Bottom-Up and Top-Down Reasoning with Hierarchical Rectified Gaussians

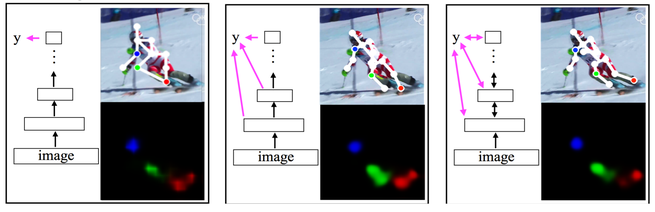
Peiyun Hu, Deva Ramanan

Carnegie Mellon
THE ROBOTICS INSTITUTE

IEEE 2016 Conference on Computer Vision and Pattern Recognition
CVPR 2016

Core ideas

- Human vision suggests that feedforward processing suffices for *vision-at-a-glance* tasks (scene recognition), but feedback is needed for *vision-with-scrutiny* (fine-grained spatial understanding) [Hochstein and M. Ahissar, Neuron 02]
- We incorporate feedback into CNNs by unrolling inference on a hierarchical probabilistic model
- Our model adds top-down feedback "for-free", *without increasing the number of parameters* in standard networks.
- Near state-of-the-art results for facial and human landmark localization



Predicting heatmaps with CNNs

Fully-convolutional (skip) CNNs
Long et al, CVPR 2015

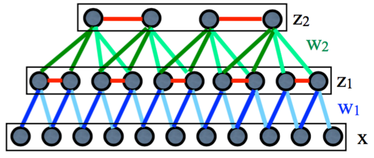
Our top-down model
No increase in parameters

Heatmap predictions improve by adding lower-level features (through skip connections) and top-down feedback (that fixes poor localization of knee and left/right ankle ambiguity).

Top-down feedback is particularly helpful when localizing occluded landmarks

Approach

- We make use of Rectified Gaussian models [Socci et al, NIPS 98], a marriage of Boltzmann machines and Gaussians.
- Socci et al demonstrate that MAP inference on such models corresponds to a quadratic program (QP).
- We show that coordinate QP updates can be unrolled into a rectified recurrent net *that naturally incorporates bottom-up and top-down processing*.
- We train our inference engine with backprop using standard CNN packages.



$$S(z) = \frac{1}{2} z^T W z + b^T z$$

$$P(z) \propto e^{S(z)}$$

Boltzmann: $z_i \in \{0, 1\}, w_{ii} = 0$

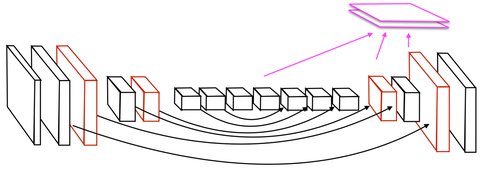
Gaussian: $z_i \in R, -W$ is PSD

Rect. Gaussian: $z_i \in R^+, -W$ is copositive

$$z_i[u] = \max(0, b_i + top_i[u] + bot_i[u]) \quad \text{where}$$

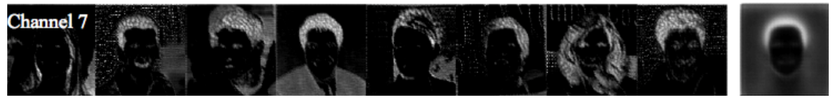
$$top_i[u] = \sum_\tau w_{i+1}[\tau] z_{i+1}[u - \tau]$$

$$bot_i[u] = \sum_\tau w_i[\tau] z_{i-1}[u + \tau]$$



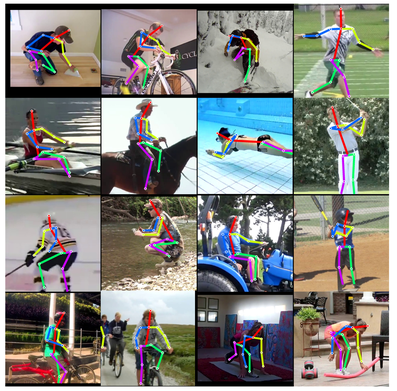
Feedforward activations for layer 2, channel 7

avg



Activations after feedback (appear to focus more on hair)

avg

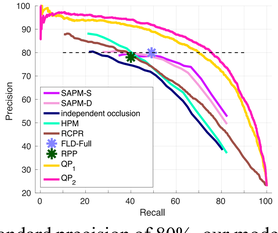


MPII Human Pose

Caltech Occluded Faces (red for predicted invisibility)

	Head	Shou	Elb	Wri	Hip	Kne	Ank	Upp	Full
GM [10]	-	36.3	26.1	15.3	-	-	-	25.9	-
ST [37]	-	38.0	26.3	19.3	-	-	-	27.9	-
YR [52]	73.2	56.2	41.3	32.1	36.2	33.2	34.5	43.2	44.5
PS [35]	74.2	49.0	40.8	34.1	36.5	34.4	35.1	41.3	44.0
TB [45]	96.1	91.9	83.9	77.8	80.9	72.3	64.8	**84.5**	82.0
$\mathbf{QP}_1$	94.3	90.4	81.6	75.2	80.1	73.0	68.3	82.4	81.1
$\mathbf{QP}_2$	95.0	91.6	83.0	76.6	81.9	74.5	69.5	83.8	**82.4**

Our top-down model (QP_2) on par with approaches that add specialized modules for top-down spatial constraints [TB]



At standard precision of 80%, our model (QP_2) nearly doubles the recall of prior art [FLD] for visibility prediction

AFLW

Visit our project site at: https://peiyunh.github.io/rg-mpii/