Human Pose Estimation

Matteo Anedda, Christoph Schröter, Masoud Taghikhah

***Abstract*—later**

I. Introduction

II. Benchmarks

Huge data sources are part of the benchmarking process for human posture estimation approaches. The datasets used specifically for this purpose contain images showing one or more individuals in different poses, as well as other information about joint and limb positions. This information is acquired by motion capture using markers on the body or individual *IMU* units attached to the body to determine the position regardless of obstacles, can be used as well. Existing video footage can also be manually annotated. Some datasets focus on different features within its content to ensure the quality of a model in relation to that aspect. Commonly used datasets are explained in more detail in this section.

A. Max Planck Institute datasets

Perceiving Systems[1] is a department of the *Max Planck Institute for Intelligent Systems*[2] that is specialized in computer vision and, in addition to scientific publications, also provides datasets[3] for e.g. pose estimation approaches. The listed datasets can be subdivided according to the following aspects:

1) Clothing extension: The individuality of a person is expressed by his clothing. To account for this feature, a model called *CAPE* is built from 4d posture sequences of 8 men and 3 women. This dataset consists of about 80000 frames. [7] Further work by Qianli et al. has generated a synthetic dataset on specific *CAPE* subjects and published it as *ReSynth* for researchers [8].

2) Full-body scans: Acquiring 3d scans and data from multiple people in an outdoor environment is challenging because the markers are difficult to track. Timo et al. have shown in their publication that capturing sufficient data in a scene is possible with 6-17 *IMU* units attached to each person, combined with a single hand-held camera. The recorded 51000 images are available for research. [15] A similar approach is followed by Yinghao et al. with 17 *IMU* units for 10 subjects in 64 sequences, resulting in 330000 time instances [6]. Human-environment interaction is mainly covered in the datasets of Mohamed et al. which consist of three parts in different scenes [3]. The *GRAB* dataset, on the other hand, targets the relationship between full-body models and object manipulation. It contains motion data of 10 individuals interacting with 51 objects in 4 different contexts, e.g., lifting, transferring, hand-to-hand transfer, and using [12]. In contrast to Grab, the Lea et al. datasets include all of a person's interactions with themselves [10].

3) Hand scans: The hand contributes to communication, e.g., the hand gesture is used to confirm a statement in conversation. To incorporate this expressiveness into existing full-body models, Javier et al. developed the *MANO* model from approximately 1000 3d scans of 31 subjects in 51 poses. These scans showed female and male hands, both left and right, interacting with primitives. [11] Yana et al. also published a synthetically generated hand dataset *obman* that focuses on the manipulation of grasped primitives [4].

4) synthetic data: A much more cost-effective approach is to create realistic body data from existing motion capture sources. An prominent example is SURREAL by Gül et al. which consists of 6 million frames [14]. David T. et al. also published their data set with pure synthetic and more realistic mixed material [5].

5) Generalization of datasets: Many different 3d scans are based on markers and motion capture software. Unfortunately, the number of markers varies from dataset to dataset, so their use as a data source for a body model leads to inaccuracies and further adjustments. A common solution to this problem is provided by the *MoSh++* algorithm, a descendant of the earlier motion capture software, and its resulting *AMASS* dataset. It consists of 11265 motions from 344 subjects with 40 hours of content. [9] [1]

III. Criteria

The quality of the estimated poses and thus of the applied algorithm is evaluated with the help of metrics. The common approach is to calculate the body parts or joint positions and compare them with particular values from the ground truth data, e.g. the lengths between key-points or frames in a specific region. For this analysis, the previously mentioned datasets from section II are used as reference. The accuracy of the results can be adjusted by a threshold value for the respective metric.

A. Percentage of Correct Parts PCP

This criterion encompasses a comparison of the recognized and recognizable body parts. The definition of a correctly recognized limb includes both the distances l_1, l_2 of its endpoints from those contained in the dataset and its total length L. Figure 1 illustrates the values explained earlier in an intuitive way, with the left (green) line indicating information from the dataset that is compared to the estimated right line. Another factor p multiplied by L defines the threshold value to which

[1] https://ps.is.mpg.de/
[2] https://is.mpg.de/
[3] https://ps.is.mpg.de/research_fields/datasets-and-code

l_1 and l_2 are compared. If l_1 or l_2 exceed the threshold, the body part is not detected correctly, resulting in a lower PCP score. The smaller p, the stricter the evaluation and thus higher the accuracy. [2]

