RAMAIAH UNIVERSITY OF APPLIED SCIENCES

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

Contents iii

List of Figures iv

1 Question 1 5

1.1 Executive Summary 3M 5

1.2 Background and Objectives 4M 6

1.3 Comparative analysis of state-of-the-art methods 7M 9

1.3.1 Evaluation Metric 13

1.3.2 An Example of Bottom Up Approach 14

1.4 Conclusion and Recommendation 6M 16

1.5 Presentation 5M 19

Bibliography 20

List of Figures

Figure 1-1 HPE Examples 5

Figure 1-2 HPE Dataset Comparison 9

Figure 1-3 DeepPose Framework 10

Figure 1-4 Heatmap representaiton of joints 11

Figure 1-5 Example of Hourglass-ResNet model 12

Figure 1-6 UniPose-LSTM architecture for pose estimation in videos 13

Figure 1-7 The input Image to the model 14

Figure 1-8 Illustration of two SOTA networks, and the simple baseline model 14

Figure 1-9 Heat Map of the various layers of output 15

Figure 1-10 Heat Map of Output 16

Figure 1-11 Connected Joint HPE 16

Figure 1-12 OpenPose pipeline 17

Figure 1-13 Human Pose from Aerial Video Model 18

Figure 1-14 Aerial Gait Dataset 18

1 Question 1

Solution to Question No. 1

1.1 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)

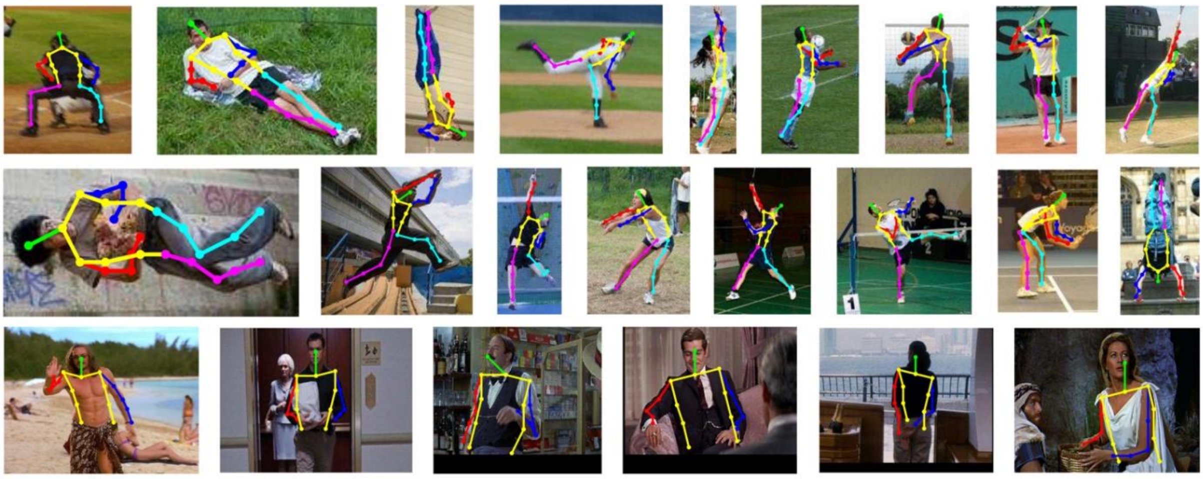


Figure 1-1 HPE Examples

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.

1.2 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.

2. 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.

3. 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.

