# Deep Fully-Connected Part-Based Models for Human Pose Estimation

**(SUPPLEMENTARY MATERIAL)**

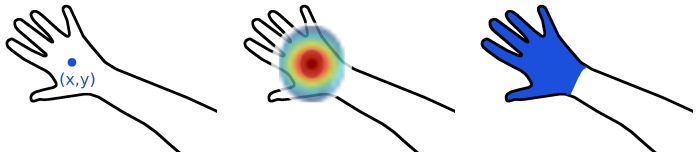
**Rodrigo de Bem** RODRIGO@ROBOTS.OX.AC.UK
*University of Oxford, UK and Federal University of Rio Grande, Brazil*
**Anurag Arnab** AARNAB@ROBOTS.OX.AC.UK
*University of Oxford, UK*
**Stuart Golodetz** STUART.GOLODETZ@ENG.OX.AC.UK
*University of Oxford, UK*
**Michael Sapienza** M.SAPIENZA@SAMSUNG.COM
*Think Tank Team, Samsung Research America, Mountain View, USA*
**Philip Torr** PHILIP.TORR@ENG.OX.AC.UK
*University of Oxford, UK*

**Editors:** Jun Zhu and Ichiro Takeuchi

Here we present further details of our method and additional results on the MPII set.

## 1. CNN Part Detector and the Conditional Random Field

Among several options to represent body parts, we consider that heatmaps with Gaussian representations rest between two extremes, namely discrete 2D points (pose vector) and per-pixel annotations, as illustrated in Fig. 1. Such representation convey more information than discrete 2D coordinates, yet still are much cheaper to obtain than per-pixel part labelling. We proposed our simple weakly supervised and part type-specific strategy, based on annotated 2D joint locations (Johnson and Everingham, 2010; Andriluka et al., 2014; Gong et al., 2017), for constructing heatmaps. The proposed CRF fully connects all body parts in the multi-level appearance model. The pairwise relations are based on Gaussian kernels that measure the likelihood of a given displacement between each pair of parts in the model. The means $\bar{\mathbf{x}}_{p,p'}$ and the standard deviations $\Sigma_{p,p'}$ of these Gaussian kernels are defined, prior to the network training, with maximum likelihood estimation over the training set. In Fig. 2 we show some of the learned parameters for the MPII dataset. In each graph, a sample position for one given part is defined to be at the centre of a $256 \times 256$ frame. The continuous




(a) (b) (c)

Figure 1: Range of body parts representations: (a) single 2D discrete point representation; (b) dense Gaussian representation; (c) dense per-pixel representation. Heatmaps with Gaussian representation stand on the central position of the spectrum. It takes into account inherent uncertainties about the location of the body parts and conveys much more spatial information that the point representation. However, with our weak annotation strategy, it is much cheaper to obtain in terms of annotation cost than the per-pixel labelling.

lines show the normalized expected displacements calculated between each other body part and the sample part in the center of the frame, whereas the dashed ellipses show the computed normalized standard deviations. These parameters, and consequently the Gaussian kernels, relate directly to the message-passing step of the mean-field inference, which is interpreted as follows; given a sample body part (e.g. the central ones in Fig. 2), at each mean-field iteration it receives messages from all the other parts conveying their expectations about its position. All these expectations are combined with the unary energies through the learnable weights and compatibility matrices. It is important to notice that Fig. 2 only shows the part in the central position of the frames "receiving messages", however for each body part at each possible location of the image there is a corresponding binary random variable, thus the messages are in fact exchanged efficiently between all these random variables at each mean-field iteration.

