Chapter 1

INTRODUCTION

This section of the thesis first highlights the introduction, motivation, and problem statement. Then, at last, this chapter winds up with the objective, contribution, and thesis outline.

1.1 Background

Human beings are very much capable of identifying human poses and their behavior. The basis of this capability is based on visual interpretation and observations. Therefore, making a machine recognize and predict the human posture and behavior is one of the most challenging computer science tasks. However, computer science, which aims to enable the computer to have such a thorough understanding of visual content, is called computer vision (CV) in literature. The most complex subject in the field of CV are people because of their deformable capability. Human detection, pose estimation, and activity recognition from images are essential high-level CV tasks addressed mainly in literature.

In the detection task, humans are generally localized with the help of a bounding box. In the human pose estimation (HPE) task, the goal is to recover the human body pose in the 2D or 3D space from an RGB image. In human activity recognition (HAR), the HAR's purpose is to understand what actions are performed in the video automatically. While many strong techniques have already been introduced for human detection in real-world environments, pose estimation and activity recognition remain an open problem. Automating HPE and HPR has become a crucial step towards several other essential purposes, such as human-computer interfaces, automatic surveillance systems, sports performance analysis, 3D scene understanding, augmented reality, and content-based image and video indexation, among many others.

In this thesis, we address the 2D and 3D HPE and Activity Recognition problems. The scope of this thesis is to solve 2D HPE problem from a single view and the 3D problem from both single and multiple perspectives. At last, the Activity recognition problem from a single view.

1.2 Motivation

Modern advancements in digital image technology, storage capacity, computational speed, networking, and constant rise in computational capacities of hardware at an economical expense have caused it to capture, store, transmit, and manipulate videos and images. Because of this, videos and images are highly available in our everyday life routine, and are utilized in many applications, medicine, science, security, education, and entertainment.

The principal task of computer vision (CV) is to replicate and model the ways humans see. This CV also comprises the reasoning regarding interactions among objects and humans. To do the aforementioned reasoning, the device needs first to access the interpretations of activity and pose obtained by the visual data. So, the role of determining the activity and pose of the human body in videos or images are referred to as activity recognition and pose estimation.

2D and 3D human pose estimation and activity recognition are critical topics of machine learning, image processing too. It has been widely utilized in many applications like human-robot interaction, animation, surveillance, cloth parsing, translating sign languages, posture correction, and many more.

This thesis handles 2D and 3D human pose estimation and activity recognition problems with machine learning. These tasks have primarily correlated because of the common subject understanding and the great relevance of human pose estimation for activity recognition. This view heads toward three axioms. First is the 2D Human Pose Estimation over

RGB images. The second is the 3D Human pose estimation over RGB images. At last, the third is activity recognition, which has benefitted from the above tasks.

Determining the human pose is a complex and challenging problem. The human body shows enormous variability in size, texture, and shape. In spite of this, due to the articulated joints of the human skeleton, it shows high degrees of freedom (DOF), which provides a huge range of motion for all body parts. HPE is also a challenging problem owing to the existence of changes in viewpoints, and complex backgrounds. Some of these difficulties have been addressed in constrained settings by motion capture (MoCap) systems with reflective markers [1] [2], depth sensors [3] [4], or inertial measurement units (IMUs) [5] [6]. However, such systems require particular hardware that is often expensive and exclusive, while the data acquisition process often restricts the range of human movement. Therefore, as a more accessible approach, estimating human pose from RGB images captured by regular cameras is gathering the attention of researchers.

2D HPE methods estimate the 2D position or spatial location of human body keypoints from images or videos. Traditional 2D HPE methods adopt different hand-crafted feature extraction techniques [7] [8] for body parts, and these early works describe the human body as a stick figure to obtain global pose structures. Significant progress has been made in this area in recent years by leveraging the power of deep learning [9] [10], which has shown remarkable performance especially in detecting 2D key points in RGB images. So motivated by this fact, we try to solve the 2D HPE problem using deep learning techniques in two ways. First, a detection-based deep learning model has been proposed. Here the model first detect the body location using a bounding box then over that, estimation has been performed to know the body joint coordinates. The second is 2D HPE model directly estimates the 2D body joint coordinates fron input image using deep learning based method.

