Latent Pyramidal Regions for Recognizing Scenes

Fereshteh Sadeghi and Marshall F. Tappen

University of Central Florida, Orlando, Florida
{fsadeghi,mtappen}@eecs.ucf.edu

Abstract. In this paper we propose a simple but efficient image representation for solving the scene classification problem. Our new representation combines the benefits of spatial pyramid representation using nonlinear feature coding and latent Support Vector Machine (LSVM) to train a set of Latent Pyramidal Regions (LPR). Each of our LPRs captures a discriminative characteristic of the scenes and is trained by searching over all possible sub-windows of the images in a latent SVM training procedure. Each LPR is represented in a spatial pyramid and uses non-linear locality constraint coding for learning both shape and texture patterns of the scene. The final response of the LPRs form a single feature vector which we call the LPR representation and can be used for the classification task. We tested our proposed scene representation model in three datasets which contain a variety of scene categories (15-Scenes, UIUC-Sports and MIT-indoor). Our LPR representation obtains state-of-the-art results on all these datasets which shows that it can simultaneously model the global and local scene characteristics in a single framework and is general enough to be used for both indoor and outdoor scene classification.

1 Introduction

In this paper, we propose a new approach for representing images for recognition systems and particularly scene recognition. The scene recognition problem can be viewed as a type of whole-image classification problem where the goal is to view the entire image and assign it a label identifying the type of scene depicted in the image. It can be argued that the dominant approaches used for solving whole-image classification problems have evolved to rest on three fundamental components:

- Dense extraction of gradient-based image descriptors, such as SIFT [1] or HoG[2]
- Representing descriptors using some form of coding, such as k-means, sparse coding [3] or Locality-Constrained Linear Coding (LLC) [4,5].
- Pooling feature descriptors together by different pooling methods[5,6], and using a spatial pyramid representation[4,5,7].

A. Fitzgibbon et al. (Eds.): ECCV 2012, Part V, LNCS 7576, pp. 228–241, 2012.

The spatial pyramid representation is popular because it captures the spatial aspects of images. Using this approach, a feature vector describing the image is created by using the image descriptors to create a feature vector describing the whole image. In [7], this feature vector is a histogram describing the frequency of various quantized image descriptors in the image. The image is then partitioned, usually into equal quadrants, and a feature vector is created to describe each sub-region. Each of these new vectors is concatenated with the original feature vector, describing the whole image. Each region can then be recursively subdivided with the feature vectors for each new sub-region being concatenated to the overall image's feature vector.

This results in a multi-scale representation where different parts of the feature vector represent both different scales and different spatial locations in the image. The intuition behind this representation for scene recognition is that it captures regularities in image appearance, such as the sky appearing at the top of images and roads tending to appear in the bottom half.

We propose that although this intuition is valid, the spatial pyramid representation's ability to model images is limited by the fixed grid that is typically used to create the feature vectors. Depending on the composition of the scene and the pose of the camera, scene elements can occur in a wide variety of configurations and may correspond poorly to the 4×4 and 16×16 grids that are typically used for pooling.

1.1 Paper Contribution

Inspired by the success of recent latent variable approaches [8,9] and Spatial Pyramids (SP) representation [4,7], we propose a new image representation designed to be used for discriminating between image classes. In our new representation, each feature value expresses a particular type of scene region that is present in the images of one category. To make this representation robust to different spatial configurations of images, the position of each scene region is treated as a latent variable that is optimized as part of the representation.

To capture the structure within a region, each region is represented with a spatial pyramid. Thus, we refer to these regions as **Latent Pyramidal Regions** and refer to this representation as the Latent Pyramidal Regions represention, or LPR representation. The power of this representation will be demonstrated in Section 6, where experiments on three different data sets will show that this representation out-performs other representations and models, including the recent work of Pandey et al. [8] that also incorporates latent-variable models into scene recognition. Below, we will introduce the Latent Pyramidal Region Representation, and then in Section 5 we will compare this representation with related work.

2 The Latent Pyramidal Region Representation

The fundamental unit in the Latent Pyramidal Regions, or LPR, representation is an image region detector that is parameterized to find image regions with a

specific appearance. Given an input image I, the vector $\boldsymbol{v}$ will denote the LPR representation of the image I.

