How Robust is 3D Human Pose Estimation to Occlusion?

István Sárándi[1], Timm Linder[2], Kai O. Arras[2] and Bastian Leibe[1]

[1]Visual Computing Institute, RWTH Aachen University {sarandi,leibe}@vision.rwth-aachen.de
[2]Robert Bosch GmbH, Corporate Research {timm.linder,kaioliver.arras}@de.bosch.com

Abstract

Occlusion is commonplace in realistic human-robot shared environments, yet its effects are not considered in standard 3D human pose estimation benchmarks. This leaves the question open: how robust are state-of-the-art 3D pose estimation methods against partial occlusions? We study several types of synthetic occlusions over the Human3.6M dataset and find a method with state-of-the-art benchmark performance to be sensitive even to low amounts of occlusion. Addressing this issue is key to progress in applications such as collaborative and service robotics. We take a first step in this direction by improving occlusion-robustness through training data augmentation with synthetic occlusions. This also turns out to be an effective regularizer that is beneficial even for non-occluded test cases.

Figure 1. Example of partial occlusions in the context of shared human-robot workspaces. Note how easily we humans can guess the rough pose of the person behind the occlusion. Can current 3D human pose estimation methods do that as well?

1. Introduction

To collaborate with humans and to understand their actions, collaborative and service robots need the ability to reason about human pose in 3D space. An important challenge in realistic environments is that humans are often only seen partially, *e.g.*, standing behind machine parts or carrying objects in front of the body (see Fig. 1). Robust robotics solutions need to handle such disturbances gracefully and make use of the visual cues still present in the scene to reason around the occlusion.

Although recent years have brought significant advances in 3D human pose estimation, as measured on standard computer vision benchmarks such as Human3.6M [11][2], the behavior of models under occlusion remains largely unexplored, as the benchmarks do not systematically model occlusion effects.

To our knowledge, we present the first systematic study of various types of test-time (synthetic) occlusions in 3D human pose estimation from a single RGB image. As we will see, ignoring the aspect of occlusions may cause model accuracy to rapidly deteriorate, even under mild occlusion levels, despite the good benchmark performance. Such sudden and unexpected failures in the robot's perception would prevent smooth and comfortable human-robot interaction and may lead to safety hazards. Furthermore, we demonstrate that simple occlusion data augmentation during training increases model robustness. This augmentation also improves performance even for non-occluded test images. Our approach is efficient and suitable for high frame-rate applications.

2. Related Work

3D Human Pose Estimation 3D human pose estimation has seen rapid progress in recent years. For a thorough overview of approaches, we refer the reader to Sarafianos *et al.*'s survey [20]. Current state-of-the-art methods use deep neural networks, either directly on the input image or on the output of a 2D pose estimator. Based on the sweeping success of heatmap-based representations in 2D human pose estimation (*e.g.*, [15]), heatmaps have recently been also adopted in 3D methods, including volumetric [18][23] and marginal heatmaps [16].

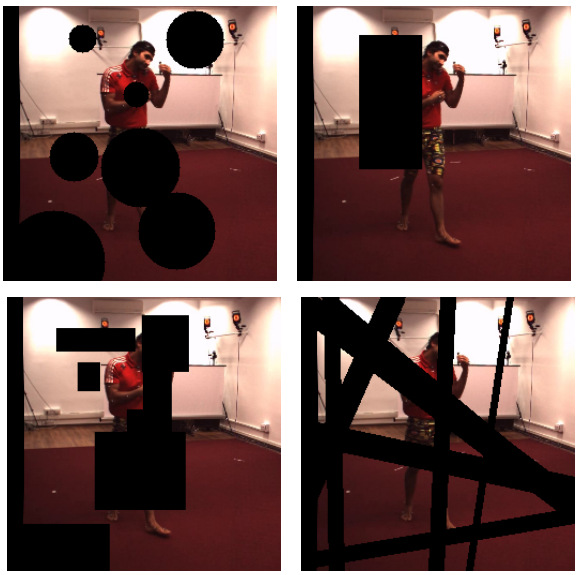

Figure 2. Examples of the applied geometric occlusions: circles, a single rectangle [26], rectangles, oriented bars. See Fig. 3 for an example with Pascal VOC objects.

