movies was reported by Li et al. [25]. Convolutional Neural Network (CNN) architecture was used, and benchmark datasets were used to compare the outcomes.

Most of the early systems used handcrafted features with edges and corners and later developed with learning approaches. In the deep learning process, millions of images and videos are used for feature generation and description of extracted features along the network and classify the activity. The dynamic motion encoder approach in [2] captured temporal information of skeleton body joint movements and classified the activity with CNN. They applied Mask R-CNN as a pose extractor to extract body joint key points. As the experimental result, the combination of dynamic motion with the spatial-temporal convolution method best achieved a mean accuracy of 87.2% on JHMDB, 84.2% on HMDB, and 98.4% on the UCF-101 dataset.

Tasnim et al. [53] explored human action recognition by spatial information and temporal changes on 3D skeleton-joint sequences. They used three fusion mechanisms by using joint and line mapping on spatial-temporal changes in two consecutive frames and achieved the performance on three deep learning networks. The multimodal feature fusion with skeleton and RGB modalities in [67] improved the classification of human activity. The system used the spatial-temporal part from the region of interest and applied the joint weight attention feature from the skeleton model. In this system, the fixed attention mechanism achieved performance and improved the system.

Zhu et al. [73], applied an attention mechanism on three spatial-temporal channels (I3D-CSTA) to recognize fine-gain action with 3D convolution. As a mentioned result, they achieved 95.76% on UCF101 and 73.97% on HMDB51 for the RGB frame. Khan et al. in [28], used a feature fusion mechanism with DNN and multi-view by using pre-trained VGG-19. They combine gradient information which has a high probability consisting of joint probability distribution of observed random variable information, and classify the activity using Naïve Bayes classifier. As a test result, the system achieved high accuracy as 93.7% on HMDB51, 98% on UCF sports, 99.4% on YouTube, 95.2% on IXMAS, and 97% on KTH datasets.

Kani et al. in [26] proposed AWFN algorithm with key frame selection using Optical Flow with adjacent pixels' movement. And, getting 96% performance on the VGG-16 network using UCF101[50] and HMDB51[31] datasets and increasing 0.3% to 7.88% accuracy on the existing method. There are many attempts that have been made in past research fields and methods were developed year by year. However, the

recognition results can be affected by human occlusion with one another, self-occlusion, and ambiguity on input situation have remained and computational processing time is a considerable factor. Human motion has complicated structures in segmentation, modeling and occlusion handling. Our idea is to provide current trends and open problems for human detection, tracking and activity recognition for video sequences with developing of deep learning framework and using with skeletal sequence generation to recognize similar pose, but different activities and to know overlapping area that can achieve misrecognition results.

2.1 Intelligent Sequences of Video Content

A video is a sequence of images (called frames) and displayed at a given frequency. Frame processing from video sequence is an effective way for visual source of information in static and dynamic motion capture system. The collection of sequential video images with a constant time interval can provide more information that object changing with respect to time. In Figure 2.1 shows the description of scene from video sequences and we can know contents of information such that

- Who are present or not?
- How many humans in this area?
- What are they doing?

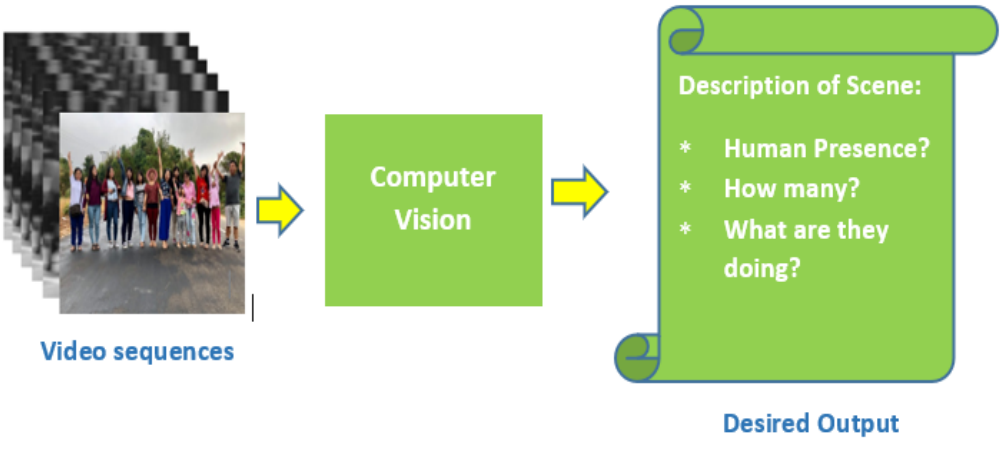


Figure 2.1 Description of Scene from Video Sequences

Depending on application, motion capture system can be different such as RGB videos, infrared sensor and depth camera with different sources like wide angle time of

flight technology and structure light camera such as RGB-D camera, Kinect camera, Asus PROLIVE etc. The conventional cameras perceive the object in two dimensions, but real-physical world is three-dimensional space and object perceive in three-dimensional form. The captured data is used to process in motion detection and recognition. An important phase of video sequence execution is to segment region of interest and that can be analyzed for later process.

There are many algorithms for motion detection and recognition such as adaptive background modeling, detection of image region, robust tracking strategies, and learning based activity recognition. Learning System can learn moving of human activities from raw videos and quickly recognize the capability of human action. A deep understanding of human activity can be represented by skeleton data which is resistant to noise and robust enough to extract relevant features from the motion of an image over time and a semantic configuration over an image space.

2.1.1 Representation of Human Body Structure

Human Body Structure can be portrayed by the ways that they look and shape. There are various illustrations of human figure, which is used for detection and position of address representations as follows:

- **Points:** A point, which could be a group of points or the centroid, is used to symbolize the bodily portion. It is suitable to use the point representation to draw attention to certain objects, such as small areas inside an image.
- **Geometric Shape:** Geometric shape means the object shape is represented by rectangle or ellipse. Basic geometric forms can be used to represent both non-rigid and simple rigid things.
- **Object silhouette and Contour:** The edge of a body portion is its contour. The area enclosed by the contour is referred to as the object's silhouette. For tracking, these representations are helpful.
- **Articulated shape models:** The bodily pieces that make up articulated objects are joined via relative joints. The human body, for instance, is an articulated item that is kept together by the hands, legs, and feet. Motion models guide the connection between these components.
- **Skeletal model:** Human skeleton is ridge line of object that can be used to extract interest skeleton feature points based on the silhouette of the object. This

model is used to depict shapes in order to identify items. Both stiff and articulated body parts can be used with skeleton models.

2.1.2 Body Model and Feature Extraction

Human body has large number of degree of freedoms and variability of articulated shapes. In early theoretical study, 3D shape is considered from cylindrical human model. Karl Rohr [9] proposed top-down approaches that consists of hierarchy body model from 3D cylinder to 2D configuration. In bottom-up approach, first track body parts in 2D and lifting 2D into 3D configuration by matching of projected motion capture data into image data through motion frames [21].

Basic transformation can modify the human body which increase or decrease the size of object, rotate the object around 2D or 3D, or translate the object to new place. A translation is a transformation that moves an object relative to its current position and easy to visualize. Rotation of an object means a transformation that changes the orientation of an object. Scaling is from the size of the object changes with respect to its current size. The spatial-temporal arrangement of human body movements can characterize an action. Most of the human model is used for analysis-by-synthesis detection and recognition along the human movements.

2.1.3 Human Joint and Pose Estimation

For action recognition tasks, defining human joint information from skeleton models has demonstrated remarkable success and that can extract overlapping area in hidden human body parts and can express more accurate information on body pose.

The localization of human joints in pictures or movies is known as pose estimation. Pose information was typically acquired from motion capture systems. Early works in recognition of human motions depend on articulated poses from frame to frame and connecting together with the pose-derived features into the spatial-temporal trajectory.

The key points of human pose are usually line up with skeleton joints relative to associate limbs along the structure of human body parts. Figure 2.2 is an example of twenty labeled key points corresponding to three main body parts consists of head, trunk and limbs; each component be associated with human pose.

Du et al. [11] proposed RPAN that forecasts relevant characteristics in human posture. In contrast to earlier research on pose-related action recognition, RPAN is an end-to-end recurrent network that can utilize human pose's spatial temporal evolutions to support action detection inside a single, cohesive framework. By sharing attention parameters partially on the semantically associated human joints, the pose attention mechanism learns robust human-part properties.

Head:	Nose	Limbs:	Left Shoulder
	Right Eye		Left Elbow
	Left Eye		Left Wrist
	Right Ear		Right Shoulder
	Left Ear		Right Elbow
Trunk:	Neck		Right Wrist
	Chest		Left Hip
	Waist		Left Knee
			Left Ankle
			Right Hip
			Right Knee
			Right Ankle

Figure 2.2 Key Points of Human Body Parts

2.1.4 Skeleton Model and Tracking Strategies

The skeleton is made up of several joints and segments that correspond to various bodily parts and limbs. Depending on the distance an object is from the camera, the pixel resolution can change, affecting the number of skeletal joints. The human skeleton is made up of stiff components joined together by joints. having distinct human bodies, but with about the same body part proportions.

The digitalized human motion data is composed of the position of skeleton point and segment coordinates, which allow for the estimation of path and trajectory movement. For example, the rotation of the ankle, knee, and hip joints can be defined using data from the joint angles. We are able to determine human behaviors and evaluate movements thanks to the relative joint skeleton data.

2.1.5 Position Feature Extraction

It is possible to categorize the feature information as global or local features to produce semantic based ability and to determine meaningful motions. The global features are useful to provide meaningful searches as well as identification. These characteristics include things like the object's size, border, center, descriptions, moments, and so forth. The local features, which are defined as features of interest to define criteria of object, are derived from the object's local neighboring regions.

Human body consists of a set joint and can consider for motion with the spatial body part movements over a series of frames by temporal sets, as well as body part arrangement in a single frame by spatial sets. That can consecutively configure the model between motion of human joints and body parts. The joint feature representation may include object size and location of image. In order to define for object position, correspondences of object information have been needed that contain the region of interest determined by the segmentation algorithm.

Joint Interest features are selected as the quantity of pixels found in every object. The window was the only place where similar elements in later frames were looked for. Since humans have articulated structure and changes may be from frame to frame on the overall shape and relative position of the object rotation causes to new exposure region as well as occlusion of previous ones.

The study of huge regions may contain extra points from articulated objects, but non-rigid faces will not organize for relative positioning. In order to be better understand the sequence, 2D to 3D representation can define more accurate result for appearance of object shape, object size, and object position over the sub sequence changes.

2.2 Human Detection Strategies

The main challenge in representing human mobility from perceptual data in a frame is human detection. The basic concept of motion-detecting objects and tracking is to detect the target object in the video sequence and track object over consecutive frames. The structure of the object may change in shape, size, location and orientation over subsequent video frames. An accurate and efficient way of detection method depend on the use of feature points. Object detection method is based on motion segmentation with frame differencing, temporal differencing, background subtraction

and optical flow method. In object detection, most of early system used adaptive background model to get foreground pixels, and detect the silhouette shape of human. Object detection in a video stream involves pre-processing, segmentation, foreground extraction, and feature processing, which proceed together lead to the final detection result. Various algorithms and techniques is mainly focused on shape-based, feature-based, motion-based, and learning-based.

2.2.1 Shape-based Approach

Shape-based approach has been designed using hierarchical shape template matching and human shapes are directly modeled as a set of global shape templates organization in a hierarchical tree manner. Global shape models are generated by part synthesis and decompose each of boundary into region with part-templates model. These methods work well in cluttered scenes with partial occlusion of humans.

Lin et al. [35] used combination of Local part and global shape-template to detect humans for inter-occlusion purpose. The different parts of templates are obtained from the global shape models with top-town decomposing structure. Each part in the structure has a default location/size and is categorized with shape and region information.

Beleznai et al. [7] introduced efficient detection with contour-based template matching and analyzed subsequent occlusion. They used an approximated form of shape context descriptor to get reliable hypotheses in presence of occlusions and estimated object parts. The combination of local and global shape cues and demonstrates accurate detection performance in complex scenes.

The template matching method is a useful for various pattern matching system, the existing have measurement of uncertainties with respect to translation, rotation, scaling, and other variations. Another approach is using of sparse local features or visual parts collection to detect body parts. These approaches learn body parts based on sparse interest points and descriptors.

Wu et al. [63] pointed out part detectors based on edgeless features. In this system, responses from part detectors are cooperative to form a joint likelihood model and that includes an analysis information with occlusions. Trajectory initialization and termination are automatic and rely on the confidences computed from the detection responses.

2.2.2 Feature-based Approach

Image Features can be represented by object properties such as colors, edges, corners, interest points, regions and ridges. High level feature might be centroid, entire range or interested point of object. Low level feature such as color representation, slope and movements.

A verity of methodologies for detection of moving objects are focused on feature extraction techniques. David G. Lowe [36] described Scale Invariant Feature Transform (SIFT) with reliable to invariant. SIFT extracts key points relative to local feature by using Hough Transform and recognize pose estimation with least square determination that are workable to changes in illumination, scaling, noise and variations in sense of perspective.

There are different feature extraction methods and another robust method is Speed Up Robust Features (SURF) [5]. That is local feature descriptor that is working on three steps as detection, description, and matching steps. The matching score is computed based on Euclidean distance between vectors and it is reliable for local variation in illumination changes and orientation.

Next fast feature extractor is Histograms of Oriented Gradients (HoG) approach, a gradient is computed on each pixel where orientation is defined by direction of the gradient and the magnitude of the gradient as weight. The histograms of all cells are configured and train to a machine learning discriminator to classify current detection window whether the state of the result or not [10].

And improved efficient feature descriptor with Haar-like feature that is faster extraction process than HOG and that can perform on low-resolution images under various illumination situation [40]. For 3D skeleton sequences, to get the benefit of the temporal relationship between joint movement which Hand-crafted features by using multiple covariance matrices for skeleton joint locations over time achieve better performance for activity recognition [20].

2.2.3 Motion-based Approach

Motion-based approach is based on changes pixel detection and most of the system use with background modeling due to simplicity and computational efficiency. This approach is applied the periodic property of the captured images to recognize moving objects from other. In literature, there are four different approaches that can be

employed as statistical method, background subtraction, temporal differencing and optical flow.

Statistical methods can be used to extract change regions by the basic background subtraction method. The characteristics of individual pixels or group of pixels have been used to construct more advance background models which can be updated dynamically [1].

Background subtraction is the widely used technique for moving objects detection. It involves absolute difference between current image and background image. The background image needs to be updated regularly that can adjust influence of dynamic scenes [43]. Migdal et al. proposed the novel method for background model that develops the spatial and temporal dependencies in objects motion through with Markov random fields of binary segmentation. That approach produces more accurate and less prone to noise and handles the complex motion. [39].

Temporal differencing [49] is by taking of pixel-by-pixel difference between two or more consecutive frames to extract moving objects in an image sequence. The presence of moving objects is determined by the difference between two consecutive images. Frame differencing technique is very simple and easy to define but it is difficult to achieve sufficient working with moving object as result of moving object detection is not accurate.

Optical flow [19] based on motion detection uses optical flow estimating, background modeling and foreground extracting to denote independently of moving objects even in the presence of camera motion. The background model is considered on transforming of optical flow utilizing and a dual-mode judge mechanism to improve the system's variation to different situations. The advantage of Optical Flow is quick calculations and good result for object tracking. But that have been required for occlusion and lighting changes.

Kim et al. [29] proposed a motion compensation method to estimate background motion through feature matching. These methods compensate for the camera movement in dynamic background modeling. A recent shift of focus towards part-based representations has resulted in detection methods capable to detect parts of humans and perform occlusion reasoning based on the part-detection results.

Motion detection results depend on foreground detection. Zhaoyang et al. [70] introduced background model to detect the moving area by convolution neural network

to get effective detection result. This method improves the detection rate for small moving object.

2.2.4 Learning-based Approach

Learning-based representation, and in particular deep learning, has introduced the concept of end-to-end learning by using the trainable feature extractor followed by a trainable classifier. Deep learning to extract the best features learning mainly employs two approaches:

1) preserving pre-trained network and updating the weights based on the new training dataset
2) using the pre-trained network for feature extraction and representation, followed by a generic classifier

In order to train a learning model, the first task is to collect training images with label data. The next task is to extract the features and add it to the classification model. Many recent works use convolutional neural networks to learn feature representations for obtaining the score maps. One main problem for detection model is that the positions are estimated from low resolution score maps.

The success of the training process depends on the feature extraction, classifier selection and training step (i.e. the iteration process). Among these steps, feature extraction is very important to get the desired outcome. Feature extraction minimizes the data dimensionality by extracting the redundant data, thus improves the inference and training speed.

Heo et al. [16] proposed deep learning to achieve a robust performance against in a dynamic background. They considered appearance features in addition to motion features on deep learning architecture by utilizing the appearance of the target object and the motion difference. Eddy Ilg et al. [22] introduced end-to-end FlowNet2 with optical flow estimation and made it work really well on a large variety of scenes and applications.

2.3. Human Tracking Strategies

The purpose of the object-based tracking strategy is to create a temporal object consisting of collection of objects at sequential time frames. This result in a collection

of pixels that have been grouped in space and time. Initially this is performed using object size and location to determine corresponding objects. Tracking would be more robust if features could be correlated with corresponded objects identified from frame to frame.

