Blocks that Shout: Distinctive Parts for Scene Classification

Mayank Juneja[1] Andrea Vedaldi[2] C. V. Jawahar[1] Andrew Zisserman[2]

[1] Center for Visual Information Technology, International Institute of Information Technology, Hyderabad, India

[2] Department of Engineering Science, University of Oxford, United Kingdom

{mayank.juneja@research.,jawahar@}iiit.ac.in {vedaldi,az}@robots.ox.ac.uk

Abstract

The automatic discovery of distinctive parts for an object or scene class is challenging since it requires simultaneously to learn the part appearance and also to identify the part occurrences in images. In this paper, we propose a simple, efficient, and effective method to do so. We address this problem by learning parts incrementally, starting from a single part occurrence with an Exemplar SVM. In this manner, additional part instances are discovered and aligned reliably before being considered as training examples. We also propose entropy-rank curves as a means of evaluating the distinctiveness of parts shareable between categories and use them to select useful parts out of a set of candidates.

We apply the new representation to the task of scene categorisation on the MIT Scene 67 benchmark. We show that our method can learn parts which are significantly more informative and for a fraction of the cost, compared to previous part-learning methods such as Singh et al. *[4]. We also show that a well constructed bag of words or Fisher vector model can substantially outperform the previous state-of-the-art classification performance on this data.*

1. Blocks that shout: learning distinctive parts

In characterizing images of particular scene classes, *e.g.* a computer room, library, book store, auditorium, theatre, etc., it is not hard to think of distinctive parts: chairs, lamps, doors, windows, screens, etc., readily come to mind. In practice, however, a distinctive part is useful only if it can be detected automatically, preferably by an efficient and simple algorithm.

Learning a distinctive part means identifying a localized detectable entity that is informative for the task at hand (in our example discriminating different scene types). This is very challenging because (i) one does not know if a part occurs in any given training image or not, and (ii) when the part occurs, one does not know its location. We solve this problem of learning parts incrementally, which allows detecting occurrences of them before their appearance is fully learned. Our strategy for part-learning combines three ideas: seeding, expansion, and selection. Fig. 1 shows examples of the learned parts detected on the test set.

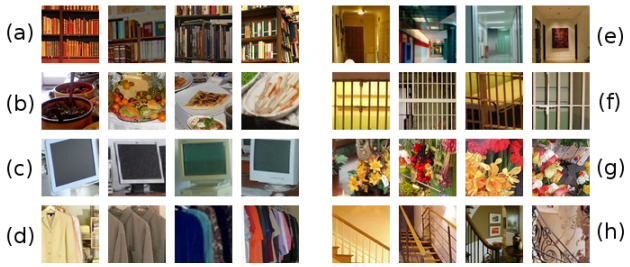


Figure 1. Example of occurrences of distinctive parts learned by our method from weakly supervised image data. These part occurrences are detected on the test data. (a) bookstore, (b) buffet, (c) computerroom, (d) closet, (e) corridor, (f) prisoncell, (g) florist, (h) staircase.

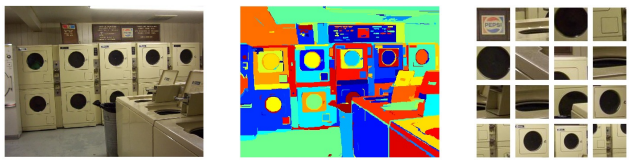
1.1. Seeding: proposing an initial set of parts

We use low-level image cues to identify a subset of image locations that are more likely to be centered around distinctive parts. In order to find promising locations we use image over-segmentations, and segment each training image into super-pixels at multiple scales. Superpixels of intermediate sizes, defined as the ones whose area is in the range 500 to 1,500 pixels, are retained (Fig. 2).