Each element in the vector $\boldsymbol{v}$ is computed by finding the maximum response of a cost function applied to different sub-windows in the image. If v_i is the ith element of $\boldsymbol{v}$, we formally denote it as

$$v_i = \max_{w \in I} \theta_i^\top \boldsymbol{f}(I, w), \tag{1}$$

where $\boldsymbol{f}(I, w)$ is a function that returns a vector of features extracted from sub-window w in the image I. This max operation occurs over the set of all possible sub-windows in the image and thus w is the latent variable of our model. As will be discussed in Section 4, we represent these regions using the coding scheme proposed in [4].

The vector θ_i is a set of parameters that defines what type of image region each detector selects for. In our current implementation, these are trained discriminatively using methods described in Section 3.

In our implementation, each region detector examines sub-windows of a size that is fixed relative to the size of the image. For example, some detectors operate on windows with a size that is 40% of the size of the image, while other detectors may use a window that is 60% of the input image. As will be shown in the experiments, using multiple window scales improves the performance of the model.

The LPR representation has an intuitive explanation in terms of an object model with parts, like [9]. The vector $\boldsymbol{v}$ is created from a bag of part detectors by applying each detector to the image, then recording the highest score found by the detector.

3 Learning the Parameters for Region Detectors

The key parameters in the LPR representation are the region detector parameters, denoted as θ_i in Eq. (1). Because this representation will eventually be used to discriminate between different scene categories, the parameters of the region detectors can be found through a discriminative training process based on one-versus-all training of a structural Latent SVM.

The underlying idea behind the training process is to build the set of region detectors optimized for separating each scene category from the others. A detector defined by parameters θ_k is created by first choosing a particular scene category k. Each training image, I can then be assigned a label $y \in \{-1, +1\}$, with y taking the label $+1$ if the I belongs to category k. Otherwise, y takes the value -1.

The goal in training is to learn a prediction rule of the form:

$$\mathcal{F}_\theta(I) = \operatorname*{argmax}_{k,w} [\theta_k^\top \boldsymbol{f}(I, w)], \tag{2}$$

where k will be the predicted label and w will be the sub-window with the highest detection score. As in Eq. (1), the function $\boldsymbol{f}(I, w)$ evaluates to a vector of features extracted from sub-window w.

The parameter vector θ is found by minimizing the cost function

$$f(\theta) = \frac{\lambda}{2}\|\theta\|^2 + \sum_{j=1}^{N} R_j(\theta), \tag{3}$$

with λ balancing between the quadratic regularizer $\|\theta\|^2$ and the risk function $R_j(\theta)$, which is summed over the N training images.

The risk function $R_j(\theta)$ is structured to penalize the prediction function when it predicts an incorrect label. Following [10], it is formulated as:

$$R_j(\theta) = \max_{y,w} \left[\theta^\top \Phi(I_j, y, w) + \Delta(y, y_j)\right] - \max_{w} \theta^\top \Phi(I_j, y_j, w) \tag{4}$$

The recognition loss $\Delta(y, y_j)$ is a 0/1 loss that measures the difference between the ground-truth label y_j and the predicted label y, i.e. $\Delta_{0/1}(y, y_j) = 1$ if $y \neq y_j$, and 0 otherwise. The feature vector Φ is defined to reflect the statistics of image I local to the sub-window w, i.e. $\Phi(I, y, w) = \boldsymbol{f}(I, w)$ and following [11] we define Φ as zero vectors for the images of other categories (i.e. $y_j = -1$) during the training.

For solving the non-convex optimization problem of Eq. (3), we use non-convex bundle optimization in [12] which is a recent variant of bundle methods for regularized risk minimization [13,14]. This method iterates between finding the best sub-window w and the optimal discriminative parameter vector θ, until convergence. In each iteration, a new linear cutting plane is found via a subgradient of the objective function which will finally build up a piecewise quadratic approximation. For finding the latent sub-window w^* in each image I_j, we define the inference function $\mathcal{I}(I_j, w, y, \theta) = \theta^\top \Phi(I_j, y, w)$ where:

$$w^* = \underset{w \in I_j}{\operatorname{argmax}} [\mathcal{I}(I_j, w, y, \theta)]. \tag{5}$$

