Spatial pyramid matching

Svetlana Lazebnik, Cordelia Schmid, Jean Ponce

1

Spatial Pyramid Matching

Svetlana Lazebnik (lazebnik@cs.unc.edu)
Department of Computer Science
University of North Carolina at Chapel Hill

Cordelia Schmid (Cordelia.Schmid@inrialpes.fr)
INRIA Grenoble, Rhône-Alpes, France

Jean Ponce (ponce@di.ens.fr)
LIENS (CNRS/ENS/INRIA UMR 8548)
Ecole Normale Supérieure, Paris, France

This chapter deals with the problem of whole-image categorization. We may want to classify a photograph based on a high-level semantic attribute (e.g., indoor or outdoor), scene type (forest, street, office, etc.), or object category (car, face, etc.). Our philosophy is that such global image tasks can be approached in a *holistic* fashion: It should be possible to develop image representations that use low-level features to directly infer high-level semantic information about the scene without going through the intermediate step of segmenting the image into more "basic" semantic entities. For example, we should be able to recognize that an image contains a beach scene without first segmenting and identifying its separate components, such as sand, water, sky, or bathers. This philosophy is inspired by psychophysical and psychological evidence that people can recognize scenes by considering them in a "holistic" manner, while overlooking most of the details of the constituent objects (Oliva and Torralba, 2001). It has been shown that human subjects can perform high-level categorization tasks extremely rapidly and in the near absence of attention (Thorpe et al., 1996; Fei-Fei et al., 2002), which would most likely preclude any feedback or detailed analysis of individual parts of the scene.

Renninger and Malik (2004) have proposed an orderless texture histogram model to replicate human performance on "pre-attentive" classification tasks. In the computer vision literature, more advanced orderless methods based on *bags of features* (Csurka et al., 2004) have recently demonstrated impressive levels of performance for image classification. These methods are simple and efficient, and they can be made robust to clutter, occlusion, viewpoint change, and even non-rigid deformations. Unfortunately, they completely disregard the spatial layout of the features in the image, and

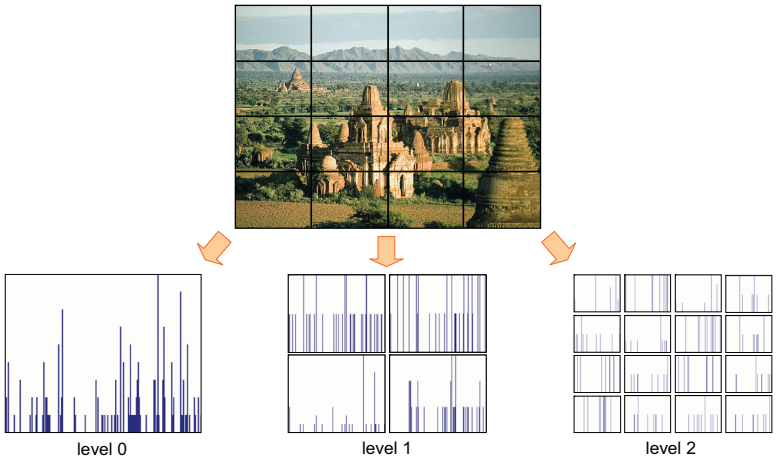

Fig. 1.1. A schematic illustration of the spatial pyramid representation. A spatial pyramid is a collection of orderless feature histograms computed over cells defined by a multi-level recursive image decomposition. At level 0, the decomposition consists of just a single cell, and the representation is equivalent to a standard bag of features. At level 1, the image is subdivided into four quadrants, yielding four feature histograms, and so on. Spatial pyramids can be matched using the *pyramid kernel*, which weights features at higher levels more highly, reflecting the fact that higher levels localize the features more precisely (see Section 1.2).

thus cannot take advantage of the regularities in image composition and the spatial arrangement of the features, which can make very powerful cues for scene classification tasks. Therefore, an important research direction is to augment bags of features with global spatial relations in a way that significantly improves classification performance, yet does not compromise the simplicity and computational efficiency that makes them so attractive for real-world applications.

In Lazebnik et al. (2006), we have proposed to extend bags of features to *spatial pyramids* by partitioning the image into increasingly fine sub-regions and concatenating histograms of local features found inside each sub-region (Figure 1.1). This representation is combined with a kernel-based *pyramid matching* scheme (Grauman and Darrell, 2005) that efficiently computes approximate global geometric correspondence between sets of features in two images. While the spatial pyramid representation sacrifices the geometric invariance properties of bags of features, it more than compensates for this loss with increased discriminative power derived from the global spatial information. This has allowed the spatial pyramid method to significantly outperform bags of features on challenging image categorization tasks, in our original experiments (Lazebnik et al., 2006), as well as in several subse-

quent publications (Bosch et al., 2007a,b; Chum and Zisserman, 2007; Liu et al., 2007; Marszałek et al., 2007; Varma and Ray, 2007).

The rest of this chapter is organized as follows. In Section 1.1, we discuss relevant previous work on global image representations and scene recognition. In Section 1.2, we review pyramid matching as introduced by Grauman and Darrell (2005) and then describe our adaptation of this framework to the spatial domain. Section 1.3 presents our original experimental results on a fifteen-category scene dataset and on the standard Caltech-101 benchmark. Finally, Section 1.4 surveys recent extensions and applications of the technique that have appeared in the literature since our original publication.

