2020 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)

HuTrain: a Framework for Fast Creation of Real Human Pose Datasets

Ricardo R. Barioni*
Voxar Labs, UFPE

Willams L. Costa*
Voxar Labs, UFPE

José A. C. Neto*
Voxar Labs, UFPE

Lucas S. Figueiredo*
Voxar Labs, UFPE

Veronica Teichrieb*
Voxar Labs, UFPE

Jonysberg P. Quintino*
P&D CIn/Samsung, UFPE

Fabio Q. B. da Silva*
Centro de Informática, UFPE

André L. M. Santos*
Centro de Informática, UFPE

Helder Pinho†
SiDi, Campinas

Abstract

Image-based body tracking algorithms are useful in several scenarios, such as avatar animations and gesture interaction for VR applications. In the last few years, the best-ranked solutions presented on the state of the art of body tracking (according to the most popular datasets in the field) are intensively based on Convolutional Neural Networks (CNNs) algorithms and use large datasets for training and validation. Although these solutions achieve high precision scores while evaluated with some of these datasets, there are particular tracking challenges (for example, upside-down cases) that are not well-modeled and, therefore, not correctly tracked. Instead of lurking an all-in-one solution for all cases, we propose HuTrain, a framework for creating datasets quickly and easily. HuTrain comprises a series of steps, including automatic camera calibration, refined human pose estimation, and known dataset formats conversion. We show that, with our system, the user can generate human pose datasets, targeting specific tracking challenges for the desired application context, with no need to annotate human pose instances manually.

Index Terms: Motion capture—Camera calibration—Reconstruction—Image processing

1 Introduction

Human Pose Estimation (HPE) is the task of localizing a set of human body keypoints that matches with the spatial arrangement of the individuals presented in an image. Many applications require this computer vision-based task, such as activity recognition, model animation, and natural interaction by performing in-air gestures. This problem leverages several challenges, such as the uncertainty of the number of people, occlusions, illumination, or the number of possible poses people can assume.

Existing HPE benchmarks vary in respect to which challenge they are supposed to aim: some cover as many scenarios as possible [1,7], while others focus on sports tasks [6], in-the-wild activities [8] or even people wearing head-mounted displays (HMDs) [9]. Although most cases are well-covered by the available HPE datasets, there are still scenarios that the techniques fail to address due to the lack of fed samples during the training step [2]. For that, the solution relies on creating new training samples. Nonetheless, the process of conceiving an HPE dataset may raise some obstacles.

For HPE benchmarks assembled by manual annotations from humans, the desired number of samples defines the financial [8] and time costs [6,7]. Also, humans may have different understandings regarding the precise location of body parts, which can leverage ambiguities [7]; annotations must be carefully analyzed so that the desired quality is maintained.

Concerning automatically generated HPE datasets, these have overcome the laborious and error-prone manual annotation process. Some assemble a controlled environment by using multiple cameras in an outside-in fashion [4], producing pose estimations with higher quality. In this case, the accuracy of the algorithm defines the quality of the conceived dataset. Although many works were devoted to forging new benchmarks concerning a series of individual HPE scenarios, standard users cannot generate datasets automatically according to the specific context of HPE usage they want to address, due to its challenging technical nature.

Another approach is to generate synthetic HPE benchmarks. In this case, a dataset generator receives motion capture recordings and takes care of creating new HPE samples by generating photorealistic images that describe the received data [9]. Even though these approaches provide sufficient realism, they do not tackle the task of conceiving samples of unseen human poses.

Although many benchmarks meeting the most varied HPE scenarios exist, there will still be circumstances where the user needs to create its own, which will, in short, demand time, money and prior technical knowledge. In this work, we propose HuTrain, a semi-automatic framework that aims to help the generation of 3D HPE datasets by covering the camera calibration, pose estimation, and annotation of samples from synchronized recordings performed by the user, alongside an easy and fast way to perform the setup with only RGB cameras. We also propose an automatic Human-Based Camera Poses Calibration (H-BCPC) algorithm, which calculates the adopted cameras' poses by using an off-the-shelf HPE method output as the image feature extractor for scenarios with sparse cameras and low-quality feature descriptors.

2 Method

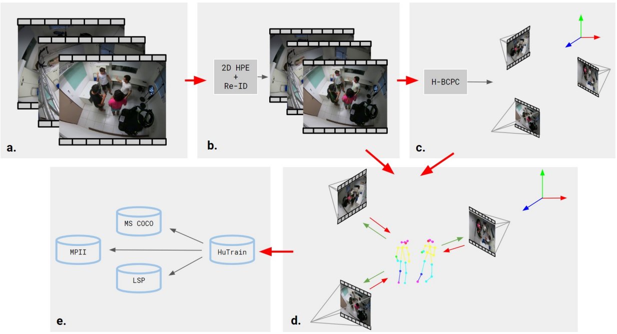

Figure 1: Pipeline of our framework. **a.** Input RGB videos; **b.** Initial 2D human pose estimation followed by cross-view person re-identification; **c.** Human-based camera poses calibration; **d.** 3D human pose triangulation (red arrows) and refined 2D human pose estimation (green arrows); **e.** Conversion to known datasets format.