Then we predict the label y^* by using w^*,

$$y^* = \underset{y \in \mathcal{Y}}{\operatorname{argmax}} [\mathcal{I}(I_j, w^*, y, \theta) + \Delta(y, y_j)]. \tag{6}$$

The optimization method of [12] improves the approximation of the quadratic lower bound of the objective function by iteratively adding a new cutting plane using the sub-gradient of the cost function, which is

$$\delta_\theta f = \lambda\theta + \sum_{j=1}^{N} [\Phi(I_j, y^*, w^*) - \Phi(I_j, y_j, w^*)]. \tag{7}$$

In practice, it is possible to learn multiple region detectors in one cycle of one-versus-all training by introducing multiple latent w variables and additively extending the feature function $\boldsymbol{f}(\cdot)$ to accommodate multiple windows.

As mentioned earlier, the image is represented by the vector $\boldsymbol{v}$ which is the concatenation of the maximum response of each detector in the image. If the problem contains K classes and L region detectors are learned per class, this will result in LK response values in the vector $\boldsymbol{v}$.

4 Features and Image Representation

The image content inside each region is represented using the locality constraint coding scheme proposed in [4]. The actual image descriptors were computed using the dense SIFT routine from the LabelMe toolbox. For quantizing the SIFT features and constructing the codebook we use a sparse-coding algorithm proposed in [3] to find an over-complete basis that will be used to represent the descriptors. We use the sparse-coding algorithm because the ℓ_1 regularization in the optimization produces sparse bases that are robust to irrelevant features and noise in the data. Using this codebook, descriptors are then encoded using the LLC coding in [4]. As mentioned in the introduction, each region is represented using a spatial pyramid representation with 3 levels. The image was represented with 16×16 SIFT descriptors that were computed at every 8 pixels. The features inside a cell in the spatial pyramid are pooled using a max-pooling operation, similar to [4]. Following [5] our codebook is generated with 1024 entries and is optimized from a set of 200,000 randomly sampled descriptors.

5 Related Work

Our work is related to recent efforts to combine both local and global information for scene recognition. As mentioned in the introduction, a number of systems have been proposed that use a spatial pyramid combined with a coding of image descriptors [4,5,7]. Other approaches including [15] propose training a classifier with both global scene labels and salient regions manually segmented by humans.

In [16,17], Li et al. propose another unique object-centered approach called the Object Bank(OB). In this model, each image is represented as the response map of a large number of pre-trained generic object detectors. The key challenge in this approach is the need to have thousands of object detectors, which is computationally expensive.

Our proposed model also has similarities to the recent work of Pandey and Lazebnik [8]. The recognition system in [8] utilizes the deformable object detector made available as part of the work in [9] to classify scenes without requiring human-segmented regions [8]. While both our work and [8] use latent variable models, the key differences in this work lie in how the models are constructed. Imposing the object model from [9] on scenes has two significant limitations:

1. This imposes a strong spatial structure on the scene with parts being tied to specific locations relative to the root filter.
2. This combines the localization of the parts with the classification of the object category. In deformable parts model [9], the same parameters used to position the part locations are also used to compute the classification score.

The significance and novelty of our approach lie in how we handle these two issues differently:

1. Responding to the varied appearance of scenes, our model removes spatial constraints and focuses on finding image regions that characterize scene appearance.

2. Our approach separates the scene categorization classifier from the model for identifying image regions. This allows the classifier to optimize the weights for distinguishing between classes without having to balance how the weight values will affect which image regions are chosen.

These design decisions are validated by our experiments in Section 6.4. Our representation significantly raises recognition accuracy on the MIT Indoor dataset to over 44%, compared with an accuracy rate of 30.4% reported in [8].

5.1 Limitations of This Approach

The primary limitation of this approach is that it is discriminatively trained for a particular scene recognition problem. This somewhat blurs the line between a classifier and image representation. However, in defense of this approach, we submit that most recognition systems rely on a codebook that has been optimized on the same images that will be used for training the system. While the LPR representation is more directly optimized for specific categories, codebook based methods are often implicitly tied to the dataset through the codebook generation process.