1.1 Survey of Related Work

The origin of many of today's image classification systems can be traced to empirical *appearance-based methods* for recognition, including subspace methods (Turk and Pentland, 1991; Murase and Nayar, 1995) and histograms (Swain and Ballard, 1991; Schiele and Crowley, 2000). Many of the early appearance-based approaches required registered training images, and did not tolerate unmodeled photometric or geometric transformations. For these reasons, global appearance-based representations were superceded by *local invariant features* (see Schmid and Mohr (1997); Lowe (2004) for two important examples), which have much better tolerance to clutter, occlusion, lighting changes, and geometric deformations. In the last few years, local features have been successfully incorporated into many state-of-the-art recognition systems (Csurka et al., 2004; Opelt et al., 2004; Grauman and Darrell, 2005; Lazebnik et al., 2005; Sivic et al., 2005; Zhang et al., 2007).

Today, local features continue to enjoy a great degree of success, but at the same there is a notable resurgence of interest in global appearance-based methods (Oliva and Torralba, 2001; Hays and Efros, 2007; Russell et al., 2007; Torralba et al., 2007, 2008). There are two reasons for this revival. One is the availability of large-scale training datasets gathered from the Web (Ponce et al., 2006). Instead of having to factor out geometric and photometric variations with local invariant features, we can essentially use large datasets to sample all possible variations by brute force. The second reason is an improved understanding of *context* (Hoiem et al., 2005; Oliva and Torralba, 2007), or the way that global image appearance and geometry influence the perception of individual objects in the scene. A good contextual description of an image may be used to inform the subsequent search for specific objects. For example, if the image, based on its context, is likely to be a highway, we have a high probability of finding a car, but not

a toaster, and we should adjust the prior probabilities of the different types of objects accordingly.

The spatial pyramid method can be viewed as an updated version of a global appearance-based method, or as a hybrid of local and global representations. In any case, it can prove useful for efficient scene recognition in large datasets, as well as for capturing contextual information. It follows the strategy of "subdivide and disorder" — i.e., partition the image into subblocks and compute orderless statistics of low-level image features in these subblocks. This strategy has been practiced numerous times in computer vision, for global image description (Gorkani and Picard, 1994; Szummer and Picard, 1998; Vailaya et al., 1998; Squire et al., 1999; Torralba et al., 2003), as well as for description of image sub-windows (Dalal and Triggs, 2005) and keypoints (Lowe, 2004). Existing methods have used a variety of different features (raw pixel values, gradient orientations, or filter bank outputs), orderless statistics (means or histograms), and different spatial subdivision schemes (including regular grids, quadtrees, as well as "soft" windows). The spatial pyramid method attempts to replace ad-hoc implementation choices by a clean overarching framework. The framework itself is independent of the choice of features (as long as the features can be quantized to a discrete vocabulary) and the spatial decomposition is determined by the goal of approximate geometric matching. In this way, it is not necessary to find the single "best" level of spatial subdivision, and the different levels are combined in a principled way to improve performance over any single level.

A possible unifying theory underlying the seemingly disparate variety of subdivide-and-disorder techniques in the literature is offered by the concept of *locally orderless images* of Koenderink and Van Doorn (1999). This concept generalizes histograms to histogram-valued scale spaces. For each Gaussian aperture at a given location and scale, the locally orderless image returns the histogram of image features aggregated over that aperture. Our spatial pyramid approach can be thought of as a more restricted kind of a locally orderless image, where instead of a Gaussian scale space of apertures, we define a fixed hierarchy of rectangular windows. Koenderink and Van Doorn (1999) have argued persuasively that locally orderless images play an important role in visual perception. The practical success of spatial pyramids observed in our experiments as well as subsequent work suggests that locally orderless matching may be a powerful mechanism for estimating overall perceptual similarity between images.

Additional hints as to the importance of locally orderless representations may be gleaned from a few recent publications in the machine learning literature. For example, Lebanon et al. (2007) have proposed *locally weighted*

bags of words for document analysis. This is essentially a locally orderless representation for text documents, although it does not include any explicit multi-scale structure. Cuturi and Fukumizu (2006) have described a very general and abstract framework for kernel-based matching of nested histograms, which is potentially applicable to many different data types. An important direction for future work is extending the insights from these publications into the domain of visual learning, and unifying them with existing image-based theories of locally orderless representations.

1.2 Spatial Pyramid Matching

In this section, we describe the general pyramid matching framework of Grauman and Darrell (2005), and then introduce our application of this framework to create a *spatial pyramid* image representation.