We firstly perform 2D human pose estimation of each video independently with the OpenPose method [2], followed by the re-identification of people between different cameras with the MVPose approach [3]. After that, we estimate each cameras' pose with an adaptation of Structure from Motion pipeline by using the previously

*{rrb, wlc2, jacn, lsf, vt, jpq, fabio, alms}@cin.ufpe.br
†helder.p@sidi.org.br

978-1-7281-7675-8/20/$31.00 ©2020 IEEE
DOI 10.1109/ISMAR-Adjunct51615.2020.00031

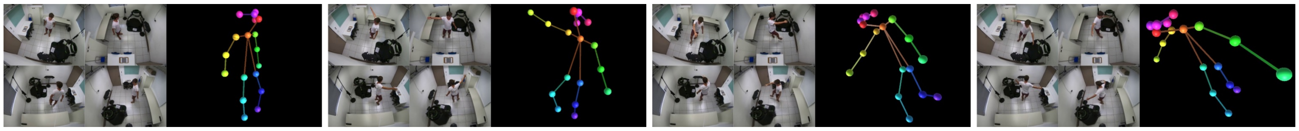

Figure 2: Qualitative results of the generated dataset. Although the body hides some of the keypoints in a camera view, the skeleton is still consistent because of other cameras.

estimated keypoints as the features correspondence among cameras. In this case, the keypoints with the best confidence values globally (considering its occurrences at all points of view) are chosen as the best features, in order to obtain the relative poses between cameras in scenarios where the baseline between points of view are sparse.

With that, it is possible to estimate the 3D human poses from multi-view information with an algebraic triangulation approach [5]. In this approach, the triangulation is performed by valuing its 2D occurrences whose confidence values (both location [2] and re-identification [3]) are higher. At last, we convert the human poses to the formats of some of the most used human pose datasets [1, 6, 7], so that it can be integrated to a training pipeline as transparently as possible.

3 Experiments and Results

To evaluate our framework, we generated sequences by positioning 8 top-down GoPro Hero cameras around a 3,72m x 3,27 room and pointing to the center. We recorded 68 seconds of a human randomly performing actions, at 60 frames per second (fps). In this test, the recorded human performs rotations, which address self-occluding scenarios. The generated videos were manually synchronized with a video editing software and then fed to the HuTrain framework.

3.1 Runtime Comparison

We used a 16GB RAM laptop with one NVIDIA GeForce GTX-1060 GPU. The videos' size is 1920×1440 (in pixels). We compare the elapsed time to generate our dataset with other existing benchmarks, as shown in Table 1.

Table 1: Time cost of the generation of existing HPE datasets, including ours.

Dataset	# Samples	Total Time Cost	Time Cost per Sample
MS COCO	± 250,000	+ 70,000h	± 16min 48s
LSP	2,000	20h	36s
LSP Extended	10,000	25h	9s
H3D	2,000	± 166h	5min
MPII Cooking	1,277	± 21h	± 1min
xR-EgoPose	383,000	± 25500h	4min
SURREAL	6,500,000	± 3600 - 9000h	2 - 5s
HuTrain	**32,488**	**1h 43min 10sec**	**0.19s**

3.2 Quantitative Results

We quantitatively evaluated our results by manually annotating 100 random frames from arbitrary views of the dataset, calculating the mean Euclidean distance error of the estimated 3D keypoints when projected into the cameras' view, and comparing it with only using the initially adopted off-the-shelf 2D HPE [2] (Table 2).

4 Conclusion

We developed a framework to facilitate the process of generating human pose datasets to improve the quality of HPE approaches for specific scenarios of use. Our key contribution is the capability of fastly creating a HPE dataset without prior technical knowledge needed for creating one. Also, we proposed a camera calibration algorithm for scenarios with sparse cameras and poor in feature descriptors. Our method shows promising results in self-occlusion scenarios.

Table 2: Mean Euclidean distance error (in pixels) for each body part, of our framework and the method of [2].

Body Part	[2]	Ours
Nose	309.0	**10.2**
Neck	**26.5**	27.2
Shoulders	14.7	**13.7**
Elbows	27.1	**16.8**
Wrists	96.7	**48.0**
Hips	**24.5**	26.1
Knees	33.2	**18.6**
Ankles	25.7	**14.0**
Eyes	442.8	**12.1**
Ears	300.2	**11.9**

Acknowledgments

The authors would like to thank Maria Euzebio for the support related to the equipment used in the project and Arlindo Gomes for helping with the visual design.

References

[1] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, pp. 3686–3693, 2014.

[2] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7291–7299, 2017.

[3] J. Dong, W. Jiang, Q. Huang, H. Bao, and X. Zhou. Fast and robust multi-person 3d pose estimation from multiple views. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7792–7801, 2019.

[4] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2014.

[5] K. Iskakov, E. Burkov, V. Lempitsky, and Y. Malkov. Learnable triangulation of human pose. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 7718–7727, 2019.

[6] S. Johnson and M. Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *BMVC*, vol. 2, p. 5, 2010.

[7] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshik, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

[8] B. Sapp and B. Taskar. Modec: Multimodal decomposable models for human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3674–3681, 2013.

[9] D. Tome, P. Peluse, L. Agapito, and H. Badino. xr-egopose: Egocentric 3d human pose from an hmd camera. *arXiv preprint arXiv:1907.10045*, 2019.