4. 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,

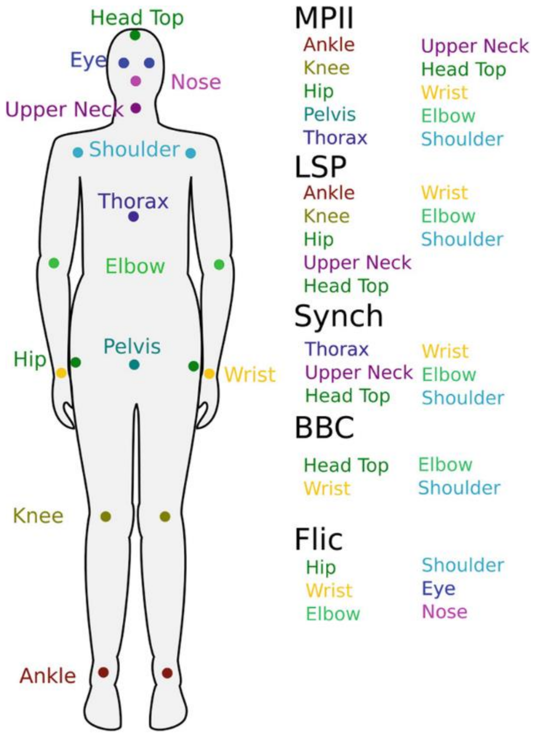


Figure 1-2 HPE Dataset Comparison

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.

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

For the sake of keeping this assignment simple and understandable we'll only compare 2D Single Person Human Pose Estimation papers that have been out in the last 2-3 years. For further reading, Monocular Human Pose Estimation: A Survey of Deep Learning-based Models by Yucheng Chen, Yingli Tian and Mingyi He can be referred.

2D Single Human Pose Estimation is to localize the body join points of a single person in an input image, for images with more than one person has to be cropped for each person and in the pre-processing stage. Based on the different formulations of HPE task, the proposed models using CNN can be majorly classified as Regression based and Detection based.

Regression based methods attempt to map the input human image directly into pose joint coordinates, these coordinates are the actual regressed coordinates that the model has learnt over the training data. While on the other hand detection-based models tend to predict the approximate location of each of the body joints, in a form of 2D Heatmap of probabilities, these are then joined together to for the pose.

Direct regression learning of only one single point is a difficulty since it is a highly nonlinear problem and lacks robustness, while heatmap learning is supervised by dense pixel information which results in better robustness. Compared to the original image size, heatmap representation has much lower resolution due to the pooling operation in CNNs, which limits the accuracy of joint coordinate estimation.

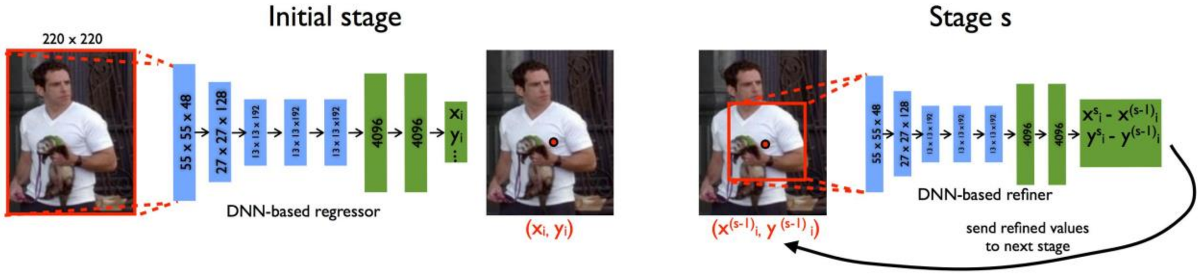


Figure 1-3 DeepPose Framework

Toshev and Szegedy (2014) firstly attempted to train an AlexNet-like deep neural network to learn joint coordinates from full images in a very straightforward manner without using any body model or part detectors as shown in above figure. Moreover, a cascade architecture of multi-stage refining regressors is employed to refine the cropped images from the previous stage and show improved performance. Alex-Net was then also further used on sequence of concatenated frames as input to predict the human pose from videos. The issue here is that regression for HPE is highly non-linear and we require huge models like AlexNet to get some descent results.

Gkioxari et al. (2014) used a R-CNN architecture to detect a person, estimate pose, and classify action, Fan et al. (2015) proposed a dual-source deep CNNs which take image patches and full images as inputs and output a heatmap of each of the joints. As shown in the below image, each of the joint shows a 2D Gaussian distribution centered at the target joint location. Since heatmap representation is more robust than coordinate representation, most of the recent research is based on the heatmap representation.

