Learning discriminative spatial representation for image classification

Gaurav Sharma, Frédéric Jurie

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

Abstract

Spatial Pyramid Representation (SPR) [7] introduces spatial layout information to the orderless bag-of-features (BoF) representation. SPR has become the standard and has been shown to perform competitively against more complex methods for incorporating spatial layout. In SPR the image is divided into regular grids. 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 the BoF representation. We define a space of grids where each grid is obtained by a series of recursive axis aligned splits of cells. We cast the classification problem in a maximum margin formulation with the optimization being over the weight vector and the spatial grid. In addition to experiments on two challenging public datasets (Scene-15 and Pascal VOC 2007) showing that the learnt grids consistently perform better than the SPR while being much smaller in vector length, we also introduce a new dataset of human attributes and show that the current method is well suited to the recognition of spatially localized human attributes.

1 Introduction

Image representation is a fundamental aspect of computer vision. Recent works have established, somewhat surprisingly, the bag-of-features (BoF) representation [3] as being an effective representation for various computer vision tasks *e.g.* object recognition, object detection, scene classification etc. It has been recently shown [7, 8, 12, 14, 15, 18] that adding spatial information to the standard BoF (which doesn't use any spatial information) can improve performance. Among these representations one of the most popular is the Spatial Pyramid Representation (SPR) [7] which incorporates spatial layout information of the features by dividing the image into uniform grids at different scales and then concatenating the BoF features from the different grid cells with appropriate normalizations. SPR, with discriminative maximum margin based classifiers, has become the standard representation and has been shown to perform competitively with more complex representations and models for many tasks [1, 13, 17] including action recognition [9].

The spatial partitioning in the SPR is, however, taken to be a uniform grid (at different scales *i.e.* 2×2, 4×4). The choice of partition of space does not have any particular theoretical or empirical motivation *i.e.* there have been no systematic exploration of the space of partitions and the grids have been derived out of practitioners' experience.

However, the choice of partitioning is expected to be important for the task *e.g.* a partition with prominently horizontal cells for 'coastal scene' (with beach, sea and sky) and one with prominently vertical cells for 'tall buildings' (both of these classes are part of the public benchmark Scene-15 dataset). Also, for cases where the discriminative information is localized, the grids could be finer in the important regions. We believe it is especially important for the recognition of human attributes in human centred images: for instance, in case of the 'wearing shorts' attribute, the partitioning in the middle part of the human is expected to be discriminant.

In the present paper we propose to learn the spatial partitioning for a given classification task. We define the space of grids (Sec. 2.2) as the set containing grids generated by recursive splitting of grid cells by axis aligned cuts (starting with the full image as the only cell). We then formulate the classification problem (Sec. 2.1) in the maximum margin framework and perform the optimization over both the weight vector *and* the grid parameters. We propose an efficient approximate algorithm (Sec. 2.3, Alg. 1) to perform the optimization and show experimentally (Sec. 3) that the learnt grids perform better than the classic SPR while leading to vectors smaller (as much as half) in length to the SPR. We also introduce a challenging dataset (Sec. 3.2) of human attributes (based on age, sex, appearance and pose) containing real world images collected from image sharing site Flickr.com. We demonstrate the relevance of learning the grids on such cases where the discriminating information is spatially localized.

1.1 Related works

The current state-of-the-art methods for object/scene recognition are built upon the bag-of-features (BoF) representation of Csurka *et al.* [3]. The representation works by extracting local features (*e.g.* SIFT [9]) from the images, vector quantizing them (*e.g.* using k-means clustering) and then representing images as histograms over the *visual words*. Thus, in the BoF representation the spatial layout is completely discarded.

Various methods have been proposed to incorporate spatial layout into the BoF representation. Two types of spatial information have been considered, whether the positions of the features are related to other features or to their absolute positions in the image.

Among those using pairwise positions of features, Savarese *et al.* [15] propose to form a bag-of-word representation over spatially neighbouring image regions, Liu *et al.* [8] use a feature selection method based on boosting which progressively mines higher-order spatial features, while Morioka *et al.* [12] propose joint feature space clustering to build a compact local pairwise codebook and in another work [11] incorporate the spatial orders of local features. Quack *et al.* [14] suggest to find distinctive spatial configurations of visual words by using data mining techniques, such as frequent itemsets.

