Real-Time Human Pose Recognition in Parts from Single Depth Images: Supplementary Material

Jamie Shotton Andrew Fitzgibbon Mat Cook Toby Sharp Mark Finocchio
Richard Moore Alex Kipman Andrew Blake
Microsoft Research Cambridge & Xbox Incubation

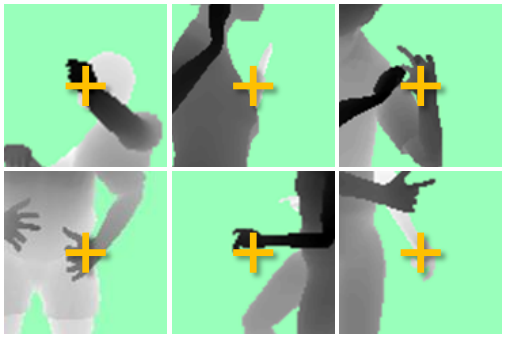

Figure 1. **Differences in local appearance within one body part.**

This supplementary material accompanies the paper "Real-time Human Pose Recognition in Parts from Single Depth Images" at CVPR 2011. It provides more detail on the randomized synthetic rendering pipeline and further experimental results. Please also see the supplementary video.

1. Rendering pipeline

Even within a single body part there is considerable variation in appearance due to even just pose variation. Fig. 1 gives examples for the (screen-)right hand.

To account for this variation in the training data, we built a comprehensive rendering pipeline of images of people from which we randomly sample labeled training images. The variations described below are the best approximation we could reasonably achieve to the variations one expects in the real world, including pose, clothing, camera noise, *etc*. We cannot hope to sample all possible combinations of variations. However, if samples contain independent variations (and we thus exclude artificial correlations such as thin people always wear a hat), we can expect the classifier to learn a large degree of invariance.

We now run through the variations.

Base Character We use 3D models of 15 varied base characters, both male and female, from child to adult, short to tall, and thin to fat. Some examples are shown in Fig. 2 (top row). A given render will pick uniformly at random from the characters.

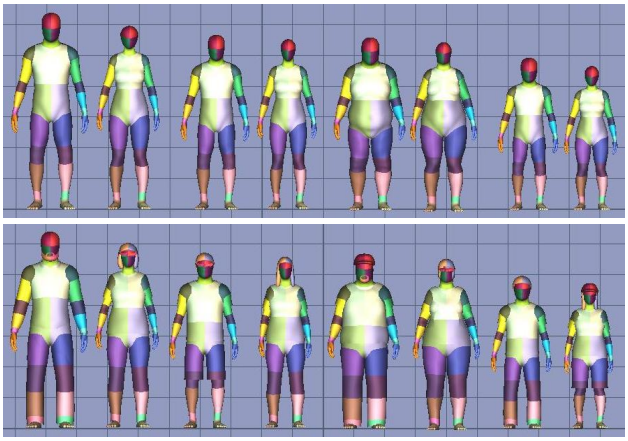

Figure 2. **Renders of several base character models.** Top row: without skinning. Bottom row: with random skinning of hair and clothing.

Pose Having discarded redundant poses from the mocap data, we retarget the remaining poses to each base character, and choose uniformly at random. The pose is also mirrored left-right with probability $\frac{1}{2}$ to prevent a left or right bias.

Rotation & Translation The character is rotated about the vertical axis and translated in the scene, uniformly at random.

Hair & Clothing We add mesh models of several hair styles and items of clothing chosen at random; some examples are shown in Fig. 2 (bottom row).

Weight & Height Variation The base characters already have a wide variety of weights and heights. To add further variety we add an extra variation in height $\pm 10\%$ and weight $\pm 10\%$. For rendering efficiency, this variation does not affect the pose retargetting.

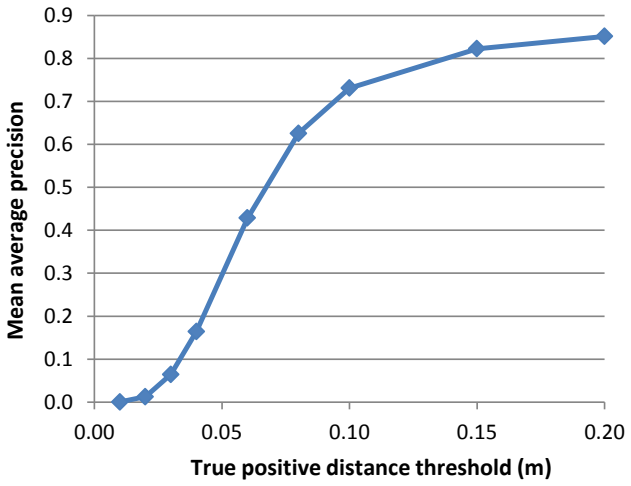

Figure 4. **True positive threshold *vs*. mean average precision.**

Camera Position & Orientation The camera height, pitch and roll are chosen uniformly at random within a range believed to be representative of an entertainment scenario in a home living room.

