Ten years of pedestrian detection, what have we learned? Supplementary material

Rodrigo Benenson   Mohamed Omran   Jan Hosang   Bernt Schiele

Max Planck Institut for Informatics
Saarbrücken, Germany
firstname.lastname@mpi-inf.mpg.de

1 Reviewing the effect of features

The idea behind the experiments in section 4.1 of the main paper is to demonstrate that, within a single framework, varying the features can replicate the jump in detection performance over a ten-year span ($2004-2014$), i.e. the jump in performance between `VJ` and the current state-of-the-art.

See figure 1 for results on INRIA and Caltech-USA of the following methods (all based on `SquaresChnFtrs`, described in section 4 of the paper):

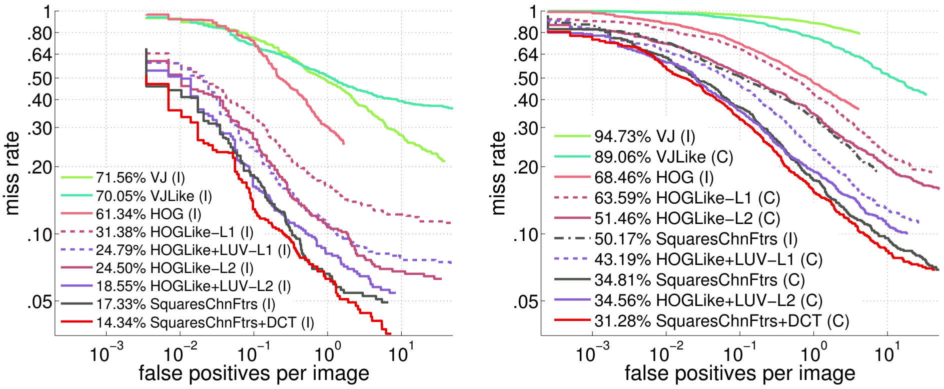
`VJLike` uses only the luminance colour channel, emulating the original `VJ` [1]. We use 8 000 weak classifiers to compensate for the weak input feature, only square pooling regions, and level-2 trees to emulate the Haar wavelet-like features used by `VJ`.

`HOGLike-L1/L2` uses 8×8 pixel pooling regions, 6 oriented gradients, 1 gradient magnitude, and level 1/2 decision trees (1/3 threshold comparisons respectively). A level-1 tree emulates the non-linearity in the original `HOG`+linear SVM features [2].

`HOGLike+LUV` is identical to `HOGLike,` but with additional LUV colour channels (10 feature channels total).

`SquaresChnFtrs` is the baseline described in the beginning of the experiments section (§4). It is similar to `HOGLike+LUV` but the size of the square pooling regions is not restricted.

`SquaresChnFtrs+DCT` is inspired by [3]. We expand the ten HOG+LUV channels into 40 channels by convolving each of the 10 channels with three DCT (discrete cosine transform) filters (7×7 pixels), and storing the absolute value of the filter responses as additional feature channels. The three DCT basis functions we use as 2d-filters correspond to the lowest spatial frequencies. We name this variant `SquaresChnFtrs+DCT` and it serves as reference point for the performance improvement that can be obtained by increasing the number of channels.



(a) INRIA test set (b) Caltech-USA reasonable test set

Figure 1: Effect of features on detection performance. (I)/(C) indicates using INRIA/Caltech-USA training set respectively.

2 Complementarity of approaches

Table 1 contains the detailed results of combining different approaches with a strong baseline, related to section 4.2 of the main paper. `Katamari-v1` combines all three listed approaches with `SquaresChnFtrs`. We train and test on the Caltech-USA dataset. It can be noticed that the obtained improvement is very close to the sum of individual gains, showing that these approaches are quite complementary amongst each other.

Table 1: Complementarity between different extensions of the `SquaresChnFtrs` strong baseline. Results in MR (lower is better). Improvement in MR percent points. Expected improvement is the direct sum of individual improvements.

Method	Results	Improvement	Expected improvement
`SquaresChnFtrs`	34.81%	-	-
`+DCT`	31.28%	3.53	-
`+SDt` [4]	30.34%	4.47	-
`+2Ped` [5]	29.42%	5.39	-
`+DCT+2Ped`	27.40%	7.41	8.92
`+SDt+2Ped`	26.68%	8.13	9.86
`+DCT+SDt`	25.24%	9.57	8.00
`Katamari-v1`	*22.49%*	12.32	13.39

References

1. Viola, P., Jones, M., Snow, D.: Detecting pedestrians using patterns of motion and appearance. In: CVPR. (2003)
2. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR. (2005)
3. Nam, W., Dollár, P., Han, J.H.: Local decorrelation for improved detection. In: arXiv. (2014)
4. Park, D., Ramanan, D., Fowlkes, C.: Multiresolution models for object detection. In: ECCV. (2010)
5. Ouyang, W., Wang, X.: Single-pedestrian detection aided by multi-pedestrian detection. In: CVPR. (2013)