6 Experiments

In this section we evaluate the performance of the proposed method on three scene datasets with diverse types of scenes including natural scenes (15-Scenes [7]), complex event and activity images (UIUC-Sports [18]), and cluttered indoor scenes (MIT-Indoor [15]). In future work, we plan to pursue scaling our approach to solve large-scale datasets like the SUN database[19].

On each of the three mentioned datasets, we report three key results, in addition to results presented in previous work:

- The accuracy computed using a linear SVM combined with the spatial pyramid representation of the image created by the software released by the authors of [4]. This representation encodes the image descriptors using a locality-constrained coding(LLC) scheme [4]. Because this representation is the concatenation of multiple feature vectors, the high dimensionality of this descriptor limits the classifier to a linear classifier. The results reported for LLC are obtained using the same codebook that we used in our system.
- The accuracy computed using the maximum response of region detectors associated with each class. If we denote V_k to be the set of all region detectors trained to respond to class k, using the training procedure in Section 3, the classification score for the class is computed by summing the reponse of those detectors. Formally, this is expressed as

$$y = \underset{k \in K}{\operatorname{argmax}} \left[\sum_{i \in V_k} v_i \right], \tag{8}$$

where there are K possible classes.

This is similar to the approach taken in [8,9] and will be referred to as LPR-MS.
– The accuracy computed using an SVM with a Radial Basis-Function(RBF) kernel and the LPR representation as well as a Linear SVM. We refer to these as LPR-RBF and LPR-LIN.

In the experiments our region detectors are initialized with a random sub-window of the image and are trained to learn a discriminative region of the scene by the procedure explained in Section 3. All accuracies are the average of the per-class recognition rates which is the mean over the diagonal values of the confusion matrix.

6.1 Key Results

While the following sections will describe results in detail, we wish to highlight the following key results:

– While the LLC representation and LPR representation use the exact same descriptors and coding scheme, the LPR representation consistently outperforms LLC.
– The LPR representation consistently outperforms other single feature accuracy results. When other systems outperform the LPR representation, this requires the fusion of multiple features. For example, by fusing five different types of features Xiao et al. report an accuracy of 88.1% on the 15-scene dataset in [19]. As reported in [19] the highest accuracy on the 15-scene dataset achieved by any single feature is 81.2%.
– Using just gradient-based descriptors, the proposed LPR representation outperforms the deformable part model approach proposed in [8] by over 14% accuracy. The approach in [8] must incorporate additional color information to be competitive.

6.2 15-Scenes Dataset

The 15-Scenes dataset contains 15 natural scene classes including 4485 images from a variety of outdoor natural scenes, outdoor man-made scenes and indoor scenes. For the test/train split we followed the setting suggested in [7]. Table 1 summarizes the performance comparison of the proposed method and state-of-the-art methods. Here we have used three LPRs with the size of 100%, 65% and 25% of the total image. Examples of the regions detected are presented in Fig. 1. As can be seen, our proposed method has outperformed the other state-of-the-art methods.

Based on the results of Table 1, the overall accuracy obtained from LPR-MS and the state-of-the-art methods are competitive. By using either an RBF or a linear kernel SVM classifier for our LPR representation, we obtain an accuracy of 85.81% and 85.75%, respectively, both of which are significantly higher than LLC (our baseline) and other state-of-the-art methods. These results show that

Table 1. The average per-class accuracy results of the proposed method compared with state of the art in the 15-Scenes dataset. The abbreviations for our approach are defined in Section 6.

Method	Accuracy	Method	Accuracy
KC [6]	76.67	LLC(baseline)	80.57
OB2 [16]	80.9	LPR-MS(our approach)	**83.29**
ScSPM [5]	80.28	LPR-LIN(our approach)	**85.72**
KSPM [7]	81.4	LPR-RBF(our approach)	**85.81**

Table 2. Per-class accuracy of the proposed method (LPR-MS and LPR-RBF) compared with our baseline (LLC [4]) and KSPM in 15-Scenes dataset. The categories are listed in decreasing order of classification accuracy of (LPR-RBF). The accuracies for (KSPM) are taken from [7].