1.2.1 Pyramid Match Kernels

Let X and Y be two sets of vectors in a d-dimensional feature space. Grauman and Darrell (Grauman and Darrell, 2005) propose *pyramid matching* to find an approximate correspondence between these two sets. Informally, pyramid matching works by placing a sequence of increasingly coarser grids over the feature space and taking a weighted sum of the number of matches that occur at each level of resolution. At any fixed resolution, two points are said to match if they fall into the same cell of the grid; matches found at finer resolutions are weighted more highly than matches found at coarser resolutions. More specifically, let us construct a sequence of grids at resolutions $0, \ldots, L$, such that the grid at level ℓ has 2^ℓ cells along each dimension, for a total of $D = 2^{d\ell}$ cells. Let H_X^ℓ and H_Y^ℓ denote the histograms of X and Y at this resolution, so that $H_X^\ell(i)$ and $H_Y^\ell(i)$ are the numbers of points from X and Y that fall into the ith cell of the grid. Then the number of matches at level ℓ is given by the *histogram intersection* function (Swain and Ballard, 1991):

$$\mathcal{I}(H_X^\ell, H_Y^\ell) = \sum_{i=1}^{D} \min\left(H_X^\ell(i), H_Y^\ell(i)\right). \tag{1.1}$$

In the following, we will abbreviate $\mathcal{I}(H_X^\ell, H_Y^\ell)$ to $\mathcal{I}^\ell$.

Note that the number of matches found at level ℓ also includes all the matches found at the finer level $\ell+1$. Therefore, the number of *new* matches found at level ℓ is given by $\mathcal{I}^\ell - \mathcal{I}^{\ell+1}$ for $\ell = 0, \ldots, L-1$. The weight

associated with level ℓ is set to $\frac{1}{2^{L-\ell}}$, which is inversely proportional to cell width at that level. Intuitively, we want to penalize matches found in larger cells because they involve increasingly dissimilar features. Putting all the pieces together, we get the following definition of a *pyramid match kernel*:

$$\kappa^L(X,Y) = \mathcal{I}^L + \sum_{\ell=0}^{L-1} \frac{1}{2^{L-\ell}} \left(\mathcal{I}^\ell - \mathcal{I}^{\ell+1}\right) \tag{1.2}$$

$$= \frac{1}{2^L}\mathcal{I}^0 + \sum_{\ell=1}^{L} \frac{1}{2^{L-\ell+1}} \mathcal{I}^\ell \,. \tag{1.3}$$

Both the histogram intersection and the pyramid match kernel are Mercer kernels (Grauman and Darrell, 2005).

1.2.2 Spatial Matching Scheme

As introduced in Grauman and Darrell (2005), a pyramid match kernel works with an orderless image representation. It allows for multiresolution matching of two collections of features in a high-dimensional appearance space, but discards all spatial information. Another problem with this approach is that the quality of the approximation to the optimal partial match provided by the pyramid kernel degrades linearly with the dimension of the feature space (Grauman and Darrell, 2007), which means that the kernel is not effective for matching high-dimensional features such as SIFT descriptors. To overcome these shortcomings, we propose instead to perform pyramid matching in the two-dimensional image space, and use standard vector quantization techniques in the feature space. Specifically, we quantize all feature vectors into M discrete types, and make the simplifying assumption that only features of the same type can be matched to one another. Each channel m gives us two sets of two-dimensional vectors, X_m and Y_m, representing the coordinates of features of type m found in the respective images. The final kernel is then the sum of the separate channel kernels:

$$K^L(X,Y) = \sum_{m=1}^{M} \kappa^L(X_m, Y_m) \,. \tag{1.4}$$

This approach has the advantage of maintaining continuity with the popular "visual vocabulary" paradigm — in fact, it reduces to a standard bag of features when $L = 0$.

Because the pyramid match kernel (1.3) is simply a weighted sum of histogram intersections, and because $c \min(a,b) = \min(ca, cb)$ for positive

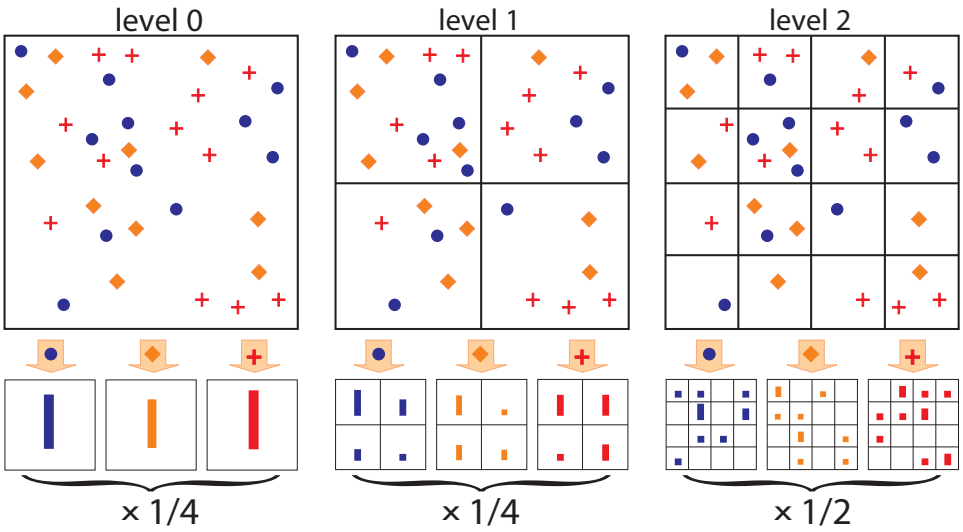

Fig. 1.2. Toy example of constructing a pyramid for $L = 2$. The image has three feature types, indicated by circles, diamonds, and crosses. At the top, we subdivide the image at three different levels of resolution. Next, for each level of resolution and each channel, we count the features that fall in each spatial bin. Finally, we weight each spatial histogram according to eq. (1.3).