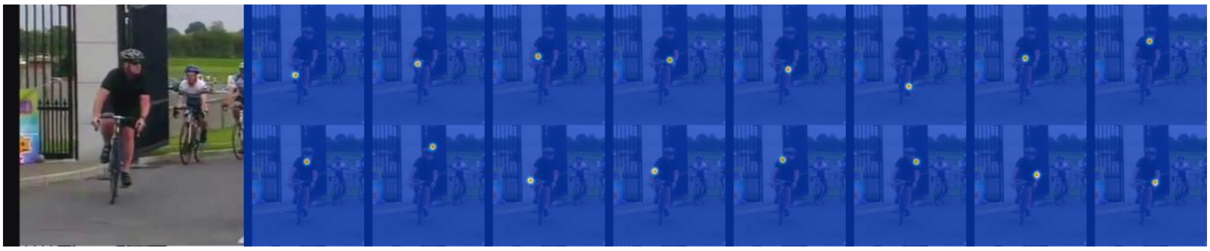


Figure 1-4 Heatmap representaiton of joints

Below is a table describing various papers (Regression-based and Detection-based) that make up for the SOTA models.

Table 1-1 Comparison of various HPE Methods

Methods	Backbone	Input size	Highlights	PCKh (%)
Regression-based				
(Toshev and Szegedy, 2014)	AlexNet	220×220	Direct regression, multi-stage refinement	-
(Carreira et al., 2016)	GoogleNet	224×224	Iterative error feedback refinement from initial pose.	81.3
(Sun et al., 2017)	ResNet-50	224×224	Bone based representation as additional constraint, general for both 2D/3D HPE	86.4
(Luvizon et al., 2017)	Inception-v4+ Hourglass	256×256	Multi-stage architecture, proposed soft-argmax function to convert heatmaps into joint locations	91.2
Detection-based				
(Tompson et al., 2014)	AlexNet	320×240	Heatmap representation, multi-scale input, MRF-like Spatial-Model	79.6
(Yang et al., 2016)	VGG	112×112	Jointly learning DCNNs with deformable mixture of parts models	-
(Newell et al., 2016)	Hourglass	256×256	Proposed stacked Hourglass architecture with intermediate supervision.	90.9
(Wei et al., 2016)	CPM	368×368	Proposed Convolutional Pose Machines (CPM) with intermediate input and supervision, learn spatial correlations among body parts	88.5
(Chu et al., 2017)	Hourglass	256×256	Multi-resolution attention maps from multi-scale features, proposed micro hourglass residual units to increase the receptive field	91.5
(Yang et al., 2017)	Hourglass	256×256	Proposed Pyramid Residual Module (PRM) learns filters for input features with different resolutions	92.0
(Chen et al., 2017)	conv-deconv	256×256	GAN, stacked conv-deconv architecture, multi-task for pose and occlusion, two discriminators for distinguishing whether the pose is 'real' and the confidence is strong	91.9
(Peng et al., 2018)	Hourglass	256×256	GAN, proposed augmentation network to generate data augmentations without looking for more data	91.5

(Ke et al., 2018)	Hourglass	256×256	Improved Hourglass network with multi-scale intermediate supervision, multi-scale feature combination, structure-aware loss and data augmentation of joints masking	92.1
(Tang et al., 2018a)	Hourglass	256×256	Compositional model, hierarchical representation of body parts for intermediate supervision	92.3
(Sun et al., 2019)	HRNet	256×256	high-resolution representations of features across the whole network, multi-scale fusion.	92.3
(Tang and Wu, 2019)	Hourglass	256×256	data-driven joint grouping, proposed part-based branching network (PBN) to learn representations specific to each part group.	92.7
(Zhihui et.al, 2019)	Hourglass-Multistage	256x256	Cascaded ResNet-50 and ResNet-101, and multi stage network	93.9
(Bruno and Andreas, 2020)	Hourglass	256x256	WASP Module, LSTM like architecture for videos	92.7
(Adrian et al, 2020)	Hourglass	256x256	Toward fast and accurate human pose estimation via soft-gated skip connections	94.1