Camera Noise Real depth cameras exhibit noise. We distort the clean computer graphics renders with dropped-out pixels, depth shadows, spot noise and disparity quantization to match the camera output as closely as possible. In practice however, we found this noise addition had little effect on accuracy, perhaps due to the quality of the cameras or the more important appearance variations due to other factors such as pose.

We use a standard graphics rendering pipeline to generate the scene, consisting of a depth image paired with its body part label image. Examples are given Fig. 3.

2. Further Experimental Results

True positive threshold. We show in Fig. 4 the effect of the threshold on true positives in our recall-precision metrics on joint prediction accuracy in the synthetic test set. Even at a tighter 6cm threshold our approach still gives an mAP of about 0.5.

Number of trees. We show in Fig. 5 test accuracy as the number of trees is increased, using 5k images for each depth 18 tree. The improvement starts to saturate around 4 or 5 trees, and is much less than the improvement by making the trees deeper. The error bars give an indication of the remarkably small variability between trees.

Definitions of body parts. The 31 parts used contain several (e.g. LU arm) that do not map directly to the body joints of interest. To investigate whether these 'redundant' parts are useful, we trained a forest on 30k images where these

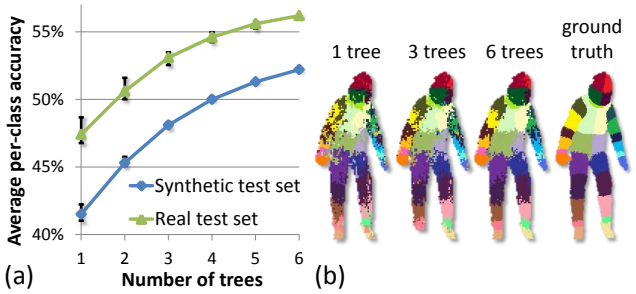

Figure 5. **Number of Trees.** (a) Quantitative results. (b) Qualitative results.

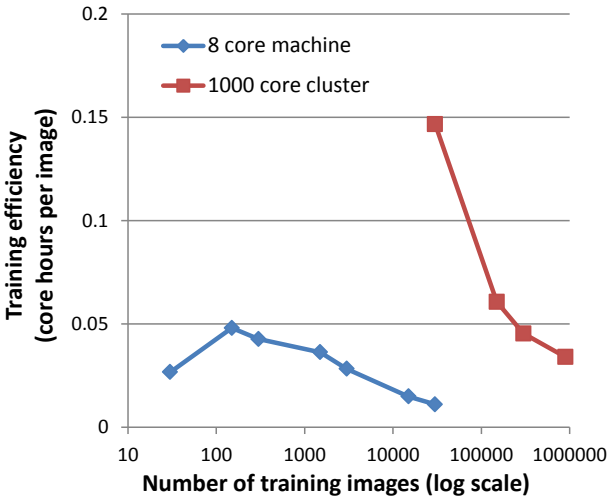

Figure 6. **Training Efficiency.** Curves shown are for 3 trees each to depth 20 with 2000 features and 50 thresholds tested per node.

parts are merged. This results in a total of 17 body parts including 16 'direct-joint' parts and 1 merged part. The forest gave an mAP of 0.554, considerably worse than the 0.685 for the original 31 parts. The reason for the drop is apparent in the resulting images: the direct-joint parts bleed a lot into the merged part, hurting localization. It appears that the standard entropy objective used by random classification forests is unable to separate the merged part from the 16 direct-joint parts, probably due to the imbalance of their contributions to the objective function. Our carefully designed body parts instead allow good localization giving high accuracy even using standard RFs.

Training efficiency. We compare in Fig. 6 the efficiency of training using a single machine and a large cluster. The efficiency of the distributed training algorithm is comparable to a single machine implementation, and improves considerably with the number of images. Various parameters of the distribution could be adjusted to give a further efficiency for a given number of images. Of course, the absolute training time is also dramatically reduced.