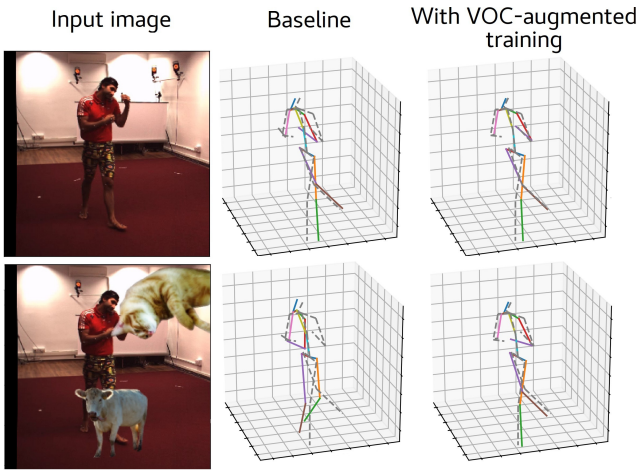

Figure 3. Prediction change in the presence of synthetic test-time occlusion. Ground truth is shown with grey dashed lines, predictions with colorful ones. The baseline model fails to predict the pose of the occluded limbs, while the model trained with occlusion augmentation behaves more robustly.

Occlusions, Erasing and Copy-Pasting In a pre-deep learning study based on silhouettes and HOG features, Huang *et al*. tackled occlusions in 3D pose estimation from RGB [10], but their analysis was limited to walking actions and occlusions with two rectangles. Occlusion effects have also been studied in 3D pose estimation from depth input [19], where exploiting semantic information from the occluder itself was found to improve predictions.

Data augmentation by erasing a rectangular block from the input has recently been concurrently investigated under the names *Random Erasing* [26] and *Cutout* [3], for image classification, object detection, and person re-identification. Similarly, synthetically placing objects into a scene by image-level *copy-pasting* has been shown to help object detection [5][4][7]. However, those methods are trained to detect these pasted objects, while in our case the task is to infer what lies behind them. Ke *et al*. [12] augment training images for 2D human pose estimation by copying background patches over some of the body joints. Research on facial landmark localization has investigated and modeled occlusions for a long time [1][8], including augmenting training images with randomly pasted occluding objects [25].

3. Approach

In this paper we study the effect of occlusion on the accuracy of 3D human pose estimation. To this end, we have devised a 3D pose estimation approach that reaches state-of-the-art benchmark performance, leading us to expect that the observations drawn from our experiments also transfer to other models.

Architecture We use a fully convolutional net to predict volumetric body joint heatmaps from the input RGB image, based on a ResNet-50 [9] backbone architecture. After discarding the global average pooling layer, we adjust the number of output channels of the ResNet to be the product of the number of joints and the number of heatmap-voxels along the depth axis. Reshaping the resulting tensor yields the volumetric heatmaps. Nominal stride and depth discretization are configured to yield heatmaps of size $16 \times 16 \times 16$ for an image of size 256×256. Given the volumetric heatmap, coordinate predictions are obtained using soft-argmax [13][17][23]. As in [18], the x and y coordinates are interpreted as image space coordinates, while z is the depth of the particular joint relative to the root (pelvis) joint depth, with the 16 voxels covering 2 meters. In order to concentrate on the aspect of articulated pose, as opposed to person localization, we assume that the true root joint depth is given by an oracle at test time. The coordinates are back-projected to camera space using the known camera intrinsics. Finally, the L^1 loss is computed on the predicted and ground truth 3D coordinates in camera space. Since all of the preceding operations are differentiable, the network can be trained end-to-end.