numbers, we can implement K^L as a single histogram intersection of "long" vectors formed by concatenating the appropriately weighted histograms of all channels at all resolutions (Fig. 1.2). For L levels and M channels, the resulting vector has dimensionality $M \sum_{\ell=0}^{L} 4^\ell = M\frac{1}{3}(4^{L+1} - 1)$. Several experiments reported in Section 1.3 use the settings of $M = 400$ and $L = 3$, resulting in $34,000$-dimensional histogram intersections. However, these operations are efficient because the histogram vectors are very sparse. In fact, just as in Grauman and Darrell (2005), the computational complexity of the kernel is linear in the number of features (more recently, Maji et al. (2008) have shown that the histogram intersection kernel is amenable to further optimizations, leading to extremely fast support vector classifiers).

The final issue is normalization, which is necessary to account for images with different numbers of local features. In our own work, we follow a very simple strategy: we normalize all histograms by the total weight of all features in the image, in effect forcing the total number of features in all images to be the same. Because we use a dense feature representation (see Section 1.3.1), and thus do not need to worry about spurious feature detections resulting from clutter, this practice is sufficient to deal with the effects of variable image size.

1.3 Experiments

1.3.1 Experimental Setup

We have conducted experiments with two types of features: "weak features" that have very small spatial support (a single pixel) and take on just a few possible discrete values; and "strong features" that are computed over larger image patches and quantized using a large vocabulary to capture more distinctive and complex patterns of local appearance. More specifically, the weak features are oriented edge points, i.e., points whose gradient magnitude in a given direction exceeds a minimum threshold. We extract edge points at two scales and eight orientations, for a total of $M = 16$ channels. We designed these features to obtain a representation similar to the "gist" (Oliva and Torralba, 2001) or to a global SIFT descriptor (Lowe, 2004) of the image. The strong features are SIFT descriptors of 16×16 pixel patches computed over a grid with spacing of 8 pixels. Our decision to use a dense regular grid instead of interest points was based on the comparative evaluation of Fei-Fei and Perona (2005), who have shown that dense features work better for scene classification. Intuitively, a dense image description is necessary to capture uniform regions such as sky, calm water, or road surface (to deal with low-contrast regions, we skip the usual SIFT normalization procedure when the overall gradient magnitude of the patch is too weak). We perform k-means clustering of a random subset of patches from the training set to form a visual vocabulary. Typical vocabulary sizes for our experiments are $M = 200$ and $M = 400$.

Next, we report results on a fifteen-category scene dataset and the standard Caltech-101 benchmark (Fei-Fei et al., 2004). We perform all processing in grayscale, even when color images are available. All experiments are repeated ten times with different randomly selected training and test images, and the average of per-class recognition rates is recorded for each run. The final result is reported as the mean and standard deviation of the results from the individual runs. Multi-class classification is done with a support vector machine (SVM) trained using the one-versus-all rule: a classifier is learned to separate each class from the rest, and a test image is assigned the label of the classifier with the highest response.

1.3.2 Scene Category Recognition

Our first dataset (Fig. 1.3) is composed of fifteen scene categories: thirteen were provided by Fei-Fei and Perona (2005) (eight of these were originally collected by Oliva and Torralba (2001)), and two (industrial and store) were

Fig. 1.3. Example images from the scene category database. The database is publicly available at `http://www-cvr.ai.uiuc.edu/ponce_grp/data`.

	Weak features ($M = 16$)		Strong features ($M = 200$)		Strong features ($M = 400$)	
L	Single-level	Pyramid	Single-level	Pyramid	Single-level	Pyramid
0 (1×1)	45.3 ±0.5		72.2 ±0.6		74.8 ±0.3	
1 (2×2)	53.6 ±0.3	56.2 ±0.6	77.9 ±0.6	79.0 ±0.5	78.8 ±0.4	80.1 ±0.5
2 (4×4)	61.7 ±0.6	64.7 ±0.7	79.4 ±0.3	**81.1** ±0.3	79.7 ±0.5	**81.4** ±0.5
3 (8×8)	63.3 ±0.8	**66.8** ±0.6	77.2 ±0.4	80.7 ±0.3	77.2 ±0.5	81.1 ±0.6

Table 1.1. Classification results for the scene category database (see text). The highest results for each kind of feature are shown in bold.

collected by ourselves. Each category has 200 to 400 images, and average image size is 300×250 pixels.

Table 1.1 shows detailed results of classification experiments using 100 images per class for training and the rest for testing (the same setup as Fei-Fei and Perona (2005)). The table lists the performance achieved using just the highest level of the pyramid (the "single-level" columns), as well as the performance of the complete matching scheme using multiple levels (the "pyramid" columns). First, let us examine the performance of strong features for $L = 0$ and $M = 200$, corresponding to a standard bag of features. Our classification rate is 72.2%. For the 13 classes inherited from Fei-Fei and Perona (2005), it is 74.7%, which is much higher than their best results

of 65.2%, achieved with an orderless method and a feature set comparable to ours. With the spatial pyramid at $L = 2$, the performance of our method goes up to 81.1% — an almost 10% improvement over a bag of features.

