Supplementary Materials for the ICCV 2017 Paper: Benchmarking and Error Diagnosis in Multi-Instance Pose Estimation

Matteo Ruggero Ronchi
www.vision.caltech.edu/~mronchi

Pietro Perona
perona@caltech.edu

California Institute of Technology, Pasadena, CA, USA

Abstract

We propose a new method to analyze the impact of errors in algorithms for multi-instance pose estimation and a principled benchmark that can be used to compare them. We define and characterize three classes of errors - localization, scoring, and background - study how they are influenced by instance attributes and their impact on an algorithm's performance. Our technique is applied to compare the two leading methods for human pose estimation on the COCO Dataset, measure the sensitivity of pose estimation with respect to instance size, type and number of visible keypoints, clutter due to multiple instances, and the relative score of instances. The performance of algorithms, and the types of error they make, are highly dependent on all these variables, but mostly on the number of keypoints and the clutter. The analysis and software tools we propose offer a novel and insightful approach for understanding the behavior of pose estimation algorithms and an effective method for measuring their strengths and weaknesses.

1. Supplementary Materials

This document accompanies the paper "Benchmarking and Error Diagnosis in Multi-Instance Pose Estimation".

We provide clarification on how to interpret some of the presented content and illustrate the results of our evaluation analysis on other datasets. Finally, we include the result of our analysis for other methods in addition to those contained in the Main Paper.

- **Human Pose and Skeleton Color Coding** (Sec. 1.1): Visualization of the color-coding of the human skeleton obtained from a pose estimation algorithm.
- **Fine-Grained Precision Recall Plots** (Sec. 1.2): In-depth explanation on how to interpret the performance plots computed by our analysis tools.
- **Correction of Localization Errors** (Sec. 1.3): Visualization of a predicted human skeleton as the localization errors it contains are progressively corrected.
- **Multi-Instance Mouse Pose Evaluation** (Sec. 1.4): Analysis of the performance of a multi-instance pose estimation algorithm on the *Caltech Resident Intruder Mouse* dataset CRIM13 [2].
- **Performance Analysis Reports** (Sec. 1.5): The performance reports obtained by running our analysis code on several algorithms.

1.1. Human Pose and Skeleton Color Coding

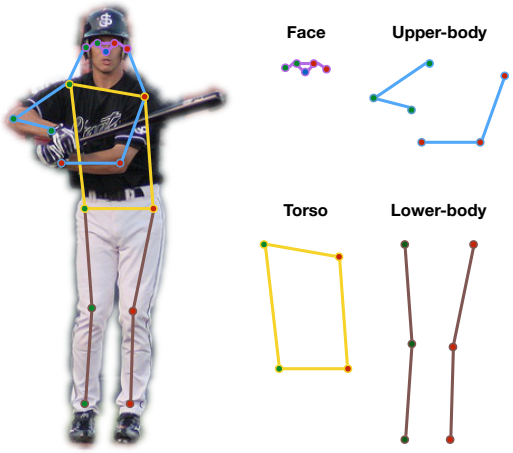

Figure 1. **Human Pose and Skeleton Color Coding.**

We adopt the following color coding to visualize algorithm's keypoint detections:

- The location of the left and right parts of the body is indicated respectively with red and green dots; the location of the nose is plotted in blue.
- Face keypoints (*nose*, *eyes*, *ears*) are connected by purple lines.
- Upper-body keypoints (*shoulders*, *elbows*, *wrists*) are connected by blue lines.

- Torso keypoints (*shoulders*, *hips*) are connected by yellow lines.
- Lower-body keypoints (*hips*, *knees*, *ankles*) are connected by brown lines.

1.2. Fine-Granined Precision Recall Plots

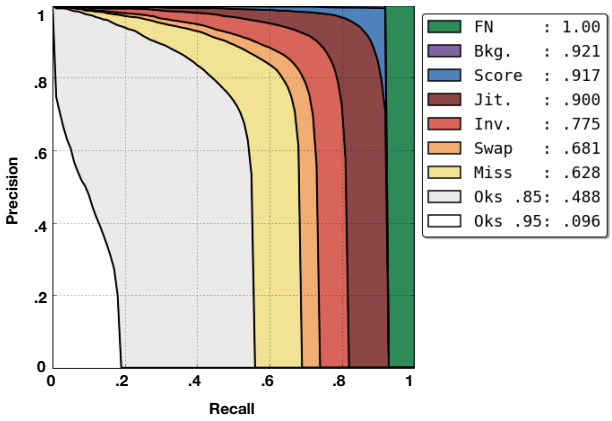

Figure 2. **Fine-Grained Error Analysis.** We study the errors occurring in multi-instance pose estimation, and provide the tools for a fine-grained description of performance, which allows to quantify the impact of each type of error at a single glance.

Fig. 2 summarizes the impact of all types of error on the performance of a multi-instance pose estimation algorithm. It is composed of a series of Precision Recall (PR) curves where each curve is guaranteed to be strictly higher than the previous as the evaluation setting becomes more permissive. The legend shows the Area Under the Curve (AUC) obtained for each of the following evaluation settings:

- **Oks .95, .85**: PR curves obtained at the OKS thresholds of .95 and .85 respectively.