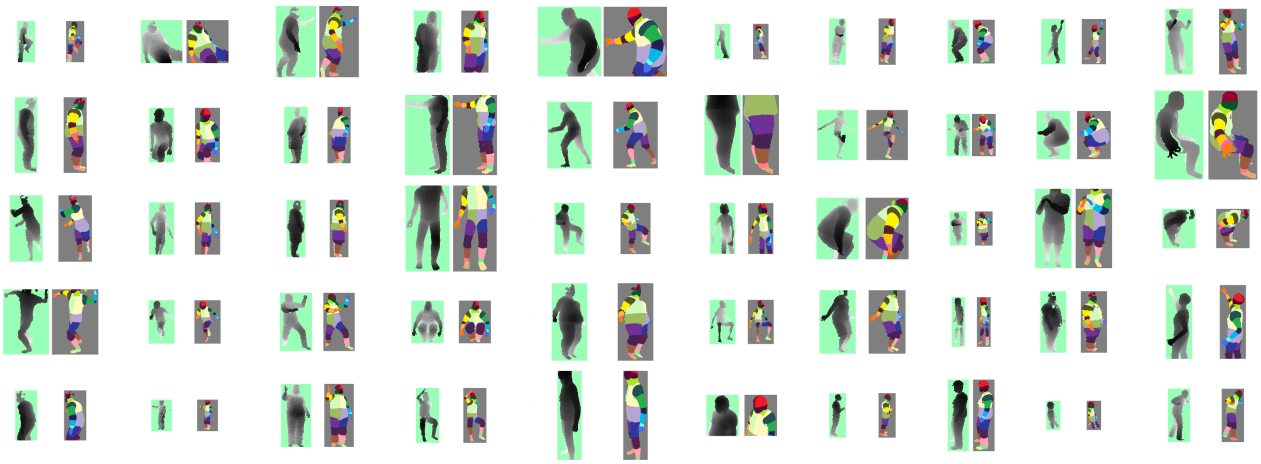

Figure 3. **Example training images.** Pairs of depth image and ground truth body parts. See also Figure 2 in the main paper.

Number of features and thresholds. Fig. 7 shows the effect of the number of candidate features θ and thresholds τ used during tree training on test classification accuracy. Most of the gain occurs up to 500 features and 20 thresholds. In the easier real test set the effects are less pronounced. These results use 5k images for each of 3 trees to depth 18.

Further qualitative results. We show more classification and joint prediction results in Fig. 8 on real images from our structured light depth camera. In Fig. 9 we show more results on the Stanford test set from their time of flight camera.

Example chamfer matches. We illustrating in Fig. 10 the quality of chamfer matches obtained for the comparison in the main paper using 130k exemplars.

Tree node visualization. In Fig. 11 we visualize how the split functions recursively partition the space of human body appearances as we descend the tree. Note how, as one descends the tree, the patches become more specific, indicating how the tree split nodes are separating different body parts by distinguishing different appearances.