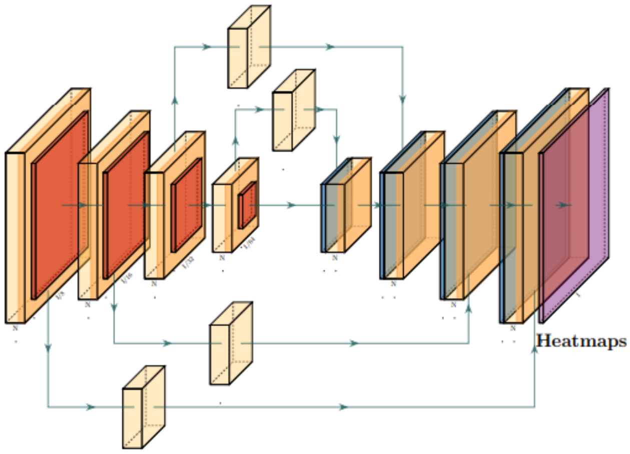


Figure 1-5 Example of Hourglass-ResNet model

The Neural Network architecture style plays a vital role to make better use of the input information to the model. Hourglass models have been proven to be better that doing that, the above table is a very good representation of how the SOTA models have model to the Hourglass models, also along with the introduction of skip networks like in ResNet, SOTA models all use ResNet like architecture as the backbone.

Some papers like that of UniPose: Unified Human Pose Estimation in Single Images and Videos (Bruno and Andreas, 2020) have used ResNet as backbone along with (Waterfall Atrous Spatial Pooling) or WASP modules. WASP is designed with the goal of reducing the number of parameters in order to deal with memory constraints and overcome the main limitation of atrous convolutions. For the pose estimation in videos, the paper proposes Unipose-LSTM, in which the joint heatmaps from their decoder network is fed to a LSTM along with the final heatmaps from the previous LSTM state. The convolution layers following the LSTM recognize the outputs into the final heatmaps used for joint localization.

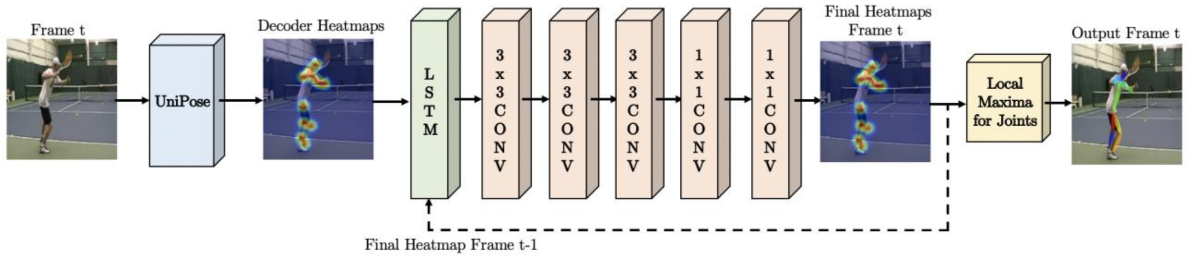


Figure 1-6 UniPose-LSTM architecture for pose estimation in videos.

1.3.1 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))

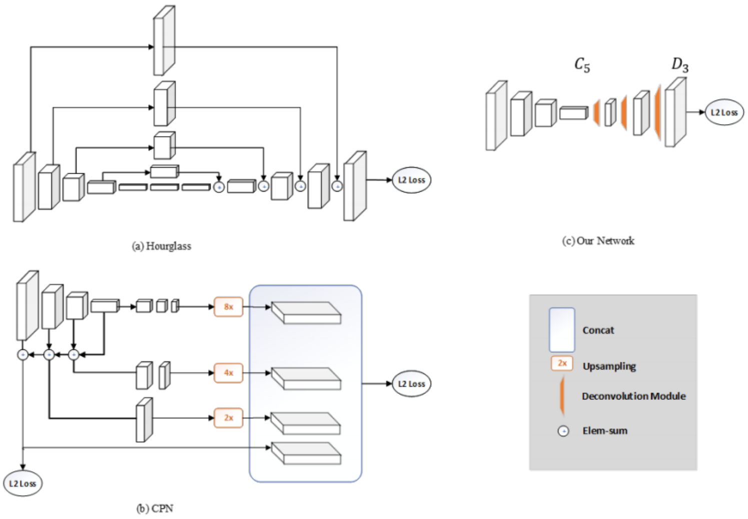
2. **The Average Precision (AP)**, For systems in which there are only joint locations but no annotated bounding boxes for human bodies/heads or number of people in the image as ground truth at testing, the detection problem must be addressed as well. Similar to object detection, an Average Precision (AP) evaluation method is proposed, which is first called Average Precision of Keypoints (APK) in (Yang and Ramanan, 2013).