In addition to pairwise relationships, images often have spatial biases *i.e.* the composition of the pictures of particular object or scene category typically share common layout properties. This is especially true for the recognition of attributes in human centred images. One of the pioneer works in the direction of exploiting spatial layout was by Lazebnik *et al.* [7] which proposed the Spatial Pyramid Representation (SPR). In SPR, the image is divided into uniform grids at different scales *i.e.* 2×2, 4×4, and the features are concatenated over all cells with appropriate normalization. SPR, working at spatial level rather than feature level, improved the BoF performance by a significant margin. More recently, Yang *et al.* [16] showed that incorporating sparse coding into the SPR improves performance. In the work of Cao *et al.* [2], local features of an image are first projected to different directions or

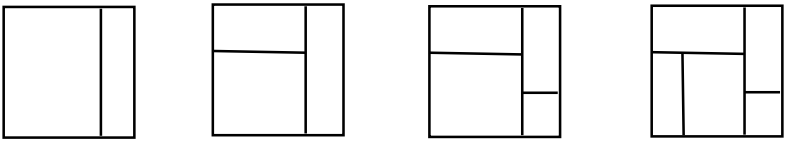

Figure 1: The formation of the spatial grid by successive splitting of cells; grid with depth 1 (left) to depth 4 (right)

points to generate a series of ordered bag-of-features. Zhou *et al*. [18] model region appearances with a mixture of Gaussian (MoG) density, and use the posterior over visual words for the image regions to generate so called "Gaussian maps", encoded by SPR. Very recently, Harada *et al*. [6] divide images by a regular grid, and learn weight maps for the grid cells.

We have seen that many state-of-the-art classification methods are based on the SPR, but the different parameters involved, the number of pyramid levels and the structure of the grid at each level, are empirically adapted to the situation *e.g*. [7, 16] use up to 4 pyramid levels with uniform grids of 1×1, 2×2, 4×4 and 8×8, while the winner of Pascal VOC 2007 competition, Marszalek *et al*. [10] followed by many others such as [17, 19], use three pyramid levels with grids of 1×1, 2×2 and 3×1. The SPR parameters are chosen in an ad-hoc manner and no work reports systematic construction of the representation.

The proposed method addresses this issue and learns a representation where the parameters are learnt for the given task.

2 Approach

As explained above, while the spatial pyramid representation [7] has been very successful in the task of object recognition, the spatial grids are fixed and are the same for all the classes, constituting a limitation of the approach. We propose here to learn the spatial partitioning for the given classification task, by defining a space of spatial grids and learning the best grid over this space for the task.

We use similar image representation as used by [7]. We represent images by extracting SIFT descriptors over dense multiscale patches and quantizing the features using a dictionary learnt using k-means. Given a spatial partitioning of the image (e.g. uniform grid of 2×2) we construct spatial histogram for the image and use it as the image representation.

Denoting the space of all possible spatial grids by $\mathscr{G}$ (Sec. 2.2 defines the space) we can write the scoring function as (denoting the dot product between vectors with '$\cdot$')

$$f(I) = w \cdot \hat{g}(I) + b, \tag{1}$$

where $\hat{g}(I) \in \Re^d$ is the histogram feature obtained by applying the best grid $\hat{g} \in \mathscr{G}$ learned for the task and to the image $I \in \mathscr{I}$. Here, g corresponds to a given spatial grid $g(I)$ represents the formation of the histogram features for the image with the given grid. Unlike spatial pyramid, we use only the final grid obtained after optimization and not a pyramid as we expect the grid to adjust its resolution depending on the spatial distribution of discriminative information for the class.

2.1 Learning the grids

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

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

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

We propose an efficient approximate solution to the problem using coordinate descent like iterations with greedy forward selection.

2.2 The space of grids

We define the space of grids $\mathscr{G}$ by construction. Starting with the full image as the grid with one cell, we recursively split the cells further, into two parts, with axis aligned lines (Fig. 1). In theory the split can occur continuously at 0 to 1 times the image height/width, while in practice the splits are quantized. Thus the usual spatial pyramid grids of 2×2, 4×4 and 8×8 partitions are members of the considered space of grids $\mathscr{G}$, and so is the finest grid possible i.e. one which isolates every pixel in the image.