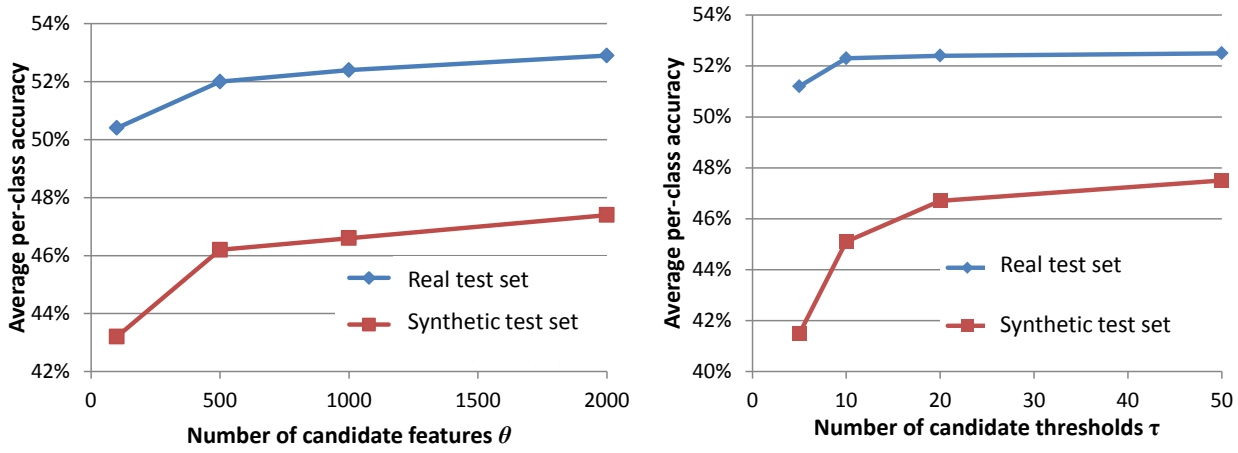

Figure 7. **Effect of the number of features and thresholds on test accuracy.**

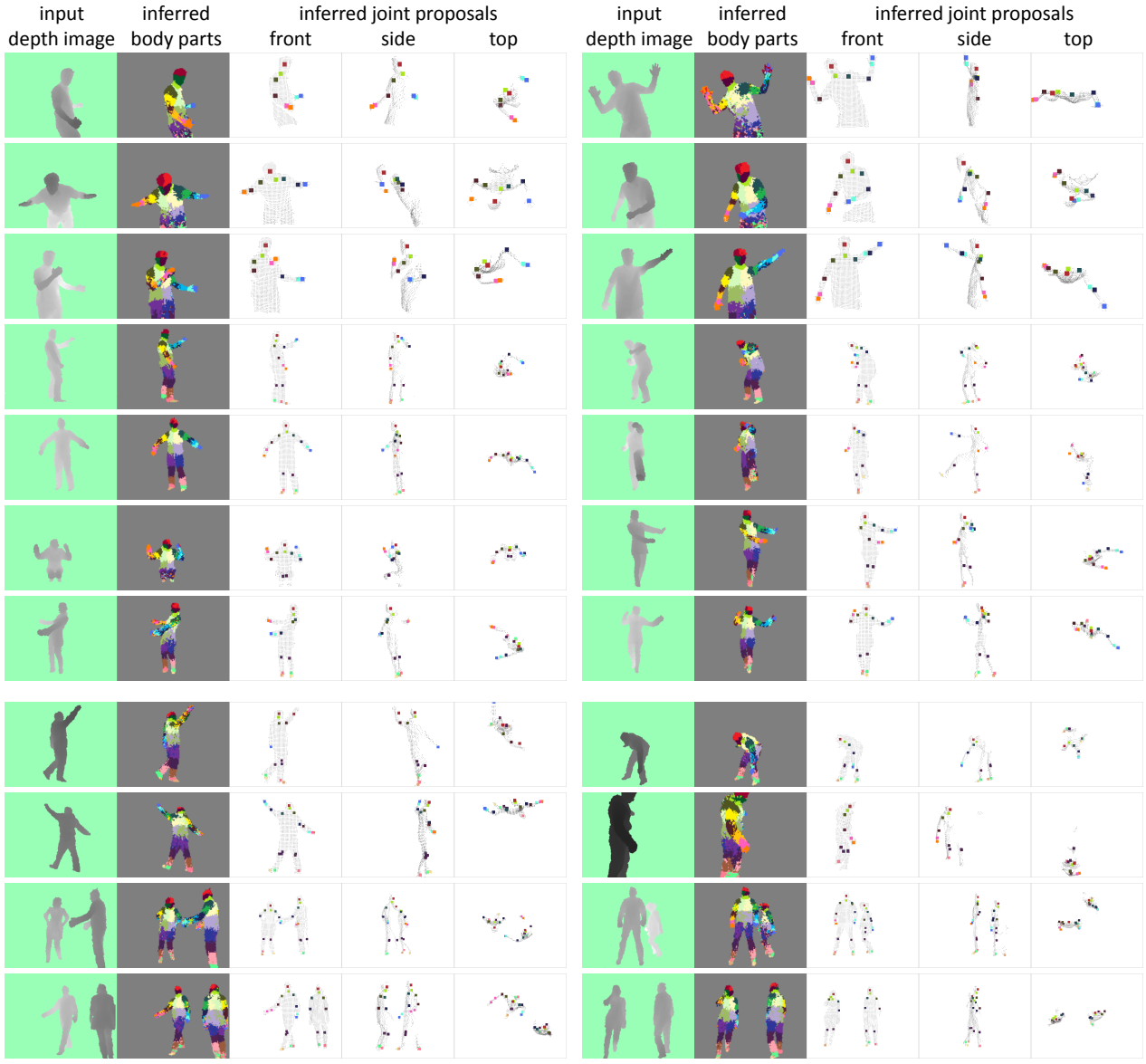

Figure 8. **Qualitative results on real data.** The top 7 rows use the $\pm 120^\circ$ classifier and show for each joint the most confident proposal, if above a fixed threshold. The bottom 4 rows use the full 360° classifier and show the top 4 proposals above the threshold for each joint. These results include six different individuals (four male, two female) wearing varied clothing at different depths. The system can even reliably classify and produce accurate joint proposals for images containing multiple people. See also the supplementary video.