4. Experimental Setup

Dataset Human3.6M [11] is the largest public 3D pose estimation dataset. It contains 11 subjects imitating 15 actions in a controlled indoor environment while being recorded with 4 cameras and a motion capture system. Following the most common experimental protocol in the literature, we use five subjects (S1, S5, S6, S7, S8) for training and

	Direct	Discuss	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD	Smoke	Wait	Walk	WalkD	WalkT	Avg
Zhou [28]	87.4	109.3	87.1	103.2	116.2	139.5	106.9	99.8	124.5	199.2	107.4	118.1	79.4	114.2	97.7	113.0
Tekin [24]	102.4	147.7	88.8	125.4	118.0	182.7	112.4	129.2	138.9	224.9	118.4	138.8	55.1	126.3	65.8	125.0
Zhou [27]	91.8	102.4	97.0	98.8	113.4	125.2	90.0	93.8	132.2	159.0	106.9	94.4	79.0	126.0	99.0	107.3
Sun [22]	90.2	95.5	82.3	85.0	87.1	94.5	87.9	93.4	100.3	135.4	91.4	87.3	78.0	90.4	86.5	92.4
Sun [23] (ArXiv)	63.8	64.0	56.9	64.8	62.1	70.4	59.8	60.1	71.6	91.7	60.9	65.1	51.3	63.2	55.4	64.1
Pavlakos [18]	67.4	72.0	66.7	69.1	72.0	77.0	65.0	68.3	83.7	96.5	71.7	65.8	59.1	74.9	63.2	71.9
Pavlakos [18] (known root depth)	59.3	64.9	59.4	61.3	65.1	69.0	57.1	60.1	75.1	91.9	64.5	59.6	66.8	53.7	56.8	64.8
Ours (no occlusion augm.)	60.2	64.1	55.9	58.3	63.8	69.5	58.8	64.4	67.7	90.8	61.9	59.2	66.0	56.9	50.8	63.3
w/ circles augm.	52.9	58.0	51.8	54.8	56.9	62.6	51.4	55.0	64.7	79.2	56.3	52.5	58.8	47.9	43.0	56.8
w/ single rectangle augm.[26]	52.0	58.6	51.0	53.5	56.1	62.6	51.5	54.2	65.7	71.2	56.1	52.9	58.2	47.8	42.9	56.1
w/ rectangles augm.	51.9	57.9	52.5	54.2	57.3	61.9	51.7	55.2	63.4	76.7	56.5	51.7	58.8	47.8	43.4	56.5
w/ bars augm.	55.0	60.1	54.1	56.4	59.9	64.9	52.4	59.5	67.7	88.7	58.5	54.2	62.4	50.0	45.4	59.6
w/ VOC objects augm.	51.2	58.7	51.7	53.4	56.8	59.3	50.7	52.6	65.5	73.2	56.8	51.4	56.6	47.0	42.4	55.8
w/ mixture augm.	51.3	57.8	52.5	53.8	55.9	58.7	50.9	52.8	66.7	77.1	56.6	51.7	56.6	47.6	42.8	56.1

Table 1. Mean per joint position error on Human3.6M for methods using no extra pose datasets in training. Methods below the line have access to the ground-truth root joint depth at test-time. (No synthetic occlusions are used on the test inputs.)

two (S9, S11) for testing. We train action-agnostic models, as opposed to action-specific ones.

Dataset Subsampling To reduce the redundancy in training poses, we adaptively subsample the frames similarly to [14], only keeping a frame when at least one body joint moves by at least 30 mm compared to the last kept frame. For the test set we follow prior work and use every 64^{th} frame.

