|  |  |
| --- | --- |
| **Assignment** | |
| **Course Code** | CSE308A |
| **Course Name** | Computer Vision |
| **Programme** | B.Tech |
| **Department** | CSE |
| **Faculty** | FET |

|  |  |
| --- | --- |
| **Name of the Student** | Satyajit Ghana |
| **Reg. No.** | 17ETCS002159 |
| **Semester/Year** | 07/2020 |
| **Course Leader(s)** | Dr. Aruna Kumar S V |



|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Declaration Sheet | | | | | | | | |
| Student Name | Satyajit Ghana | | | | | | | |
| Reg. No | 17ETCS002159 | | | | | | | |
| Programme | B.Tech | | | | | Semester/Year | 07/2020 | |
| Course Code | CSE308A | | | | | | | |
| Course Title | Computer Vision | | | | | | | |
| Course Date |  | | to |  | | | | |
| Course Leader | Dr. Aruna Kumar S V | | | | | | | |
| **Declaration**  The assignment submitted herewith is a result of my own investigations and that I have conformed to the guidelines against plagiarism as laid out in the Student Handbook. All sections of the text and results, which have been obtained from other sources, are fully referenced. I understand that cheating and plagiarism constitute a breach of University regulations and will be dealt with accordingly. | | | | | | | | |
| Signature of the Student | |  | | | | | Date |  |
| Submission date stamp  (by Examination & Assessment Section) | |  | | | | | | |
| Signature of the Course Leader and date | | | | | Signature of the Reviewer and date | | | |
|  | | | | |  | | | |

# Contents

[Declaration Sheet ii](#_Toc57759878)

[Contents iii](#_Toc57759879)

[List of Figures iv](#_Toc57759880)

[1 Question 1 5](#_Toc57759881)

[1.1 Executive Summary 3M 5](#_Toc57759882)

[1.2 Background and Objectives 4M 6](#_Toc57759883)

[1.3 Comparative analysis of state-of-the-art methods 7M 8](#_Toc57759884)

[1.4 Conclusion and Recommendation 6M 8](#_Toc57759885)

[1.5 Presentation 5M 9](#_Toc57759886)

[Bibliography 10](#_Toc57759887)

# List of Figures

[Figure 1‑1 HPE Examples 5](#_Toc57760121)

[Figure 1‑2 HPE Dataset Comparison 9](file:///D:\University-Work\University-Work-SEM-07\Assignment-01-2020\CV\ComputerVision.docx#_Toc57760122)

# Question 1

Solution to Question No. 1

## Executive Summary 3M

The problem taken us for this assignment is that of Human Pose Estimation or HPE, being one of the most challenging computer vision problems with a multitude of applications, human pose estimation has been one of the primary research areas that the computer vision community tried to solve with Deep Learning and Convolutional Neural Networks (CNNs). (Bulat et.al, 2020)