More generally, for all three kinds of features (weak features and strong features with $M = 200$ and $M = 400$), results improve dramatically as we go from $L = 0$ to a multi-level setup. Though matching at the highest pyramid level seems to account for most of the improvement, using all the levels together confers a statistically significant benefit. For strong features, single-level performance actually drops as we go from $L = 2$ to $L = 3$. This means that the highest level of the $L = 3$ pyramid is too finely subdivided, with individual bins yielding too few matches. Despite the diminished discriminative power of the highest level, the performance of the entire $L = 3$ pyramid remains essentially identical to that of the $L = 2$ pyramid. This, then, is the main advantage of the spatial pyramid representation: because it combines multiple resolutions in a principled fashion, it is robust to failures at individual levels.

It is also interesting to compare performance of different feature sets. As expected, weak features do not perform as well as strong features, though in combination with the spatial pyramid, they can also achieve acceptable levels of accuracy (note that because weak features have a much higher density and much smaller spatial extent than strong features, their performance continues to improve as we go from $L = 2$ to $L = 3$). Increasing the visual vocabulary size from $M = 200$ to $M = 400$ results in a small performance increase at $L = 0$, but this difference is all but eliminated at higher pyramid levels. Thus, we can conclude that the coarse-grained geometric cues provided by the pyramid have more discriminative power than an enlarged visual vocabulary. Of course, the optimal way to exploit structure both in the image and in the feature space may be to combine them in a unified multiresolution framework; Liu et al. (2007) is a recent example of work in this direction.

1.3.3 Caltech-101

Our second set of experiments is on the Caltech-101 database (Fei-Fei et al., 2004) (Fig. 1.4). This database contains from 31 to 800 images per category. Most images are medium resolution, i.e., about 300×300 pixels. Caltech-101 is one of the most diverse object database available today, though it is not without shortcomings. Namely, most images feature relatively little clutter, and the objects are centered and occupy most of the image. In addition, a number of categories, such as minaret (see Fig. 1.4), are affected

	Weak features		Strong features (200)	
L	Single-level	Pyramid	Single-level	Pyramid
0	15.5 ±0.9		41.2 ±1.2	
1	31.4 ±1.2	32.8 ±1.3	55.9 ±0.9	57.0 ±0.8
2	47.2 ±1.1	49.3 ±1.4	63.6 ±0.9	**64.6** ±0.8
3	52.2 ±0.8	**54.0** ±1.1	60.3 ±0.9	64.6 ±0.7

Table 1.2. Classification results for the Caltech-101 database.

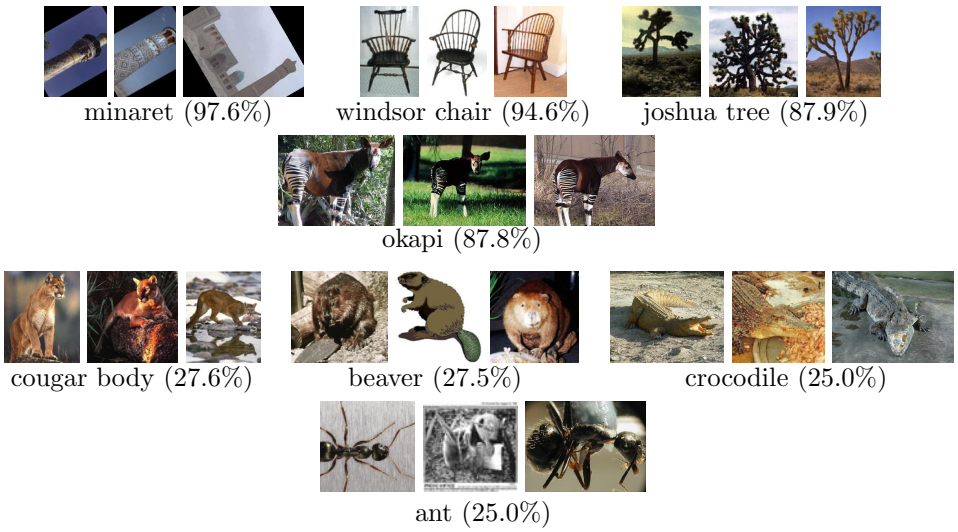

Fig. 1.4. Top two rows: some classes on which our method ($L = 2, M = 200$) achieved high performance. Bottom two rows: classes on which our method performed poorly.

by "corner" artifacts resulting from artificial image rotation. Though these artifacts are semantically irrelevant, they can provide stable cues resulting in misleadingly high recognition rates.

We follow the standard experimental setup of training on 30 images per class and testing on the rest. For efficiency, we limit the number of test images to 50 per class. Note that, because some categories are very small, we may end up with just a single test image per class. Table 1.2 gives a breakdown of classification rates for different pyramid levels for weak features and strong features with $M = 200$. The results for $M = 400$ are not shown, because just as for the scene category database, they do not bring any significant improvement. For $L = 0$, strong features give 41.2%, which is slightly below the 43% reported by Grauman and Darrell (2005). Our best result is 64.6%, achieved with strong features at $L = 2$. Thus, the

spatial pyramid improves over the bag of features by over 20%. The behavior of weak features on this database is also noteworthy: for $L = 0$, they give a classification rate of 15.5%, which is consistent with a naive graylevel correlation baseline Berg et al. (2005), but in conjunction with a four-level spatial pyramid, their performance rises to a much more respectable 54%.