The Kalman filter tracker based on the sequential model estimates state vectors via the Kalman filter equations [46]. The prediction and estimation are calculated based on measurement vector as

$$z_k = HX_{tk} + \mathrm{B}_k \tag{2.1}$$

where H is noiseless connection matrix between state vector and measurement vector, B_k is measurement error. The motion tracker based on above model and sequentially estimate state vectors via the Kalman filter equations. Estimating the current values of the state from past and current observations $\widehat{\mathrm{X}}_\mathrm{k}$ is based on forecasting subsequent values of the state as

$$\widehat{\mathrm{X}}_\mathrm{k} = \widetilde{\mathrm{X}}_\mathrm{k} + \mathrm{K}_\mathrm{k}(\mathrm{z}_\mathrm{k} - \mathrm{H}\widetilde{\mathrm{X}}_\mathrm{k}) \tag{2.2}$$

where K_k denotes Kalman gain and also forecasting subsequent values of the state as

$$\tilde{X}_k = \Phi\hat{X}_{k-1} \tag{2.3}$$

2.3.1 Detection-Based Tracking

Frame-by-Frame Detection: Objects are independently detected in each frame using object detection algorithms such as YOLO (You Only Look Once), SSD (Single Shot Multi-Box Detector), or Faster R-CNN.

Data Association: Detected objects in consecutive frames are linked based on their spatial and temporal variations. Hungarian algorithm or Kalman filters are popular methods for detections of object between frames. Kalman Filter can be used to measure observed variable over time to estimate the positions and velocities of moving objects.

2.3.2 Space Time Volume Estimation

Techniques based on space-time have gained popularity and shown promise on both basic and complicated datasets. Space-time volumes, or three-dimensional spatial-temporal cuboids, are used to represent the features in the space-time domain. This method's fundamental component is an action recognition similarity measure between two volumes.

In their proposal for an action recognition system, Bobick and Davis [8] used motion history images (MHI) and 2D binary motion-energy-images (MEI) to represent actions, and then used the volume motion template matching technique to recognize human actions.

This work was expanded by Hu et al. [18], who integrated MHI with two appearance-based features of the foreground image and the histogram of oriented gradients (HOG) for action representation. Instance learning using support vector machine classification was then performed.

A number of related works in action recognition are concerned with identifying joint movements and tracking human body components. An alternate field of study has regarded actions as space-time volumes inhabited by the body, as opposed to tracking the individual limbs. Direct classification of actions using low-level spatial-temporal information has gained greater attention recently. HAR is the foundation for many applications. The spatial-temporal arrangement of human body movements can characterize an action.

2.4 Activity Recognition Strategies

The definition of activities from still photos or video sequences is the aim of human activity recognition. The reason for this is because HAR systems classify incoming data according to the underlying activity category. There are two main types of vision-based methods for recognizing human activities. [45].

1) The conventional manual method, which uses descriptors and feature detectors like Hessian3D, Scale-Invariant Feature Transform (SIFT), and Histogram of Oriented Gradients (HOG), Local Binary Patterns (LBPs), Enhanced Speeded-Up Robust Features (ESURF) and followed by trainable classifier for action recognition.
2) The learning-based strategy, which makes use of raw data's inherent ability to autonomously extract characteristics. In contrast to the conventional manual method, it employs the idea of a trainable feature extractor, which is followed by a classifier.

2.4.1 Based on handcrafted features

In Early approaches for action recognition are mainly based on handcrafted features [59, 32]. That derived from various algorithms using the information present in the image itself which represent as a number of local descriptors. Other approaches deployed multiple covariance matrices over time as a discriminative descriptor and encode the relationship between joint movement over sub-sequences in a hierarchical manner. However, hand-crafted approaches may only capture the local contents and thus lack the discriminative power to recognize complex actions [60]. Karahoca et al. [27] proposed human activity recognition system with pattern recognition techniques. They used Motion History Images (MHI) and classify with Support Vector Machines and K Nearest Neighbors.

2.4.2 Based on Learning Features

With significant successes of CNNs in image recognition [58, 30], several efforts made to create efficient CNNs for video action identification [24, 6, 55]. Two-stream CNNs, in which spatial and temporal CNNs were created to process RGB images and capture motion information on a brief temporal scale, are among the most widely used techniques [48].

Recently, attention has been incorporated into CNN to learn detailed spatial - temporal action cues. Yang et al. [66] introduced the idea of a two branches attention architecture that focuses key stages on spatial-temporal and filters out unreliable joint predictions. Baccouche et al. [3] developed two-steps labeling scheme that utilizes spatial-temporal features over deep learning technology.

Belagiannis et al. [6] addressed the problem of multi-person 3D pose estimation. They build 3D body part state space from 2D body part hypotheses. To resolve the problem of mixed body parts of multiple humans, they used 3D pictorial structure to represent human shapes. But this method only considers on the geometric cues of 2D pose correspondences. In crowded scenarios, it is not robust due to heavy occlusion and truncation.

Bin et al. [34] proposed a spatial-temporal graph routing for skeleton based action recognition. That learns spatial connectivity by local clustering and temporal by correlation of node connectivity. They solved the weakness of predefined human structure and verify the effective classification result for action recognition.

2.5 Deeping Learning Strategies

Computational models consisting of several processing layers can acquire representations of data with various levels of abstraction through deep learning. In many computer vision applications, it has become a very popular direction in machine learning and has surpassed the conventional methods. Deep learning algorithms have the capacity to extract features from unprocessed data, obviating the requirement for manually generated feature detectors and descriptors. However, in order to train the algorithm, deep learning models need a massive amount of data.

An input layer, a fully-connected layer, one or more hidden layers, and an output layer make up a deep convolutional neural network's structure is designed as in Figure 2.3. The hidden layer applies two operations, namely convolution and pooling, which normalize the output with activation function to reduce over-fitting and under-fitting. To represent the output into the appropriate probability value in the output layer, the soft-max function is utilized. Initial weight values are chosen at random using both Gaussian and uniform distributions. To update the weight values, the stochastic gradient descent approach is applied. Up until it achieves the minimal loss for the prediction result, the network is trained both forward and backward.

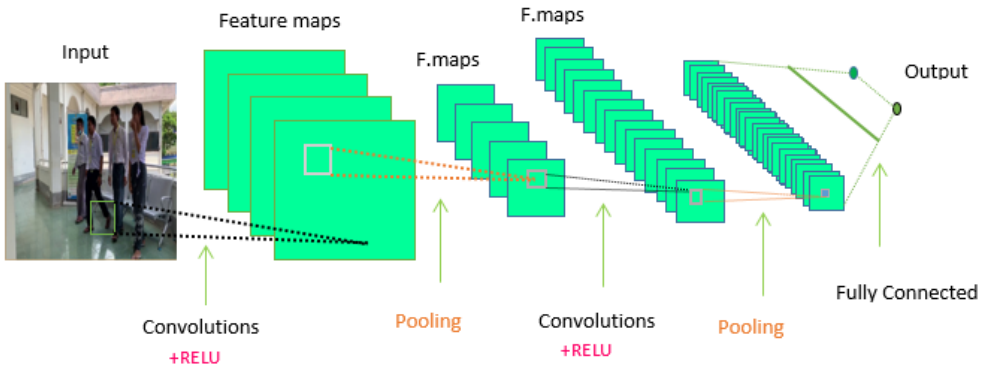


Figure 2.3 The Structure of Deep Convolutional Neural Network

Other attempts for deep learning approaches as Zhi et al. [71] applied on raw video sequences from depth sensors and extract spatial-temporal features by employing 3D-based Deep Convolutional Neural Networks and classified joint feature vector with Support Vector Machine, which are based on position and angle information between skeletal joints. Because deep learning-based techniques produce good recognition performance in several computer vision applications, they have become more popular.

Put otherwise, these achievements are mostly linked to supervised deep learning models. While recurrent networks have been used for sequential data like text and speech recognition, deep convolutional neural networks have made significant advances in the processing of pictures, videos, speech, and audio.

2.6 Summary

This chapter discusses the several recognition methods for the purpose of identifying human presence and activity with related methodologies and development results. Investigating the state-of-the-art detection and recognition will be the interesting research area that support computer vision into real-world problem. The development of sensor and vision technologies has led to the widespread usage of HAR-based systems in numerous practical applications.

Conversely, other researchers have put forth and categorized deep learning-based techniques with varying activities. Deep learning's capacity to analyze vast amounts of data makes it extremely effective for gaining a deeper comprehension of human behavior. In order to achieve strong action recognition performance, the study can more precisely define the localization of joints. This field's primary goal is to identify joint sequences and subsequently identify human actions within a movie or series of photos.

CHAPTER 3
BACKGROUND THEORY

This chapter describes the theory and background model of human detection and activity recognition for RGB video sequences. The basic background concepts of digital image processing concerned with the approach consists of image representation, segmentation method, human detection and tracking strategies, 2D to 3D configuration on human skeleton model and activity recognition on deep learning frame work are presented in detail. The development process of the working system is shown in Figure 3.1.

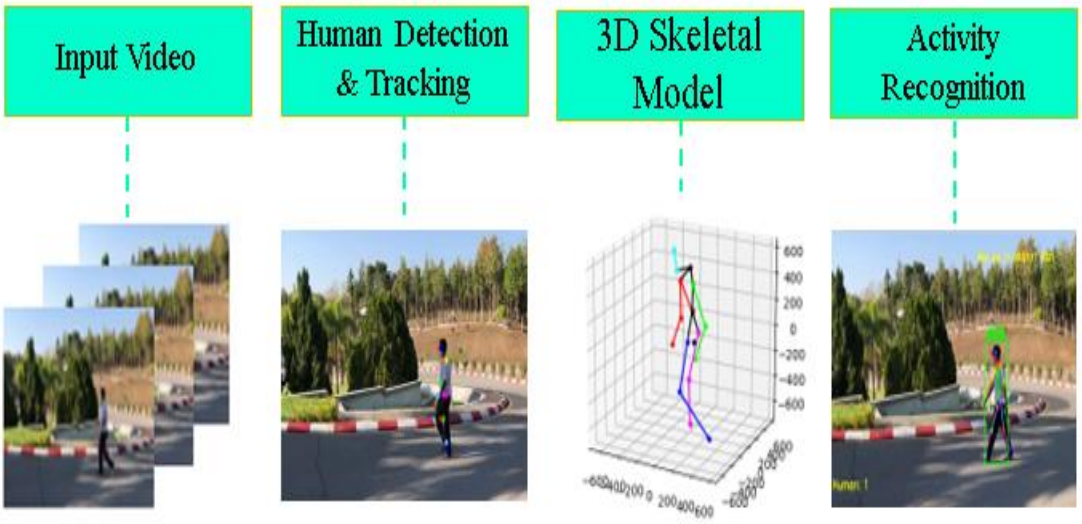


Figure 3.1 Development Process of the System

3.1 Image Representation

Image representation from the digital image, is defined as $I(r, c)$, containing rows and columns which is processing by computer imaging software. That consists of the different levels and various types of operation can be developed by the hierarchical image pyramid, as shown in Figure 3.2. In the figure, Left side of image operations are corresponding to the right side of image representation. At the lowest level, low-level preprocessing with large number of individual pixels are performed and higher levels are increasingly working for information representation [13].

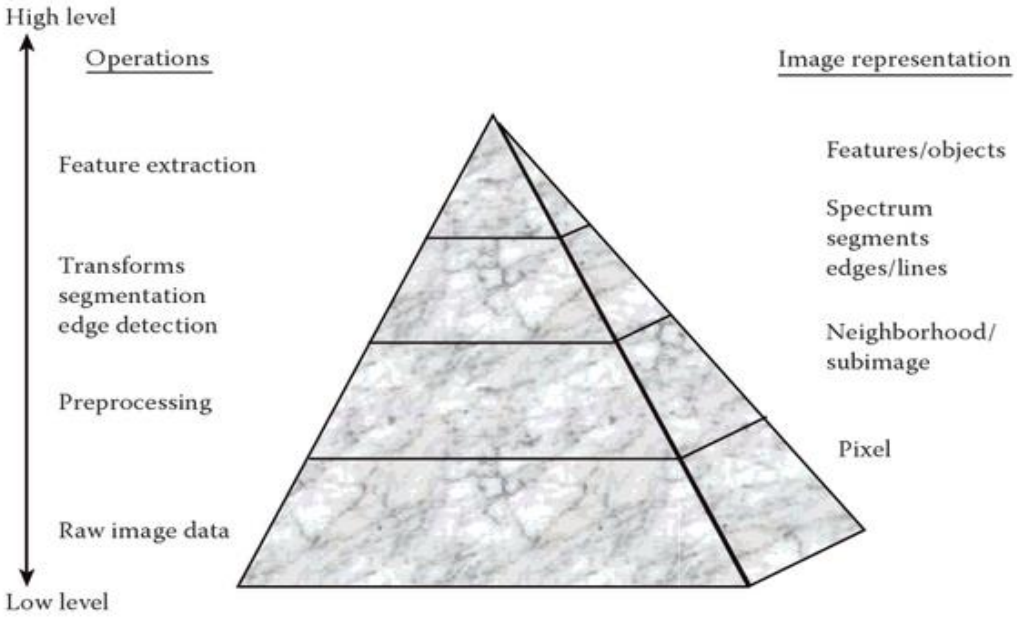


Figure 3.2 The Hierarchical Image Pyramid

The digital image, $I(r, c)$, is referred to two-dimensional array of matrix, and consists of row or column is called vector whose values represent brightness of the image corresponds to each pixel value at the point (r, c). This can be modeled with different function that corresponds to each distinct brightness information band based on three image types are binary image, gray-scale image, and RGB or color image.

Binary image is simplest types of image whose intensity ratings are limited to two. Usually, they are shown in black and white. In terms of numbers, there are two intensity values: 0 for black and 1 or 255 for white. Grayscale images can be converted to binary images by applying a threshold operation, which turns all pixels that exceed the threshold value from black to white.

Gray-scale image is single or monochrome in which the color is gray shade**.** The range of each pixel value is between 0 and 255 where value closer to 0 is darker and closer to 255 is lighter. The usual image contains 8-bit per pixel and the numer of bits for each pixel can determine the different gray levels.

RGB or color image is corresponding to the primary color of red, green, and blue light. The numeric representation of each color as 0...255 (8 bits), 0…65535 (16 bits),

extending the 24 bits, 32 bits, 48 bits, 64 bits or even larger. That denoted the intensity value for each color channel with bit depth n as $(2^n - 1)$, two raised to the power n minus one and that is range from zero to some n power of bits group. The notation of RGB (Res, Green, Blue) colors are expressed in Table 3.1.

Table 3.1 Notation of RGB Color Image Representation

Notation	RGB
Arithmetic	(0.0, 1.0, 0.0)
Percentage	(0%, 100%, 0%)
Digital 8-bits per channel	(0, 255, 0)
Digital 12-bits per channel	(0, 4095, 0)
Digital 16-bits per channel	(0, 65535, 0)
Digital 24-bits per channel	(0, 16777215, 0)
Digital 32-bits per channel	(0, 4294967295, 0)
Digital 48-bits per channel	(0,1099511627775, 0)
Digital 64-bits per channel	(0, 281474976710655, 0)

3.2 Image Segmentation

Segmentation with foreground and background plays the basis step in object detection and activity recognition system. The objective of image segmentation is to classify each pixel with a specified range in an RGB image. Accurate image segmentation method is important for human detection to support later result efficiently. That means to identify part of the image and to understand what features they belong to the requirement of object. The fundamental of image segmentation can be defined as the following:

Let R be spatial region of an image. The segmentation process consists of partitions R into n subregions $R_1, R_2, \dots, R_n$ such that

(1) $\bigcup_{i=1}^{n} R_i = R$

(2) R_i is a connected set, for $i = 0, 1, 2, \dots, n$.

(3) $R_i \bigcap R_j = \emptyset$ for all i and j, $i \neq j$.

(4) $Q(R_i)$ =TRUE for $i = 0, 1, 2, \dots, n$.

(5) $Q(R_i \cup R_j)$ = FALSE for any adjacent regions R_i and R_j

Statement (1) shows that th e segmentation of every pixel in a region must be complete in the sense. Statement (2) requires some predefined sense that points must be connected in a region (e.g the points must be 8-connected). Statement (3) means that the regions must be disjoint. Statement (4) says the condition must be satisfied by the pixels in a segmented region. For example $Q(R_i)$ =TRUE if all pixels in R_i have the same intensity. Statement (5) indicates that two adjacent regions R_i and R_j must be different in the sense of predicate Q.

There are various image segmentation methods that are frequently used in the past decade but that have less efficient than advanced deep learning techniques. Learning techniques use with filtering methods to extract image features and separate the entire image for desire result. In study, the traditional methods have been required to define additional facts and knowledge for real-life application. The general approaches for some segmentation methods are:

- Region-Based Method
- Clustering-Based Method
- Histogram-Based Method
- Edge-Based Method
- Learning- Based Method

3.2.1 Region-Based Method

This method separates the objects based on partition of an image into different regions that are similar according to a set of predefined criteria. By using local or global threshold method based on specified objects such as two regions consists of object and background can be divided by global threshold value and multiple objects along with background, can be divided by using local threshold value.

Most easily used method is Background Subtraction that defines absolute difference between current image and background reference image along the changing of adaptive background over a period of time. That have required to use a specified threshold value to separate pixels into two levels for isolated objects. In order to do, firstly, converts grayscale images into binary images and discriminates the lighter and darker pixels of a color image.