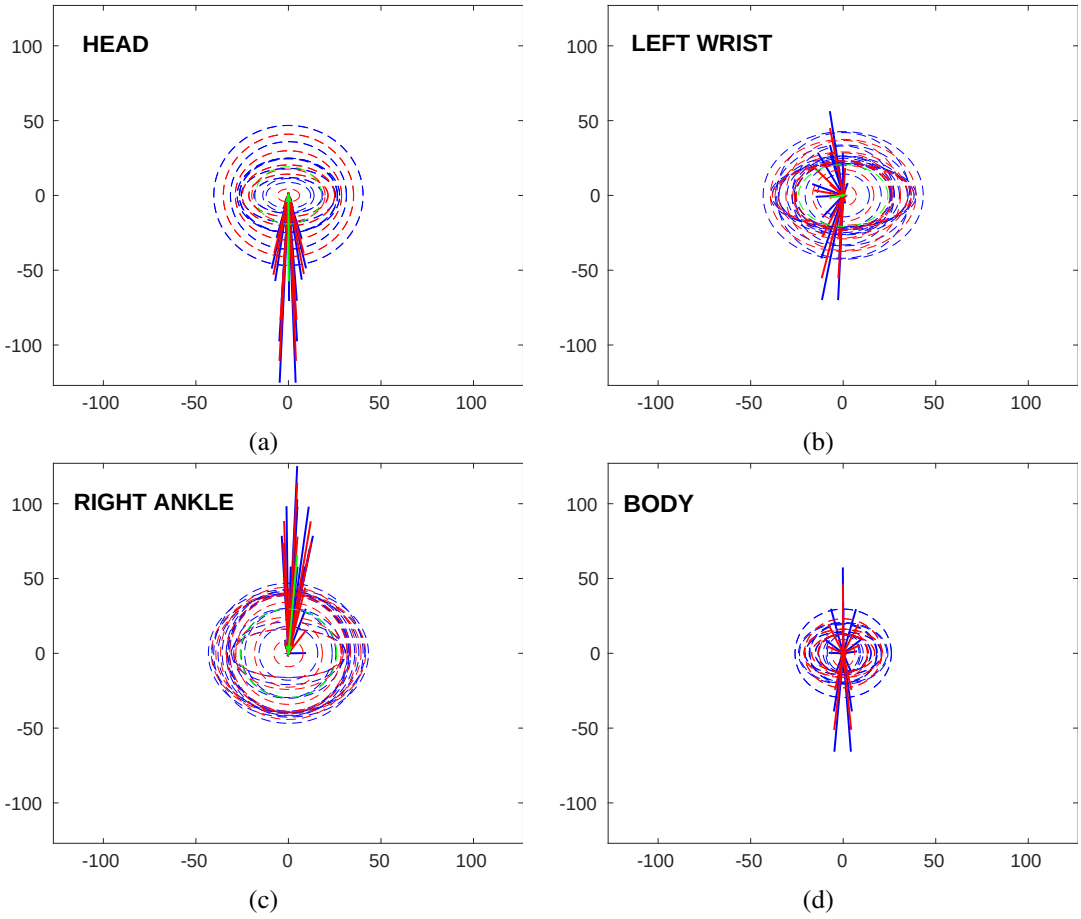



Figure 2: **Spatial priors:** mean displacements and standard deviations between sample parts (i.e. *head*, *left wrist*, *right ankle* and *body*) and all other parts in the model for the MPII dataset. Parameters learned prior to the network training through maximum likelihood estimation. Displacements and standard deviation ellipses between the sample parts and other joints are showed in blue, while the analogous parameters are showed in red for rigid parts and in green for the body. Note, for instance, that for the MPII dataset the mean displacement of the head w.r.t. all other parts corresponds to a person standing and looking towards the camera. All parameters are normalized w.r.t. the size of the frame. Best viewed in colour.

## 2. Additional Results on MPII Dataset

Figure 3: Sample pose predictions for the MPII dataset.

(a) Heatmaps predicted for a person standing in front of the camera.

Figure 4: Samples of the heatmaps predicted for images from the MPII test set. From the top left to the bottom right we have: *original image with the limbs superimposed, heat top, neck, thorax, pelvis, right shoulder, right elbow, right wrist, right hip, right knee, right ankle, left shoulder, left elbow, left wrist, left hip, left knee, left ankle, head, torso, right upper arm, right lower arm, right upper leg, right lower leg, left upper arm, left lower arm, left upper leg, left lower leg, whole body, background.*

## References

Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *CVPR*, 2014.

Ke Gong, Xiaodan Liang, Dongyu Zhang, Xiaohui Shen, and Liang Lin. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In *CVPR*, 2017.

Sam Johnson and Mark Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *BMVC*, 2010.