The remaining evaluations are performed with the lowest OKS threshold considered in the legend (.85 in this case).

- **Miss, Swap, Inv., Jit.**: PR curves after the algorithm's keypoint detections are progressively corrected to remove each type of localization error, as shown in Sec. 1.3. As keypoint localization is corrected, the OKS between a detection and ground-truth match improves, possibly exceeding the current OKS evaluation threshold and becoming an additional True Positive. We show with different colors, the AUC improvements obtained by fixing each type of localization error.
- **Score**: PR curves after the algorithm's keypoint detections have been rescored with the optimal confidence score described in the main paper.
- **Bkg.**: PR curves after all of the algorithm's background False Positive detections are removed.
- **FN**: PR curves after all the False Negative errors are ignored.

In the case of the Cmu [3] algorithm, the AUC evaluated at Oks=.95 is only .096, but improves to .488 when lowering the threshold to .85. At this threshold, correcting all the *miss* errors results in a large improvement of the AUC to .628. Smaller AUC gains are obtained when correcting *swaps*, .681, and *inversions*, .775. Another large improvement is obtained when *jitter* errors are removed, resulting in an AUC of .900. This shows what would the performance of [3] be if it had a perfect localization of keypoints. When localization is very good, the impact of scoring is not as significant, but still results in an AUC improvement of about 2%. Optimally scoring detections greatly diminishes the impact of Background False Positives, as detections rarely remain unmatched. Finally, removing background False Negatives provides the remaining AUC to obtain perfect performance. In summary, Cmu's errors are dominated by imperfect localization, mostly *miss* and *jitter* errors, and missed detections.

1.3. Sequential Correction of Localization Errors

The fine-grained PR curves shown in Fig. 2 are obtained by fixing an OKS threshold and evaluating the performance of an algorithm after progressively correcting its detections. To do so, we compute for every predicted keypoint what is the Keypoint Similarity (KS) with its corresponding ground-truth body part, and with different ground-truth body parts of the same person, and of other people in the image. This allows us to define the types of localization error, as done in Sec. 3.1 of the main paper, and correct them.

- **Miss** errors are corrected by repositioning a keypoint prediction on the .5 KS circle centered on the true location; *left-elbow* and *wrists* in Fig. 3.
- **Swap** and **Inversion** errors are corrected by repositioning a keypoint prediction at a distance from the correct ground-truth location so that the new value of KS is the same that the prediction had with the wrong body part it mistakenly detected (belonging to a different/same person for swap/inversion); in Fig. 3 the *right-elbow* is a swap, *right-knee* is an inversion.
- **Jitter** errors are corrected by repositioning a keypoint prediction on the .85 KS circle centered on the true location; *left-ankle* in Fig. 3.

Miss and *jitter* errors are corrected by bringing a prediction to a fixed distance from its true position. The new location of *swaps* and *inversions* depends instead on how good was the prediction of the wrong joint: after correction, a good/bad prediction of the wrong body part becomes a good/bad prediction (high/low KS) of the true body part.

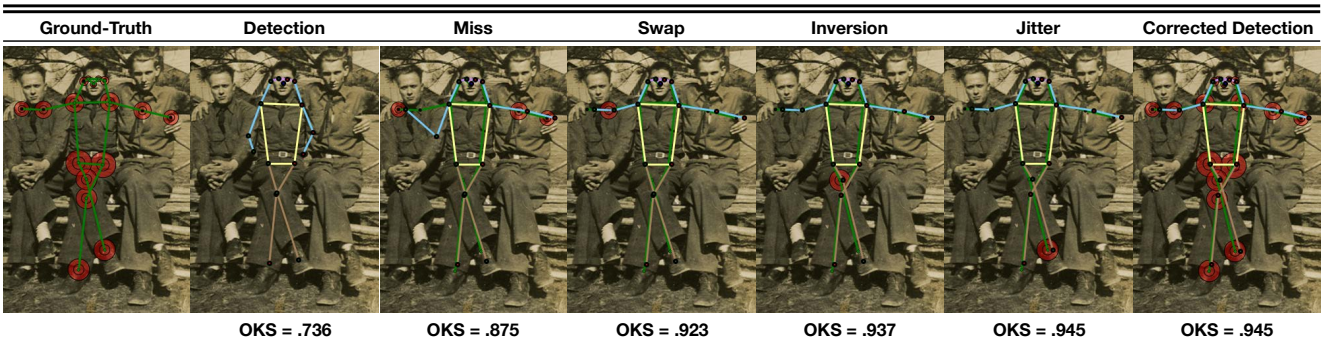

Figure 3. **Correction of Keypoint Localization Errors.** The change of a detection's keypoint positions and the resulting OKS improvement as localization errors are progressively corrected. We plot the ground-truth skeleton in green and the detection using the color coding discussed below. The red concentric circles indicate the .5 and .85 KS threshold as discussed in Fig.2 of the main paper. When visualizing the individual error types, we show the concentric circles around the ground-truth location only for the keypoints that are being corrected.