BG subtraction uses morphological technique to remove noise and shadow is removed by shape analysis. To get desired result, all incoming frames from video sequences are subtracted from reference BG modeling frame and compare the difference with denoted threshold value and segment the image between FG and BG. BG subtraction applies Gaussian Mixture Model to get better extract for FG object. The example result make use of this method is described in Figure 3.3.

Figure 3.3 Original Image and Segmentation using Background Subtraction Method with Threshold Value

3.2.2 Clustering-Based Method

Clustering is the task of dividing data points or pixels of the image into a number of groups or clusters, such that data points in the same groups are more similar than data points in these other groups. K-means clustering method is popular and that algorithm processes groups into segment of interest areas in the specific data with the K variable representing of the number of distinct groups.

The algorithm assigns K data points into each group based on similarity of features. Rather than predefined groups, the algorithm works iteratively to gradually form groups. K-means clustering is distance based algorithm and first partial stretching enhancement can be applied to improve the quality of the image. "It is a widely known and powerful unsupervised machine learning algorithm. and can provide the groups with unlabeled dataset into different clusters.

K-Means clustering algorithm can be defined as the following procedure:

- First, select the number of clusters K

- Then randomly assign K data points
- Define the closet centroid of K clusters
- Calculate and express the new centroid of each cluster
- Reassign each data point to the new closest centroid

Repeat last three steps until either the center of the clusters does not change or connected domain that reach the requirement results are obtained. But it doesn't optimize when clusters are in different shapes. The testing result of k-means clustering with number of clusters are defined as 5 and taking of random center are described in Figure 3.4.

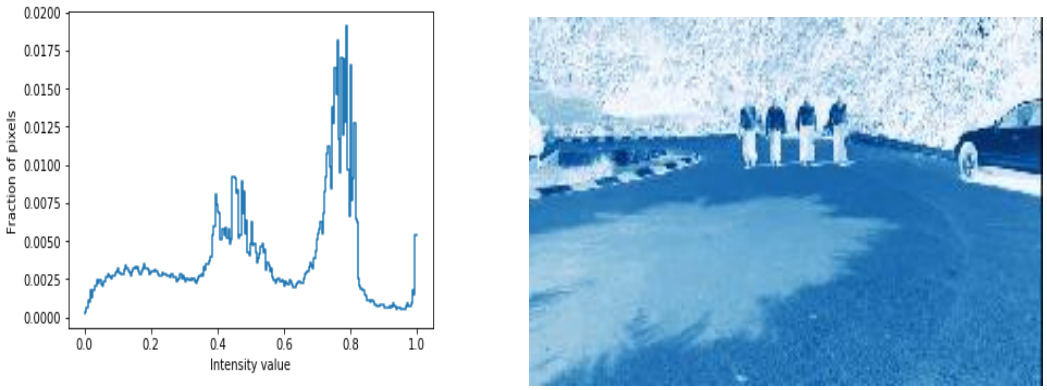
Figure 3.4 Segmentation using K-means Clustering Method

3.2.3 Histogram-Based Method

An image segmentation based on a histogram serves as a visual depiction of digital image with intensity value corresponding to the dominant peaks. The key point of histogram-based technique is the selection of specific thresholds that can distinguish objects and background pixels and that defines a histogram to a group pixel.

Examining the histogram to determine identification of image, it can possible to assess the distribution as a whole and can focus on peak value states to determine similar pixels of the image. The larger peak states the background gray level and the smaller peak represents the object. Example of image segmentation effect with intensity values and fraction of histogram pixels are expressed in Figure 3.5.

(a) Original Image **(b) Convert to GrayScale**



(c) Intensity Values of histogram **(d) Result on Intensity Values**

Figure 3.5 Segmentation using Histogram-based Technique

3.2.4 Edge-Based Method

Edge detection is a technique of image segmentation and typically involves discontinuous points with unlike local features of an image. That can observe a boundary of the object with strong gradients as corners or curved line segments or edges. That has sharp or discontinuities where occurs surface color, depth and illumination changes.

There are many techniques to find changes in intensity for edge detection by using of second order derivative or first order derivative using gradient algorithms. The first order derivative is applied to the Roberts edge detector, Sobel edge detector, Kirsh edge detector, Robinson edge detector, Marr-Hildreth edge detector, Prewitt edge detector, LoG edge detector, and Canny edge detector. And Laplacian Operator with

Zero Crossing based edge detection and Laplacian of Gaussian methods are used second order derivative.

Basic process of edge detection technique includes:

- Filtering- Apply a filter to the image to enhance the edge detector's noise-reduction capabilities that eliminate noise and other artifacts and locate all true edges.
- Enhancement- Highlight the pixels whose local intensity has changed significantly. that precisely pinpoint the edge and determine the edge's orientation and defines edges as close as possible to get true edges. And process each edge point to return a single response.
- Detection-Identify edges by different threshold method

Example of edge detection with Sobel edge detector is depicted in Figure 3.6. firstly, it computes the picture intensity gradient at each pixel in the image to produce the desired result. Next, determine the direction and rate of change of the highest increase from light to dark. It performs a 2D spatial gradient measurement on an image and emphasizes regions of high spatial gradient which corresponds to edges.

Sobel operator works through for averaging and emphasizes on the pixel to the center of mask. It is one most popular edge detector and less affected by noise. First order derivative is more expensive than second order on computation, since they have been required to find gradients in two directions and normalization.

Figure 3.6 Segmentation using Sobel Edge Detection Technique

Another example of edge detection with Laplacian Operator by second order derivative is described in Figure 3.7. Firstly, convert the Image color into grayscale, then pass the image onto by specifying kernel-filter inside the function as an argument.

Filtering is required for smooth and use with Gaussian filter and find magnitude and orientation of gradient to define edges.

Figure 3.7 Segmentation using Laplacian Operator Edge Detection Technique

3.2.5 Learning-Based Method

Learning is a specific type of machine learning which is used to process layers of representation and different levels of abstraction that can acquire sense of data. Deep learning-based approach enables to process the raw forms of images or videos and automate the process of feature extraction, representation, and classification [33]. The method utilizes the trainable feature extractors and computational models with multiple processing layers for action representation and recognition. That can be used for segmentation into different regions, making it possible to identify specific features within video sequences. The example of segmentation between the input sequences and prediction pixels are shown in Figure 3.8.

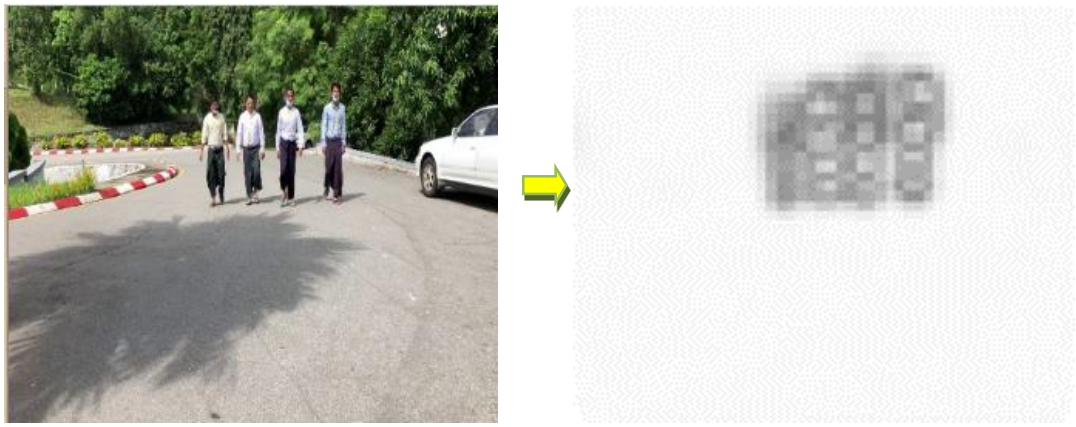


Figure 3.8 Segmentation using Deep Learning Technique

More intuitively, the idea in deep neural networks can be viewed as high levels extraction the information with features by each of the network layers. Types of network layers include feedforward, convolutional, and recurrent neural layers. These layers are utilized to design different network types based on the application, with convolutional neural networks (CNNs) being widely used in computer vision.

3.3. Human Detection and Tracking Strategies

There are three main types of human detection and tracking models to estimate human pose in 2D plane and 3D space.

Skeleton-based model: this representation, also known as the kinematic model, consists of a collection of important points, or joints, such as the ankles, knees, shoulders, elbows, wrists, and limb orientations, that are mostly used for 2D and 3D pose estimation. This adaptable and simple model, which represents the skeletal framework of the human body, is widely used to depict the relationships between various body parts.

Contour-based model: often referred to as the planar model, it is a 2D posture estimation tool that includes the general breadth and contour of the torso, limbs, and body. In essence, it depicts the form and appearance of the human body, with bodily components shown alongside limits and contour rectangles. One well-known example is the Active Shape Model (ASM), which uses principal component analysis (PCA) to capture deformations of the human silhouette and the full body graph.

Volume-based model: also known as the volumetric model, is employed in the estimate of 3D poses. It consists of several widely used 3D human body models and poses that are represented by geometric human meshes and shapes. Deep learning-based models are typically used to record and estimate 3D human stance.

3.3.1 3D Human Pose Estimation using 2D Key Points

The spatial locations of the body joints on a 2D image can be used to represent the 2D human pose. The 2D human pose matrix $W \in R^{2xJ}$ is a two-dimensional matrix in which each column reflects the 2D Cartesian coordinates (x and y) of one of the main body joints, assuming that J is the number of skeleton joints of human body parts. A

transformation function that maps the coordinates of the 2D joints to the corresponding 3D coordinates in 3D space can be used to represent the 3D human pose estimation, given that $S \in R^{3xJ}$ is defined as the 3D human pose matrix and that each column represents the 3D Cartesian coordinates (x, y, and z) of one of the main body joints.

Different projection models may be implemented to optimize the transformation result. One of the considerable methods is reducing the projection inaccuracy. In [72], they estimate unknown 3D shape model from 2D single image to detect variability of complex shape by using convex approach. That uses linear relationship between the 2D and 3D poses with camera calibration matrix and linear combination of predefined shapes which determine the prediction problem concerned with the sparse coefficients representation as well as the perspective of the camera.

3.3.2 Designing Human Pose Estimation

The process of determining the skeletal joints' position in video sequences is known as human pose estimation. The connections between the bones in a human body are defined by the main joints, which provide an overall view of the skeleton by linking the corresponding joints. The human pose with highest score value of skeleton joints on human body parts are illustrated in Figure 3.9.

Figure 3.9 Presenting the Pose using the Skeleton Joints

Pose estimation is a crucial aspect of action recognition. Partially, the previous pose is used to fit as a known reference and assists as a preprocessing step for subsequent processing. Pose estimation can handle any composition of rigidly moving

parts connected by joints. The system identifies human body parts to estimate the 3D human pose related to analyzing 2D joint locations with joint attention score.

Where joint attention score defined by α_t^I can be computed as:

$$\alpha_t^I(k) = \frac{exp\{\alpha_t^I(k)\}}{\sum_k exp\{\tilde{\alpha}_t^I(k)\}} \quad (3.1)$$

The related feature of human body parts is described as:

$$F_t^P = \sum_{I \in P} \sum_k \alpha_t^I(k) V_t(k) \quad (3.2)$$

$V_t(k)$ is the feature vector of V_t at the k^{th} spatial location ($k = 1, \ldots, K_1 \times K_2$), $\tilde{\alpha}_t^I(k)$ is the unnormalized attention score of $V_t(k)$ for each joint $I \in P$, P is the body part. That describes the human figure which has relative positions of the limbs and generate 3D pose estimation that contains frontal, lateral, backwards and forwards displacement. And then tracking algorithm is applied to track the detected object across different frame.

3.3.3 Matching 2D to 3D with Embedding Model

The matching probability from 2D joint locations to 3D pose described in Figure 3.10 are estimated via an Expectation-Maximization (EM) algorithm through the entire sequence. The EM algorithm is an approach for performing maximum likelihood estimation when certain variables are present. It assumes that all relevant variables in the problem are observable, while others unobserved or hidden variables are referred to as latent variables. EM algorithm effectively handles missing data and latent variables, making it well-suited for tasks involving incomplete or partially observed datasets, such as 2D to 3D pose estimation.

The process begins by estimating the values of the latent variables, followed by optimizing the model, and iterating between two steps, known as the E-step and M-step, until convergence. It is most commonly used for density estimation, which involves selecting a probability distribution function and determining the parameters of the joint probability distribution based on the observed data. Expectation maximization offers an iterative solution to maximum likelihood estimation with latent variables.

The approach of reasoning about 3D human posture using minimal representations, such as sparse 2D projections. This might involve key features of joint coordinates, angles, and distances between joints. For a given 2D pose, encode it into

the latent space. Identify the closest 3D pose in the latent space using a similarity measure, such as Euclidean distance. The retrieved 3D pose can then be decoded back into the original 3D joint positions. In 3D pose estimation, body configuration can be determined by analyzing angles and body shape. This approach ensures a robust mapping between 2D and 3D poses, enabling accurate 3D pose estimation from 2D.

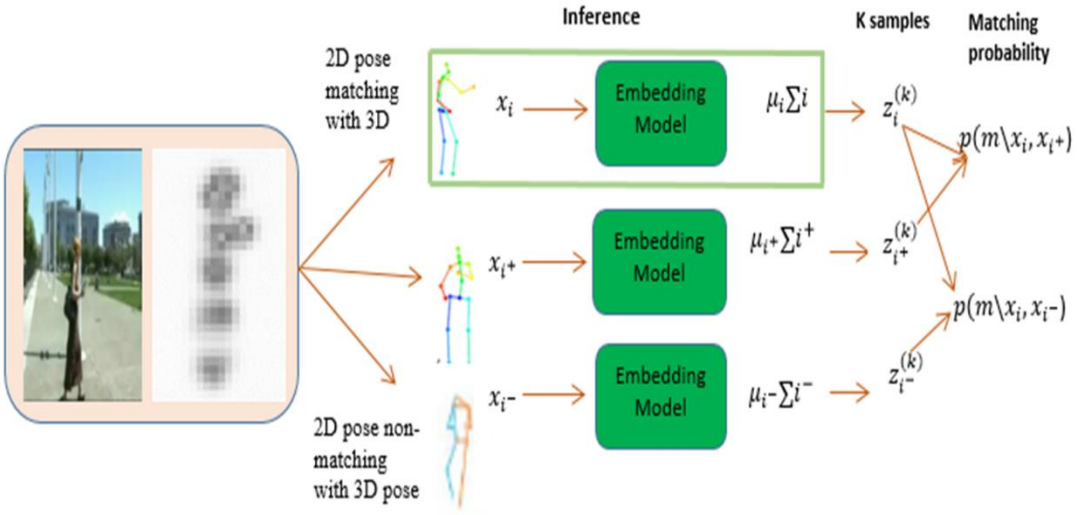


Figure 3.10 2D to 3D Pose Matching with Embedding Model

3.4 Deep Neural Network-based 2D to 3D Joints

This method involves training a network to learn the mapping between 2D joint positions and their corresponding 3D coordinates. The system closely aligns with recent advancements that map 2D to 3D coordinates using deep neural networks. Authors in [64] proposed a deep convolutional neural network utilizing the stacked hourglass architecture, which maps 2D joint probability heatmaps to 3D probability distributions. Another approach in [41] predicts a pairwise distance matrix from 2D to 3D space. The invariance of distance matrices ensures robustness against transformations and the system achieves more accurate and reliable predictions.

3.4.1 Graph Modeling

Spatial-temporal graph modeling offers a robust framework for 2D to 3D recognition by effectively capturing the intricate spatial and temporal dependencies of human motion. This approach not only enhances the accuracy of 3D pose estimation

but also provides a deeper understanding of human activities. The system uses a novel adaptive dependency matrix for modeling spatial-temporal graphs, which is learned by node embedding across interrelated human body components. The latent spatial dependencies included in the data can be captured by it. Its receptive field grows exponentially with the number of layers and can accommodate extremely long sequences for pose estimation.

3.4.2 Supervised Learning

The method of deep supervised learning is teaching a computer to identify images into groups such as people, automobiles, houses, and pets. This process begins by assembling a large dataset of labeled images for each category. During training, the system processes images, generating a vector of scores indicating probabilities for each category. Until the model's parameters are improved by training, the desired category does not start out with the highest score.

Measuring the pattern of difference between the target score and the output scores is an objective function. The system then optimizes to reduce difference its internal parameters, called weights. These weights serve as adjustable factors that determine how inputs are transformed into outputs within the system. In deep-learning systems, weight values are optimized using gradient descent algorithm. The algorithm computes gradients for each weight, showing how small adjustments affect the error, and updates the weights to minimize error effectively.

3.4.3 Weight Initializations

Numerous neural network gradient descent algorithms call for tiny, arbitrary initial weight values. The weights were initialized by the system using a Gaussian distribution with mean zero and standard deviation 0.1, or a uniform random distribution in the interval [-0.1, 0.1]. Nevertheless, it was discovered that neither the distribution nor the range had a significant impact on the outcomes. Random beginning conditions have the side effect of requiring multiple replications of each experiment in order to assess significance.

3.4.4 Gradient Descent Method

Gradient descent method is a widely employed optimization algorithm used to determine the weights or coefficients of learning algorithms like artificial neural networks and logistic regression. The process involves making predictions on training data and adjusting based on the prediction errors. The primary goal of gradient descent is to determine model parameters, such as weights or coefficients, that minimize the error on the training dataset. This is achieved by iteratively adjusting the model parameters in the direction that reduces the error, following the gradient or slope of errors towards the minimum error value. The configuration of learning model is illustrated in Table 3.2.