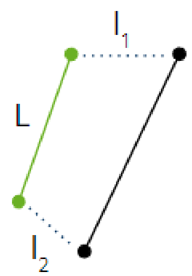

Fig. 1. Visualisation of estimated values for PCP calculation

B. Percentage of Detected Joints PDJ

The metric *PCJ* addressed in this section is similar to the already established *PCP* methode. While *PCP* depends on individual limb lengths, *PCJ* uses the torso length T as a global reference, so that a body part is correctly detected if both l_1 and l_2 do not surpass the threshold given by T and a factor p. [13]

C. Percentage of Correct Key-points PCK

For this particular criteria, the maximum bounding box length B must be calculated from the existing dataset information. The *PCK* score is calculated analogous to the previously mentioned algorithms in subsection III-A and subsection III-B from l_1, l_2 and a further threshold defined by $B \times p$. [1], [16]

1) Head-normalized Probability of Correct Key-points PCKh: In this variant of *PCK*, the threshold is chosen depending on the size of the individuals. It should be between a fraction of the reference segment determined by a constant $\alpha \in \mathbb{R}$ multiplied by the head size, which is 60% of the diagonal length of the head frame

D. Area Under the Curve AUC

The measurement of the *PCK* under variation of the threshold value results in a curve. This analysis provides information on how the model is able to distinguish the individual joints of the body. A large curve area defines a qualitative model

E. Object Key-point Similarity OKS

Equation 1 illustrates the simplified OKS score[4], which is a sum of $n \in \mathbb{N}$ detected and ground truth key-points. The parameters $d \in \mathbb{R}$ are the Euclidean distance between the corresponding ground truth value and the detected key-point, while $s \in \mathbb{R}$ denotes the object segment area and $k \in \mathbb{R}$ is a constant describing a falloff.

$$OKS = \sum_{i=1}^{n} e^{-\frac{d_i^2}{2 \times s^2 \times k_i^2}} \tag{1}$$

[4]https://cocodataset.org/#keypoints-eval

Optimal predictions have a high OKS value, while low values indicate poor predictions.

IV. 2D Pose Estimation

V. 3D Pose Estimation

VI. Conclusion

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[2] Marcin Eichner, Manuel Marin-Jimenez, Andrew Zisserman, and Vittorio Ferrari. 2d articulated human pose estimation and retrieval in (almost) unconstrained still images. *International journal of computer vision*, 99(2):190–214, 2012.

[3] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J. Black. Resolving 3D human pose ambiguities with 3D scene constraints. In *Proceedings International Conference on Computer Vision*, pages 2282–2292. IEEE, October 2019.

[4] Yana Hasson, Gül Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 11807–11816, June 2019.

[5] David T. Hoffmann, Dimitrios Tzionas, Michael J. Black, and Siyu Tang. Learning to train with synthetic humans. In *German Conference on Pattern Recognition (GCPR)*, pages 609–623. Springer International Publishing, September 2019.

[6] Yinghao Huang, Manuel Kaufmann, Emre Aksan, Michael J. Black, Otmar Hilliges, and Gerard Pons-Moll. Deep inertial poser: Learning to reconstruct human pose from sparse inertial measurements in real time. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 37:185:1–185:15, November 2018. Two first authors contributed equally.

[7] Qianli Ma, Jinlong Yang, Anurag Ranjan, Sergi Pujades, Gerard Pons-Moll, Siyu Tang, and Michael J. Black. Learning to dress 3D people in generative clothing. In *Computer Vision and Pattern Recognition (CVPR)*, pages 6468–6477. IEEE, June 2020.

[8] Qianli Ma, Jinlong Yang, Siyu Tang, and Michael J. Black. The power of points for modeling humans in clothing. In *Proc. International Conference on Computer Vision (ICCV)*, pages 10974–10984, October 2021.

[9] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *Proceedings International Conference on Computer Vision*, pages 5442–5451. IEEE, October 2019.

[10] Lea Müller, Ahmed A. A. Osman, Siyu Tang, Chun-Hao P. Huang, and Michael J. Black. On self-contact and human pose. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 9990–9999, June 2021.

[11] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6):245:1–245:17, November 2017.

[12] Omid Taheri, Nima Ghorbani, Michael J. Black, and Dimitrios Tzionas. Grab: A dataset of whole-body human grasping of objects. In *Computer Vision – ECCV 2020*, volume LNCS 12355, pages 581–600, Cham, August 2020. Springer International Publishing.

[13] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[14] Gül Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2017*, pages 4627–4635, Piscataway, NJ, USA, July 2017. IEEE.

[15] Timo von Marcard, Roberto Henschel, Michael J. Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3D human pose in the wild using IMUs and a moving camera. In *European Conference on Computer Vision (ECCV)*, volume Lecture Notes in Computer Science, vol 11214, pages 614–631. Springer, Cham, September 2018.

[16] Yi Yang and Deva Ramanan. Articulated human detection with flexible mixtures of parts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):2878–2890, 2013.