1.3.2 An Example of Bottom Up Approach

In the example we are going to see the results from Simple Baselines for Human Pose Estimation and Tracking (Bin Xiao, Haiping Wu, and Yichen Wei, 2018), the paper aims to provide a simple baseline model since the network architecture and experiment practice have steadily become more complex. This makes the algorithm analysis and comparison more difficult.

Their method simply adds a deconvolution network on top of an existing ResNet Head. This architecture is arguably the simplest to generate heatmaps from deep and low-resolution features and also adopted in the state-of-the-art Mask R-CNN.

Figure 1-8 Illustration of two SOTA networks, and the simple baseline model



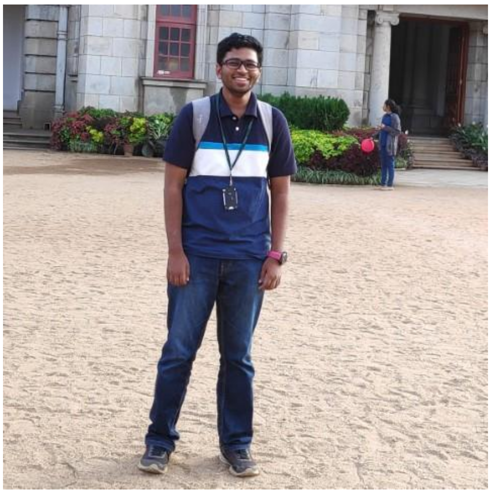


Figure 1-7 The input Image to the model

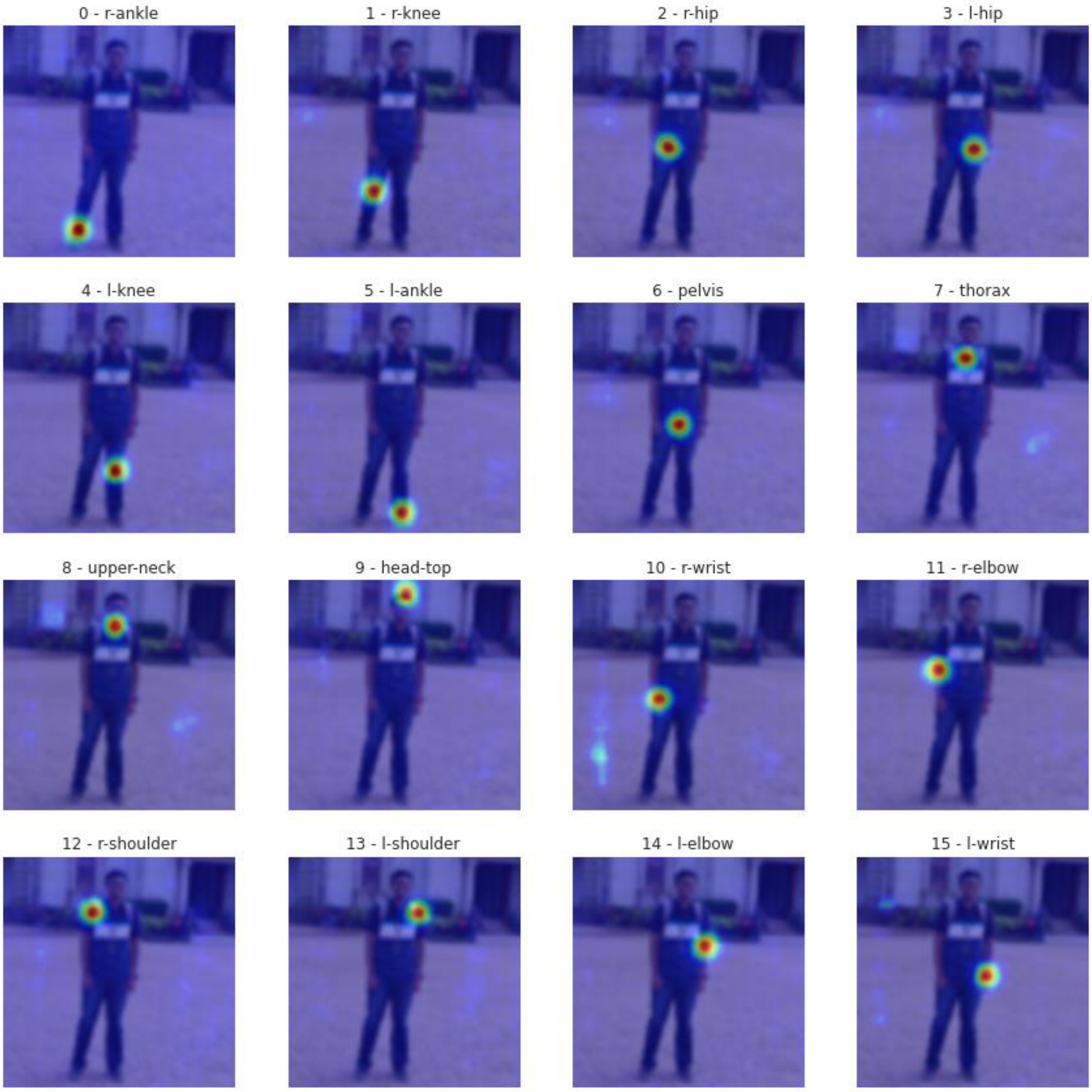


Figure 1-9 Heat Map of the various layers of output