Table 3.2 Learning Model Configuration

```
1. model = initialization(...)
2. n_trainings
3. train_data
4. for i in n_epochs:
5. train_data = shuffle (train_data)
6. x, y = split (train_data)
7. predictions = predict (x, model)
8. error = calculate_error (y, predictions)
9. model = update_model (model, error)
```

3.4.5 Multi-Directional Networks

The network mentioned above has access to all points $(P_1, \ldots\ldots, P_n)$ in the input sequence at some point $(p_1, \ldots\ldots, p_n)$ such that $p_d \leq p_d \forall d \in (1, \ldots\ldots, n)$. This describes the entire sequence's n-dimensional context region. When the object to be recognized is fully enclosed within the context region, the network processes the image in accordance with the ordering and outputs the object label. Nonetheless, it is ideal for the network to be able to access the external context from any angle, and accurate localization is needed to specify segmentation.

Computational models with several processing layers can learn different levels of abstraction for data representations thanks to deep learning. The state-of-the-art in many other domains, including speech recognition and genomics, as well as visual

object recognition and object detection, has been greatly advanced by these techniques. Using the backpropagation algorithm which determines how a machine should modify its internal parameters to calculate representations in each layer based on the previous layer and deep learning reveals complex structures in massive datasets.

These techniques use multi-layered computational models for action recognition and representation, as well as trainable feature extractors. A deep learning research study claims that deep learning techniques can handle differences in an object's position, orientation, illumination, and other characteristics. Multiple processing layers, each representing a distinct level of abstraction, are used in deep learning models. More logically, the goal of deep neural networks is to achieve more degrees of abstraction by using the features that are retrieved from each layer as the input for a subsequent neural layer.

In practice, deep neural networks are composed of multiple layers. These networks are trained to create a discriminative representation of data. Depending on the use case, deep neural networks have different kinds of layers. These layers are used to design various network architectures tailored to specific applications, with convolutional neural networks (CNNs) are widely used in computer vision.

3.4.6 Convolutional Neural Network (CNN)

CNNs are made to handle data in the form of numerous arrays, like a color image made up of two-dimensional arrays that show the intensities of the pixels in each of the three RGB color channels. As numerous arrays, many data modalities exist (see Figure 3.11 for 1D signals and sequences, 2D pictures or audio spectrograms, and 3D video or volumetric images). A typical CNN's architecture is divided into several stages.

Convolutional and pooling layers are the two types of layers present in the early stages. A filter bank, which is a collection of weights, connects each unit in a convolutional layer to local patches in the feature maps of the layer above. The units in a convolutional layer are organized into feature maps. After that, the local weighted sum result is run through a non-linearity, like a ReLU. While separate feature maps within a layer utilize different filter banks, all units in a feature map share the same filter bank.

Convolutional Layers: A convolutional layer consists of a set of trainable parameters (weights) which are called filter banks. The filter banks do not change spatially and are

kept the same for all inputs of each layer. Convolutional layers apply a convolution operation between the input of the layer and the filter bank. The convolution operation in Figure 3.12 helps in extracting features from the input by utilizing local combinations of features within pixel neighborhoods. In the first layer, the feature extraction process applied directly to an image can be viewed as a type of edge detection. In the rest of the layers, the extracted features correspond to higher levels of abstraction.

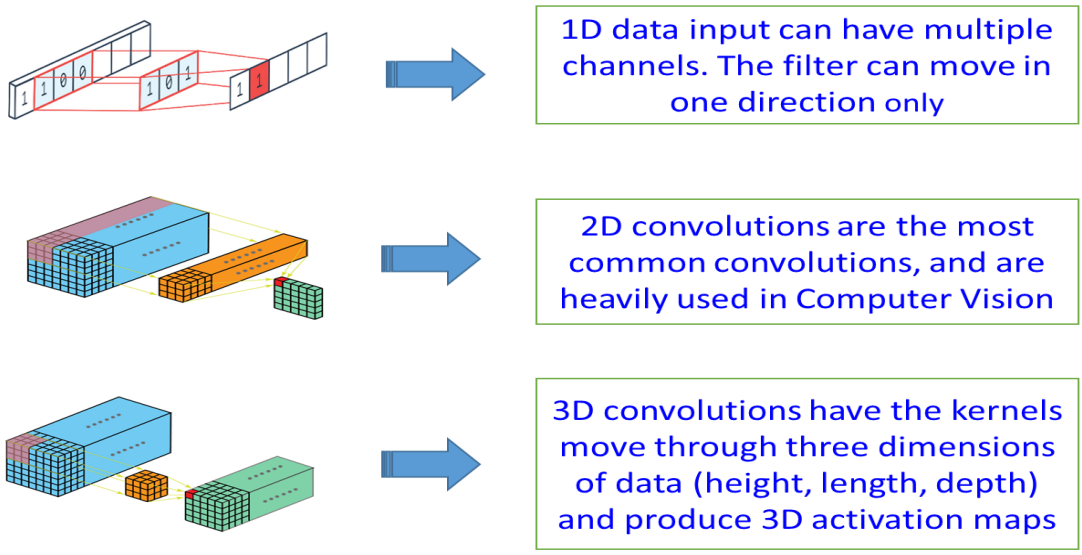


Figure 3.11 1D, 2D and 3D Convolutions

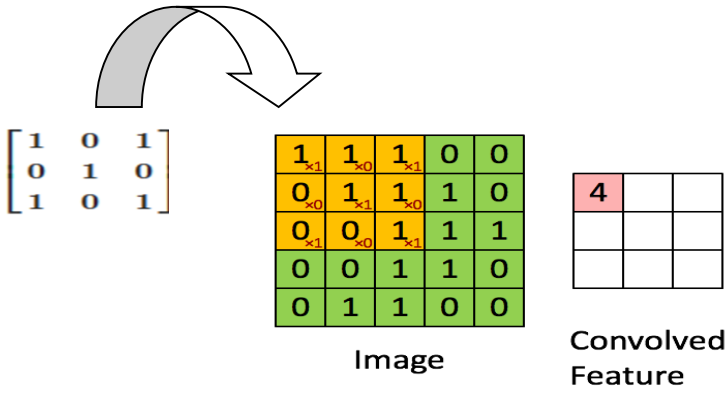


Figure 3.12 Operation of Convolution

Pooling Layers: It reduces the dimensionality of each feature and merges similar features into one. Types of pooling operations are: described in Figure 3.13.

- Maximum Pooling
- Average Pooling
- Minimum Pooling
- Adaptive Pooling

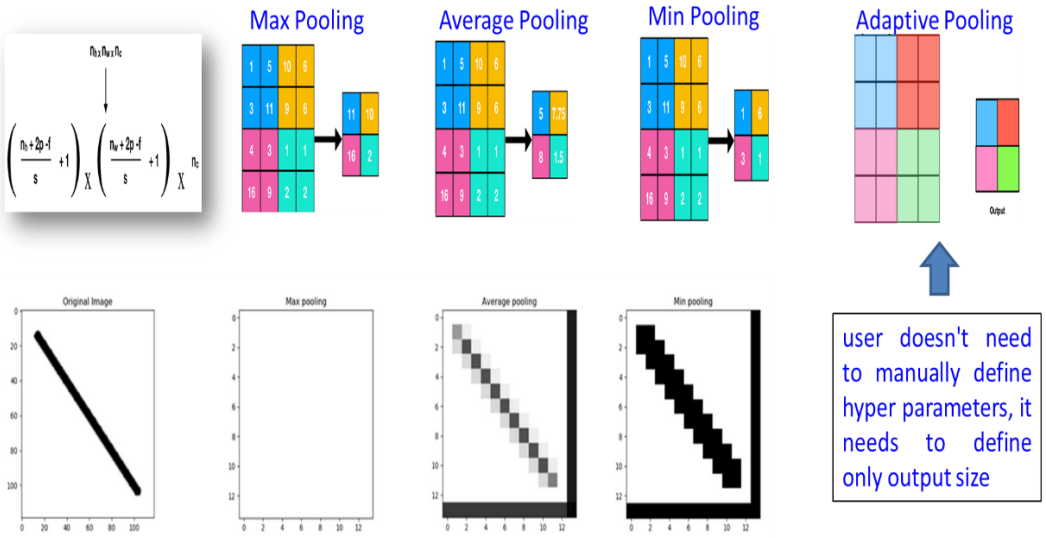


Figure 3.12 Types of Pooling Operations

Fully Connected Layer: After convolution and pooling, the model analyzes the logit scores. The softmax activation function is used for making multi-class predictions, classifying probabilities, and producing an efficient result for the output.

$$softmax(y)_i = \frac{e^{y_i}}{\sum_{j=1}^{N} e^{y_j}} \tag{3.3}$$

where

- e is the base of the natural logarithm
- y_i is the i[th] element of the input vector y
- $\sum_{j=1}^{N} e^{y_j}$ is the sum of the exponentials of all elements in the input vector

3.5 Summary

Following an examination of background theory concepts, an effective approach for human motion detection, tracking, and activity recognition was developed. that is capable of defining relevant data and doing extensive environmental analysis. The various variations in human position, shape, illumination changes, and background look present numerous obstacles. The suggested system uses a deep neural network architecture to recognize human movement in both indoor and outdoor environments with a high degree of accuracy.

CHAPTER 4
THE PROPOSED SYSTEM ARCHITECTURE

Skeleton-based human detection and activity recognition are greatly affected to computer vision and available in real-world applications such as security, physical condition of human movement, robotic representation, smart home system and visual reality applications, etc. Most of the system occur wrong recognition on human activity for similar pose but different activities such as standing may be recognize as walking and other direct and indirect dependency of joint movements. Currently, several approaches have been found by using skeleton data that is directly taken from sensor devices. That is cost effective and constraints on lighting condition, distance measurement and requirement of apparatus devices for indoor and outdoor applications.

Human action is harder to define on different conditions, and many attempts have been described in the literature. Despite encouraging development in the past decade, occlusion of body part detection and recognition of similar pose but different activities are still required on various situations. Under existing solutions for similar activities such as walking is similar to standing and occur miss recognition result for human activity. The human skeleton involves several joints and segments, representing of body parts and limbs. The location of skeletal joints can vary on different people and can depend on how far separate an individual from the camera.

The digital images from the video sequences are in the form of matrices and require mathematical operation to extract valuable information. Firstly, the proposed system detects regions of interest (ROI) to segment region-by-region and by finding the ratio between the intersection of target mask and prediction pixels of human body parts. And extracts geometric features consist of joint indices and distances and coordinate features of 2D joint locations. The timeline of 2D skeleton point, and the segmentation result forms the digitized human motion and that can track from which paths and trajectories are moving.

In the proposed system, to get better performance on occlusion of body part detection and human skeleton is considered from 2D joint to 3D space. The system is defined on the necessary joint with Joint Collection Distance and that can classify more accurate result on joint sequence activity. The spatial-temporal arrangement of human body movements can characterize an action. An image over time serves as the temporal

component of the procedure, while semantic similarity throughout image space serves as the spatial aspect. The approach is configured 3D skeletal model with Deep Learning approach to recognize more reliable information for overlapping area between unseen body joints and human activity changes over time.

In deep learning process, millions of images and videos are used to generate feature extraction process on network and classify the activities. Advance in development of learning method, human detection is more reliable and 3D skeleton view of motion recognition is more effective for appearance information. The method is used with OpenPose detector to extract effective results for 2D key points of body parts and reconstructs 3D skeleton model to recognize more accurate result on human activities in different environment.

This chapter discusses contributions of the research work, the architecture of the proposed system and the details of each step are developed to provide accurate result for activity recognition in real life environment. To clarify the process of the proposed method, the overview of the proposed motion detection and activity recognition architecture design is described in Figure 4.1.

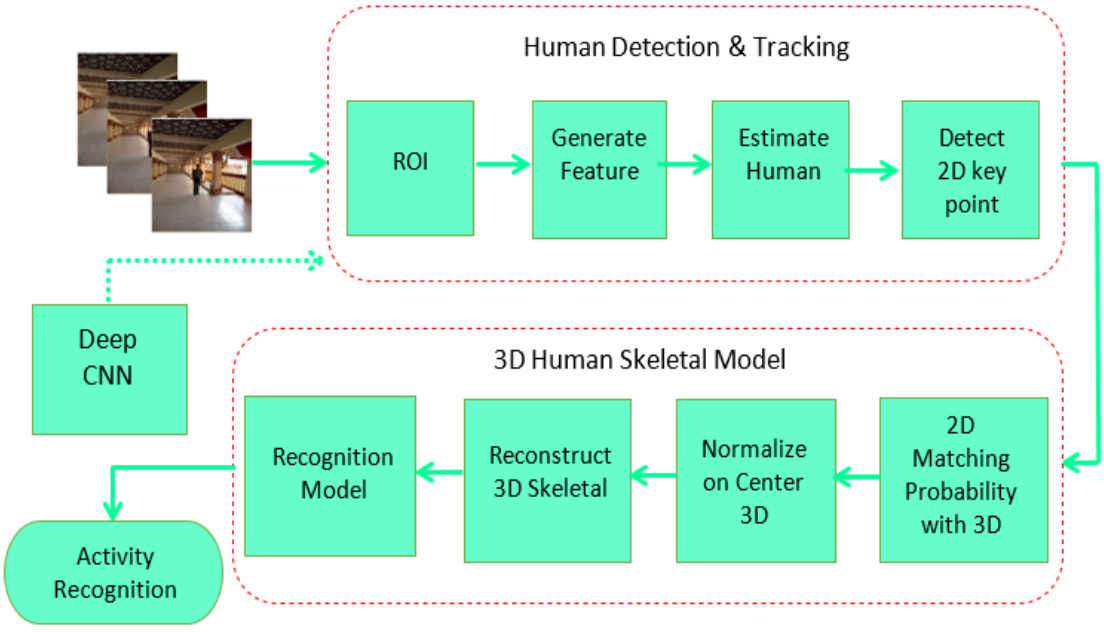


Figure 4.1 Overview of the Proposed System Design

4.1 Human Detection and Tracking

The processing steps of human detection in a video stream is performed by pre-processing, region of interest extraction, feature generation, human estimation and tracking on moving state. Human body is detected by OpenPose detector that define key points of human body parts and extract 2D joint locations. OpenPose can effectively detect to view point changes and can run in real-time.

In order to attain superior outcomes, interest points are chosen using a non-maximum suppression technique, with the highest-scoring outcome being referred to as the estimated position. This method enhances interactions from a previously unknown joint sequence captured alongside video and supports partial occlusions as well. On the human skeleton, every coordinate is referred to as a component (or a joint, or a critical point). A pair is a legitimate joining of two components (or a limb).

Top-down approach: This incorporates a person detection by estimating the parts and then calculate the parts for each person.

Bottom-up approach: This method finds all of the image's components, or each person's component parts, and then associates those components with specific individuals.

4.1.1 Region of Interest Extraction

The region where the segment can be found is contained within Region of Interest (ROI) mask. The Intersection-over-Union (IoU) is applied to get the performance of object segmentation and to extract ROI. If detection of target is less than IoU minimum, that is defined as reject assessment. The bounding box is computed maximum value between the last detection of an active track and candidate detections in the next frame. The IoU mask is computed by area of intersection of target mask over the area of all prediction pixels in both masks.

$$IoU = \frac{Area\ of\ Intersection}{Area\ of\ Union} \tag{4.1}$$

That is achieved by running on sequential with region-by-region and by finding the ratio between the intersection of target mask and prediction pixels based on ROI mask is shown in Figure 4.2.

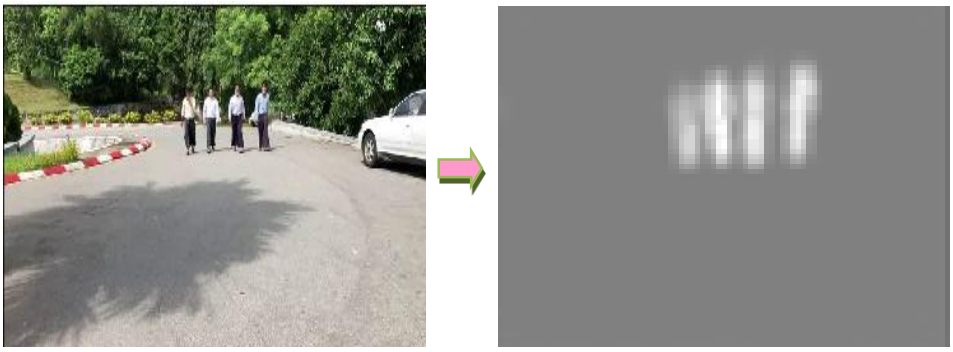


Figure 4.2 Input Image and Result of Segmentation between the Intersection of Target Mask and Prediction Pixels based on ROI Mask

4.1.2 Feature Generation

The input image's feature maps (F) are created using the first 10 layers of convolution in order to obtain the effective result for 2D key points. The feature maps are processed on two branches convolutional neural network in Figure 4.3 and continue to produce 2D key points. A convolutional neural network that has been trained on the picture first analyzes it to create a set of feature mappings F.

