Single Person 2-D Pose Estimation from RGB images using Convolutional Neural Networks

Akhil Bommadevara
abommad@ncsu.edu

Ishan Bhatt
ivbhatt@ncsu.edu

Aishwarya Seth
aseth@ncsu.edu

I. MOTIVATION.

In human pose estimation (HPE), we try to develop systems to localize body key points to recognize the posture of an individual in an image. In the general sense, an HPE system may be expected to report postures of multiple people (in a single frame), work with different imaging techniques (such as 3D depth-sensing cameras) and deal with occlusions as they may arise in the real world. However, for this project, we propose to develop an HPE that:

- Takes RGB images as input.
- Assumes each image to have a subject person without significant occlusions.
- Detects the visible key points of the subject person.

HPE systems have applications in areas such as activity recognition, VR gaming, physical therapy/rehabilitation monitoring, person-tracking, animation, posture improvement, assisted living, etc. [1][2][3].

Fig 1: *Sample output (image for illustration only)*

As shown in Figure 1, our system will be able to locate key points from a picture of a person. Using those key points, we will be able to trace the pose of the person.

The main challenges we might face in our project include:

- Occluded body parts (We will remove such cases from the dataset.)
- Truncated body parts (System should be able to handle cases where only a part of the subject person is visible in the picture)
- Other people in the background (If other background people are sufficiently blurred, a robust HPE system should be able to filter them out)

II. DATA

For HPE, the popular datasets are listed below:

1) *MS COCO -2020 dataset* [4]: This dataset contains approximately 150K (out of ~330K) images containing only one person along with annotated keypoints.[cite how to use coco for HPE blog] In this dataset, curators have annotated 19 key points of the human body.
2) *MPII Human Pose dataset* [5] : This dataset contains approximately 25K images of about 40K people. This dataset contains images containing both single and multiple persons in a single frame. However, curators have provided 24K bounding boxes having sufficiently separated persons[6]. In this dataset, curators have annotated 15 key points in the human body.

We plan to use a combination of datasets 1 and 2 for training and testing our model. We do not want to include other popular datasets such as FLIC[7] and LSP-extended[8] due to their questionable annotation quality (annotated on Amazon's Mechanical Turk). We do not include LSP[9] because of its small size(only 2K images)

III. METHODOLOGY

For a Single Person 2-D HPEs, there are two competent state-of-the-art approaches we plan to compare.

1. *Convolutional Pose Machines (CPMs)* [10]
2. *Stacked Hourglass Networks (SHNs)* [11]

Both the approaches use heatmaps to localize the body key points. Neither of them (nor any other architectures) make use of regression to predict the coordinates of the key points. For both of the approaches, the basic building block resembles a U-Net [12]. CPMs use a large, robust building-block while SHNs stack multiple smaller units to achieve the same goal.

IV. EVALUATION

The main evaluation metric for MPII dataset is Percentage of Correct Key-points (PCK). Additionally, we use mean normalised distance for sanity checks. "mAP" is the most popular rubric to report performance of CV models and we use it to compare our performance with the benchmark models. Another popular evaluation metric is Area Under Curve(AUC) of Precision-Recall graphs which can help visualize model performance graphically.[13]

REFERENCES

[1] A. Yao, et. al., ''Coupled action recognition and pose estimation from multiple views,'' Int. J. Comput. Vis., Oct 2012
[2] M. B. Holte, et. al., ''Human pose estimation and activity recognition from multi-view videos: Comparative explorations of recent developments,'' IEEE J. Sel. Topics Signal Process., Sep 2012
[3] Spagnolo and Paolo, ''Proceedings IEEE conference on advanced video and signal based surveillance. AVSS 2003,'' in Proc. IEEE Conf. Adv. Video Signal Based Surveill., Jul 2003
[4] M. Faber, "How to analyze the COCO dataset for pose estimation", Towards Data Science, Dec 2020
[5] M. Andriluka, et. al., "2D Human Pose Estimation: New Benchmark and State of the Art Analysis", CVPR, 2014
[6] Documentation available at: https://dbcollection.readthedocs.io/en/latest/datasets/mpii_pose.html
[7] Documentation available at: https://www.tensorflow.org/datasets/catalog/flic
[8] Documentation available at: https://dbcollection.readthedocs.io/en/latest/datasets/leeds_sports_pose_extended.html
[9] Documentation available at: https://dbcollection.readthedocs.io/en/latest/datasets/leeds_sports_pose.html
[10] S.-E. Wei, et. al.. "Convolutional pose machines" in CVPR, 2016
[11] A. Newell, 'Stacked Hourglass Networks for Human Pose Estimation", arxiv, July 2016
[12] O. Ronneberger, et. al. "U-Net: Convolutional Networks for Biomedical Image Segmentation" arXiv, May 2015
[13] T. L. Munea, et. al., "The Progress of Human Pose Estimation: A Survey and Taxonomy of Models Applied in 2D Human Pose Estimation", IEEE Access, July 20, 2020