Combination Features and Models for Human Detection

Yunsheng Jiang and Jinwen Ma
Department of Information Science, School of Mathematical Sciences and LMAM, Peking University, Beijing, 100871, China.

In the past several years, many existing features/models have achieved impressive progress for human detection, like the person Grammar model [4], Poselet model [1], etc. However, their performances are still limited by the biases rooted in their self-structures, that is, a particular kind of feature/model may work well for some types of human bodies, but not for all the types. To tackle this problem, we try to combine certain complementary features/models together with effective organization/fusion methods.

HOG-III features

We extend the first-order gradients in the HOG features [3] to a collection of gradients with three different orders, augmented with zero-order and second-order gradients which correspond to color and bar-shape information respectively. The resultant features are denoted as *HOG-III* features.

Color features: The zero-order gradient is the image itself. We convert the RBG image to HSI color space, to extract the pure color information (H and S). We map the *(hue, saturation)* to the *(orientation, magnitude)* of the first-order gradient, and follow the entire HOG computation process to obtain the so-called Histograms of Color (HoC) features.

Bar-shape features: The second-order gradient is related to bar-shape information [2], which may be valuable for human detection, since: (1) the mammalian visual system seems to have bar-like receptive fields [6]; (2) articulated objects like human bodies can be modelled as connected bar-and-blob structures [7]. By replacing the first-order gradients in HOG with the second-order gradients, we get the Histograms of Bar-shape (HoB) features.

Both HoC and HoB have similar structures (cell-based histograms) with HOG. We concatenate HOG, HoC and HoB together to form our HOG-III features. Note that the contrastive-sensitive components in HOG features are excluded from HOG-III features.

***Weighted*-NMS based model fusion method**

We combine the detections from different models with our newly proposed *weighted*-NMS fusion algorithm, which enhances the probable true activations as well as suppresses the overlapped detections. The entire procedure of model fusion includes calibration step and fusion step.

Calibration step: The same detection score may have very different confidence levels in different models. This causes difficulties for the comparison between models. We need to calibrate the scores into the same criterion first. For each model, we plot the *threshold-precision* curve on the validation set, and measure the confidence level of a threshold score with its corresponding precision. Therefore, using the *precision* as a bridge, we can calibrate the scores from different models into the same criterion.

Fusion step: It is foreseeable that the two different models may output many overlapped detections. We eliminate these overlaps with the so-called *weighted*-NMS algorithm. First, we merge the detections from these two models and normalize their calibrated scores to the interval $[0,1]$ with sigmoid function. After that, we take into account each detection greedily, from high score to low score. If $(p_h, \tilde{s}_h)$ is a high-scored detection, and there exists a lower-scored detection $(p_l, \tilde{s}_l)$ which has *enough* overlap with $(p_h, \tilde{s}_h)$, then $(p_l, \tilde{s}_l)$ will be deleted, **AND**, the score $\tilde{s}_l$ will be *partially* absorbed into the score $\tilde{s}_h$ with a decay weight w_{hl} :

$$\tilde{s}_h \leftarrow \tilde{s}_h + w_{hl} \cdot \tilde{s}_l. \quad (1)$$

The score of the retained detection p_h is enhanced because that, if a "hypothesized object" can be detected by two different/complementary models, it is more likely to be a "true object". By this way we can enhance the probable true detections as well as eliminate the redundant overlaps. The decay weight should belong to $[0,1]$, and we simply set it as the *overlap* between the two corresponding detections: $w_{hl} = overlap(p_h, p_l) = \frac{area(p_h \bigcap p_l)}{area(p_h \bigcup p_l)}$. Note that if we fix this weight as $w_{hl} \equiv 0$, then the *weighted*-NMS algorithm degenerates to general NMS algorithms.

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

Model	VOC2007	VOC2010	VOC2012
Grammar	45.8	47.6	47.9
Poselet	47.0	48.5	48.1
LSVM-MDPM(V5)	43.2	45.2	44.5
Boosted-HOG-LBP	44.6	46.5	–
HSC	41.4	–	–
DDSMM	44.8	49.2	–
CN-HOG	44.0	43.3	–
Regionlet	43.4	43.5	–
(Grammar, HOG-III)	51.3	52.2	52.1
(G-P, HOG)	52.3	54.1	53.7
(G-P, HOG-III)	**55.5**	**57.2**	**57.0**

Table 1: Full results (AP%) on PASCAL VOC dataset for person-class. All the results here are obtained without using large auxiliary datasets or contextual information about other object categories.

The performance for human detection

We fuse the person Grammar model [4] and Poselet model [1] with our *weighted*-NMS algorithm, and denote the fusion model as *G-P Model*. Further, if we apply our HOG-III features to this G-P model, we obtain the integrated framework (G-P, HOG-III). Table 1 shows the full results on PASCAL VOC datasets. Our integrated framework (G-P, HOG-III) has gained a substantial advantage over the best model excluding ours, *e.g.*, an improvement of 8.5% over Poselet model on VOC2007 testset, 8.0% over DDSSM on VOC2010 testset, and 8.9% over Poselet model on VOC2012 testset.

Recently, the R-CNN model [5] has obtained impressive detection performances based on the deep CNN features. However, without regard to the large auxiliary datasets, high-level hardware or the long training & prediction time caused by R-CNN, we can also fuse R-CNN and Grammar model with our *weighted*-NMS algorithm. The resultant fusion model shows very good performance. For example, the AP of person class on VOC2007 is 51.3% for Grammar(HOG-III), 58.7% for R-CNN, and 65.2% for the fusion model of Grammar and R-CNN, which is indeed a significant improvement.

Extension to the whole VOC 20 classes

To investigate the generalization ability of the HOG-III features and *weighted*-NMS fusion algorithm, we extend them to the detection of the whole VOC 20 object categories. We use DPM [3] and R-CNN [5] for experiments, as they are are applicable to the whole object categories.

First, the HOG-III features still show good performances on the whole 20 classes, though not as impressive as that for person class. For example, in the framework of DPM, the mean AP on VOC2007 testset is 33.7% for HOG, 34.3% for HOG-LBP, 34.3% for HSC, 34.8% for CN-HOG, while 35.0% for the proposed HOG-III features.

Second, we fuse DPM and R-CNN with the *weighted*-NMS algorithm, and the fusion model also gains competitive improvements on the whole 20 classes. Specifically, the mean AP on VOC2007 testset is 33.7% for DPM, 58.4% for R-CNN, while 60.5% for the fusion model of DPM and R-CNN.

[1] L. Bourdev, S. Maji, T. Brox, and J. Malik. Detecting people using mutually consistent poselet activations. In *Proc. ECCV*, pages 168–181. 2010.

[2] N. Dalal. *Finding people in images and videos*. PhD thesis, Institut National Polytechnique de Grenoble-INPG, 2006.

[3] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE T-PAMI*, 32(9): 1627–1645, 2010.

[4] R. B. Girshick, P. F. Felzenszwalb, and D. A. McAllester. Object detection with grammar models. In *Proc. NIPS*, pages 442–450, 2011.

[5] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proc. CVPR*, 2014.

[6] D. H. Hubel. *Eye, brain, and vision.* Scientific American Library/Scientific American Books, 1995.

[7] S. Ioffe and D. Forsyth. Mixtures of trees for object recognition. In *Proc. CVPR*, volume 2, pages 180–185, 2001.