In the initial stage, feature maps with the highest active value of the joint locations between corresponding limbs and the related joints of human body parts are extracted using the function variables ρ^1 and Φ^1. At the next Stage, the prediction of previous stage along with image features F are concatenated and produce more refined predictions. The process continues to define the confidence maps and part affinity fields, and concatenated to produce 2D key points of all people in the image on two branches multi-stage CNN is described as follows:

Stage 1: $S^1 = \rho^1(F)$ (4.2)

$L^1 = \Phi^1(F)$ (4.3)

Where $\rho^1 and\ \Phi^1$ are function variable which represent input (F) to output confidence maps (S^1) and part affinity fields (L^1).

:

.

Stage t: The prediction of confidence maps and affinity fields are being processed

and concatenated to produce 2D key points for all people in the image

$$S^t = \rho^t(F, S^{t-1}, L^{t-1}), \forall\, t \geq 2 \quad (4.4)$$

$$L^t = \Phi^t(F, S^{t-1}, L^{t-1}), \forall\, t \geq 2 \quad (4.5)$$

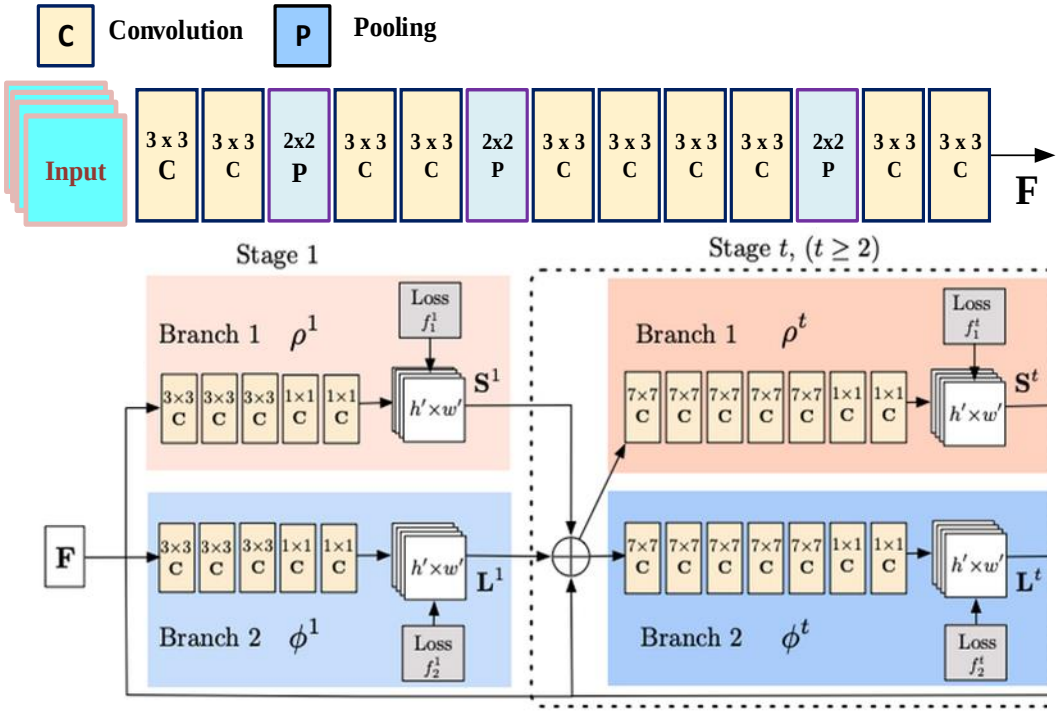


Figure 4.3 Feature Generation on Two Branches Multi-Stage CNN

4.1.3 Human Estimation

To estimate the human structure, human pose skeleton is more effective type to detect the human shapes and more robust to occlusions. Because human joints can vary greatly in their relative locations to one another. During the feature processing stage, each pixel's estimated confidence level has a unique joint score and spatial dependence value. The joint detection model defines the spatial layout of the human body in order to refine the confidence level of joints. Pose skeleton detection is more advantageous since it can show a specific body component, regardless of whether an item is occluded.

4.1.3.1 Non-Maximum Suppression

Non-maximum suppression means with maximal probability classification. That close to one is result. For detection of an object, non-maximum suppression can run independently. It can also ensure and have not any redundant or extraneous

bounding boxes and keep only the most positive result. It works with the following steps:

- Each output prediction contains $[P_r\ b_x\ b_y\ b_w\ b_h]$
- Discard all boxes with $P_r \leq$ threshold
- While there are any remaining boxes:
- Pick the box with the largest P_r as output prediction

Where P_r is probability and $b_x\ b_y\ b_w\ b_h$ are coordinates of bounding boxes.

4.1.3.2 OpenPose Detector

OpenPose is a multi-person, real-time, 2D pose detector. It can generate human poses as a root-centered graph using two-dimensional coordinates. Multiple people's skeletons can be recognized in the same scene thanks to OpenPose. After identifying important locations in the image that correspond to the human body, OpenPose assigns specific body parts to different people. This uses non-maximum suppression to locate joints, and it then connects two associated joints that are part of each person's leg. It may be able to locate all legitimate joints with a high confidence score and estimate each person's 2D joints in every picture.

Two parallel branches of convolutional layers receive the features extracted from an image by the first ten layers of the OpenPose network. A collection of confidence maps, each of which represents a distinct region of the human skeleton, are predicted by the first branch. The degree of relationship between bodily parts is represented by a set of Part Affinity Fields (PAFs) that are predicted by the second branch.

4.1.3.3 Spatial-Temporal Information Analysis

Human action recognition still struggles with efficiently representing spatial-temporal skeleton sequences. To generate effective sequences for each human body, the primary approach involves measuring the minimum distance between the target pose and the OpenPose detector across frames.

$$D_{V_1,J_*}(t) = \min_{v \in V_1} d_{J_*}(p_i(t), v) \tag{4.6}$$

the meaningful sequences that can be extracted are defined as follows:

$$\mathrm{Seq_i(V_l)} = \{\ \mathrm{D_{V_l,J}(t), D_{V_l,J_a}(t), \ldots,\ D_{V_l,J_d}(t)}\}\forall \mathrm{l} \in \mathcal{L} \tag{4.7}$$

where target pose is $\mathrm{p_i(t)}$ and $\mathrm{V_l}$ be prototypes for action l.

4.1.4 Tracking on Moving State

Aim of tracking is to capture the trajectory of multiple objects in video sequence. Motion tracking depends on object detection accuracy and association of previously tracked objects. New detection and tracking is considered based on Euclidean distances between centroids and bounding boxes as expressed in Figure 4.4.

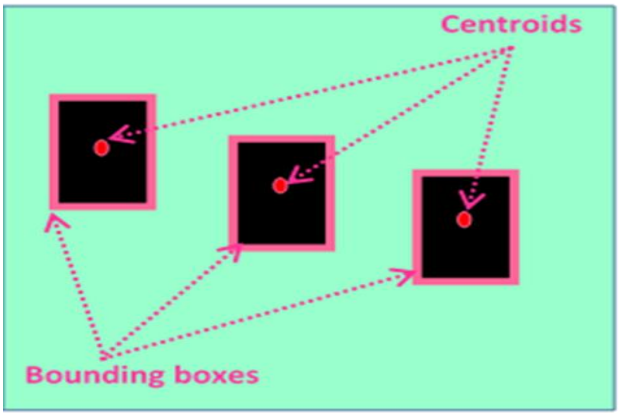


Figure 4.4 Tracking Based on Centroids and Bounding Boxes

4.1.4.1 Dynamic Measurement Model

For motion tracking, dynamic measurement model with Kalman filter and tracking is used to track position and velocity on state-space. That is a state estimation method based on linear dynamical system. State vector is defined by

$$X_t = (x_t v_t)^T \tag{4.8}$$

where x_t and v_t are position and velocity of target moving state.

$$X_{tk} = \Phi X_{tk-1} + A_k \tag{4.9}$$

where X_{tk} denote true state at time T, t∈T is time interval, k is moving state, A_k is noise on matrix and Φ is transition matrix from kT to (k+1) T as

$$\Phi = \begin{pmatrix} 1 & T \\ 0 & 1 \end{pmatrix}$$

Depending on state vector, the measurement vector is determined with

$$z_k = HX_{tk} + \mathrm{B}_k \tag{4.10}$$

where H is noiseless connection matrix between state vector and measurement vector, B_k is measurement error.

The motion tracker based on above model and sequentially estimate state vectors via the Kalman filter equations:

- Forecasting subsequent values of the state is based on past observation values:

$$\tilde{X}_k = \Phi\hat{X}_{k-1} \tag{4.11}$$

- Estimating the current values of the state from past and current observations:

$$\hat{X}_k = \tilde{X}_k + K_k(z_k - H\tilde{X}_k) \tag{4.12}$$

- Kalman gain:

$$K_k = \tilde{P}_k H^T(H\tilde{P}_k H^T + R) \tag{4.13}$$

where R is covariance matrix on noise, and $\tilde{P}_k$ is covariance matrix error.

4.1.4.2 Normalization on Gradient Vector

Normalization on gradient vector can make invariant to illumination changes. By using normalization on image gradient, that predict new location and more robust for tracking. It can provide an updating direction and helps to control the change of the solution through a well-designed step length. In Figure 4.5 shows how can change the result when normalized to original data.

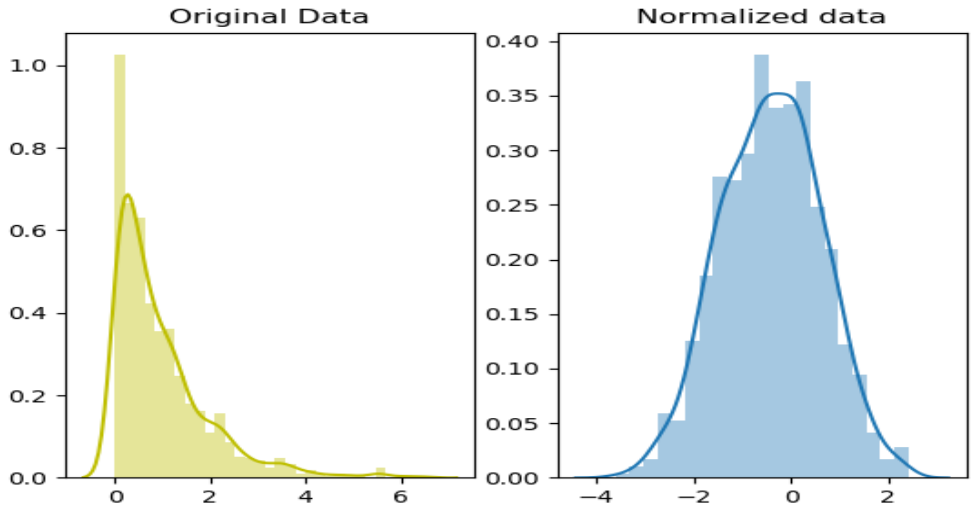


Figure 4.5 Example of Original Data and Normalized Data

That can be computed by two steps:

- First calculate length of components x and y as:

$$norcVec = \sqrt{dx^2 + dy^2} \tag{4.14}$$

- Second each of its components is divided by its length

$$V_x = \frac{dx}{normVec} \ , \ V_y = \frac{dy}{normVec} \tag{4.15}$$

4.2 3D Human Skeleton Model

A human model based on skeletons is more accurate at identifying human activities, and a 3D image may identify joint sequence gaps. Given that skeletal data can capture many views for the purpose of varied activity recognition and contains compact information about the primary body joints. The pose model generated by the relative positions of the limbs can be utilized to manipulate a 3D motion model. One particular process of skeleton configuration is evolving the skeleton feature points in the body components.

By using a graph model to record movements over a series of frames and the spatial-temporal representation of body parts in a single frame [12]. The graph model, which represents human mobility using 3D skeleton joints, captures the geometry and appearance variations of humans at each frame. Pairwise features, which relate to the positions of each pair of joints in relation to one another, are used to represent the structural data. Two joints, i and j, have their orientation computed, as in:

$$\theta(i,j) = arcsin\left(\frac{i_x - j_x}{dist(i,j)}\right)/2\pi \tag{4.16}$$

where $dist(i,j)$ is define for the geometry distance between two joints i and j in 3D space.

4.2.1 Joint Estimation on 3D space

3D reconstruction from 2D key points offers advantages such as viewpoint invariance and significantly impacts performance. Figure 4.6 illustrates changes between 2D and 3D pose structures to mitigate ambiguity related to physical constraints on limb lengths. The 3D pose is estimated using an Expectation-Maximization (EM) algorithm across the entire video sequence.

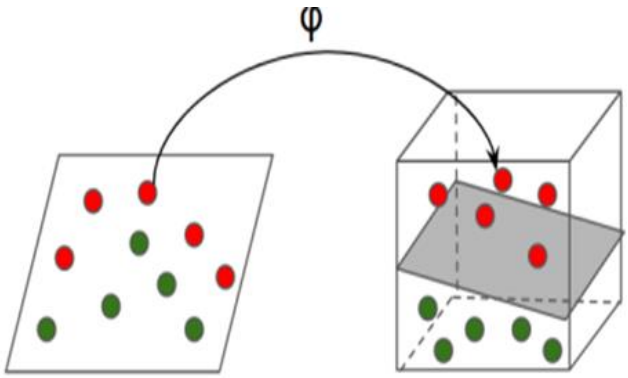


Figure 4.6 Changes from 2D to 3D Mapped Space

4.2.2 Expectation-Maximization (EM) Algorithm

Anticipation-Maximization an algorithm is a method for estimating maximum likelihood when latent variables are present. Latent variables are a broader term for the hidden variables. Estimating the values of the latent variables comes first, followed by model optimization and two iterations known as the E-step and M-step until convergence. It is a useful method for estimating the joint probability density.

Density estimation comprises selection of a probability distribution function and the parameters of the joint probability distribution with the observed data. Expectation maximization provides an iterative solution to posterior distribution on each variables. The matching probability from 2D joint locations to 3D pose are estimated through the entire sequence and obtain the probability distribution of 2D joint location heatmap values, compute the mean, and normalize the nearest 3D poses. EM-Algorithm is worked by the following process.

The expectation “E-step” is defined by $q(z_i)$ and estimates the nearest 3D pose and maximizing the expected likelihood to update the model parameter w by maximization "M-step", that estimate weight value with the sum of prior distribution and inverse of the combined covariance matrix. Iteration continue between the E-step and M-step until convergence to describe point estimation and posterior distribution on each z_i.

Process of EM-Algorithm

Input: Data $x_{1:n}, x_i \in R^2$ *and model* $x_i \sim N(W_{z_i}, \sigma^2 I)$, $z_i \sim N(0, I), z_i \in R^3$

Output: Point estimate of W and posterior distribution on each z_i

Expectation-Step: Set each $q(z_i) = p(z_i \backslash x_i, W) = N(z_i \backslash \mu_i, \sum i)$ where

$$\sum i = (I + \frac{W^T W}{\sigma^2})^{-1}, \mu_i = \sum i \, W^T x_i / \sigma^2$$

Maximization-Step: Update W by maximizing the objective

$$W = \left[\sum_{i=1}^{n} x_i \mu_i^T\right] \left[\sigma^2 I + \sum_{i=1}^{n} (\mu_i \mu_i^T + \sum i)\right]^{-1}$$

Iterate Expectation and Maximization steps until increase in $\sum_{i=1}^{n} \ln p(x_i \backslash W)$ is "converge".

4.2.3 2D Matching Probability on 3D Space

For a pair of input 2D poses (x_i, x_j), the proposed system defines the probability $p(m \backslash x_i, x_j)$ and their corresponding 3D poses as matching of (y_i, y_j) are visually similar. i.e., $p(m \backslash x_i, x_j) \sim (y_i, y_j)$. By mapping 2D pose to probabilistic embedding is defined as: $x \rightarrow p(z \backslash x)$

$$p(m \backslash x_i, x_j) = \int p(m \backslash z_i, z_j) p(z_i \backslash x_i) p(z_j \backslash x_j) dz_i \, dz_j \quad (4.17)$$

Each distribution defined by K samples as:

$$p(m \backslash x_i, x_j) \approx \frac{1}{K^2} \sum_{k_1=1}^{K} \sum_{k_2=1}^{K} p(z_i^{(k_1)}, z_j^{(k_2)}) \quad (4.18)$$

4.2.4 3D Point Reconstruction

The 3D point reconstruction from 2D images has focused on the development of 2D joint information and acquisition techniques from scenes and objects. That consists of projecting to 3D scene from 2D image. The probability of 2D heat maps

from the detected 2D joints and that is trained convolutional neural networks to estimate 3D pose. During this process, the points visible on 3D space are the projections of the real points on the image. The 3D point on space is defined by

$$X_i = [x_i \; y_i \; z_i]^T \tag{4.19}$$

$$X = [X_1^T \ldots X_s^T] \approx \theta_1\beta_1 + \cdots + \theta_{3k}\beta_{3k} = \theta\beta \tag{4.20}$$

where $\theta_j \in R^{3s}$ is trajectory basis vector and $\beta = [\beta_1 \ldots \beta_{3k}]^T \in R^{3k}$ is coefficient of point trajectory. 3D point trajectory reconstruction is defined as

$$\eta = \frac{\|\theta^{\perp}\beta_{\hat{X}}^{\perp}\|}{\|\theta^{\perp}\beta_{X}^{\perp}\|} \tag{4.21}$$

Where $\hat{X} = \text{estimated point trajectory}$. The reconstruction of human joint point on 3D space is shown in Figure 4.7.