Fig. 1.4 shows a few of the "easiest" and "hardest" object classes for our method. The successful classes are either dominated by rotation artifacts (like minaret), have very little clutter (like windsor chair), or represent coherent natural "scenes" (like joshua tree and okapi). The least successful classes are either textureless animals (like beaver and cougar), animals that camouflage well in their environment (like crocodile), or "thin" objects (like ant).

At the time of its initial publication, the results of our method have exceeded previously published state-of-the-art orderless methods (Grauman and Darrell, 2005; Zhang et al., 2007) and methods based on precise geometric correspondence (Berg et al., 2005). Concurrently, two other methods (Wang et al., 2006; Zhang et al., 2006) were published reporting classification rates similar to ours. Since then, a few more approaches have established new records on the Caltech-101 dataset, and these will be discussed in Section 1.4. Finally, it must also be noted that a re-implementation of the spatial pyramid method has received good baseline performance for the Caltech-256 dataset (Griffin et al., 2007), which is the "next generation" version of Caltech-101.

1.3.4 Discussion

In summary, our experiments have shown that the spatial pyramid method does very well on global scene classification tasks, or on object recognition tasks in the absence of clutter with most of the objects assuming "canonical" poses, as in the Caltech-101 dataset. However, because the spatial pyramid method relies on a non-invariant spatial decomposition of an image, it may seem susceptible to heavy clutter and geometric deformations. In practice, though, this is not the case. As discussed in Section 1.3.2, spatial pyramid matching tends to "zero in" on the scale that contains the most discriminative spatial information. If a dataset happens to be so highly variable that global position of features yields no useful cues at all, the matching scheme will simply "fall back" on level 0, which is equivalent to an orderless bag of features. To test how well spatial pyramid matching performs under highly variable conditions, we have performed another set of experiments on the Graz dataset (Opelt et al., 2004), which features people and

bikes in varying positions and at varying scales, against heavily cluttered backgrounds. Even in this relatively adverse setting, the spatial pyramid was still able to achieve about a 4% improvement over an orderless bag of features (see Lazebnik et al. (2006) for details). These results underscore the surprising and ubiquitous power of global scene statistics: even in highly variable datasets like the Graz, they can still provide useful discriminative information.

1.4 Applications and Extensions

This section surveys the extensions and applications of spatial pyramid matching that have appeared since its original publication (Lazebnik et al., 2006). Major themes and areas of improvement include (1) learning adaptive weights for different levels of the spatial pyramid; (2) extending the weighted kernel framework to combine multiple feature "channels"; (3) applying the spatial pyramid within an image sub-window for more precise object localization; and (4) extending pyramid matching to video.

Bosch et al. (2007a) have generalized the spatial pyramid kernel in two ways. First, they do not restrict themselves to the histogram intersection kernel for single-level comparisions, but consider other kinds of kernels for comparing bags of features, including kernels based on χ^2 distance. Second, instead of using fixed weights determined by the approximate geometric matching formulation, they select class-specific weights for each level using a validation set. Moreover, this adaptive weighting scheme extends not only across different pyramid levels, but also across different feature types. Specifically, Bosch et al. (2007a) use histograms of gradient orientations (Dalal and Triggs, 2005), as well as SIFT descriptors computed either on the grayscale image or the three color channels. This approach has achieved 77.8% accuracy on Caltech-101. In Bosch et al. (2007b), spatial pyramid matching is further generalized to find a region of interest containing the object, which increases the performance level to 81.3%. This work also introduces a *random forest* classification approach that achieves slightly lower classification accuracy than support vector machines, but is much more computationally efficient for the task. Varma and Ray (2007) use the spatial pyramid kernel with the same features as (Bosch et al., 2007a,b), as well as geometric blur descriptors (Berg and Malik, 2001). Instead of selecting the adaptive kernel weights by cross-validation, they introduce a convex optimization framework to learn them automatically. This method achieves 87.82% accuracy on Caltech-101.

Besides Caltech-101, another major benchmark in the recognition com-

munity is the PASCAL Visual Object Classes Challenge (Everingham et al., 2006). One of the top-performing methods for this challenge is by Chum and Zisserman (2007), who use a spatial pyramid inside an image sub-window to automatically learn regions of interest containing instances of a given object class, without requiring training data annotated with bounding boxes. Marszałek et al. (2007) present another high-performing method on the PASCAL challenge. Like Bosch et al. (2007a,b); Varma and Ray (2007), this method learns adaptive weights to combine different feature "channels," but it uses a genetic algorithm to accomplish this task.

Finally, a few recent methods apply the spatial pyramid kernel to video. Liu et al. (2007) have combined spatial pyramid kernels with feature space pyramid kernels, for an approach that performs simultaneous multi-scale partitioning of the high-dimensional feature space and the two-dimensional image space. The resulting *feature and space covariant kernel* has been shown to perform better than either the methods of Grauman and Darrell (2005) or Lazebnik et al. (2006). While this work has been successfully applied to video indexing on the TRECVID dataset, it does not include any explicit matching over the temporal domain. By contrast, Xu and Chang (2007) develop a scheme for *temporally aligned pyramid matching* to explicitly capture multi-scale temporal structure in establishing correspondence between video clips. Laptev et al. (2008) also generalize pyramid matching to the spatio-temporal domain for the application of human action recognition in movies.

