

## PREFACE

---

Due to technological advancement, massive multimedia data is usually available in image and video forms. Many demanding applications currently use pictures and videos like human-computer interaction, automatic surveillance systems, 3D scene understanding, sports performance analysis, etc. This thesis addresses Human Pose Estimation (HPE) and Human Activity Recognition (HAR) problem. These tasks are highly utilized in the above-mentioned applications. The HPE problem is divided into two categories based on the application requirement. The first one is 2D HPE, and the second is 3D HPE. Many state-of-the-art techniques have already been introduced for HPE and HAR problem solutions, but there are many challenges yet to be solved.

This thesis first presents a detailed survey for 2D HPE, 3D HPE, and HAR problems, followed by research gaps and challenges. Second, this thesis addresses the problem of 2D HPE from the single-view perspective. The task has further been classified into a direct 2D HPE and detection-based 2D HPE.

For detection-based 2D HPE, a machine and deep learning-based model has been proposed that detects the region of interest having a human as a subject and then estimates the respective body joints' locations. For the detection part, the HOG and LBP features have been extracted and then fused. The resultant fused features have been utilized with lib-SVM for classification purposes. A CNN-based deep model has been used to make

the detected proposals more accurate. Finally, a deep network has been proposed for estimation of the joint locations over the detected part. The model has been examined on publically available datasets like the INIRIA person dataset, LSP, and MPII pose estimation dataset. The proposed method gives a remarkable performance on these datasets as compared to other state-of-the-art methods.

The direct 2D HPE method directly estimates the 2D body joint coordinates from an input image. The model consists of three consecutive parts: first, a two-branch input DCNN network has been used for feature extraction. Second, a cascaded feature integration technique has been used for feature refinement that is based on the stacked hourglass method. Finally, the fusion has been performed over the detected part heatmap and the context heatmap to make the system more accurate towards occlusion. To check the effectiveness of the model, the evaluation has been performed on two standard datasets MPII and LSP. The proposed method improves pose prediction results on the PCK metric and has been compared to other state-of-the-art techniques.

Later on in this thesis, we have extended our research towards estimating 3D human pose, which includes single view and multi-view data inputs.

For 3D HPE, three models have been proposed, where the first two work on single view data input and the third works on multi-view data input. The first model uses spatial data information for 3D HPR. For this, a two-stage deep learning model has been introduced. In the first stage, the model estimates the 2D joint key points, and in the next step, a densely connected deep architecture has been used for 3D HPE.

The second model over a single view data input uses spatial and temporal information to estimate 3D human pose. In this method, a three-stage deep learning model has been proposed. First, the model evaluates the 2D Human pose. Then, in the second stage, basic deep learning concepts like convolution, batch normalization, ReLU, and dropout have been utilized for spatial 3D pose estimation. In the last step, LSTM deep architecture has been used for temporal refinement. The third model is from multi-view data input. This model proposes a two-stage multi-view 3D HPR technique using deep learning that utilizes both the direct and 2D pose information in a multi-view scenario to estimate the 3D pose. All the proposed 3D HPE techniques give outstanding results on MPJPE, P-MPJPE, and N-MPJPE metrics over the Human3.6M dataset compared to other state-of-the-art approaches.

Finally, in this thesis, a conventional and machine learning-based technique has been proposed that uses RGB data with its 3D skeleton information for HAR. The model takes 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: first, motion representation images like motion history images (MHI) and motion energy images (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 first analyzed by a convnet, which generates the set of feature maps 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 suggested approach privileges the perfect utilization of RGB and skeleton data available from an RGB-D sensor. The proposed approach has been tested on three famous and challenging multimodal datasets: UTD-MHAD, CAD-60, and NTU-RGB + D120. Results have shown the method gives a satisfactory outcome as compared to the other state-of-the-art.