Image Preprocessing Before feeding an image to the network, we center and zoom it on the person, at a resolution of 256×256 px. To ensure correct perspective (with the principal point at the image center), we reproject the image onto a virtual camera pointed at the center of the person's bounding box, as provided in the dataset. Scaling is applied so that the larger side of the person's bounding box covers about 80% of the image side length. Common data augmentation techniques are used in training, including random rotation, scaling, translation, horizontal flipping, as well as image filtering such as color distortions and blurs.

Evaluation Metrics Following standard practice on Human3.6M, we evaluate prediction accuracy by the so-called mean per joint position error (MPJPE), which is the mean Euclidean error of all joints after skeleton alignment at the root (pelvis) joint. Procrustes alignment is not used.

Synthetic Occlusions for Robustness Analysis We consider solid black shapes and some more realistic object segments from the Pascal VOC 2012 dataset [6] as occluders in this study (see Fig. 2 and 3). The number, position and size of the objects are generated at random. We define the *degree of occlusion* as the percentage of occluded pixels inside the person's bounding box and vary this quantity between 0% and 70%.

Occlusion-Augmented Training We hypothesize that synthetic occlusion data augmentation during training can improve test-time occlusion-robustness. To verify this, we use the same kinds of occlusions as described in the previous section, with an additional *mixture* variant, which uses one of the other types at random for each frame. We make sure to strictly separate the VOC objects used for training and testing. Furthermore, we try the RE-0 variant of Zhong *et al*.'s *random erasing* [26], generating a single occluding black rectangle of random size according to their pseudocode. We refer to this mode as *single rectangle* in this paper.

To make these strategies comparable, we parameterize them such that the distribution of the number of occluded pixels is similar. Notably, we only apply these augmentations with 50% probability for each frame. This was found important in prior work on occlusion augmentation [3].

Implementation Details We use the implementation of ResNet-50v1 and the corresponding ImageNet-pretrained initial weights from the TensorFlow-Slim library [21]. Training is done with the Adam optimizer and a mini-batch size of 64, for 40 epochs, taking approximately 24 hours on an NVIDIA GeForce Titan X (Pascal) GPU.

5. Results

We start presenting our results by showing that our baseline model has state-of-the-art performance. We then show how performance deteriorates with test-time occlusions and that this can be mitigated using occlusion data augmentation. The augmentations are then shown to help even when the test images do not contain synthetic occlusions.

Baseline Performance The current state-of-the-art among published methods which use no extra 2D pose datasets for training is by Pavlakos *et al*. [18] (see Table 1). Since our evaluation assumes knowledge of the root joint depth at test time, we compare with Pavlakos *et al*.'s performance under the same conditions, for which the results can be found in their supplementary material. Our baseline's MPJPE of 63.3 mm is already better than Pavlakos *et al*.'s 64.8.

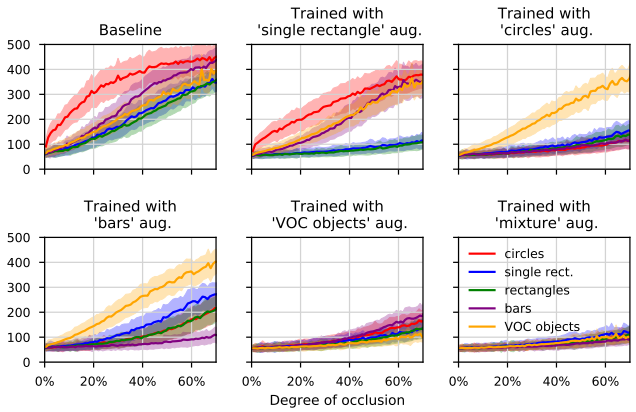

Figure 4. Assessing occlusion-robustness on Human3.6M. Each subplot shows the performance when training with a particular augmentation method. Within a subplot, each line shows the mean and standard deviation of MPJPE under increasing degrees of occlusion of a particular type.