We split the space of grids into disjoint parts as $\mathscr{G} = \cup_k\{\mathscr{G}_k\}$, with each subset $\mathscr{G}_k$ containing the grids obtained with exactly k successive splits. 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 $g = (g_1, g_2, \ldots g_{k+1})$ 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. Here $x, y \in [0,1]$ are fractional multiples of the image width and height respectively. We write g^k for a grid g where we want to make explicit the depth k (note that a grid with depth k has $k+1$ cells, the full image is obtained with a grid of depth 0).

2.3 Dual form

The dual form of the optimization problem (2), in terms of Lagrange's multipliers, $\alpha = \{\alpha_i\}$, is given by,

$$\max_{\alpha} \sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j \;\; g(I_i) \cdot g(I_j) \tag{3}$$

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

The dual formulation allows us to propose an efficient approximate optimization strategy (Alg. 1) 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. To increase the depth of the grid we resort to greedy forward selection, computing the next best split using gradient based optimization. The numerical gradient is computed efficiently using integral histograms and matrix dot products.

2.4 Gradient based optimization

While doing the alternating coordinate descent iterations, in the step where we keep α constant, we use numerical gradient to optimize efficiently. Given the grid at depth k, we want

Algorithm 1 Computing the grid $g^K \in \mathcal{G}^K$ for the given classification task

```
Initialize the grid g^0 ∈ G^0 to be the full image.
for k=1 ... K do
    for all grid cells do
        Initialize {s, α} by choosing s randomly and optimizing (3)
        while convergence do
            Optimize (4) w.r.t. s (keeping α fixed) using the gradient in (6)
            Optimize w.r.t. α (using efficient linear SVM solvers) keeping s fixed
        end while
    end for
    g^{k+1} ← (g^k, s*), s* being the best grid cell split
end for
```

to derive the next grid of depth $k+1$. We have α for the current iteration and we hold them constant. The coordinate descent step becomes an unconstrained optimization problem,

$$F(g) = -\frac{1}{2}\sum_{i,j} t_{ij}\ \ g(I_i)\cdot g(I_j), \tag{4}$$

where $t_{ij} = \alpha_i \alpha_j y_i y_j$ are constants depending on the current α and the training labels y.

When a further split is made the new histogram differs only in the part of the image which was split (Fig. 2). The gradient of the objective function depends on the gradient of the linear kernel function parametrized by the split parameter s. Considering two images I_i and I_j with histograms $g(I_i) = x$ and $g(I_j) = y$, we can write the kernel function between them as

$$k_s(x,y) = \begin{pmatrix} x_s \\ x_o \end{pmatrix} \cdot \begin{pmatrix} y_s \\ y_o \end{pmatrix} = x_s \cdot y_s + x_o \cdot y_o, \tag{5}$$

where for histogram $x = (x_s x_o)$, x_s is the part of the histogram which is affected by the split while x_o is the part which isn't. Thus, we calculate the numerical gradient for the kernel function when we take a step from s to s' (Fig. 2) as

$$\begin{aligned}
\Delta_s k_s(x,y) &= x_{s'} \cdot y_{s'} - x_s \cdot y_s \\
&= \frac{1}{N_x}\begin{pmatrix} c_1^x - c_\Delta^x \\ c_2^x + c_\Delta^x \end{pmatrix} \cdot \frac{1}{N_y}\begin{pmatrix} c_1^y - c_\Delta^y \\ c_2^y + c_\Delta^y \end{pmatrix} - \frac{1}{N_x}\begin{pmatrix} c_1^x \\ c_2^x \end{pmatrix} \cdot \frac{1}{N_y}\begin{pmatrix} c_1^y \\ c_2^y \end{pmatrix} \\
&\propto (c_1^x - c_\Delta^x)\cdot(c_1^y - c_\Delta^y) + (c_2^x + c_\Delta^x)\cdot(c_2^y + c_\Delta^y) - (c_1^x \cdot c_1^y + c_2^x \cdot c_2^y) \\
&\propto (c_1^x \cdot c_1^y - c_1^x \cdot c_\Delta^y - c_\Delta^x \cdot c_1^y + c_\Delta^x \cdot c_\Delta^y) + \\
&\quad (c_2^x \cdot c_2^y + c_2^x \cdot c_\Delta^y + c_\Delta^x \cdot c_2^y + c_\Delta^x \cdot c_\Delta^y) - (c_1^x \cdot c_1^y + c_2^x \cdot c_2^y) \\
&\propto 2\, c_\Delta^x \cdot c_\Delta^y - c_1^x \cdot c_\Delta^y - c_\Delta^x \cdot c_1^y + c_2^x \cdot c_\Delta^y + c_\Delta^x \cdot c_2^y
\end{aligned} \tag{6}$$

where, c_1 c_2 and c_Δ are the histograms (un-normalized raw counts) of different parts involving the split as shown in Fig. 2. The gradient for objective function in (4) is the sum of these gradients, for all pairs, weighted by t_{ij}.