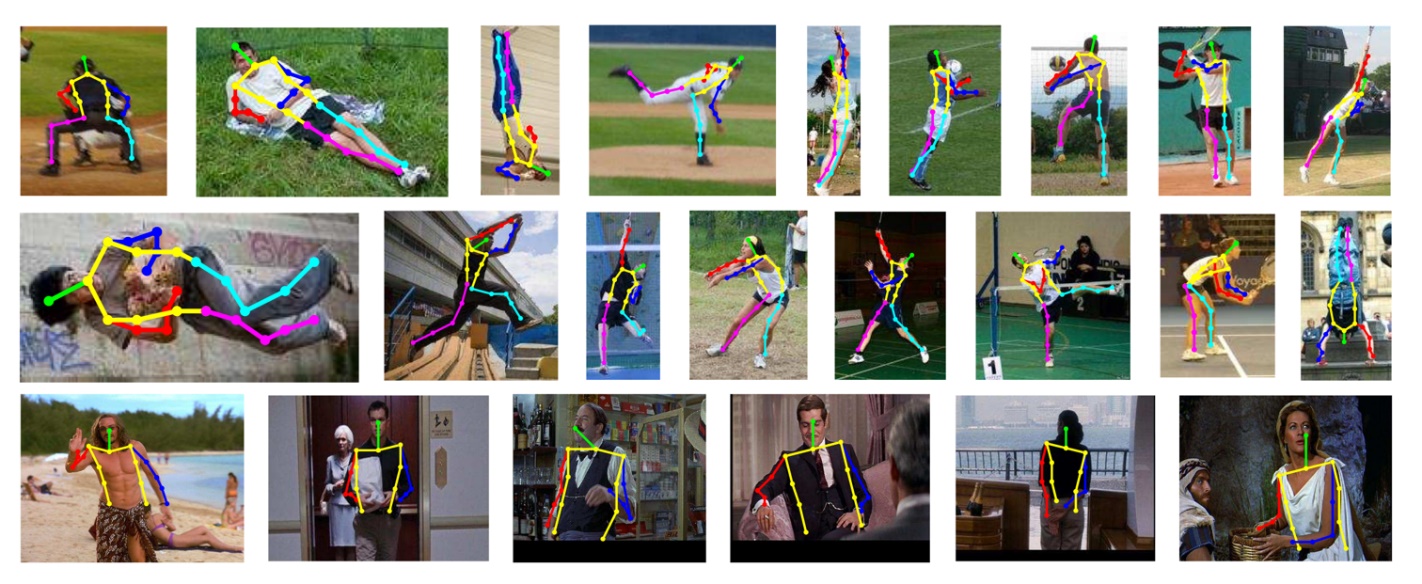
Human Pose Estimation refers to the task of recognizing the human body landmarks (Head, Shoulder, Elbow, Wrist, Hip, Knee, Ankle), from a single monocular image, it can be applied to various applications such as activity recognition, action recognition, human tracking, movies and animations, virtual reality, human-computer interaction, video surveillance, medical assistance, self-driving cars, motion analysis, etc. (Chen et.al, 2020)

Figure ‑ HPE Examples

Monocular Human pose estimation has some unique characteristics and challenges,

* Flexible body configuration indicates complex independent joints and degree-of-freedom limbs, which may cause self-occlusions or rare/complex poses.
* Diverse body appearance includes different clothing and self-similar parts.
* Complex environment may cause foreground occlusion, occlusion or similar parts from nearby persons, various viewing angles, and truncation in the camera view.

In this assignment we are going to go through various categories that there are for HPE and Human Body Models, and compare various methods that have been applied to achieve the state-of-the-art results on some chosen standard datasets. And thus, summarizing the challenges, main frameworks, benchmarks, evaluation metrics, performance comparison, and discuss some promising future research directions.

## Background and Objectives 4M

HPE Methods are broadly classified into these 4 buckets

* Generative and Discriminative (3D Single Person)
* Top Down and Bottom Up (Multi-Person)
* Regression and Detection Based (Single Person)
* One-Stage and Multi-Stage

Now, let’s go through each of them specifically and see the differences between them

1. Generative vs Discriminative

The main difference between generative and discriminative methods is whether a method uses human body models or not. Based on the different representations of human body models, generative methods can be processed in different ways such as prior beliefs about the structure of the body model, geometrically projection from different views to 2D or 3D space, high-dimensional parametric space optimization in regression manners.

Discriminative methods directly learn a mapping from input sources to human pose space (learning-based) or search in existing examples (example-based) without using human body models. Discriminative methods are usually faster than generative methods but may have less robustness for poses never trained with.

1. Top-Down vs Bottom-Up

For multi-person pose estimation, HPE methods can generally be classified as top-down and bottom-up methods according to the starting point of the prediction: high-level abstraction or low-level pixel evidence. Top-down methods start from high-level abstraction to first detect persons and generate the person locations in bounding boxes. Then pose estimation is conducted for each person.

In contrast, bottom-up methods first predict all body parts of every person in the input image and then group them either by human body model fitting or other algorithms. Note that body parts could be joints, limbs, or small template patches depending on different methods. With an increased number of people in an image, the computation cost of top-down methods significantly increases, while keeps stable for bottom-up methods. However, if there are some people with a large overlap, bottom-up methods face challenges to group corresponding body parts.

1. Regression vs Detection

Based on the different problem formulations, deep learning-based human pose estimation methods can be split into regression-based or detection-based methods. The regression-based methods directly map the input image to the coordinates of body joints or the parameters of human body models. The detection-based methods treat the body parts as detection targets based on two widely used representations: image patches and heatmaps of joint locations.