3D HPE, which aims to predict locations of body joints in 3D space, has been attracted much interest in recent years since it can provide extensive 3D structure information related to the human body. It can be applied to various applications (e.g., 3D movie and animation industries, virtual reality, and online 3D action prediction). Although significant improvements have recently been achieved in 2D HPE, 3D HPE still remains as a challenging task. Most existing research works tackle 3D HPE from monocular images or videos [11] [12], which is an ill-posed and inverse problem due to projection of 3D to 2D where one dimension is lost. When multiple viewpoints are available, or other sensors such as IMU and LiDAR are deployed, 3D HPE can be a well-posed problem employing information fusion techniques. Another limitation is that deep learning models are data-hungry and sensitive to the data collection environment. Unlike 2D human datasets, where accurate 2D pose annotation can be easily obtained, collecting accurate 3D pose annotation is time-consuming, and manual labeling is not practical. Also, datasets are usually collected from indoor environments with selected daily actions. Recent works [13] [14] have validated the poor generalization of models trained with biased datasets by cross-dataset inference [15]. So, we have proposed three 3D HPE methods that handle the above-mentioned projection ambiguity problem and less availability of 3D ground-truth data problems very well. The first two models focus on single-view monocular image, and then the third model covers 3D HPE from multi-view data input.

The goal of a Human Activity Recognition (HAR) system is to predict the label of a person's action from an video. This interesting topic is inspired by many useful real-world applications, such as simulation, visual surveillance, understanding human behavior, etc. Action recognition through videos is a well-known and established research problem. There are various methods present for activity recognition [16] [17]; every technique has its advantages and disadvantages. Despite being a lot of research work, recognizing activity is still a complex and challenging task. By monitoring the movement of your skeletal joints, Kinect is then able to accurately interpret your actions. In this work, we proposed an approach for human activity recognition that uses 3D skeleton person poses data to recognize its activity.

1.3 Problem Statement

The problem that this thesis is addressing is **"Study of Conventional and Deep Learning models for human pose estimation and activity recognition and Implementation of frameworks to identify human poses and activities in images or video by using convention machine learning and deep learning approaches"**.

Given an input RGB image or video, the objective is to estimate the 2D and 3D skeleton joints locations of the human and its activity that are inside the camera's field of view, in order to extract the body poses of all the humans.

The pose estimation and recognition process must be robust towards the self-occlusions, partial/cropped views, non-frontal subjects, non-typical (i.e. top-down) viewpoints, cluttered scenes, projection ambiguity and less availability of ground-truth data. Additionally the algorithm design should allow the seamless and efficient implementation of the developed 2D and 3D human pose estimation and human activity recognition algorithm to a variety of platforms.

1.4 Thesis Objective

The objective of the thesis is to introduce different deep learning approaches for 2D and 3D HPE and human activity recognition with better accuracy compared to other state-of-the-art techniques. Firstly, two deep learning models have been introduced for 2D HPE that are robust to above-mentioned challenges. The three 3D HPE methods have been introduced based on deep learning, out of which two ways work for single view data input and third for multi-view data input. Subsequently, at last, one deep learning HAR technique has been proposed using RGB and 3D skeleton data. These all given models have been evaluated on publicly available datasets like INIRIA person, MPII, and LSP datasets for 2D HPE, Human3.6M and HumanEva-I datasets for 3D HPE, and UTD-MHAD, CAD-60, and NTU-RGB + D120 datasets for HAR.

The thesis objectives are:

- To study the existing literature of 2D HPE, 3D HPE, and HAR techniques extensively, including classical approaches, conventional machine learning methods, and deep learning methods.

- To propose deep learning based techniques for 2D HPE that are more accurate than other existing state-of-the-art techniques and its implementation and performance evaluation.

- To propose deep learning networks for 3D HPE from a single and multi-view perspective that are robust and perform more accurately than other state-of-the-art techniques and their implementation and performance evaluation.

- To propose deep learning-based approaches for HAR and their implementation and performance evaluation to address the limitation of the existing methods.

The contribution to the thesis are:

- An extensive study of the existing literature on experiments and research performed under Human Pose Estimation and Activity Recognition using conventional as well deep learning approaches to identify research gaps are presented.

- Proposed a 2D HPE framework using conventional machine and deep learning networks. Here the input is an RGB image; first, we detect the location of the human body using feature extraction and SVM classifier. Over the resultant output, we used the proposed deep network that estimates 2D body joint coordinates.

- Proposed a 2D HPE method totally based on Deep neural networks. This method is direct 2D pose estimation approach, this has three consecutive parts named as feature extraction, feature refinement, and the fusion of detection and contextual maps. This method takes input as RGB image and produce 2D body joint coordinates as output.

- Proposed a deep neural network based 3D HPE method for single-view data input. This method has two consecutive stages as, first estimates the 2D pose, second estimate 3D body joint coordinates using deep neural network architecture. This method uses only spatial information for 3D body joint coordinates estimation.

- Proposed a three stage 3D HPE method based on deep neural network. The method works for single view data input. Here the method uses both spatial and temporal information and have three consecutive stages, first estimates the 2D body poses using simple Deep convolutional neural network (DCNN), second estimates the 3D human body joint coordinates using DCNN, and at last LSTM has been used to refine the 3D body joint locations.

- Proposed a 3D HPE method from multi-view input data. The method uses both 2D pose information data for 3D HPE and also image data for direct 3D HPE. By this strategy, the model delivers both of the merits concurrently in a multi-view scenario.