The first cost associated in computing the gradient is the calculation of the histogram c_Δ for the changing part of the grid. The step sizes are quantized and hence the calculation of the count histogram is fast using integral histograms for the grid induced by the quantized

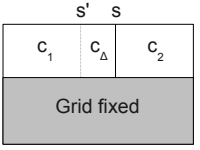

Figure 2: The histograms (raw counts, not normalized) when a step is taken from the current split s to a new split s'. c_1 and c_2 are the histograms for the two parts generated by split s and c_Δ is the histogram for the part between split s and s'.

step sizes. The second is the computation of the dot products of matrices, which is also performed efficiently using optimized matrix algebra libraries.

The split can occur in any of the cells of the grid i.e. a depth d grid has $d+1$ cells and the further split can occur in any of them. Since, we can't consider the splits in the different grid cells as being instances of a single variable, we run the one dimensional optimization separately for every grid cell and take the cell split increasing the objective function the most.

The other coordinate descent step involves training linear SVM which, owing to recent progress, is also achieved efficiently. Hence, both the steps are fast and the overall optimization is quite efficient in practice.

3 Experiments and results

The motivation of these experiments is twofold: first, we show that optimizing the spatial representation lead consistently to better results that the standard SPR, as demonstrated on two popular databases (Scene 15 and Pascal VOC). Second, we show that the proposed representation is especially suited for the recognition of human attributes, problem for which we introduce a new database (presented in Sec. 3.2).

3.1 Implementation details

We use the standard bag of features (BoF) with spatial pyramid representation (SPR) as the baseline [7]. We use multiscale SIFT features extracted at 8 scales separated by a factor of 1.2 and a step size of 8 pixels. We randomly sampled 200,000 SIFT vectors from the training images and quantized them using k-means into 1000 clusters. Finally we use nonlinear SVM with histogram intersection kernel in a one-vs-all setting (Scene 15 only) to perform the classification. Note that we use formulation (3) to learn the grids which is equivalent to using a linear kernel but we use nonlinear kernel for training the final SVM to compare with the SPR baseline.

3.2 Datasets

Scene 15 database[1] contains 15 classes of different scenes e.g. *kitchen, coast, highway*. Each class has 260 to 410 images and the database has a total of 4492 grayscale images. The problem is of multiclass categorization and, like previous works, we train on 100 random images per class and test on the rest. We do so 10 times and report the mean and standard deviation of the mean class accuracy.

[1] http://www.featurespace.org/data.htm

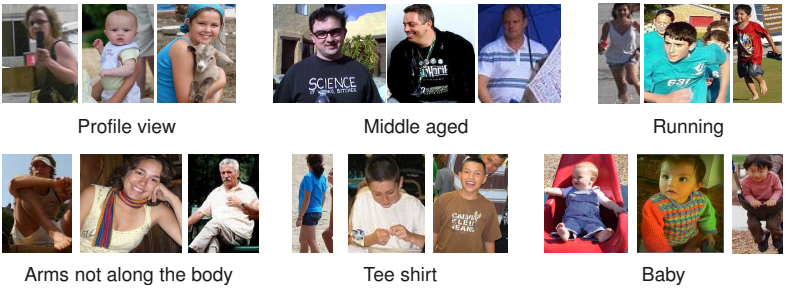

Figure 3: Example images for some attributes from our database. The images are scaled to the same height for better visualization.

Pascal VOC 2007 database[2] has 20 object categories. It is a challenging database of images downloaded from internet, containing 9963 images split into train, val and test sets. We use the train and val sets to learn our models and report the mean average precision for the 20 classes on the test set as the performance measure, following the standard protocol for this database.