Direct mapping from images to joint coordinates is very difficult since it is a highly nonlinear problem, while small-region representation provides dense pixel information with stronger robustness. Compared to the original image size, the detected results of small-region representation limit the accuracy of the final joint coordinates.

1. One Stage vs Multi Stage

The deep learning-based one-stage methods aim to map the input image to human poses by employing end-to-end networks, while multi-stage methods usually predict human pose in multiple stages and are accompanied by intermediate supervision. For example, some multi-person pose-estimation methods first detect the locations of people and then estimate the human pose for each detected person.

Other 3D human pose estimation methods first predict joint locations in the 2D surface, then extend them to 3D space. The training of one-stage methods is easier than multi-stage methods, but with less intermediate constraints. (Chen et.al, 2020)

And while there are many datasets for HPE,

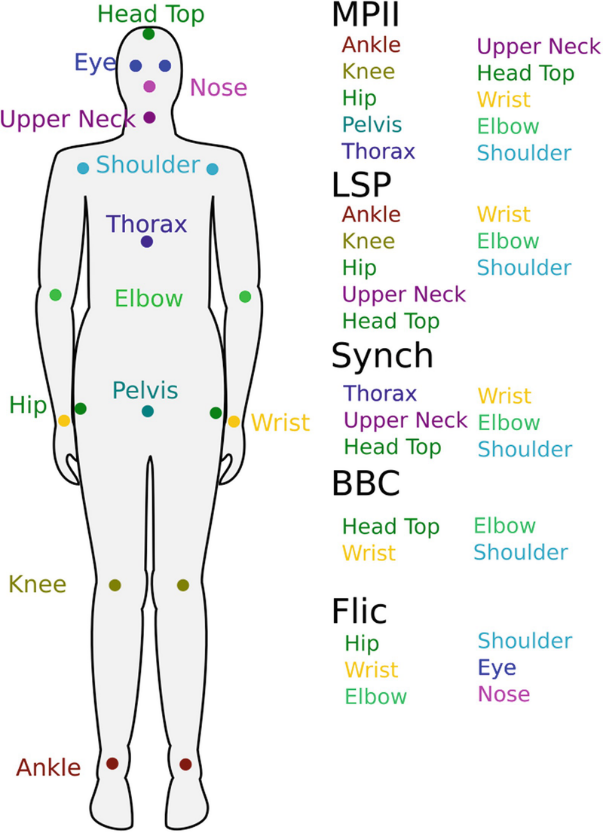
MPII Human Pose dataset is a state-of-the-art benchmark for evaluation of articulated human pose estimation. The dataset includes around 25K images containing over 40K people with annotated body joints. The images were systematically collected using an established taxonomy of every day human activities. Overall, the dataset covers 410 human activities and each image is provided with an activity label. Each image was extracted from a YouTube video and provided with preceding and following un-annotated frames. In addition, for the test set we obtained richer annotations including body part occlusions and 3D torso and head orientations.

Figure ‑ HPE Dataset Comparison

## Comparative analysis of state-of-the-art methods 7M

### Evaluation Metric

1. **Percentage of Correct Keypoints (PCK)** (Yang and Ramanan, 2013) measures the accuracy of the localization of the body joints. A candidate body joint is considered as correct if it falls within the threshold pixels of the ground-truth joint. The threshold can be a fraction of the person bounding box size (Yang and Ramanan, 2013), pixel radius that normalized by the torso height of each test sample (Sapp and Taskar, 2013) (denoted as Percent of Detected Joints (PDJ) in (Toshev and Szegedy, 2014)), 50% of the head segment length of each test image (denoted as PCKh@0.5 in (Andriluka et al., 2014))
2. **The Average Precision (AP)**, For systems in which there are only joint locations but no annotated bounding boxes for human bodies/heads or number of people in the image as ground truth at testing, the detection problem must be addressed as well. Similar to object detection, an Average Precision (AP) evaluation method is proposed, which is first called Average Precision of Keypoints (APK) in (Yang and Ramanan, 2013).

