Learning discriminative spatial representation for image classification

Gaurav Sharma
gaurav.sharma@{inrialpes,unicaen}.fr
LEAR
INRIA Grenoble Rhône-Alpes

Frederic Jurie
frederic.jurie@unicaen.fr
GREYC
University of Caen

Spatial Pyramid Representation (SPR) [1] introduces spatial layout information to the orderless bag-of-features (BoF) representation and performs competitively against more complex methods for incorporating spatial layout. In SPR the image is divided into regular grids at different scales (2×2, 4×4, etc.). However, the grids are taken as uniform spatial partitions without any theoretical motivation. In this paper, we address this issue and propose to learn the spatial partitioning with BoF representation.

We formulate the learning problem in a maximum margin framework, with slack variables,

$$\min_{w,g} \frac{1}{2}||w||^2 + C\sum \xi_i \qquad (1)$$

$$\text{s.t.} \quad y_i(w \cdot g(I_i) + b) \geq 1 - \xi_i$$

where $g(I)$ is the histogram feature obtained by applying the grid $g \in \mathscr{G}$ on the image I. The optimization is on both the weight vector *and* the grid for the task.

We define the space of grids $\mathscr{G}$ by construction. Starting with the full image, we recursively split the cells, into two parts, with axis aligned lines. We call the number of splits taken to obtain the grid as the *depth* of the grid and represent the grid as a set of cells $g = (g_1, g_2, \dots g_{k+1})$ with $g_i = (x_1^i, y_1^i, x_2^i, y_2^i) \in \Re^4$ representing the i^{th} cell in the grid. $x, y \in [0,1]$ are fractional multiples of the image width/height.

The dual form of the optimization problem (1), in terms of Lagrange's multipliers, $\alpha = \{\alpha_i\}$, is given by,

$$\max_{\alpha} \sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j \;\; g(I_i) \cdot g(I_j) \qquad (2)$$

$$\text{s.t.} \quad 0 \leq \alpha_i \leq C \text{ and } \sum_i \alpha_i y_i = 0$$

The dual formulation allows us to propose an efficient approximate optimization strategy based on two popular methods, coordinate descent like iterations and greedy forward selection. We treat the SVM parameters α and the grid parameters g as two sets of variable on which we do alternating coordinate descent like iterations to find the best grid for a fixed depth. The numerical gradient is computed efficiently using integral histograms and matrix dot products.

We experiment on two public databases, Scene 15 and VOC 2007, and we introduce a new large database of human attributes (HAT).

On Scene 15 (Fig. 1) the learnt grids achieve higher performance (mean class accuracy) with comparable vector sizes and outperform the SPR at depth as low as 4. The difference in vector sizes translates directly into computational savings.

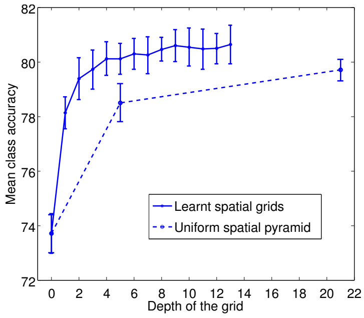

Figure 1: The performances of SPR and learnt grid at comparable vector lengths for Scene 15 dataset

On the more challenging VOC 2007 database with objects at diverse scales, locations and poses, the learnt grids again outperform SPR at lower grid depths and perform comparably at higher grid depths. The performance (avg. precision) of most of the classes, and on an average is higher (50.8 vs. 49.5) for the learnt grids at depth 4 (Fig. 2).

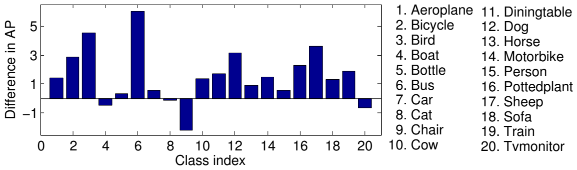

Figure 2: The difference in AP for VOC 2007 dataset at a grid depth of 4 with the learnt grid and the uniform SPR

Our database of human attributes (HAT) contains 9344 images. We used an automatic program to download the top ranked images from the popular image sharing site Flickr for manually specified queries. We then run a state-of-the-art person detector to obtain the human images and removed the few false positives manually. The images are annotated for 27 attributes based on gender, pose, age, appearance etc.

Table 1 shows the AP of the the learnt grids for some of the attributes at grids of depth 0 and 4. The learnt grids perform better than the SPR on most of the classes and also on a average. On some of the classes the improvement is quite high e.g. 49.9 vs. 42.7 for 'Female wearing a long skirt' and 62.1 vs. 51.9 for 'Female in wedding dress'. On an average also the learnt grids are better than the SPR (53.8 vs. 52.3).

Table 1: Classwise AP for selected human attributes with the learnt grids at depth 0 and 4 (see paper for all 27 attributes)

No.	Attribute	Depth 0	Depth 4
1	Female	72.5	82.0
4	Turned back	49.5	67.4
7	Running/walking	61.3	67.6
11	Elderly	21.9	29.3
14	Teen aged	25.2	29.1
18	Wearing Tee shirt	54.8	59.1
27	Bermuda/beach shorts	31.6	39.3

The grids learnt are interpretable in terms of spatial distribution of visual discriminant information. Fig. 3 shows the grids from two classes of VOC 2007 and two classes of the human attributes database overlayed on representative images. The grid learnt for bicycle class seems to focus on the wheels with square cells in the middle and the bar with horizontal cells towards the top. The cells for the cow class are predominantly horizontal capturing the contour of the cow. The grids for the bent arm and running classes seem to focus on the pose of the hands and feet respectively.

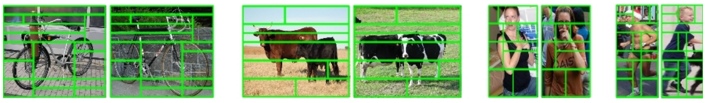

Figure 3: Learnt grids for VOC 2007 classes 'bicycle' and 'cow' and human attributes 'arms bent' and 'running' overlayed on representative example images.

Thus we have built upon the SPR of [1] and addressed one of its fundamental limitation i.e. the fixed structure. We have proposed to adapt the shape of the spatial partition to the classification tasks considered. Furthermore, we have experimentally showed that our representation significantly outperforms the standard SPR. Future work will investigate the possibility of optimizing the average precision instead of accuracy and of optimizing multichannel grids with different types of features.

[1] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.