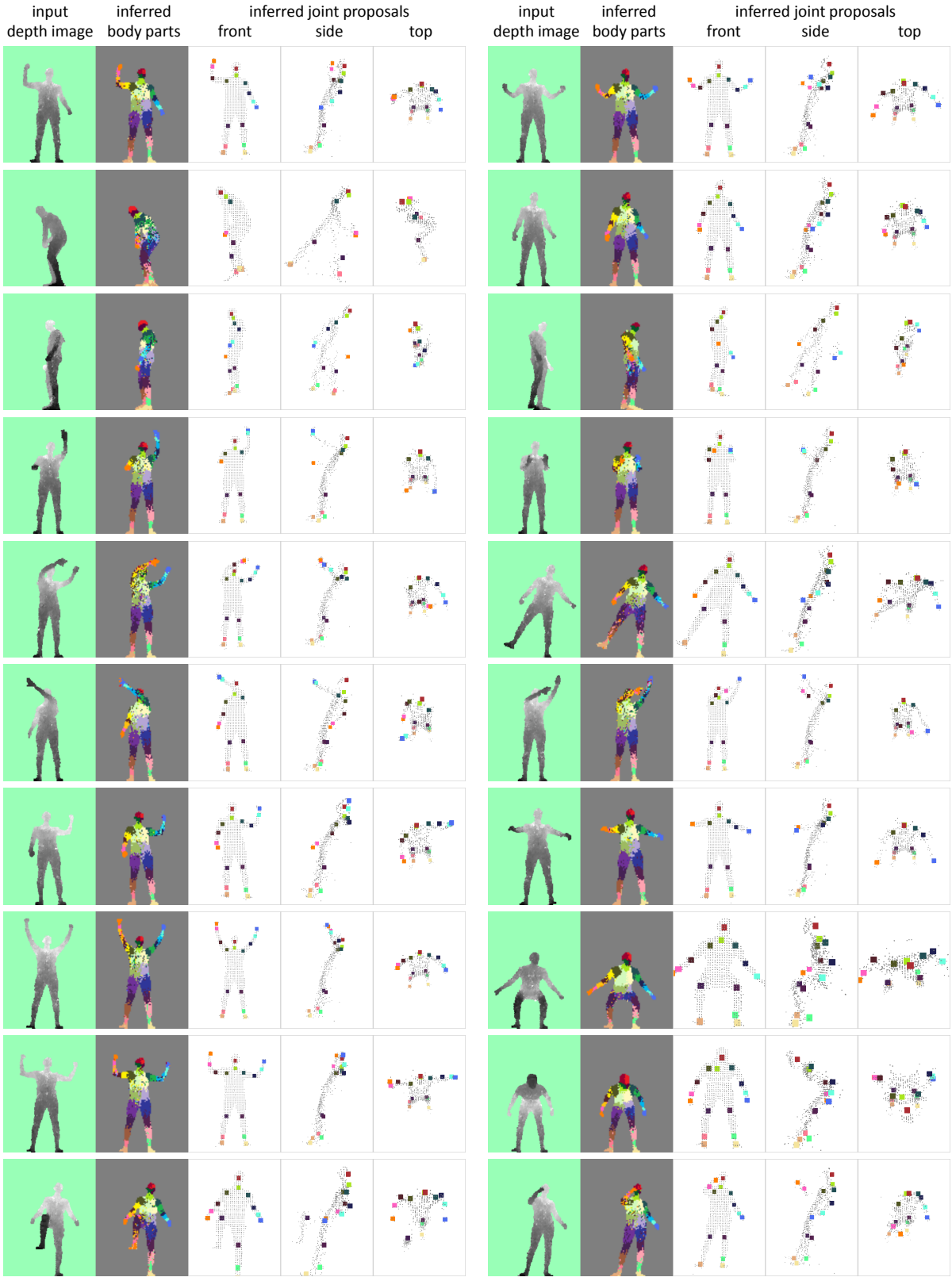

Figure 9. **Example results from the Ganapathi *et al*. test set.** This test set contains one actor from a low fixed camera viewpoint (which makes the actor appear to be leaning backwards). Our system can accurately localize body joints despite being designed for a different depth camera. Quantitative comparison in main paper.

