Finding Tiny Faces
Peiyun Hu, Deva Ramanan
Carnegie Mellon University

How to handle extreme scales? Since pre-trained networks are tuned for objects of characteristic scales, how do we extend them to extreme scales?

a) A detector tuned for 50x40 faces is 6.3% more accurate than one tuned for 25x20 on finding 25x20 faces when applied on 2X upsampled images. A detector tuned for 125x100 faces is 5.6% more accurate than one tuned for 250x200 on finding 250x200 faces when applied on 2X downsampled images.
b) There exists a natural regime for picking which resolution to build a detector at a given target resolution. For finding high-resolution faces (taller than 140px), build detectors at 0.5X resolution; for finding low-resolution (shorter than 40px) faces, build detectors at 2X resolution. For sizes in between, build at the original resolution.

State-of-the-art results on WIDER Face and FDDB

a) Precision-recall curves on WIDER Face test set (hard only). Our approach achieves state-of-the-art performance on all subsets (easy, medium, hard). In particular, ours outperforms the prior art by 17.6% on the hardest subset.
b) ROC curves on FDD test set. Our out-of-the-box detector (HR) achieves state-of-the-art on discrete score. With post-hoc elliptical regression, our approach (HR-ER) achieves state-of-the-art on continuous score as well.

Context in human vision

a) We visualize a low-resolution (top) and a high-resolution (bottom) human face. One does not need context to recognize the high-resolution face, while the low-resolution face is dramatically unrecognizable without its context.
b) We quantify this observation with an human experiment, where users are asked to classify true and false positive faces generated by our proposed detector. Adding proportional context provides a small improvement on medium and large faces but insufficient for low-resolution (S and XS) faces. Adding a fixed contextual window of 300 pixels dramatically reduces error on low-resolution faces by 20%. This suggests that context is crucial for human to recognize low-resolution faces and it can be modeled in a scale-invariant manner.

Core ideas

a) How to model scale invariance: Instead of an “one-size-fits-all” detector, we train separate detectors with each for a different scale in an multi-task fashion.
b) How to generalize pre-trained networks: We interpolate images to extend pre-trained features tuned for objects of a typical scale to ones of novel scales.
c) How best to encode context: We encode massively large amount of context with “foveal” descriptors and demonstrate the “foveal” structure is crucial for detecting low-resolution faces.