Fig. 3, 5 provide two examples of how the keypoints belonging to a detection can be progressivily improved. The OKS increase obtained by correcting the localization errors depends both on the number of errors of that type, and the total number of visible keypoints present in an instance, see Eq. 1 of the main paper. Fixing the position of predicted keypoints impacts the overall AUC of the PR curves: the detection in Fig. 3, previously a FP for OKS evaluation thresholds above .7, has become, after correction, a TP at all thresholds between .75 and .9.

1.4. Multi-Instance Mouse Pose Evaluation

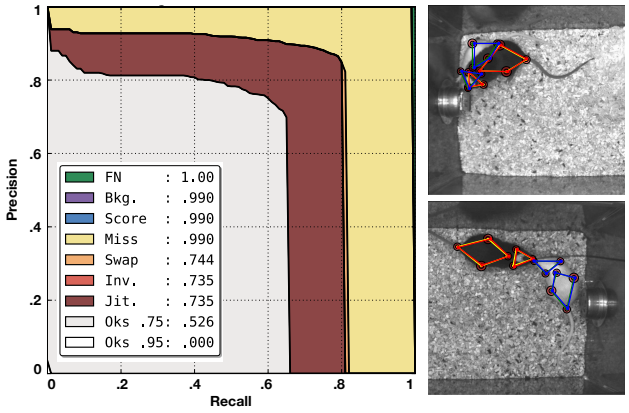

Figure 4. **Performance Breakdown on CRIM13 [2].** (Left) The PR curves for a multi-instance pose estimation algorithm trained and evaluated on the CRIM13 [2] dataset. (Right) Examples of detections containing swap errors; Top - right hip of the black mouse; Bottom - nose of both mice. The ground-truth skeleton of the black and white mice are shown respectively in yellow and green, the corresponding detections in red and blue.

The study of multi-instance pose estimation errors and performance conducted in the main paper extends beyond humans, to any object category where the location of parts is estimated along with a detection, and to situations where cluttered scenes may contain multiple object instances. This is common in fine-grained categorization, i.e. birds [1], or animal behavior analysis, i.e. mice [2] and flies [4], where part alignment is often crucial. To show the versatility of our software tools, we evaluated the performance of a top-down pose estimation algorithm on the CRIM13 [2] dataset, which consists of images of pairs of mice (a black resident and a white intruder) engaging in social behavior. For our experiment, we used 10000 images, separated into a Training, Validation and Test sets of 8500, 500 and 1000 images, for which human annotations of 7 keypoint locations (*nose*, *ears*, *neck*, *hips*, *tail*) were available. During evaluation, the Keypoint Similarity metric and the OKS between a detection and an annotation are computed in the same way described in Sec. 2.2 of the main paper. Fig. 4.(Left), shows the performance of a top-down method, composed of a Multi-Box object detector [6] to find each mouse, followed by a stacked hourglass network [7] for predicting the keypoint locations. Results indicate that the two predominant errors are *jitter* and *miss*; *swap* errors have a very limited impact and occur during interactions which results in some amount of occlusion, Fig. 4.(Right); *inversion* errors are mostly absent. The scoring of detections is not critical, as in the human case, since images always contains exactly two mice. Because of the fairly simple, clutter-free and fixed-viewpoint image capture settings, background errors (False Positives and False Negatives) are mostly absent.

Figure 5. **Correction of Keypoint Localization Errors.**

1.5. Performance Analysis Reports

In the following pages we include the performance reports[1] generated by the released analysis code[2] on the following algorithms:

- **Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields** [3]
- **Towards Accurate Multi-person Pose Estimation in the Wild** [8]
- **Mask R-CNN** [5]

[1] Some values have been obfuscated to preserve the sanctity of the *COCO test-dev* split. Check the Main Paper to find the corresponding values on the *COCO training set*.

[2] Available for download at: https://goo.gl/9EyDyN

References

[1] S. Branson, G. Van Horn, S. Belongie, and P. Perona. Bird species categorization using pose normalized deep convolutional nets. *arXiv preprint arXiv:1406.2952*, 2014. 3

[2] X. Burgos-Artizzu, P. Dollár, D. Lin, D. Anderson, and P. Perona. Social behavior recognition in continuous videos. In *CVPR*, 2012. 1, 3

[3] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. *arXiv preprint arXiv:1611.08050*, 2016. 2, 4

[4] E. Eyjolfsdottir, S. Branson, X. P. Burgos-Artizzu, E. D. Hoopfer, J. Schor, D. J. Anderson, and P. Perona. Detecting social actions of fruit flies. In *European Conference on Computer Vision*, pages 772–787. Springer, 2014. 3

[5] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. *arXiv preprint arXiv:1703.06870*, 2017. 4

[6] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In *European Conference on Computer Vision*, pages 21–37. Springer, 2016. 3

[7] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. *arXiv preprint arXiv:1603.06937*, 2016. 3

[8] G. Papandreou, T. Zhu, N. Kanazawa, A. Toshev, J. Tompson, C. Bregler, and K. Murphy. Towards accurate multi-person pose estimation in the wild. *arXiv preprint arXiv:1701.01779*, 2017. 4