The Model outputs 16 layers after the DeConv Layers, which are basically the 2D Gaussian plots of the specific joints, since the dataset it was trained on was MPII, we have 16 such layers. The 2D Gaussian center can be computed for each of the heatmap to get the join coordinate, these coordinates are then stitched together to form the estimated human pose.

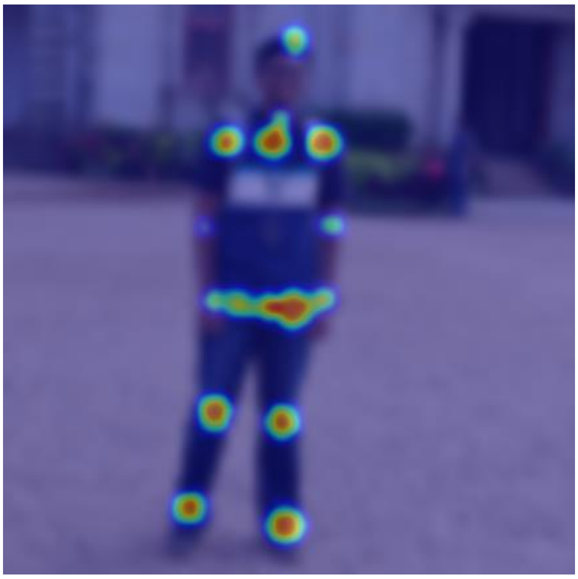


Figure 1-10 Heat Map of Output

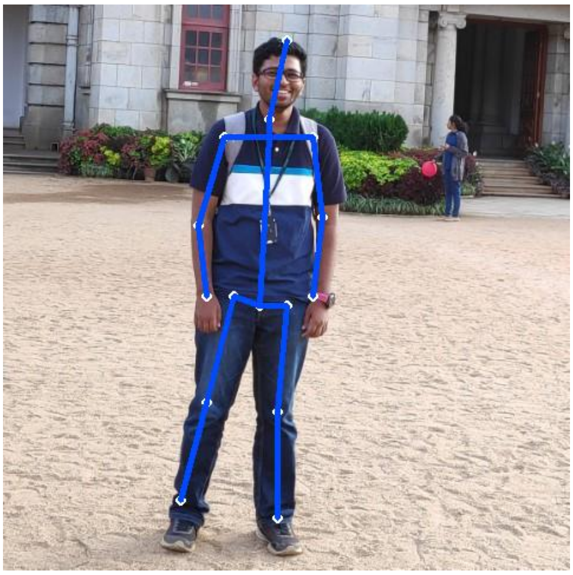


Figure 1-11 Connected Joint HPE

(Satyajit G, 2020)

1.4 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)

Future networks should explore both global and local contexts for more discriminative features of the human body while exploiting human body structures into the network for prior constraints. Current networks have validated some effective network design tricks such as multi-stage structure, intermediate supervision, multi-scale feature fusion, multi-task learning, body structure constrains. Network efficiency is also a very important factor to apply algorithms in real-life applications.