1.2. Expansion: learning part detectors

Learning a part detector requires a set of part exemplars, and these need to be identified in the training data. There is a special case in which a part detector can be learned without worrying about exemplar alignment: a training set consisting exactly of one part instance (Exemplar SVMs [2]). In practice, at each round of learning the current part model is used to rank blocks from images of the selected class and the highest scoring blocks are considered as further part oc-

Full version of the paper available at: `http://www.robots.ox.ac.uk/~vgg/publications/2013/Juneja13/`



(a) Original Image (b) Superpixels (c) Seed blocks

Figure 2. **Selecting seed blocks.** The super-pixels (b) suggest characteristic regions of the image, and blocks are formed for these.

currences. This procedure is repeated a set number of times (ten in the experiments), adding a small number of new part exemplars (ten) to the training set each time. All the part models obtained in this manner, including the intermediate ones, are retained and filtered by distinctiveness and uniqueness in Sect. 1.3. Figure 3 shows an example seed part on the left, and the additional part occurrences that are added to the training set during successive iterations of expansion. We use the LDA technique [1] to avoid the hard negative mining for each trained detector, which is very costly.

(a) Seed (b) Round 1 (c) Round 2 (d) Round 3 (e) Round 4 (f) Round 5

Figure 3. **Mining of part instances.** Seed block and the additional example blocks added to the positive training set.

1.3. Selection: identifying distinctive parts

Our notion of a discriminative block is that it should occur in many of the images of the class from which it is learnt, but not in many images from other classes. However, it is not reasonable to assume that parts (represented by blocks) are so discriminative that they only occur in the class from which they are learnt. In selecting the block classifiers we design a novel measure to capture this notion. We introduce *Entropy-Rank Curves (ER curves)* to measure the entropy of a block classifier at different ranks. An ER curve is similar to a Precision-Recall Curve (PR curve), with rank on the x-axis and entropy values on the y-axis. Analogously to Average Precision, we then take the Area Under Curve (AUC) for the ER graph as an overall measure of performance of a detector. The top scoring non-redundant detectors based on this measure are then retained.

2. Image representations

2.1. Bag of Parts

The part detectors developed in Sect. 1 are used to construct "bag of parts" image-level descriptors. In order to compute an image-level descriptor from the parts learned in Sect. 1, all the corresponding classifiers are evaluated densely at every image location at multiple scales. Part scores are then summarized in an image feature vector by using max-pooling, by retaining the maximum response score of a part in a region. The pooling is done in a spatial-pyramid fashion.

2.2. Bag of Words

A Bag of Words representation is computed using Dense RootSIFT descriptors and a number of different feature encodings are compared: (i) hard assignment (*i.e.* VQ) BoW; (ii) kernel-codebook encoding BoW; (iii) Locality-constrained Linear Coding (LLC) BoW; and (iv) Improved Fisher Vectors (IFV). Weak geometric information is retained in the descriptors by using spatial histogramming.

3. Learning and classification

Learning uses the PEGASOS SVM algorithm, a linear SVM solver. In order to use non-linear additive kernels instead of the linear one, the χ^2 explicit feature map is used. For the IFV encoding, we use square-root (Hellinger) kernel. The parameter C of the SVM (regularization-loss trade-off) is determined by 4-fold cross validation. For multi-class image classification problems, 1-vs-rest classifiers are learned. In this case, it was found beneficial to calibrate the different 1-vs-rest scores by fitting a sigmoid to them based on a validation set. Table 1 reports the performance of our methods on the MIT 67 indoor scene dataset [3].

Method	Acc. (%)	Mean AP (%)
Patches [4]	38.10	-
Patches + GIST + SP + DPM [4]	49.40	-
BoP	46.10	43.55
LLC	53.03	51.73
IFV	60.77	61.05
LLC + BoP	56.66	55.13
IFV + BoP	63.10	63.18

Table 1. Average classification performance (previous publications and this paper).

References

[1] B. Hariharan, J. Malik, and D. Ramanan. Discriminative decorrelation for clustering and classification. In *Proc. ECCV*, 2012.

[2] T. Malisiewicz, A. Gupta, and A. A. Efros. Ensemble of exemplar-svms for object detection and beyond. In *Proc. ICCV*, 2011.

[3] A. Quattoni and A. Torralba. Recognizing indoor scenes. In *Proc. CVPR*, 2009.

[4] S. Singh, A. Gupta, and A. A. Efros. Unsupervised discovery of mid-level discriminative patches. In *Proc. ECCV*, 2012.