Efficient Regression of General-Activity Human Poses from Depth Images: Supplementary Material

Ross Girshick[†⋆] Jamie Shotton[†] Pushmeet Kohli[†] Antonio Criminisi[†] Andrew Fitzgibbon[†]
[†]Microsoft Research Cambridge [⋆]University of Chicago

Please also see the supplementary video.

1. Additional Experimental Results

Vote length thresholds λ_j. After training the tree structure and leaf regression models, we automatically tuned the per-joint vote length threshold hyper-parameters λ_j on a 5k image validation set. To optimize λ_j we maximized mean average precision using grid search with a step size of 0.05m in the range $[0.05, 0.60]$m. Table 1 shows that depending on which error metric is used (i.e., *does missing an occluded joint count as a false negative or not?*), the optimized length thresholds are quite different. In the right column we see that when the model is penalized for missing occluded joints, it makes use of longer range votes to maximize mAP. In some cases, such as head, feet, and ankles, the difference is rather large. Intuitively this makes sense: occluded joints tend to be further away from visible depth pixels than non-occluded joints. This experiment used a forest trained on 30k images.

	***not* penalized** λ_j	**penalized** λ_j
Head	0.20	0.50
Neck	0.20	0.35
L. Shoulder	0.30	0.45
R. Shoulder	0.35	0.40
L. Elbow	0.15	0.15
R. Elbow	0.15	0.15
L. Wrist	0.10	0.10
R. Wrist	0.10	0.10
L. Hand	0.15	0.10
R. Hand	0.10	0.15
L. Knee	0.35	0.30
R. Knee	0.45	0.30
L. Ankle	0.15	0.45
R. Ankle	0.15	0.55
L. Foot	0.10	0.45
R. Foot	0.10	0.55

Table 1. Optimized values for the test-time vote length threshold λ_j under two different error metrics.

Tree structure objectives. To investigate whether including all joints in the regression objective's error function (main paper Eq. 3) is problematic, we experimented with separate regression forests, each tasked with predicting the location of a single joint. Following our procedure in the main paper, we tested these per-joint regression forests with three depth-20 trees each trained with 5k images. We evaluated four representative joints: head, l. elbow, l. wrist, and l. hand. With $\rho = \infty$, they achieved AP scores of 0.95, 0.564, 0.508, and 0.329 respectively. As expected, due to greater capacity (a forest for a *single* joint *vs*. shared for *all* joints), these per-joint forests yielded better results than E^{reg} with $\rho = \infty$, the green bars in Fig. 3 in the main paper, but were still far worse than the regression forests trained with the proxy classification objective.

True positive radius D. Following [1], we used $D = 0.1$m as the true positive radius in the main paper. Fig. 1 shows the effect of varying D. Note how our algorithm maintains much higher mAP scores as the radius shrinks in comparison to the obtained in [1]. For example, when $D = 0.06$m our system scores 0.612 *vs*. 0.429 in [1].

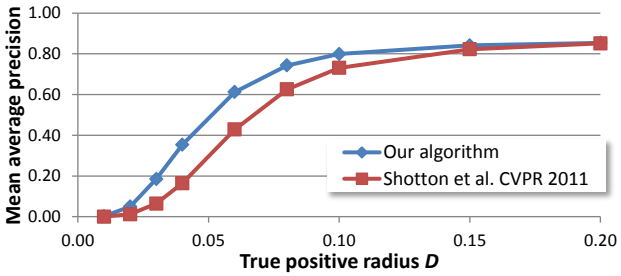

Figure 1. **Mean average precision *vs*. true positive radius.**

References

[1] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from a single depth image. In *Proc. CVPR*. IEEE, 2011. 1