class name	KSPM	LLC	LPR-MS	LPR-RBF	class name	KSPM	LLC	LPR-MS	LPR-RBF
CALsuburb	99	99	99	99	MITinsidecity	80	84	84	85
PARoffice	93	95	96	94	MITopencountry	70	66	71	83
MITforest	95	96	96	93	store	76	74	76	78
MITtallbuilding	91	93	93	93	industrial	65	65	70	77
MITstreet	90	87	89	93	kitchen	68	65	73	75
MITmountain	89	87	89	92	bedroom	68	71	68	73
MIThighway	87	84	87	92	livingroom	60	58	72	70
MITcoast	82	86	87	90					

the proposed method can perform well in all types of scene categories. To the best of our knowledge, the best result reported on the 15-Scenes dataset is 88% in [19] and [20], which is obtained by the fusion of five and eight different feature, respectively. However, LPR representation obtains higher accuracy than any single features reported in [19]. The per-class accuracy of the proposed method compared with KSPM and LLC is provided in Table 2.

6.3 UIUC-Sports Dataset

The UIUC-Sports dataset contains images from eight sport scene categories. This dataset includes 1,574 images of indoor and outdoor scenes that are highly cluttered by objects. The particular characteristic of this dataset is the presence of a structured foreground (e.g. players, sport instruments) in a highly textured background (e.g. sea, court, field). Furthermore, this dataset includes images of different activities that have similar backgrounds such as *sailing* and *rowing* or *polo*, *Bocce*, and *Croquet*. For the test/train split we followed the setting suggested in [16]. Here we have used the same setting for the number and size of LPRs as we used for the 15-Scenes dataset. According to Table 3, the proposed method could obtain improved results over both low-level features (i.e. GIST[21] and KSPM[7], which capture the global scene properties) and high-level semantical image representation methods (i.e. MM-Scene[22] and OB[17,16]). Furthermore, by using an SVM with radial basis-function kernel for classifying we obtain an accuracy of 86.25% which is a more than 4% increase over our baseline. Here, the linear kernel did not result in a significant improvement over LPR-MS but it's accuracy is still significantly better than LLC.

Table 3. The average per-class accuracy results of the proposed method compared with state of the art in the UIUC-Sports dataset

Method	Accuracy	Method	Accuracy
GIST[21]	63.88	OB1 [17]	77.88
MM-Scene [22]	71.7	LLC(our global term)	81.87
KSPM [7]	71.57	LPR-MS(our approach)	**85.0**
WWW[18]	73.4	LPR-LIN(our approach)	**85.2**
OB2 [16]	76.3	LPR-RBF(our approach)	**86.25**

Table 4. Per-class accuracy of the proposed method (LPR-MS and LPR-RBF) compared with our baseline (LLC [4]) and state-of-the-arts (KSPM and OB) on UIUC-Sports dataset. The categories are listed in decreasing order of classification accuracy of (LPR-RBF). The accuracies for (KSPM) and (OB) are taken from [17].

class name	KSPM	OB	LLC	LPR-MS	LPR-RBF
sailing	75	93	95	93	95
snow boarding	75	70	88	92	92
rowing	80	77	93	93	90
Rock climbing	93	88	93	82	90
polo	68	65	78	90	87
badminton	93	88	88	88	85
croquet	48	74	68	80	82
bocce	42	68	50	62	70

By comparing the per-class accuracy of our LPR representation with OB [17] and KSPM [7] which is summarised in Table 4, we see that the accuracy of the proposed method is significantly higher than the other two in most of the categories. This suggests that the proposed method is less confused by similar backgrounds such as sailing and rowing and similar visual structures like human players in polo and bocce.

Also, the results of Table 3 show that the accuracy obtained by the proposed method is significantly higher than the accuracy obtained by OB combined with GIST. This shows that the proposed method is capable of learning both local and global characteristics of the scenes without needing to be combined with other methods.

6.4 MIT-Indoor Dataset

MIT-Indoor is a very challenging dataset of 15,620 indoor scenes images in 67 different categories. For each category we used 80 images for training and 20 images for testing following the same test/train split as in [15].

Here, we have used four LPRs with size of 100%, 80%, 65%, and 50% of the total image size and we will show the effects of the LPRs in increasing the classification accuracy at the end of this section. In Section 6.5 and Fig. 1 we will show examples of the detected regions in this dataset.

Table 5 shows the classification results of the proposed method compared with the state-of-the-art. Based on the results, LPR-MS obtains an accuracy of 41.22% and can reach 44.41% accuracy by using and RBF kernel and 44.84% by using a linear kernel which are about 15% higher than the DPM method.