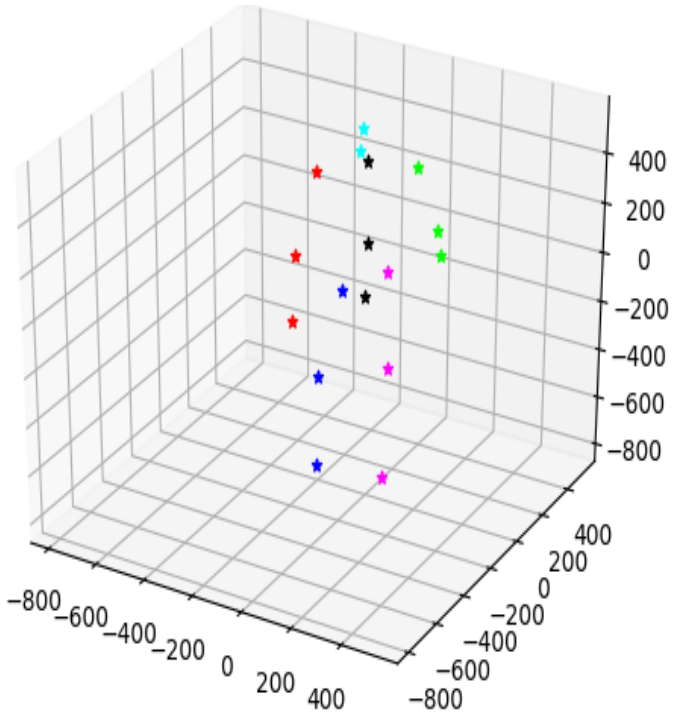


Figure 4.7 Human Joint Reconstruction on 3D Space

4.2.5 3D Joint Sequence Extraction

The proposed system reconstructed 3D human body joint to define more precise joint sequence result for activity recognition. Since, human joint information from the skeleton data has proven highly successful for action recognition tasks and that can

extract overlapping area in hidden human body parts and occlusion of one another. The extracted joint sequences are ordered according to human body parts and define the activity with the state of joint sequences movement. The order of human joint sequences on 3D space is shown in Figure 4.8.

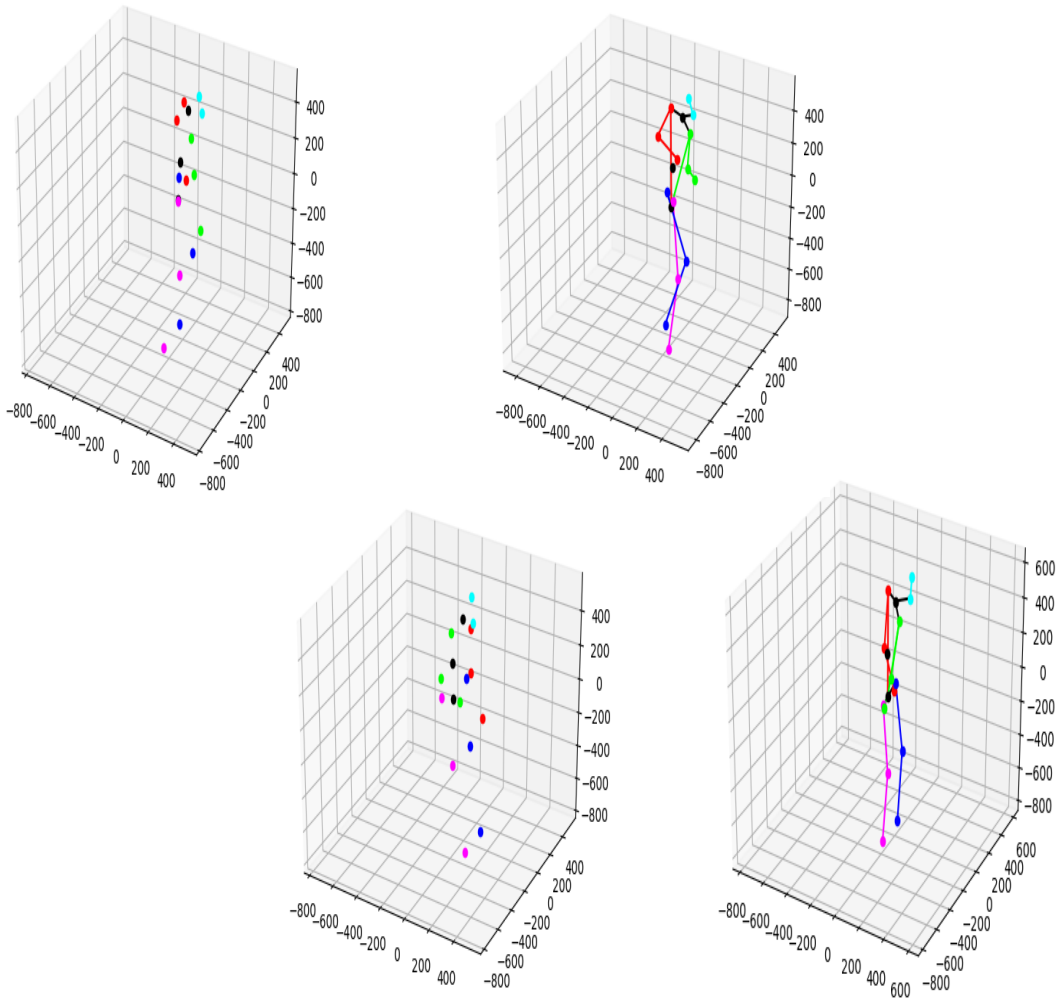


Figure 4.8 Human Joint Sequence Extraction on 3D Space

4.2.6 Joint Collection Distances

Point-of-coordinate features and geometric features are used for skeleton-based action recognition such as joint indices and location distances are important to get accurate result. The geometric features are invariant from a spatial perspective, but exhibit variability across different data sets. Actually, joint indies (i.e., the IDs of key points which consist of shoulders, elbow, wrist, etc.) can be dynamically changed in various activities. Hence, recognition errors arise, and requiring to predefine the correlation of joints by ordering of their indices. To stable each pair of active joint, by

computing NJCD with Euclidean Distance matrix along the frame sequence on different activity, and that can solve for missing joint errors and can improve the recognition result. The system is generated transformation process from 2D to 3D pose with more realistic for joint sequences and activity. Using joint collection distances can resolve various issues, and the embedding process helps minimize the impact of skeleton noise. Joint collection as defined by $S_k = \{J_1^k, J_2^k, \dots, J_N^k\}$ and JCD feature of S_k is computed as:

$$JCD^k = \begin{bmatrix} \left\|\overrightarrow{J_2^k J_1^k}\right\| & & \left\|\overrightarrow{J_2^k J_{N-1}^k}\right\| \\ \vdots & \ddots & \\ \left\|\overrightarrow{J_N^k J_1^k}\right\| & \dots & \left\|\overrightarrow{J_N^k J_{N-1}^k}\right\| \end{bmatrix} \tag{4.22}$$

where k is number of frame, $\left\|\overrightarrow{J_2^k J_1^k}\right\| (i \neq j)$ denotes Euclidean distance between $J_i^k and\ J_j^k$.

4.3 Activity Recognition

To recognize human activity, coordinate of feature point such as joint indices and location distances are important to define along the changes of human order in learning process. The system is generated from 2D to 3D pose estimation process for human activity. Human pose is represented by joints sequences. Each joint has distribution power since it is connected to nearby joints. The dependent characteristics and activity recognition are the encoded structure matrix and the model's weight matrix representation capacity is defined as:

$$\mathcal{L}_{activity} = \sum_I \sum_{t=1}^{T} \sum_{k=1}^{K_1 \times K_2} (N_t^I(k) - \alpha_t^I(k))^2 \tag{4.23}$$

If k is the spatially position, T is the total time steps, N_t^I represents the high visual data observation for each joint, and α_t^I is the joint attention score.

4.4 Deep Convolutional Neural Network

Because deep learning methods produce reliable results and automatically learn hierarchical representations from raw data, they are useful for activity recognition. The architecture of deep learning process consists of input layer, hidden layers, fully connected layer and output layer [47]. In the input layer, the normalization process is

used to reduce the impact of scale differences, and to assist the optimization process. Each hidden layer comprises three main tasks, such as convolution, max pooling and neuron activation with Rectified Linear Unit (ReLU) to speed up the convergence and to prevent the vanishing of gradient.

In the proposed system, activations to be zero-mean and unit standard deviation, which enables faster convergence, accelerates training and reduces the generalization error, experiment are described in Figure 4.9. The system performs the randomly initialized weight vectors and network is trained forward and backward multiple times until it meets minimum loss. All weight values are updated on the basis of loss value using the back propagation algorithm with gradient descent method.

The system predicts the desired format along the network and 3D skeleton information of human pose is interpreted with the distribution of joint features by fully connected layers of deep neural network. At the output layer, the probability of each possible action is classified by using soft-max function and recognizes the human activity.

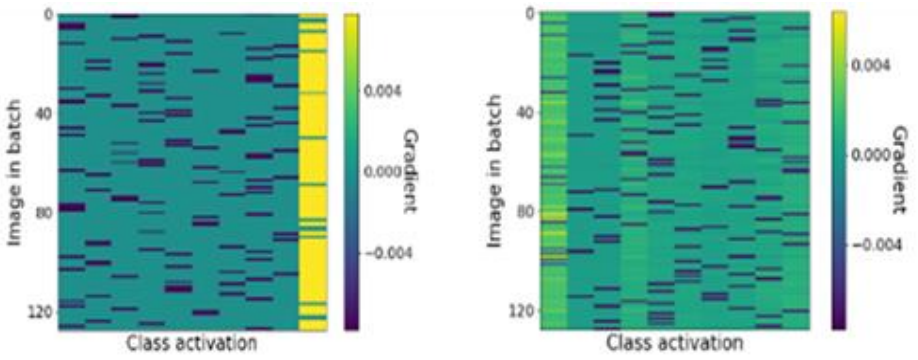


Figure 4.9 Image in Batch and Normalization on Network Process

4.4.1 Graph Modeling

This system uses the VGG net [56], a Deep Neural Network architecture for object identification, in order to achieve better results with a deeper network. The VGG net has a straightforward architecture, using 2x2 max pooling and 3x3 convolution layers across the network, each of which has pre-trained weights. On the training set, it performs better because to additional layers processing progressively smaller features.

As the network deepens, the training error decreases. This increased depth results in significant accuracy gains.

Through the development of a novel adaptive dependency matrix and its integration with node embedding, the proposed model effectively captures hidden spatial dependencies and generates a heatmap for encoding per-pixel likelihoods in joint localization of humans. Spatial and temporal data are combined into these heatmaps. Thus, by aggregating joint distributions across fully linked layers of a Deep Neural Network, a comprehensive 3D skeleton human pose is created.

4.4.2 Training on Network

During training, each sequence is processed by taking its input within a rectangular patch, which is then adjusted to RGB pixel size. Calculating the RGB channel-wise mean values and deducting them from the images is the process of data normalization. Regression training labels are composed of multi-channel heatmaps, where each channel represents the joint picture location and is characterized by its corresponding probability.

To give dense prediction for every joint, the network is composed of convolutional layers with filters and a ReLU activation function in between. Following the first three convolutional layers, a max pooling layer is added. Its output is a set of corresponding probabilities to the class labels. By reducing the loss between the open-source Cafe framework and the forecast, the network is trained.

Each layer is composed of pre-trained set of weights and adjust hyper parameters to improve the model accuracy. To accelerate the process, Adamax optimization algorithm is used on model training [51] accuracy and loss are shown in Figure. 4.10. We split size of dataset as 70% for training and 30% for testing, repeating 50 epochs with batch size of 32 and learning rate is 0.001. As a result, Adamax achieves the best results up to 94.5% on training accuracy and 93.2% on testing accuracy.

In line with earlier 3D pose techniques, testing involves assuming a subject with a bounding box, cropping the picture patch inside the bounding box in-frame, and feeding the network forward to forecast the likelihood of the joint occurring at each pixel. Lastly, acknowledge the effective comprehension of human behavior.

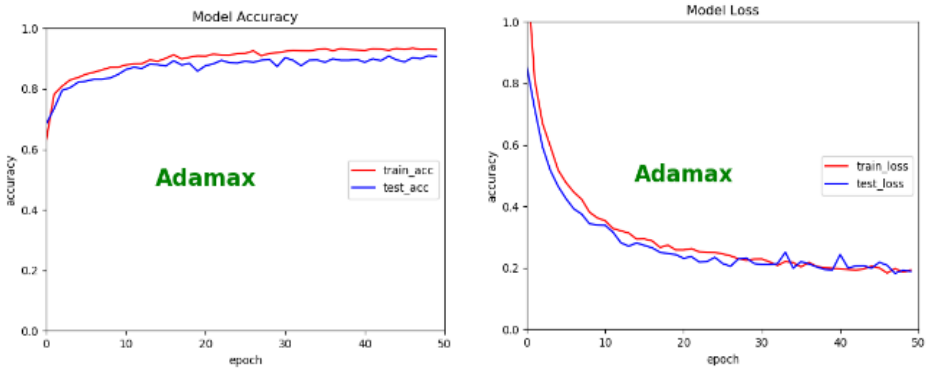


Figure 4.10 Model-accuracy and Model-Loss Graph on Training and Testing

4.5 Summary

This chapter presents the architecture of the proposed method and advanced deep learning techniques are applied to recognize human activities and to achieve better performance. The proposed method uses OpenPose detector to detect 2D skeleton key points of human body parts and reconstruct 2D to 3D key points. The proposed system recognizes overlapping area and occlusion of body parts on 3D space. In order to get more reliable result for activity recognition, Joint Collection Distance is considered to overcome the problem of similar pose but different activities. The accuracy results of experiments with testing data are discussed in chapter 5.

CHAPTER 5
DESIGN AND IMPLEMENTATION OF THE PROPOSED SYSTEM

This chapter describes about the design and implementation of proposed system along with the results obtained. The primary goal of the system across various real-life applications includes interactive systems, healthcare, and surveillance and sport analysis. The system is being developed using three approaches: the first focuses on extracting feature points from skeleton key points using the OpenPose detector; the second involves converting 2D to 3D configurations; and the third aims to optimize training time by using the EM algorithm with deep learning frameworks.

This research applies normalization on joint sequences to provide robust feature representation and generation. The purpose of 2D to 3D point consideration is to provide occlusion analysis, and viewpoint changes for target actions. The system performs skeleton extraction, which is low computing time enhances the applicability of the recognition model from video sequences, and can improve the real-time demand of the activity recognition system. The proposed method focuses on efficiently capturing and classifying human activities by utilizing skeletal data.

Recognizing human actions from videos require knowledge of each frame's geographical information as well as the time variations between the frame. Since, dynamic human recognition has increased attention due to the importance of real-time sequential video images that can provide secure information for changing objects concerning time. To generate continuous changes through the realistic sequential process, the dynamic correlation of body parts and inadequate input features is a challenging task for daily life environments.

The spatial-temporal sequence of body parts arrangement and sudden changes are still required of the detection, recognition of the object’s motion. The incorrect result has been occurred due to various circumstances such as different localization scales, occlusion of body parts, long-range of joint dependencies, background aspects, and recording situations. A deep understanding of human activity can be represented by skeleton data which is resistant to noise and robust to extract relevant features from the geometric distribution of observed motion over time and a semantic configuration over sequence of events. The continuous improvement of Artificial Intelligence (AI)

technology, deep learning with skeleton-based system is more accurate result for detection and recognition of human activity. The core concept is that the motion trajectories embedded in the dynamic alterations of joint sequences can express information about human activity. These method is robust to variations in clothing, lighting, and background appearances in real-life environments.

5.1 Data Collection and Preprocessing

The video data collection process is essential for developing robust human activity recognition system. A diverse group of participants is selected to ensure the dataset covers a wide range of body types, ages, and genders. The proposed system applies the real-life video data collected from CCTV, IP camera, and Webcams. There are a total of 1550 videos around the environment consists of four daily activities with multi-views such as forward, backward, lateral and frontal. These are recorded for indoor and outdoor conditions with different ages and different sizes including the performance of male and female actions. Example of Twenty videos from Self-Collection of Dataset are expressed in Figure 5.1.

