Efficient Object Localization Using Convolutional Networks

Jonathan Tompson, Ross Goroshin, Arjun Jain, Yann LeCun, Christoph Bregler
New York University

tompson/goroshin/ajain/lecun/bregler@cims.nyu.edu

Recent state-of-the-art performance on human-body pose estimation has been achieved with Deep Convolutional Networks (ConvNets). Traditional ConvNet architectures include pooling and sub-sampling layers which reduce computational requirements, introduce invariance, and prevent over-training. These benefits of pooling come at the cost of reduced localization accuracy. In this paper we introduce a novel architecture which includes an efficient 'position refinement' model that is trained to estimate the joint offset location within a small region of the image. This refinement model is jointly trained in cascade with a state-of-the-art ConvNet model [3] to achieve improved accuracy in human joint location estimation. An overview of the detection archiecture is shown in Figure 1.

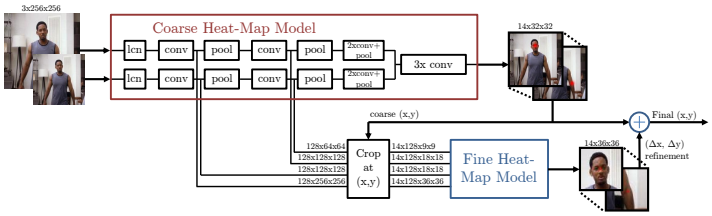


Figure 1: Overview of our Cascaded Architecture

Inspired by the work of Tompson et al. [3], we use a multi-resolution ConvNet architecture (Figure 2) to implement a sliding window detector with overlapping contexts to produce a coarse heat-map output. This network outputs a low resolution, per-pixel heat-map, which represents the likelihood of a joint occurring in each spatial location.

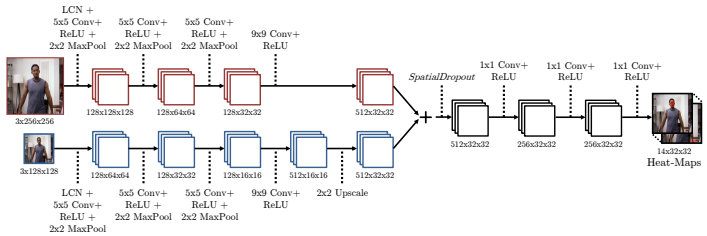


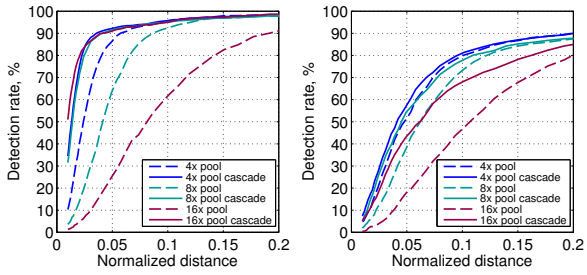
Figure 2: Multi-resolution Sliding Window Detector With Overlapping Contexts

If adjacent pixels within feature maps are strongly correlated (as is normally the case in early convolution layers) then we show that i.i.d. dropout will not regularize the activations and will otherwise just result in an effective learning rate decrease. Instead we introduce *SpaitalDropout* - a modified dropout implementation - which allows us to improve upon the model of [3] by promoteing activation independence across feature maps.

We use the architecture of Figure 1 as a platform to discuss and empirically evaluate the role of Max-pooling layers in convolutional architectures for dimensionality reduction and improving invariance to noise and local image transformations. The results in Figure 3 shows that our cascaded archiecture is able to recover spatial accuracy on the face (Figure 3a), and to a lesser extent on the wrist joint (Figure 3b), even in the presence of large pooling and sub-sampling in the coarse heat-map model. For this evaluation we use the standard PCK measure [2] on the FLIC dataset [2].

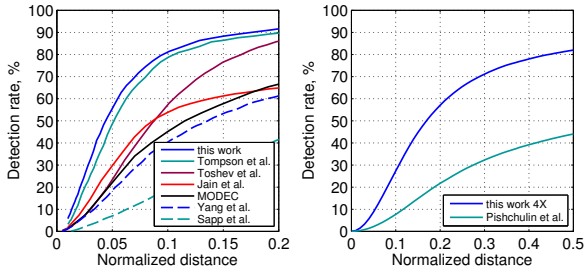
Our descriminative archiecture is able to outperform existing state-of-the-art on the FLIC and MPII [1] datasets. Figures 4a and 4b shows the PCK and PCKh results on the FLIC and MPII datasets respectively. Figure 5 shows a selection of joint predictions on the MPII test-set.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.



(a) Face (b) Wrist

Figure 3: Performance improvement from cascaded model



(a) FLIC (b) MPII

Figure 4: Our model performance on two standard datasets compared to state-of-the-art

Additionally, we carried out an informal user study to estimate the statistical variation of human annotators on the FLIC dataset. From this experiement we can conclude that the UV error variance of our detector approaches the variance of human annotations.

Figure 5: Our Model's Predicted Joint Positions on the MPII-human-pose database test-set[1]

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

[2] B. Sapp and B. Taskar. Modec: Multimodal decomposable models for human pose estimation. In *CVPR*, 2013.

[3] Jonathan Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler. Join training of a convolutional network and a graphical model for human pose estimation. *NIPS*, 2014.