1.5 Conclusion

This chapter has discussed a "holistic" approach for image categorization based on a modification of pyramid match kernels (Grauman and Darrell, 2005). This method, which works by repeatedly subdividing an image and computing histograms of image features over the resulting subregions, has shown promising results in the initial experiments (Lazebnik et al., 2006), and has since been extended in multiple ways by multiple researchers, as discussed in Section 1.4. Despite its simplicity and its reliance on global spatial information, spatial pyramid matching consistently achieves an improvement over orderless bag-of-features image representations. This is not a trivial accomplishment, given that a well-designed bag-of-features method can outperform more sophisticated approaches based on parts and relations (Zhang et al., 2007). The computational efficiency of spatial pyramid matching, together with its tendency to yield unexpectedly high recognition rates on challenging data, make it a good baseline for calibrating new datasets, such

as Caltech-256 (Griffin et al., 2007), as well as a highly effective "trick" for boosting the performance of any method that combines kernel-based classification and local features.

Despite the above practical advantages, we must emphasize that by itself, the spatial pyramid method is not meant as a sufficient or definitive solution to the general problem of scene recognition or understanding. After all, global scene statistics are not capable of localizing objects or making fine-scale semantic distinctions necessary to discriminate between subtly different scenes. For example, to correctly determine whether a given scene is a living room or bedroom, it may be necessary to locate and recognize individual objects, such as beds, sofas, coffee tables, etc. Qualitatively, the spatial pyramid method seems to capture something akin to "pre-attentive" perceptual similarity, but extensive psychophysical studies are required to validate and quantify this conjecture (see Oliva and Torralba (2007) for some initial insights on the relationship between context models in human and computer vision). In the future, in addition to pursuing connections to computational models of human vision, we are also interested in developing a broad theoretical framework that encompasses spatial pyramid matching and other locally orderless representations in the visual and textual domains (Koenderink and Van Doorn, 1999; Lebanon et al., 2007).

Acknowledgments

The majority of the research presented in this chapter was done while S. Lazebnik and J. Ponce were with the Department of Computer Science and the Beckman Institute at the University of Illinois at Urbana-Champaign, USA. This research was supported in part by the National Science Foundation under grant IIS-0535152 and the INRIA associated team Thetys.

Bibliography

A. Berg and J. Malik. Geometric blur for template matching. In *Proc. CVPR*, volume 1, pages 607–614, 2001.

A. Berg, T. Berg, and J. Malik. Shape matching and object recognition using low distortion correspondences. In *Proc. CVPR*, volume 1, pages 26–33, 2005.

A. Bosch, A. Zisserman, and X. Muñoz. Representing shape with a spatial pyramid kernel. In *CIVR '07: Proceedings of the 6th ACM international conference on Image and video retrieval*, pages 401–408, 2007a.

A. Bosch, A. Zisserman, and X. Muñoz. Image classification using random forests and ferns. In *Proc. ICCV*, 2007b.

O. Chum and A. Zisserman. An exemplar model for learning object classes. In *Proc. CVPR*, 2007.

G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *ECCV Workshop on Statistical Learning in Computer Vision*, 2004.

M. Cuturi and K. Fukumizu. Kernels on structured objects through nested histograms. In *Advances in Neural Information Processing Systems*, 2006.

N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. CVPR*, volume II, pages 886–893, 2005.

M. Everingham, A. Zisserman, C. K. I. Williams, and L. Van Gool. The PASCAL Visual Object Classes Challenge 2006 (VOC2006) Results. http://www.pascal-network.org/challenges/VOC/voc2006/results.pdf, 2006.

L. Fei-Fei and P. Perona. A Bayesian hierarchical model for learning natural scene categories. In *Proc. CVPR*, 2005.

L. Fei-Fei, R. VanRullen, C. Koch, and P. Perona. Natural scene categorization in the near absense of attention. *Procedings of the National Academy of Sciences USA*, 99(14):9596–9601, 2002.

L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories. In *IEEE CVPR Workshop on Generative-Model Based Vision*, 2004. http://www.vision.caltech.edu/Image_Datasets/Caltech101.

M. Gorkani and R. Picard. Texture orientation for sorting photos "at a glance". In *Proc. ICPR*, volume 1, pages 459–464, 1994.

K. Grauman and T. Darrell. Pyramid match kernels: Discriminative classification with sets of image features. In *Proc. ICCV*, 2005.

K. Grauman and T. Darrell. The pyramid match kernel: Efficient learning with sets of features. *Journal of Machine Learning Research*, 8:725–760, April 2007.

G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. Technical Report 7694, California Institute of Technology, 2007. URL `http://authors.library.caltech.edu/7694`.

J. Hays and A. Efros. Scene completion using millions of photographs. In *SIGGRAPH*, 2007.

D. Hoiem, A. Efros, and M. Hebert. Geometric context from a single image. In *Proc. ICCV*, 2005.

J. Koenderink and A. Van Doorn. The structure of locally orderless images. *IJCV*, 31(2/3):159–168, 1999.

I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *Proc. CVPR*, 2008.