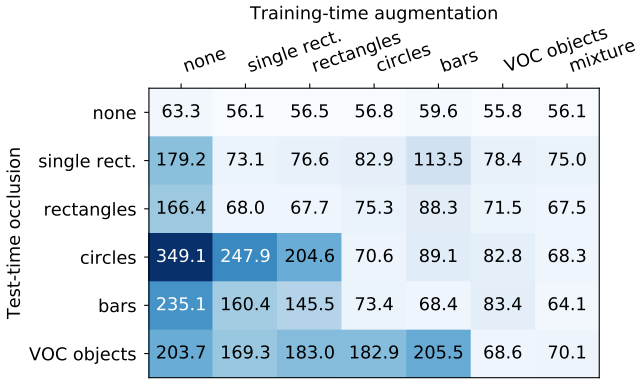

Figure 5. Exploring how much each type of training-time data augmentation protects against each type of test occlusions. The numbers are the MPJPE averaged for degrees of occlusion between 10% and 50%.

Sun *et al.*'s (unpublished) method achieves an MPJPE of 64.1 mm [23], but it is unclear whether they use the known root joint depth or resolve scale ambiguity by other means.

Robustness Analysis under Occlusion We evaluate the robustness of our baseline model using different degrees and types of occlusions (see the top left plot of Fig. 4). We observe that circular occlusions cause by far the largest increase in error, the reason for which needs further investigation. Occlusions with oriented bars, VOC objects and rectangles lead to comparable performance loss. We note that rectangles are the least problematic type of occlusion, despite being a widely used test case in the literature.

Fig. 3 visualizes an example. The baseline network gives good predictions for the unoccluded case, but when we paste two Pascal VOC objects onto the image, prediction visibly fails for the affected limbs.

Augmentation Improves Occlusion-Robustness We now turn to the evaluation of occlusion augmentation at training time for increased test-time occlusion-robustness. Fig. 4 and 5 show the results. Erasing a single rectangle (as in [26]) results in robustness against multiple rectangles at test time, but is much less effective for the other types of occlusions, being most sensitive to circles. Using several rectangles during training works slightly better than single-rectangle random erasing, but it, too, has difficulty in generalizing to other types of occlusion structures. Circular occlusion augmentation generalizes to all other simple geometric occlusion shapes, but barely helps when more realistic VOC objects are used as occluders at test time. VOC-augmentation, however, does generalize to both simple geometric shapes and other VOC objects (the objects used in training and testing are strictly separated). The qualitative difference in robustness when using this augmentation type is illustrated in Fig. 3. The network learned to use context cues and gives good prediction even for the almost fully-occluded lower left leg. Finally, the combination of all these strategies proves to be effective against all of the analyzed occlusion types together.

The Regularizing Effect of Occlusion Augmentation In the previous section we have seen that training-time occlusion augmentation is helpful when evaluating on occluded test examples. Let us now look at the effect of these augmentation schemes when evaluating on the original test data without synthetic occlusions (see Table 1). All occlusion augmentation strategies are found to improve upon the baseline result, with the *VOC objects* performing the best and *bars* the worst.

Runtime Inference of the whole pipeline runs at 64, 165, and 204 fps for batch sizes of 1, 8, and 64 images, respectively, on a single NVIDIA GeForce Titan X (Pascal) GPU. This makes the method suitable for high frame rate applications.

6. Conclusion

We presented a systematic study of occlusion effects on 3D human pose estimation from a single RGB image, using an efficient ResNet-based test model. We found that despite producing state-of-the-art benchmark results, the network's performance quickly drops when synthetic occlusions are added. Circular structures turned out to be particularly problematic, the reason of which needs further study. We then showed that training-time occlusion data augmentation is effective in reducing occlusion-induced errors, while also improving the performance without test-time occlusions.

Future experiments should also target other datasets besides Human3.6M and it remains to be seen how well our findings about synthetic occlusions generalize to real ones.

Acknowledgments

This project has been funded by a grant from the Bosch Research Foundation and by ILIAD (H2020-ICT-2016-732737).