Another issue that wasn't addressed in this assignment was that of multi-human pose detection, to address this, (Cao, et.al, 2017, 2018, 2019) came up with OpenPose, a framework that uses PAF (Part Affinity Fields) for Multi-Person 2D Pose Estimation.

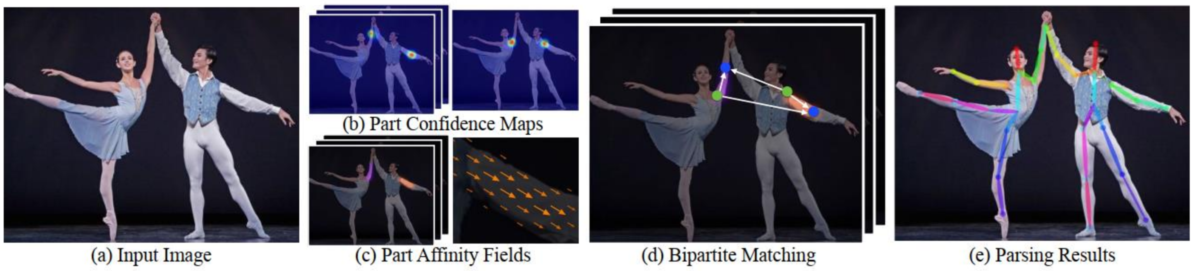


Figure 1-12 OpenPose pipeline

PAF is a set of 2D vector fields that encode the location and orientation of limbs over the image domain. Simultaneously inferring these bottom-up representations of detection and association encode global context sufficiently well to allow a greedy parse to achieve high-quality results, at a fraction of the computational cost.

Diversity data can improve the robustness of networks to handle complex scenes with irregular poses, occluded body limbs and crowded people. Data collection for specific complex scenarios is an option and there are other ways to extend existing datasets. Synthetic technology can theoretically generate unlimited data while there is a domain gap between synthetic data and real data. Cross-dataset supplementation, especially to supplement 3D datasets with 2D datasets can mitigate the problem of insufficient diversity of training data.

Many Datasets have come up recently, but they don't show up in SOTA papers, since SOTA papers need to compare their benchmarks with SOTA datasets, although this still is a good method to compare models and papers, in a real-world scenario we still need to use a real-world diverse dataset.

A good example for this is the Aerial Gait Dataset, which aim was to consider the problem of estimating human pose and trajectory by an aerial robot with a monocular camera in near real-time. This dataset accounts for the natural twists and self-occlusions of a turning human body and minimizes the false positives caused by minor variations in heading.