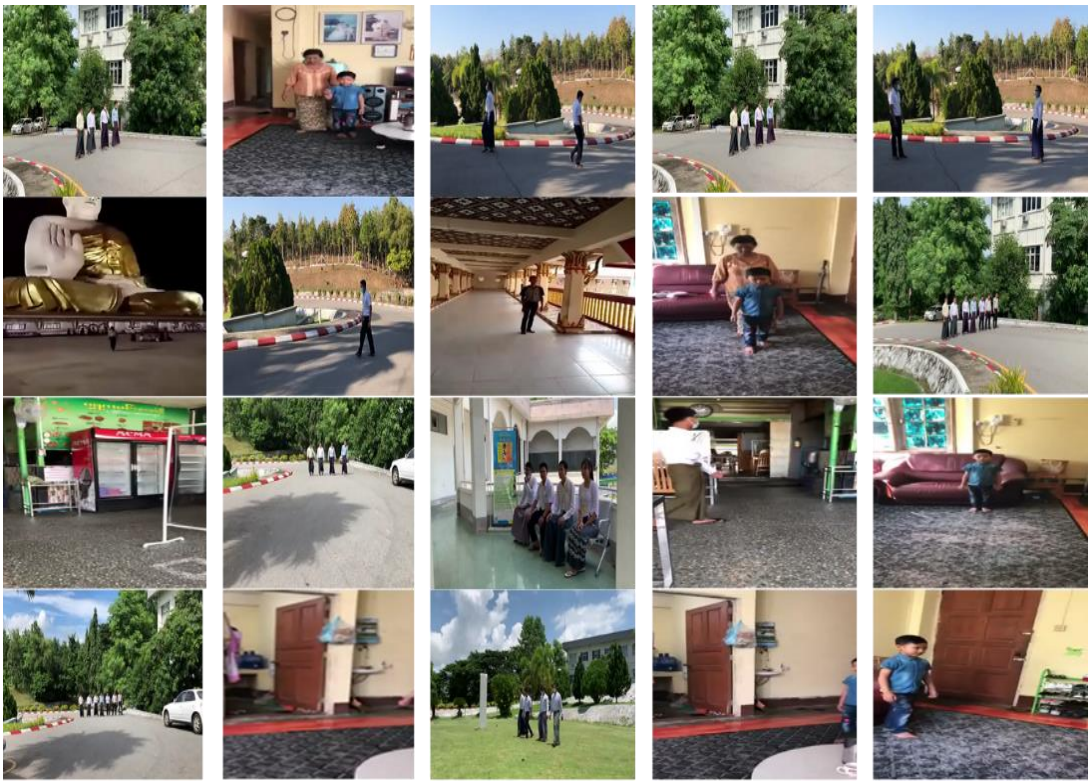


Figure 5.1 Example of Twenty Videos from Self-Collection of Dataset

The proposed system aims to achieve fast and accurate recognition results and to expand useful information from the movements of unseen skeletal joints of real-life environments. The system uses self-collection of datasets from real-life environments. Preprocessing steps include normalization of skeletal data, data augmentation, and separation of activity sequences. The design of proposed system is illustrated in Figure 5.2.

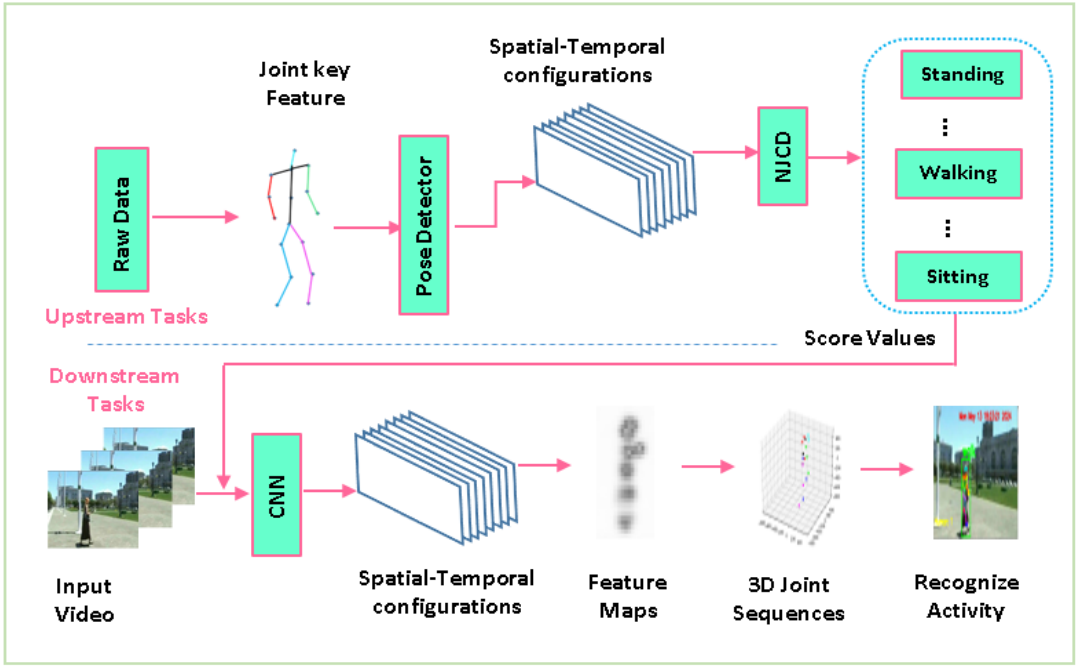


Figure 5.2 The Design of Proposed System Model

5.2 Experimental Results and Implementation

Deep learning techniques are valuable due to feature representations from raw data and more accurate result. Three primary tasks are included in each hidden layer: convolution, max pooling, and neuron activation using the Rectified Linear Unit (ReLU), which facilitates quicker training, faster convergence, and a decrease in generalization error. The system performs the randomly initialized weight vectors and network is trained forward and backward multiple times until the minimum loss. Including all weight values are updated on the basis of loss value using the back propagation algorithm.

The system predicts the desired format along the network and 2D to 3D skeleton information of human pose with the distribution of 20 joint features is interpreted by fully connected layers of Deep Convolutional Neural Network (DCNN). At the output layer, the probability of each possible action is determined by soft-max function and it

Figure 5.3 Example of Human Detection and Recognition

recognizes the human activity. Example of test result is described in Figure 5.3.

In training process, Visual Geometry Group (VGG)-19 network is used to perform easy scalability and better features processing on additional layers. It takes an input size of 224x224x3, where 224x224 is the spatial resolution, where 3 represents RGB color channels. The network consists of 3x3 convolution with stride 1 and followed by 2x2 max-pooling with stride 2. Rectified linear unit (ReLU) controls better gradient flow in convolution and that provides the dense prediction for all joints. Example of recognition result on frontal view is shown in Figure 5.4.

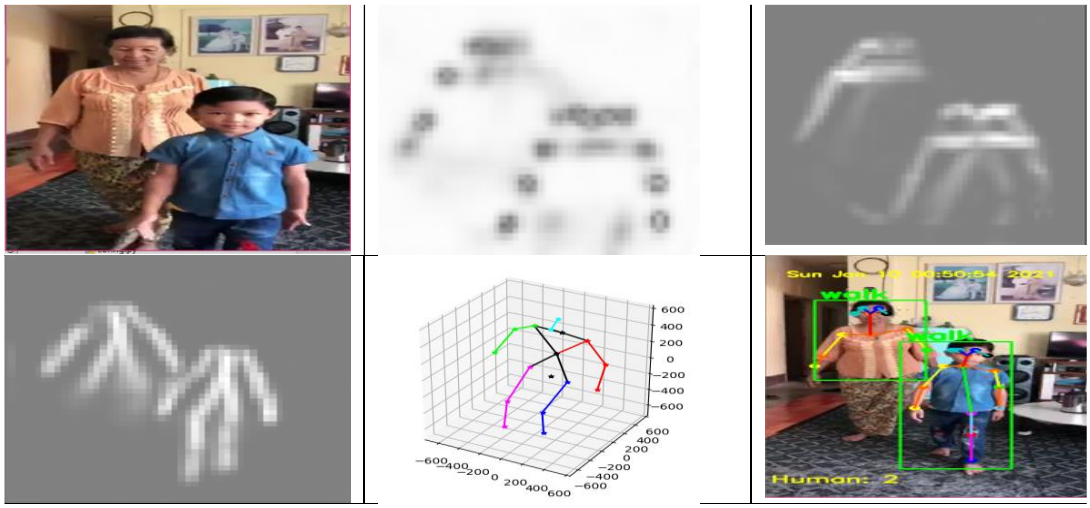


Figure 5.4 Example of Recognition Result on Frontal View

The system is measured based on True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) results in terms of Precision indicates the model is correctly classifying activity value, Recall describes true positive out of all positive, F1-Sore evaluates to define balance between precision and recall using harmonic mean and Accuracy measures the overall performance. The overall accuracy of multi-views is 94% and our system outperforms satisfactory result for human detection and activity recognition using 3D skeleton model.

$$Precision = \frac{TP}{TP+FP} \tag{5.1}$$

$$Recall = \frac{TP}{TP+FN} \tag{5.2}$$

$$F1\_Score = \frac{2*Precision*Recall}{Precision+RecallFP} \tag{5.3}$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{5.4}$$

Table 5.1 Percentages of Performance Evaluation on Multi-Views

Multi-Views	Precision	Recall	F1-Score	Accuracy
forward	94.118	88.889	91.429	89.474
backward	95.000	100.000	97.436	95.000
lateral	100.000	95.000	97.436	95.000
frontal	94.444	82.353	94.436	95.000

To know the efficient method for classification, the proposed system is tested with the experiments and it achieves 94% on real-time video. The frame width and height as 1920x1080p @ 30 fps with the average of 150 frames per video and containing the frontal, backward, forward, and side view for different actions with a few experiment results on real-life environment are described in Figure 5.5. The comparison results of different views are illustrated in Figure 5.6.

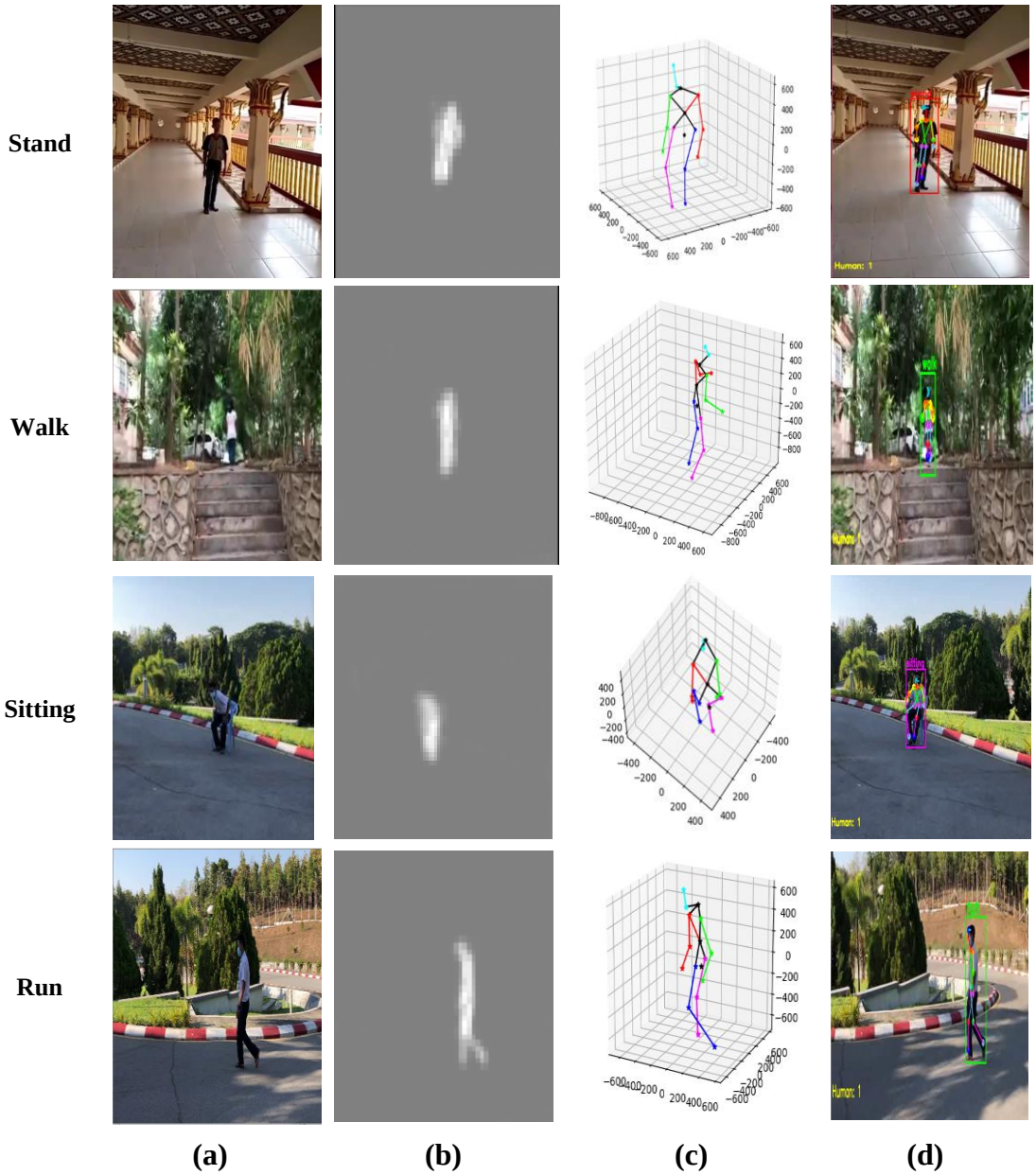


Figure 5.5 Real-time Recognition Result on Different Appearance (a) Input Video (b) Feature Maps (c) Joint Sequences on 3D (d) Activity Recognition

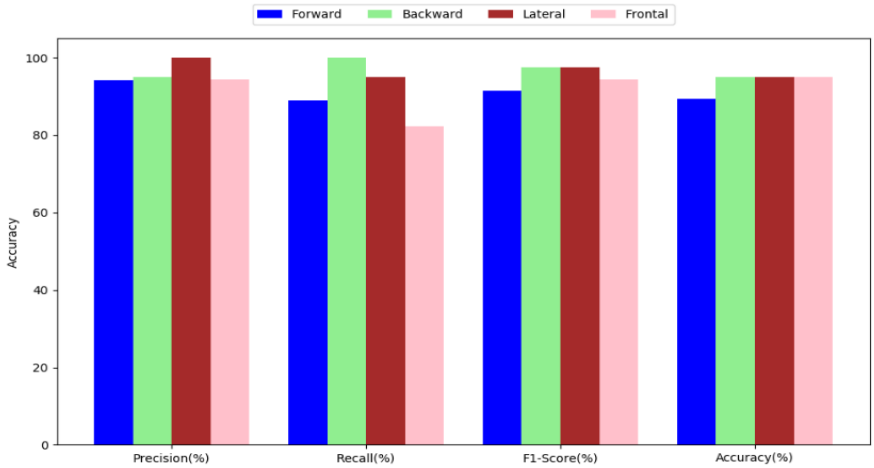


Figure 5.6 Performance Comparison Results on Different Views

5.3 Summary

This chapter demonstrates the layout and execution of the suggested human detection and activity recognition system with experimental results by our collected data. The proposed method can reduce cost and no need to human specialists. As of right now, deep learning is the most effective method for recognizing, classifying, and forecasting human behavior. Even if there has been a lot of advancement in recent years, there are still numerous obstacles to overcome before using deep learning models to the development of vision-based action detection systems and realizing their benefits in practical applications.

The main goal is to identify more accurate joint sequence results for activity recognition by detecting human body joints. As skeleton data derived from human joint information has shown remarkable performance in movement identification and it can be used to extract overlapping areas in human body parts. The system was achieved and improved with skeleton data and getting the fast and valuable information of human body parts and joint sequences movement about the target action. Chapter 6 discusses the conclusion and future work of the proposed system.

CHAPTER 6
CONCLUSION AND FUTURE WORK

There are different variations in human posture, form, variations in lighting, and backdrop appearances in real-life environment. The development of camera technologies such as RGB-Depth camera types consist of structure light and time of flight, an illuminator array and pricey wearable sensors have been utilized to analyze any visual features, it is designable to collect 3D data with the active point and that can be applied for 3D human skeleton model by analyzing of depth information in recognition systems. However, these systems are constrained by specific lighting conditions, restricted usage in outdoor environments and a restricted range.

In order to address these issues, the system is thought to and tested with different conditions on video sequences. The aim of this study is to identify efficient multi-view human activity detection systems and to obtain useful information about the target action between joint sequences and human body parts. The main idea of the proposed system is to search the state of overlapping areas and to solve the similar pose for different activities. The system is reconstructed 3D skeletal joint from detection of 2D key points and define the activities recognition for moving object.

The proposed system is considered with Joint Collection Distances and that can classify more accurate result on joint sequences activities. That also analyze the recorded information with the human motion in many scopes. Experimental results with training and testing videos are obtained from self-collection of real-time video that includes standing, walking, sitting, and running activities. There are a total of 5542 examples for training data and 1386 examples for the test data. These are recorded from indoor and outdoor conditions with different ages and different sizes consists of male and female performance. According to the result, it is satisfactory for human activity recognition.

6.1 Summary of the System

Detection and recognition of human from the video have occurred the numerous variation of the human pose, gesture and activity. This needs a well-defined method to manage behavior understanding of the different motions and different situations. Real-Time human motion detection, tracking and recognition is necessary for various

applications, including security, safety, IoT support, and human activity monitoring in both public and private areas. To provide this, the system is implemented by using OpenPose detector that achieves light weight and giving the effective results for 2D key point detection and reconstructs 3D skeletal model based on deep learning approach.

This research work aims to effectively visualize human activity recognition systems based on skeleton data and to get fast and valuable information between human body parts and joint sequences movement about the target action. The key idea is to search the state of the missing joint from 2D to 3D space and to solve the similar pose for different activities. The development is constructing to get high accuracy recognition of human movement and extracting the valuable information between skeletal joints and human movements from video sequences.

6.2 Discussion the System

This dissertation has discussed empirical research of the proposed method to support computer vision and to describe human actions and their interactions acquired from video sequences. Numerous significant applications are supported by the ability to detect and recognize human actions. Non-maximum suppression is used in the post-processing phase to recover trustworthy target detection boxes. By focusing on hidden human body parts in 2D photos, 3D skeletal models enable us to see human movement, overlapping groupings, and interpersonal interactions. Depending on the active action consists of joint sequences and corresponding limbs of the human body parts perform body-level attention score and that is a crucial feature for effective human activity recognition system.