- Proposed a two-stream model for HAR using 3D skeleton and RGB images input. The method uses convolutional neural network (Convnet) and long short-term memory (LSTM) as a recurrent neural network (RNN).

1.5 Outline of the thesis

This thesis consists of six chapters. Outline of the thesis is as follows:

Chapter 1 presents a short introduction of all the problems discussed in the thesis, accompanied by the thesis's objectives. Lastly, this section terminates with a concise description of the contributions of the idea in the thesis field of image/video processing.

Chapter 2 discusses the theoretical background and literature review for 2D and 3D HPE and HAR. In this section of the thesis, we first discuss the background and literature of 2D and 3D HPE, and at last, we discuss the same for HAR.

Chapter 3 presents two 2D HPE approaches based on the deep learning approach. The objective of the first approach is to make a simple and efficient system for human pose estimation with two modules: the human detector and its pose estimation. First, we propose an aggregated architecture with SVM and CNN to make an accurate human detector. Then, a modified Fully Convolutional Google network has been proposed to make the final prediction. We have examined both the architecture modules on the popular publicly available dataset like the INRIA person dataset, LSP, and MPII pose estimation dataset. The method gives an impressive performance on these datasets as compared to other states of the art methods.

In the second approach, the proposed network has three consecutive parts: (1) Deep Convolutional Neural Network architecture (DCNN) based feature extraction, (2) feature refinement, and (3) the fusion of detection and context information. During the feature extraction phase, we have proposed a fusion of two DCNN modules, which have been inspired by VGG-19 and Inception-v4 deep learning architectures. In the feature refinement, a cascaded feature integration technique has been proposed over the stacked hourglass, to make the system efficiently locate the challenging joints. At last, a fusion of context information with the detected prediction is performed, which makes the system accurate towards occlusion. In this way, pose with difficult joints coordinates can be reliably estimated even in the presence of occlusion or severe distracting factors. The successful testing of the proposed method has been done on popular MPII and LSP datasets. Based on the experimental results and the analysis of the selected datasets, it is found that the proposed framework is more accurate compared to other state-of-the-art methods in terms of the PCK metric.

Chapter 4 presents two 3D HPE approaches that estimate the 3D human pose from single view. The first method is over single view image, here we propose a three-stage deep network having the workflow of 2D Human Pose Estimation (HPE) followed by 3D HPR; which utilizes the proposed Frame Specific Pose Estimation (FSPE), Multi-Stage Cascaded Feature Connection (MSCFC) and Feature Residual Connection (FRC) Sub-level Strategies. In the second stage, the basic deep learning concepts have been utilized with the FRC Strategy for spatial 3D reconstruction. In the last stage, LSTM deep architecture has been used for temporal refinement. The effectiveness of the technique has been demonstrated on MPII, Human3.6M, and HumanEva-I datasets. From the experiments, it has been observed that the proposed method gives competitive results to the recent state-of-the-art techniques.

The second method is also over single image that introduces a two stage architecture, in the first stage we generate the 2D keypoints and in the next stage, the densely connected deep architecture has been beneficial to estimate human 3D pose. To check the effectiveness of the introduced network, the experimental evaluation was performed on the MPII and Human3.6M dataset. The comparison is also performed along the state-of-the-art techniques and it proves that the given approach is better in terms of MPJPE and P-MPJPE error metric.

Chapter 5 presents one 3D HPE approache that estimate the 3D human pose from multi-view image. Here we have tried to combine the merits of such traditional techniques with that of a deep architecture model. By this strategy, the model delivers both of the merits concurrently in a multi-view scenario and also fuses this knowledge on the upcoming step with early and late fusion strategies. Experimental results show that the proposed method achieves comparable performance to the state-of-the-art methods on MPII, and Human3.6M datasets.

Chapter 6 presents a technique for HAR that utilizes the RGB and skeleton information with the help of a convolutional neural network (Convnet) and long short-term memory (LSTM) as a recurrent neural network (RNN). The proposed method has two parts: frst, motion representation images like motion history image (MHI) and motion energy image (MEI) have been created from the RGB videos. The convnet has been trained, using these images with feature-level fusion. Second, the skeleton data have been utilized with a proposed algorithm that develops skeleton intensity images, for three views (top, front and side). Each view is frst analyzed by a convnet that generates the set of feature maps, which are fused for further analysis. On top of convnet sub-networks, LSTM has been used to exploit the temporal dependency. The softmax scores from these two independent parts are later combined at the decision level. The proposed approach has been tested on three famous and challenging multimodal datasets which are UTD-MHAD, CAD-60 and NTU-RGB + D120. Results have shown that the stated method gives a satisfactory result as compared to the other state-of-the-art systems.

Chapter 7 presents the principal conclusions of the proposed works done. This section also presents few feasible future perspectives of the thesis.