### An Example of Bottom Up Approach



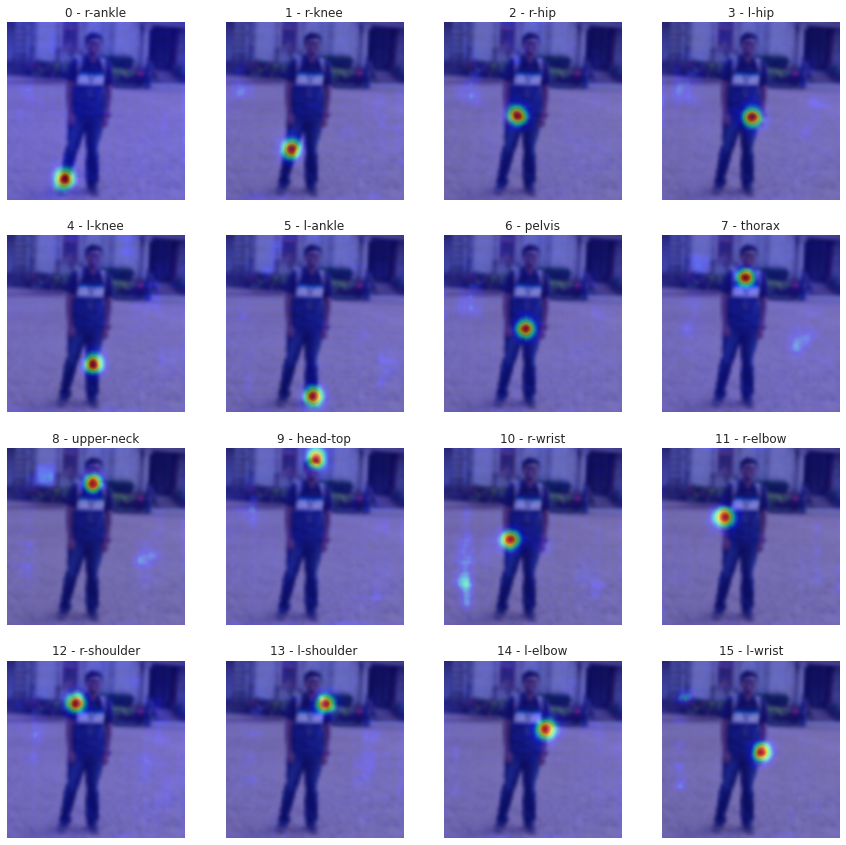


Figure ‑ Heat Map of the various layers of output

|  |  |
| --- | --- |
| Figure ‑ Heat Map of Output | Figure ‑ Connected Joint HPE |

(Satyajit G, 2020)

## Conclusion and Recommendation 6M

Despite much progress in the field, pose estimation remains a challenging and still largely unsolved task. Progress has been made in estimating the configurations of mostly un-occluded and isolated subjects. Open problems include dealing with multiple, potentially interacting people, and tolerance to unexpected occlusions. Future research is also likely to expand on the types of postures and imaging conditions that the current algorithms can handle.

Finally, there is significant evidence suggesting that successfully estimating pose independently at every frame is a very ill-posed problem. Spatio-temporal models that aggregate information over time are emerging as a way to regularize performance obtained in individual frames and smooth out the noise in the estimates. Leveraging all sources of generic prior knowledge, such as spatial layout of the body and temporal consistency of poses, and rich image observation models is critical in advancing the state-of-the-art. (Sigal L, 2014)

## Presentation 5M

# Bibliography

1. Bulat, A., Kossaifi, J., Tzimiropoulos, G. and Pantic, M., 2020. Toward fast and accurate human pose estimation via soft-gated skip connections. arXiv preprint arXiv:2002.11098.
2. Chen, Y., Tian, Y. and He, M., 2020. Monocular human pose estimation: A survey of deep learning-based methods. Computer Vision and Image Understanding, 192, p.102897.
3. Sigal L. (2014) Human Pose Estimation. In: Ikeuchi K. (eds) Computer Vision. Springer, Boston, MA. <https://doi.org/10.1007/978-0-387-31439-6_584>
4. Satyajit G., 2020. Human Pose Estimation and Quantization of PyTorch to ONNX Models - A Detailed Guide. Satyajit Ghana. Available at: https://satyajit.tensorclan.tech/2020/08/pose-estimation-onnx.html [Accessed December 1, 2020].