Mostly, the location and body size dependence of the input data are the cause of the failure occurrences. Deep learning framework can provide to perform body-level attention and can improve visual object recognition. Using the backpropagation technique, the network may be trained with the given sequences to achieve the desired outcome during the training phase. This can support the gradient of a loss function with respect to all of the network's weights. The proposed network estimates 2D to 3D location of a target subject and generates relative 3D human pose. The system can recognize moving object with better performance result on processing time and transferred time, that may improve an efficient activity recognition system on deep learning frame work.

6.3 Advantages and Limitations of the Proposed System

The proposed system can develop intelligent visual surveillance system to support human-computer interaction and occlusions of body parts detection in computer vision. It can provide information about real time interpretation and robustness to the recognition of human activities changes overtime. The deep learning architecture is flexible to capture the hidden spatial dependency and it can be learned by reducing the difference in movement between the prediction and recognition of human body parts and joint displacement. It is reliable to identify the desired target actions for all visible key points moving in the scene and improves the system performance. The proposed system can run in real-time and stable to view point changes.

However, the proposed system has required to define many joint displacements for daily activities. It has to train more features for more activities recognition with several people. It does not consider low resolution image and away from video camera. Detection and recognition errors have been occurred when background and foreground are same pixel values and loss of minimum joint information. Deep learning model is mapping of continuous geometric transformations and there is no standard theory to choose right deep learning technique and it has required to make knowledge of topology, training methods and other parameters optimization.

6.4 Future Work

The future research direction will continue to study more recognition result on multiple human. Many techniques based on skeletal sequences have been developed for human recognition. Effective multi-view human activity detection systems are still necessary, though, in order to support computer vision and satisfy the safety requirements of actual users. The development of effective real-time system can record the important information and conduct a variety of environmental analyses. The proposed system will detect human movement and provide recognition results in real-life environments.

This research work will try to solve missing of similar pose with different activities and continue to work the configuration of more joint sequences for human activities recognition changes over time. Moreover, the proposed system will continue multiple human on 3D space and their activities to develop the recognition system.

Another considerable factors are object detection, tracking, and activity recognition are hampered by abrupt changes in the direction and speed of human movements as well as by shifting camera states.

AUTHOR'S PUBLICATIONS

[P1] Sandar Win and Thin Lai Lai Thein, "Real-Time Human Motion Detection and Tracking with Learning based Representation", In Proceedings of The 17th International Conference on Computer Applications (ICCA), pp. 207-212, 2019.

[P2] Sandar Win and Thin Lai Lai Thein, "Human Motion Detection and Tracking Direction using HOG and Optical Flow", In Proceedings of The 2nd Joint International Conference on Science, Technology and Innovation (IcSTI), pp. 284-288, 2019.

[P3] Sandar Win and Thin Lai Lai Thein, "Real-time human motion detection, tracking and activity recognition with skeletal model", In Proceedings of The 18th IEEE Conference on Computer Applications (ICCA), pp. 151-156, 2020.

[P4] Sandar Win and Thin Lai Lai Thein, "Effective Multi-View for Human Activity Recognition on Skeletal Model", In Proceedings of The 12th International Conference on Future Computer and Communication (ICFCC)/ The 10th International Workshop on Computer Science and Engineering (WCSE), pp. 79-83, 2020.

[P5] Sandar Win and Thin Lai Lai Thein, "MULTI-VIEWS HUMAN DETECTION AND ACTIVITY RECOGNITION USING 3D SKELETON MODEL", Indian Journal of Computer Science and Engineering (IJCSE), Vol. 15, No. 3, pp. 292-300, 2024.

Bibliography

[1] Allili, Mohand Said, Nizar Bouguila, and Djemel Ziou, "A robust video foreground segmentation by using generalized gaussian mixture modeling", In IEEE Fourth Canadian conference on computer and robot vision (CRV'07), pp. 503-509, 2007.

[2] S. Asghari-Esfeden, M. Sznaier, O. Camps, “Dynamic motion representation for human action recognition”, Proceedings of the IEEE/CVF winter conference on applications of computer vision, Snowmass, Colorado, pp. 557-566, 24 June 2020.

[3] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt, “Sequential deep learning for human action recognition”, In Proc. Int. Workshop Human Behav. Understanding, Amsterdam, The Netherlands, pp. 29–39, Nov. 2011.

[4] Banerjee, Prithviraj, and Somnath Sengupta, "Human motion detection and tracking for video surveillance", In Proceedings of the national Conference of tracking and video surveillance activity analysis, pp. 88-92, 2008.

[5] Bay, Herbert, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool, "Speeded-up robust features (SURF)", Computer vision and image understanding 110, No. 3 pp. 346-359, 2008.

[6] Belagiannis, Vasileios, Sikandar Amin, Mykhaylo Andriluka, Bernt Schiele, Nassir Navab, and Slobodan Ilic, "3D pictorial structures for multiple human pose estimation", In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1669-1676. 2014.

[7] Beleznai, Csaba, and Horst Bischof, "Fast human detection in crowded scenes by contour integration and local shape estimation", In IEEE Conference on Computer Vision and Pattern Recognition, pp. 2246-2253, 2009.

[8] Bobick, Aaron F., and James W. Davis, "The recognition of human movement using temporal templates", IEEE Transactions on pattern analysis and machine intelligence 23, No. 3 , pp. 257-267, 2001.

[9] Bux, Allah., “Vision-based human action recognition using machine learning techniques”, Lancaster University (United Kingdom), 2017.

[10] Dalal, Navneet, and Bill 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., 2005.

[11] Du, Wenbin, Yali Wang, and Yu Qiao, "Rpan: An end-to-end recurrent pose-attention network for action recognition in videos", In Proceedings of the IEEE international conference on computer vision, pp. 3725-3734, 2017.

[12] Du, Yong, Wei Wang, and Liang Wang, "Hierarchical recurrent neural network for skeleton based action recognition", In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1110-1118, 2015.

[13] Gonzalez, Rafael C. Digital image processing. Pearson education india, 2009.

[14] Grahn, Josef, and H. Kjellstromg, "Using SVM for efficient detection of human motion", In IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, pp. 231-238. IEEE, 2005.

[15] Haritaoglu, Ismail, David Harwood, and Larry S. Davis, "W/sup 4: real-time surveillance of people and their activities", IEEE Transactions on pattern analysis and machine intelligence 22, No. 8 , pp. 809-830, 2000.

[16] Heo, Byeongho, Kimin Yun, and Jin Young Choi, "Appearance and motion based deep learning architecture for moving object detection in moving camera", In IEEE International Conference on Image Processing (ICIP), pp. 1827-1831, 2017.

[17] Hou, Jie, Baolong Guo, Wangpeng He, and Jinfu Wu, "A 3D Human Skeletonization Algorithm for a Single Monocular Camera Based on Spatial–Temporal Discrete Shadow Integration", Applied Sciences 7, No. 7, pp. 685, 2017.

[18] Hu, Yuxiao, Liangliang Cao, Fengjun Lv, Shuicheng Yan, Yihong Gong, and Thomas S. Huang, "Action detection in complex scenes with spatial and temporal ambiguities", In 12th IEEE Conference on Computer Vision, pp. 128-135, 2009.

[19] Huang, Junjie, Wei Zou, Jiagang Zhu, and Zheng Zhu, "Optical flow based real-time moving object detection in unconstrained scenes", arXiv preprint arXiv:1807.04890, 2018.

[20] Hussein, Mohamed E., et al., “Human action recognition using a temporal hierarchy of covariance descriptors on 3d joint locations", Twenty-third international joint conference on artificial intelligence, 2013.

[21] Ikizler, Nazli, and David Forsyth, "Searching video for complex activities with finite state models", In IEEE Conference on computer vision and pattern recognition, pp. 1-8, 2007.

[22] Ilg, Eddy, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox, "Flownet 2.0: Evolution of optical flow estimation with deep networks", In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2462-2470, 2017.

[23] U. Iqbal, A. Doering, H. Yasin, B. Krüger, A. Weber, and J. Gall, “A Dual-Source Approach for 3D Human Pose Estimation from a Single Image”, International Journal of Computer Vision and Image Understanding, Vol. 172, No. 7, pp. 37-49, 2018.

[24] Ji, Shuiwang, Wei Xu, Ming Yang, and Kai Yu, "3D convolutional neural networks for human action recognition", IEEE transactions on pattern analysis and machine intelligence 35, No. 1, pp. 221-231, 2012.

[25] He Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition", In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778, 2016.

[26] G. Kani, P. Geetha, “Adaptive Weighted Flow Net Algorithm for Human Activity Recognition UsingDepth Learned Features”, Computer Systems Science and Engineering, Vol. 46, No. 2, pp. 1447-1469, 2023.

[27] Karahoca, Adem, and Murat NurullaSURFlu, "Human motion analysis and action recognition", In WSEAS International Conference. Proceedings. Mathematics and Computers in Science and Engineering, Vol. 7. World Scientific and Engineering Academy and Society, 2008.

[28] M. A. Khan, K. Javed, S. A. Khan, T. Saba, U. Habib, J. A. Khan, A. A. Abbasi, “Human action recognition using fusion of multiview and deep features: an application to video surveillance”, Multimedia tools and applications, Vol. 83, No. 5, pp. 14885-14911, 2024.

[29] S. W. Kim, K. Yun, K. M. Yi, S. J. Kim, and J. Y. Choi, “Detection of moving objects with a moving camera using non-panoramic background model”, Machine Vision and Applications, Vol. 24, pp. 1015–1028, 2013.

[30] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton, "Imagenet classification with deep convolutional neural networks", Advances in neural information processing systems 25, 2012.

[31] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: a large video database for human motion recognition", Proceeding of the IEEE Conference on Computer Vision, Barcelona, Spain, pp. 2556-2563, 06-13 November 2011.

[32] I. Laptev, T. Lindeberg, "On space-time interest points", International journal of computer vision 64, No. 2, pp.107-123, 2005.

[33] LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton, "Deep learning", nature 521, No. 7553 , pp.436-444, 2015.

[34] Li, Bin, Xi Li, Zhongfei Zhang, and Fei Wu, "Spatio-temporal graph routing for skeleton-based action recognition", In Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, No. 01, pp. 8561-8568. 2019.

[35] Lin, Zhe, and Larry S. Davis, "Shape-based human detection and segmentation via hierarchical part-template matching", IEEE transactions on pattern analysis and machine intelligence 32, No. 4 , pp. 604-618, 2010.

[36] Lowe, David G., "Distinctive image features from scale-invariant key points", International journal of computer vision 60, No. 2 , pp. 91-110, 2004.

[37] D. C. Luvizon, D. Picard, H. Tabia., “2D/3D Pose Estimation and Action Recognition using Multitask Deep Learning”, Proceedings of IEEE Conference on Computer Vision Foundation, Salt Lake, Utah, pp. 5137-5146, 18-22 June 2018.

[38] Martinez, Julieta, Rayat Hossain, Javier Romero, and James J. Little, "A simple yet effective baseline for 3d human pose estimation", In Proceedings of the IEEE international conference on computer vision, pp. 2640-2649, 2017.

[39] Migdal, Joshua, and W. Eric L. Grimson, "Background subtraction using markov thresholds", In Seventh IEEE Workshops on Applications of Computer Vision (WACV/MOTION'05), Vol. 2, pp. 58-65, 2005.

[40] Park, Ki-Yeong, and Sun-Young Hwang, "An improved Haar-like feature for efficient object detection", Pattern Recognition Letters 42 , pp. 148-153, 2014.

[41] Pavlakos, Georgios. "Learning to Reconstruct 3D Humans." PhD diss., University of Pennsylvania, 2020.

[42] Plagemann, Christian, Varun Ganapathi, Daphne Koller, and Sebastian Thrun, "Real-time identification and localization of body parts from depth images", In 2010 IEEE International Conference on Robotics and Automation, pp. 3108-3113. IEEE, 2010

[43] Rakibe, Rupali S., and Bharati D. Patil, "Background subtraction algorithm based human motion detection", International Journal of scientific and research publications 3, No. 5 pp. 2250-3153, 2013.

[44] S. Ranasinghe, F. Al Machot, and H. C. Mayr, “A review on applications of activity recognition systems with regard to performance and evaluation”, International Journal of Distributed Sensor Networks, Vol. 12, No. 8, p. 1550147716665520, 2016.

[45] Rohr, Karl., "Towards model-based recognition of human movements in image sequences", CVGIP: Image understanding 59, No. 1, pp. 94-115, 1994.

[46] Saho, Kenshi, "Kalman filter for moving object tracking: Performance analysis and filter design", Kalman Filters-Theory for Advanced Applications , pp.233-252, 2017.

[47] Shrestha, Ajay, and Ausif Mahmood, "Review of deep learning algorithms and architectures", IEEE access 7, pp. 53040-53065, 2019.

[48] Simonyan, Karen, and Andrew Zisserman, "Two-stream convolutional networks for action recognition in videos", Advances in neural information processing systems 27, 2014.

[49] Singla, Nishu, "Motion detection based on frame difference method", International Journal of Information & Computation Technology 4, No. 15 , pp. 1559-1565, 2014.

[50] K. Soomro, A. R. Zamir and M. Shah, “UCF101: A dataset of 101 human actions classes from videos in the wild,” arXiv preprint arXiv: 1212.0402, 2012.

[51] D. Soydaner, "A comparison of optimization algorithms for deep learning", International Journal of Pattern Recognition and Artificial Intelligence, Vol. 34, No. 13, 2020, pp. 2052013.

[52] T. Sztyler, H. Stuckenschmidt, andW. Petrich, “Position-aware activity recognition with wearable devices,” Pervasive and mobile computing, 2017.[18]

[53] N. Tasnim, M. K. Islam, J. H. Baek, “Deep learning based human activity recognition using spatio-temporal image formation of skeleton joints”, Applied Sciences, Vol. 11, No. 6, pp.2675, 2021.

[54] Toshev, Alexander, and Christian Szegedy, "Deeppose: Human pose estimation via deep neural networks", In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1653-1660, 2014.

[55] Tran, Du, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri, "Learning spatiotemporal features with 3d convolutional networks", In Proceedings of the IEEE international conference on computer vision, pp. 4489-4497, 2015.

[56] A. Vedalai, and A. Zisserman, "VGG convolutional neural networks practical", Department of Engineering Science, University of Oxford, 66, 2016.

[57] Viola, Paul, Michael J. Jones, and Daniel Snow, "Detecting pedestrians using patterns of motion and appearance", International Journal of Computer Vision 63, No. 2 , pp. 153-161, 2005.

[58] Vosoughi, Saeid, and Maria A. Amer, "Deep 3D human pose estimation under partial body presence", In 25th IEEE International Conference on Image Processing (ICIP), pp. 569-573. IEEE, 2018.

[59] Wang, Heng, and Cordelia Schmid, "Action recognition with improved trajectories", In Proceedings of the IEEE international conference on computer vision, pp. 3551-3558, 2013.

[60] Wang, Limin, Yu Qiao, and Xiaoou Tang, "Action recognition with trajectory-pooled deep-convolutional descriptors", In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4305-4314. 2015.

[61] Wang, Limin, Yuanjun Xiong, Zhe Wang, and Yu Qiao, "Towards good practices for very deep two-stream convnets", arXiv preprint arXiv:1507.02159 , 2015.

[62] Wu, Di, and Ling Shao, "Leveraging hierarchical parametric networks for skeletal joints based action segmentation and recognition", In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 724-731, 2014.

[63] Wu, Bo, and Ram Nevatia, "Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors", International Journal of Computer Vision 75, No. 2, pp. 247-266. 75(2):247–266, 2007.

[64] Xu, Tianhan, and Wataru Takano. "Graph stacked hourglass networks for 3d human pose estimation." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 16105-16114. 2021.

[65] Yang, Fan, Yang Wu, Sakriani Sakti, and Satoshi Nakamura, "Make skeleton-based action recognition model smaller, faster and better", In Proceedings of the ACM multimedia asia, pp. 1-6, 2019.

[66] Yang, Zhengyuan, Yuncheng Li, Jianchao Yang, and Jiebo Luo, "Action recognition with visual attention on skeleton images", In 24th IEEE Conference on Pattern Recognition (ICPR), pp. 3309-3314, 2018.

[67] B. X. Yu, Y. Liu, K. C. Chan, “Skeleton focused human activity recognition in rgb video”, arXiv preprint arXiv: 2004.13979, 2020.

[68] O. Yurur, C. Liu, and W. Moreno, “A survey of context-aware middleware designs for human activity recognition,” IEEE Communications Magazine, Vol. 52, No. 6, pp. 24–31, 2014.

[69] Zarka, Nizar, Ziad Alhalah, and Rada Deeb, "Real-time human motion detection and tracking", In IEEE 3rd international conference on information and communication technologies: from theory to applications, pp. 1-6, 2008.

[70] Zhaoyang, Chen, Gao Haolin, Wang Kun, “A motion based object detection method”, In 2020 2nd International Conference on Information Technology and Computer Application (ITCA).

[71] L. Zhi, C. Zhang, and Y. Tian, “3D-based deep convolutional neural network for action recognition with depth sequences”, Image Vis. Comput., Vol. 55, No. 2, pp. 93–100, Nov. 2016.

[72] Zhou, Xiaowei, Menglong Zhu, Spyridon Leonardos, and Kostas Daniilidis, "Sparse representation for 3D shape estimation: A convex relaxation approach", IEEE transactions on pattern analysis and machine intelligence 39, No. 8, pp.1648-1661, 2016.

[73] Y. Zhu, G. Liu, “Fine-grained action recognition using multi-view attentions”, The Visual Computer, Vol.36, No. 9, pp. 1771-1781, 2020.