References

[1] X. P. Burgos-Artizzu, P. Perona, and P. Dollár. Robust face landmark estimation under occlusion. 2013.

[2] C. S. Catalin Ionescu, Fuxin Li. Latent structured models for human pose estimation. 2011.

[3] T. DeVries and G. W. Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv:1708.04552*, 2017.

[4] N. Dvornik, J. Mairal, and C. Schmid. Modeling visual context is key to augmenting object detection datasets. *arXiv:1807.07428*, 2018.

[5] D. Dwibedi, I. Misra, and M. Hebert. Cut, paste and learn: Surprisingly easy synthesis for instance detection. 2017.

[6] M. Everingham and J. Winn. The pascal visual object classes challenge 2012 (voc2012) development kit. *Pattern Analysis, Statistical Modelling and Computational Learning, Tech. Rep*, 2011.

[7] G. Georgakis, A. Mousavian, A. C. Berg, and J. Kosecka. Synthesizing training data for object detection in indoor scenes. *arXiv:1702.07836*, 2017.

[8] G. Ghiasi and C. C. Fowlkes. Occlusion coherence: Localizing occluded faces with a hierarchical deformable part model. 2014.

[9] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. 2016.

[10] J.-B. Huang and M.-H. Yang. Estimating human pose from occluded images. 2009.

[11] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. 2014.

[12] L. Ke, M.-C. Chang, H. Qi, and S. Lyu. Multi-scale structure-aware network for human pose estimation. *arXiv:1803.09894*, 2018.

[13] S. Levine, C. Finn, T. Darrell, and P. Abbeel. End-to-end training of deep visuomotor policies. 17(1):1334–1373, 2016.

[14] D. Mehta et al. Vnect: Real-time 3d human pose estimation with a single rgb camera. *ACM Trans. Gr.*, 36(4):44, 2017.

[15] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. 2016.

[16] A. Nibali, Z. He, S. Morgan, and L. Prendergast. 3d human pose estimation with 2d marginal heatmaps. *arXiv:1806.01484*, 2018.

[17] A. Nibali, Z. He, S. Morgan, and L. Prendergast. Numerical coordinate regression with convolutional neural networks. *arXiv:1801.07372*, 2018.

[18] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis. Coarse-to-fine volumetric prediction for single-image 3d human pose. 2017.

[19] U. Rafi, J. Gall, and B. Leibe. A semantic occlusion model for human pose estimation from a single depth image. In *CVPR Workshops*, 2015.

[20] N. Sarafianos, B. Boteanu, B. Ionescu, and I. A. Kakadiaris. 3d human pose estimation: A review of the literature and analysis of covariates. 152:1–20, 2016.

[21] N. Silberman and S. Guadarrama. Tensorflow-slim image classification model library. https://github.com/tensorflow/models/tree/master/research/slim, 2016.

[22] X. Sun, J. Shang, S. Liang, and Y. Wei. Compositional human pose regression. 2017.

[23] X. Sun, B. Xiao, S. Liang, and Y. Wei. Integral human pose regression. *arXiv:1711.08229*, 2017.

[24] B. Tekin, A. Rozantsev, V. Lepetit, and P. Fua. Direct prediction of 3d body poses from motion compensated sequences. 2016.

[25] K. Yuen and M. M. Trivedi. An occluded stacked hourglass approach to facial landmark localization and occlusion estimation. *IEEE Trans. Intel. Veh.*, 2(4):321–331, 2017.

[26] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang. Random erasing data augmentation. *arXiv:1708.04896*, 2017.

[27] X. Zhou, X. Sun, W. Zhang, S. Liang, and Y. Wei. Deep kinematic pose regression. 2016.

[28] X. Zhou, M. Zhu, S. Leonardos, K. G. Derpanis, and K. Daniilidis. Sparseness meets deepness: 3d human pose estimation from monocular video. 2016.