a more direct optimization function would attempt to measure the $argmax$ of both heat-maps and therefore directly minimize the final (x, y) prediction. However, since the $argmax$ function is not differentiable we instead reformulate the problem as a regression to a set of target heat-maps and minimize the distance to those heat-maps.

5. Results

Our ConvNet architecture was implemented within the Torch7 [6] framework and evaluation is performed on the FLIC [20] and MPII-Human-Pose [1] datasets. The FLIC dataset consists of 3,987 training examples and 1,016 test examples of still scenes from Hollywood movies annotated with upper-body joint labels. Since the poses are predominantly front-facing and upright, FLIC is considered to be less challenging than more recent datasets. However the small number of training examples makes the dataset a good indicator for generalization performance. On the other-hand the MPII dataset is very challenging and it includes a wide variety of full-body pose annotations within the 28,821 training and 11,701 test examples. For evaluation of our model on the FLIC dataset we use the standard PCK measure proposed by [20] and we use the PCKh measure of [1] for evaluation on the MPII dataset.

Figure 9 shows the PCK test-set performance of our coarse heat-map model (Section 3.1) when various amounts of pooling are used within the network (keeping the number of convolution features constant). Figure 9 results show quite clearly the expected effect of coarse quantization in (x, y) and therefore the impact of pooling on spatial precision; when more pooling is used the performance of detections within small distance thresholds is reduced.

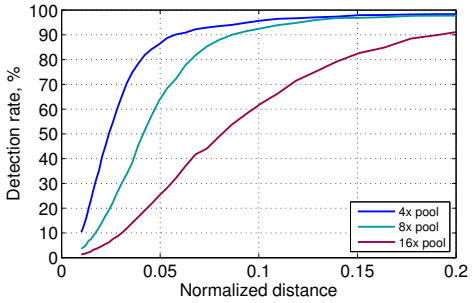

Figure 9: Pooling impact on FLIC test-set Average Joint Accuracy for the coarse heat-map model

For joints where the ground-truth label is ambiguous and difficult for the human mechanical-turkers to label, we do not expect our cascaded network to do better than the expected variance in the user-generated labels. To measure this variance (and thus estimate the upper bound of performance) we performed the following informal experiment:

	Face	Shoulder	Elbow	Wrist
Label Noise (10 images)	0.65	2.46	2.14	1.57
This work 4x (test-set)	1.09	2.43	2.59	2.82
This work 8x (test-set)	1.46	2.72	2.49	3.41
This work 16x (test-set)	1.45	2.78	3.78	4.16

Table 1: σ of (x, y) pixel annotations on FLIC test-set images (at 360×240 resolution)

we showed 13 users 10 random images from the FLIC training set with annotated ground-truth labels as a reference so that the users could familiarize themselves with the desired anatomical location of each joint. The users then annotated a consistent set of 10 random images from the FLIC test-set for the face, left-wrist, left-shoulder and left-elbow joints. Figure 10 shows the resultant joint annotations for 2 of the images.

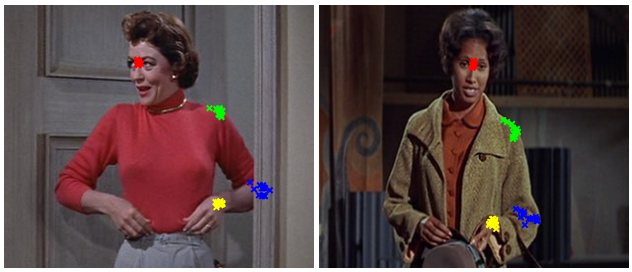

Figure 10: User generated joint annotations

To estimate joint annotation noise we calculate the standard deviation (σ) across user annotations in x for each of the 10 images separately and then average the σ across the 10 sample images to obtain an aggregate σ for each joint. Since we down-sample the FLIC images by a factor of 2 for use with our model we divide the σ values by the same down-sample ratio. The result is shown in Table 1.

The histogram of the coarse heat-map model pixel error (in the x dimension) on the FLIC test-set when using an 8x internal pooling is shown in Figure 11a (for the face and shoulder joints). For demonstration purposes, we quote the error in the pixel coordinates of the input image to the network (which for FLIC is 360×240), not the original resolution. As expected, in these coordinates there is an approximately uniform uncertainty due to quantization of the heat-map within -4 to +4 pixels. In contrast to this, the histogram of the cascaded network is shown in Figure 11b and is close to the measured label noise[1].

PCK performance on FLIC for face and wrist are shown in Figures 12a and 12b respectively. For the face, the per-

[1]When calculating σ for our model, we remove all outliers with error $>$ 20 and error < -20. These outliers represent samples where our weak spatial model chose the wrong person's joint and so do not represent an accurate indication of the spatial accuracy of our model.