Database of human attributes (HAT) is a new database[3] for learning semantic human attributes. Our database contains 9344 images, with annotations for 27 attributes. To obtain a large number of images we used an automatic program to query and download the top ranked result images from the popular image sharing site Flickr, with manually specified queries. We used more than 320 queries, chosen so as to retrieve predominantly images of people (e.g. 'soccer kid' cf. 'sunset'). A state-of-the-art person detector [5] was used to obtain the human images with the few false positives removed manually.

The database contains a wide variety of human images in different poses (standing, sitting, running, turned back etc.), of different ages (baby, teen, young, middle aged, elderly etc.), wearing different clothes (tee-shirt, suits, beachwear, shorts etc.) and accessories (sunglasses, bag etc.) and is, thus, rich in semantic attributes for humans. It also has high variation in scale (only upper body to the full person) and size of the images. The high variation makes it a challenging database. Fig. 3 shows some example images for some of the attributes (the images in the figures are scaled to the same height for visualization).

The database has train, val and test sets. The models are learnt with the train and val sets while the average precision for each attribute on the test set is reported as the performance measure. The overall performance is given by the mean average precision over the set of attributes.

3.3 Results

Comparison with standard SPR on Scene 15 and Pascal VOC07. Fig. 4 shows the performance of the learnt grid and the uniform spatial pyramid representation on the Scene 15 database. With our implementation, the spatial pyramid representation achieves a mean class accuracy of 73.7 ± 0.7 at pyramid level 0 (full image *i.e.* 1×1), 78.5 ± 0.4 at level 1 (1×1

[2] http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2007/
[3] http://users.info.unicaen.fr/~gsharma/hatdb/

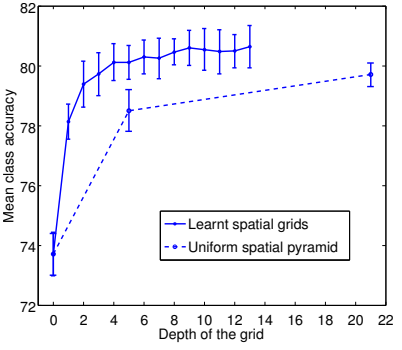

Figure 4: The performances of SPR and learnt grid at comparable vector lengths for Scene 15 database

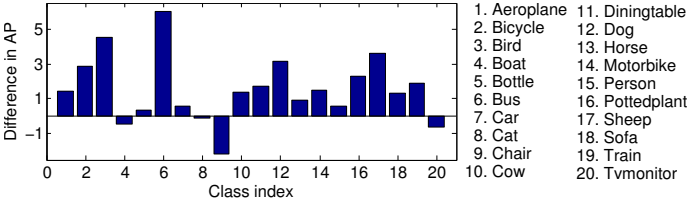

Figure 5: The difference in AP for all the classes of the VOC 2007 database at a grid depth of 4 with the learnt grid and the uniform spatial pyramid

and 2×2) and 79.6 ± 0.6 at level 2 (1×1, 2×2 and 4×4). The performance decreases for SPR if we go higher than this level. As shown in Fig. 4 the learnt grids achieve higher performance with comparable vector sizes and outperform the SPR at depth as low as 4 (80.1 ± 0.6). The performance of the learnt grids increases quickly with the depth and saturates at around a depth of 8 which is 0.4 times the length of best SPR (depth 21). Note that the vectors are computed similarly for both representations and hence have similar sparsities *i.e.* the difference in vector sizes translates directly into computational savings.

On the more challenging VOC 2007 database where objects appear 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 of most of the classes, and on a average is higher (50.8 vs. 49.5) for the learnt grids at depth 4 (Fig. 5). Owing to the unconstrained nature of the images in the database, the structural information is limited in this database. Also, the metric used is average precision while we are optimizing on accuracy in the maximum margin formulation. It would be interesting to formulate the problem with average precision being maximized directly. We hope to pursue this further.