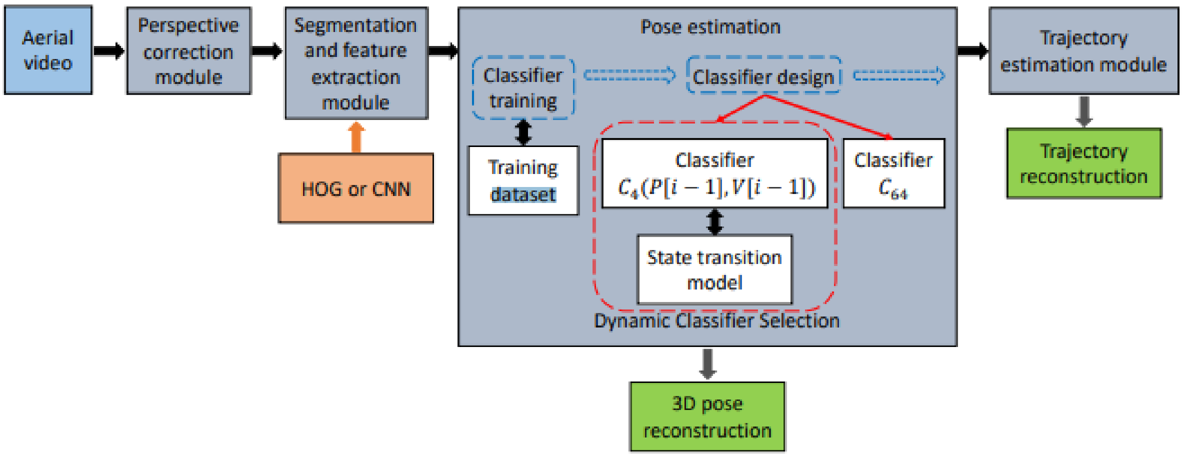

Aerial video
Perspective correction module
Segmentation and feature extraction module
HOG or CNN
Pose estimation
Classifier training
Classifier design
Training dataset
Classifier $C_4(P[i-1], V[i-1])$
Classifier C_{64}
State transition model
Dynamic Classifier Selection
3D pose reconstruction
Trajectory estimation module
Trajectory reconstruction

Figure 1-13 Human Pose from Aerial Video Model

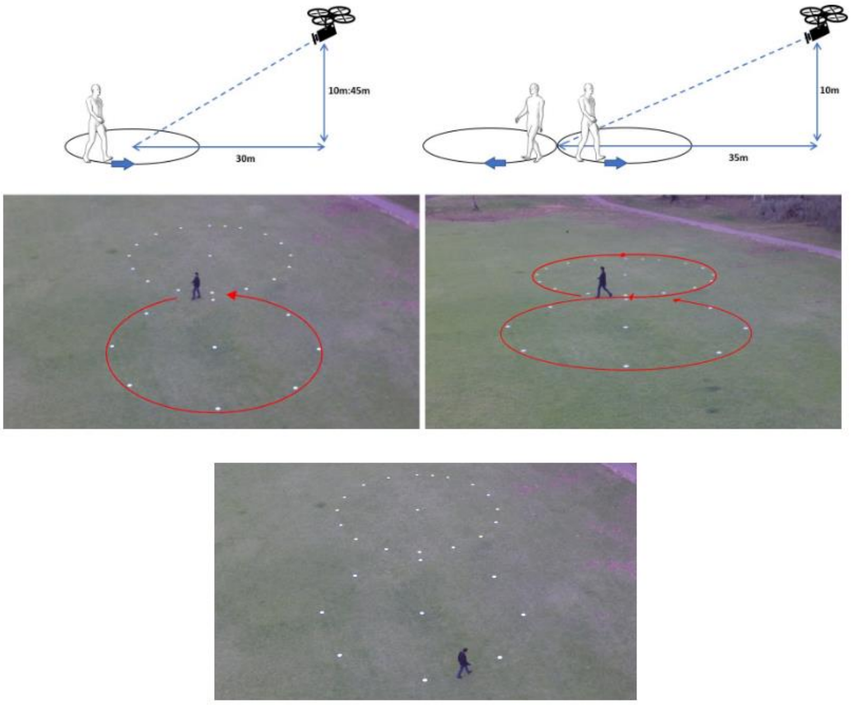


Figure 1-14 Aerial Gait Dataset

1.5 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].
5. Xiao, B., Wu, H. and Wei, Y., 2018. Simple baselines for human pose estimation and tracking. In Proceedings of the European conference on computer vision (ECCV) (pp. 466-481).
6. Cao, Z., Hidalgo, G., Simon, T., Wei, S.E. and Sheikh, Y., 2018. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. arXiv preprint arXiv:1812.08008.
7. Kumar, S. A., Yaghoubi, E., Das, A., Harish, B. S., & Proença, H. (2020). The P-DESTRE: A Fully Annotated Dataset for Pedestrian Detection, Tracking and Short/Long-term Re-Identification from Aerial Devices. IEEE Transactions on Information Forensics and Security.
8. Zhu, P., Wen, L., Du, D., Bian, X., Hu, Q. and Ling, H., 2020. Vision Meets Drones: Past, Present and Future. arXiv preprint arXiv:2001.06303.