S. Lazebnik, C. Schmid, and J. Ponce. A sparse texture representation using local affine regions. *IEEE Trans. PAMI*, 27(8):1265–1278, 2005.

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

G. Lebanon, Y. Mao, and J. Dillon. The locally weighted bag of words framework for document representation. *Journal of Machine Learning Research*, 8:2405–2441, October 2007.

X. Liu, D. Wang, J. Li, and B. Zhang. The feature and spatial covariant kernel: adding implicit spatial constraints to histogram. In *CIVR '07: Proceedings of the 6th ACM international conference on Image and video retrieval*, pages 565–572, 2007.

D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

S. Maji, A. Berg, and J. Malik. Classification using intersection kernel support vector machines is efficient. In *Proc. CVPR*, 2008.

M. Marszałek, C. Schmid, H. Harzallah, and J. van de Weijer. Learning object representations for visual object class recognition. In *ICCV 2007 Visual Recognition Challange workshop*, 2007.

H. Murase and S. K. Nayar. Visual learning and recognition of 3D objects from appearance. *IJCV*, 14(1):5–24, 1995.

A. Oliva and A. Torralba. Modeling the shape of the scene: a holistic representation of the spatial envelope. *IJCV*, 42(3):145–175, 2001.

A. Oliva and A. Torralba. The role of context in object recognition. *Trends in Cognitive Sciences*, 11(12):520–527, December 2007.

A. Opelt, M. Fussenegger, A. Pinz, and P. Auer. Weak hypotheses and boosting for generic object detection and recognition. In *Proc. ECCV*, volume 2, pages 71–84, 2004. http://www.emt.tugraz.at/˜pinz/data.

J. Ponce, T. L. Berg, M. Everingham, D. A. Forsyth, M. Hebert, S. Lazebnik, M. Marszałek, C. Schmid, B. C. Russell, A. Torralba, C. K. I. Williams, J. Zhang, and A. Zisserman. Dataset issues in object recognition. In J. Ponce, M. Hebert, C. Schmid, and A. Zisserman, editors, *Toward Category-Level Object Recognition*. Springer-Verlag Lecture Notes in Computer Science 4170, 2006.

L. Renninger and J. Malik. When is scene identification just texture recognition? *Vision Research*, 44:2301–2311, 2004.

B. C. Russell, A. Torralba, C. Liu, R. Fergus, and W. T. Freeman. Object recognition by scene alignment. In *Advances in Neural Information Processing Systems*, 2007.

B. Schiele and J. Crowley. Recognition without correspondence using multidimensional receptive field histograms. *IJCV*, 36(1):31–50, 2000.

C. Schmid and R. Mohr. Local greyvalue invariants for image retrieval. *IEEE Trans. PAMI*, 19(5):530–535, 1997.

J. Sivic, B. Russell, A. Efros, A. Zisserman, and W. Freeman. Discovering objects and their location in images. In *Proc. ICCV*, 2005.

D. Squire, W. Muller, H. Muller, and J. Raki. Content-based query of of image databases, inspirations from text retrieval: inverted files, frequency-based weights and relevance feedback. In *Proceedings of the 11th Scandinavian conference on image analysis*, pages 143–149, 1999.

M. Swain and D. Ballard. Color indexing. *IJCV*, 7(1):11–32, 1991.

M. Szummer and R. Picard. Indoor-outdoor image classification. In *IEEE International Workshop on Content-Based Access of Image and Video Databases*, pages 42–51, 1998.

S. Thorpe, D. Fize, and C. Marlot. Speed of processing in the human visual system. *Nature*, 381:520–522, 1996.

A. Torralba, K. P. Murphy, W. T. Freeman, and M. A. Rubin. Context-based vision system for place and object recognition. In *Proc. ICCV*, 2003.

A. Torralba, R. Fergus, and W. T. Freeman. 80 million tiny images: a large dataset for non-parametric object and scene recognition. Technical report, MIT, 2007.

A. Torralba, R. Fergus, and Y. Weiss. Small codes and large databases for recognition. In *Proc. CVPR*, 2008.

M. Turk and A. Pentland. Face recognition using eigenfaces. In *Proc. CVPR*, pages 586–591, 1991.

A. Vailaya, A. Jain, and H.-J. Zhang. On image classification: city vs. landscape. In *Proceedings of the IEEE Workshop on Content-Based Access of Image and Video Libraries*, pages 3–8, 1998.

M. Varma and D. Ray. Learning the discriminative power-invariance trade-off. In *Proc. ICCV*, 2007.

G. Wang, Y. Zhang, and L. Fei-Fei. Using dependent regions for object categorization in a generative framework. In *Proc. CVPR*, 2006.

D. Xu and S.-F. Chang. Visual event recognition in news video using kernel methods with multi-level temporal alignment. In *Proc. CVPR*, 2007.

H. Zhang, A. Berg, M. Maire, and J. Malik. SVM-KNN: Discriminative nearest neighbor classification for visual category recognition. In *Proc. CVPR*, 2006.

J. Zhang, M. Marszałek, S. Lazebnik, and C. Schmid. Local features and kernels for classifcation of texture and object categories: A comprehensive study. *IJCV*, 73(2):213–238, June 2007.