Figure 10. **Chamfer Matches.** In each pair, the left is the test image, and the right is the same test image overlaid with the outline of the nearest neighbor exemplar chamfer match. Note the sensible, visually similar matches obtained using 130k exemplars. However, as demonstrated in the main paper, joint localization is imprecise since the whole skeleton is matched at once.

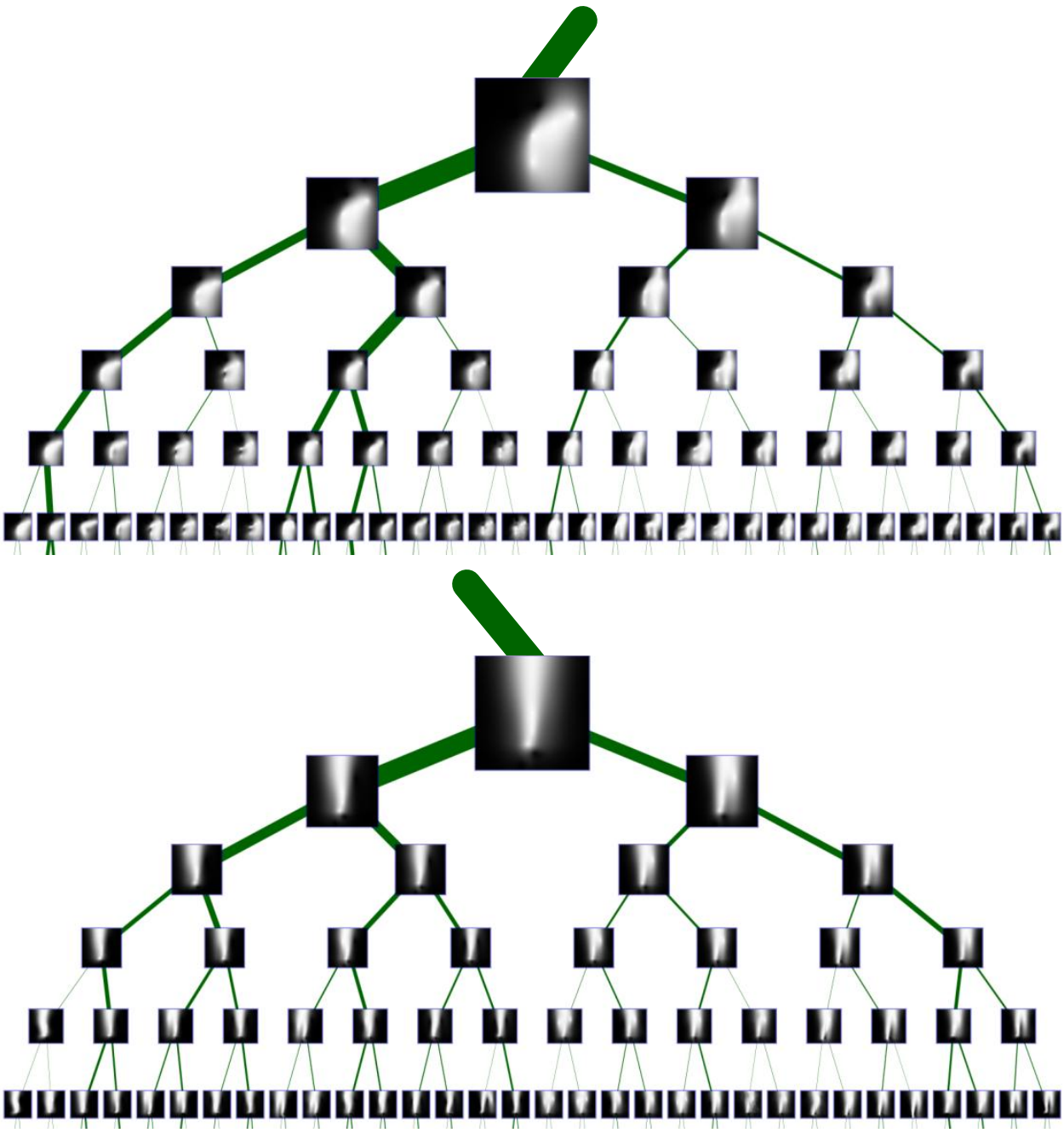

Figure 11. **Visualization of a trained decision tree.** Two separate subtrees are shown. A depth image patch centered on each pixel is taken, depth normalized, and binarized to a foreground/background silhouette. The patches are averaged across all pixels that reached any given tree node. The thickness of the edges joining the tree nodes is proportional to the number of pixels, and here shows fairly balanced trees. All pixels from 15k images are used to build the visualization shown.