Recognizing human attributes Table 1 shows the average precision of the the learnt grids for the different 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: Table showing the classwise average precision for the human attributes with the learnt grids at depth 0 and 4

No.	Attribute	Depth 0	Depth 4
1	Female	72.5	82.0
2	Frontal pose	90.1	91.3
3	Side pose	48.8	61.0
4	Turned back	49.5	67.4
5	Upper body	83.1	92.4
6	Standing straight	95.6	96.0
7	Running/walking	61.3	67.6
8	Crouching/bent	21.6	20.7
9	Sitting	52.2	54.6
10	Arms bent/crossed	92.0	91.9
11	Elderly	21.9	29.3
12	Middle aged	63.2	66.3
13	Young (college)	59.0	59.4
14	Teen aged	25.2	29.1
15	Small kid	33.5	43.7
16	Small baby	12.6	12.2
17	Wearing tank top	29.2	33.2
18	Wearing tee shirt	54.8	59.1
19	Wearing casual jacket	31.3	35.3
20	Formal men's suit	44.4	48.2
21	Female long skirt	23.1	49.9
22	Female short skirt	27.3	33.7
23	Wearing short shorts	38.6	42.7
24	Low cut top	47.8	55.6
25	Female in swim suit	29.0	28.2
26	Female wedding dress	51.3	62.1
27	Bermuda/beach shorts	31.6	39.3
	mAP	47.8	53.8

Visualizing the learnt grids The grids learnt are interpretable in terms of spatial distribution of visual discriminant information. Fig. 6 shows the grids from two classes of VOC 2007 and two classes of the human attributes database overlayed on representative example images from the database. 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.

4 Conclusions

This paper builds on the Spatial Pyramid Representation of [7] and addresses one of the fundamental limitation of this approach, i.e. the fixed structure of the SPR. We have proposed an efficient algorithm, based on a maximum margin framework, allowing to adapt the spatial partitioning to the classification tasks considered. Furthermore, we have experimentally showed that our representation significantly outperforms the standard SPR.

Figure 6: Learnt grids for VOC 2007 classes 'bicycle' and 'cow' and human attributes 'arms bent' and 'running' overlayed on representative example images.

Future work will investigate the possibility of optimizing the average precision instead of the accuracy and of optimizing multichannel grids with different types of features.

Acknowledgements The present work was supported by the ANR SCARFACE project.

References

[1] A. Bosch, A. Zisserman, and X. Munoz. Representing shape with a spatial pyramid kernel. In *CIVR*, 2007.

[2] Y. Cao, C. Wang, Z. Li, L. Zhang, and L. Zhang. Spatial bag-of-features. In *CVPR*, 2010.

[3] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *Intl. Workshop on Stat. Learning in Computer Vision*, 2004.

[4] V. Delaitre, I. Laptev, and J. Sivic. Recognizing human actions in still images: a study of bag-of-features and part-based representations. In *BMVC*, 2010.

[5] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *PAMI*, 32:1627–1645, 2010.

[6] T. Harada, H. Nakayama, and Y. Kuniyoshi. Improving local descriptors by embedding global and local spatial information. In *ECCV*, 2010.

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

[8] D. Liu, G. Hua, P. Viola, and T. Chen. Integrated feature selection and higher-order spatial feature extraction for object categorization. In *CVPR*, 2008.

[9] D. Lowe. Distinctive image features form scale-invariant keypoints. *IJCV*, 18:91–110, 2004.

[10] M. Marszalek, C. Schmid, H. Harzallah, and J. van de Weijer. Learning object representations for visual object class recognition. In *Visual recognition challenge workshop*, 2007.

[11] N. Morioka and S. Satoh. Learning directional local pairwise bases with sparse coding. In *BMVC*, 2010.

[12] N. Morioka and S. Satoh. Building compact local pairwise codebook with joint feature space clustering. In *ECCV*, 2010.

[13] F. Perronnin, J. Sanchez, and T. Mensink. Improving the fisher kernel for large-scale image classification. In *ECCV*, 2010.

[14] T. Quack, V. Ferrari, B. Leibe, and L. van Gool. Efficient mining of frequent and distinctive feature configurations. In *ICCV*, 2007.

[15] S. Savarese, J. Winn, and A. Criminisi. Discriminative object class models of appearance and shape by correlatons. In *CVPR*, 2006.

[16] J. Yang, K. Yu, Y. Gong, and T. Huang. Linear spatial pyramid matching using sparse coding for image classification. In *CVPR*, 2009.

[17] J. Yang, K. Yu, and T. Huang. Efficient highly over-complete sparse coding using a mixture model. In *ECCV*, 2010.

[18] X. Zhou, N. Cui, Z. Li, F. Liang, and T. Huang. Hierarchical Gaussianization for image classification. In *ICCV*, 2009.

[19] X. Zhou, K. Yu, T. Zhang, and T. S. Huang. Image classification using super-vector coding of local image descriptors. In *ECCV*, 2010.