Table 5. The average per-class accuracy results of the proposed method compared with state-of-the-art on the MIT-indoor dataset. Our approach outperforms previous approaches without the need to use color information.

Method	Accuracy	Method	Accuracy
HOG [2]	22.8	DPM+GIST-color [8]	39.0
ROI+gist [15]	26.5	DPM+KSPM [8]	40.5
MM-scene [22]	28.0	DPM+KSPM+GIST-color [8]	43.1
GIST-color [2]	29.7	LLC(our global term)	37.32
DPM [8]	30.4	LPR-MS(our approach)	**41.22**
KSPM [7]	34.4	LPR-LIN(our approach)	**44.84**
CENTRIST [23]	36.9	LPR-RBF(our approach)	**44.41**
OB2 [16]	37.6		

When DPM is combined with GIST-color, KSPM and both of them, accuracies of 39%, 40.5% and 43.1% are obtained, respectively. These accuracies, though competitive, are still less than the accuracy obtained by our method. This shows that, as opposed to our method which can simultaneously learn global and local features of the scene, DPM can only capture local features and thus needs to be combined with other feature types to obtain satisfying results. Table 6 contains the per-class accuracies obtained by LPR, our baseline(LLC) and the best results obtained by DPM after fusion with both GIST and KSPM features.

To examine the contribution of each region in our LPR model, we tested our performance in the MIT-indoor dataset with different number of regions. Similar experiments are carried out on the other two datasets, and similar results are obtained. Table 7 contains the accuracy obtained by changing the numbers of region detectors from one to five. The regions cover 35%, 50%, 65%, 80% and 100% of the whole image, respectively. The obtained results show that our LPRs contributed significantly in improving the classification accuracy. More surprisingly, we obtained our best performance in the MIT-indoor dataset by using only four regions, while adding the smallest size regions (i.e. 35%) slightly reduced the accuracy. This result shows that in this dataset, adding regions with smaller sizes can make the system overfit on the very small details of the scene. This result is in contrast with [8], where the authors obtained their best result by using eight parts. We believe that the reason for this contrast lies in the fact that in [8] each part covers a much smaller portion of the scene image than our LPR regions (i.e. about 10% of the image). As opposed to our approach, where each LPR models one characteristic of the scene including foreground objects and background context, each part in [8] models one part of a recurring object available in images of each scene category and thus only learns local discriminative pattern of the scenes.

6.5 Qualitative Analysis of LPR

For evaluating the behaviour of the region detectors in the proposed method, we visualized the regions that are discovered by LPR representation in a number

Table 6. Per-class accuracy of the proposed method (LPR-MS and LPR-RBF) compared with our baseline (LLC [4]) and (DPM) on MIT-indoor dataset. The categories are listed in decreasing order of classification accuracy of (LPR-RBF). The accuracies for (DPM) are taken from [8].

class name	DPM+ GC+KSPM	LLC	LPR-MS	LPR-RBF	class name	DPM+ GC+KSPM	LLC	LPR-MS	LPR-RBF
cloister	95	85	85	95	bathroom	56	44	39	39
elevator	86	76	86	90	hairsalon	52	24	38	38
bowling	55	70	70	90	grocerystore	48	48	43	38
greenhouse	75	80	90	85	laundromat	50	32	41	36
inside subway	52	57	57	81	fastfood restaurant	24	18	24	35
garage	56	56	67	78	prisoncell	50	30	30	35
buffet	80	75	75	75	poolinside	45	15	10	35
inside bus	57	74	74	74	library	35	25	30	35
church inside	79	68	74	74	subway	62	29	33	33
closet	72	50	56	72	dining room	56	33	39	33
classroom	61	78	78	72	restaurant kitchen	13	30	22	30
corridor	67	67	76	71	mall	20	15	15	30
concert hall	80	60	55	70	bookstore	35	25	30	30
florist	89	68	68	68	bedroom	10	33	33	29
casino	47	53	58	68	videostore	23	5	14	27
stairscase	55	55	55	65	operating room	26	16	21	26
movietheater	55	65	70	65	museum	17	22	22	26
studiomusic	63	47	63	63	lobby	35	10	30	25
trainstation	70	70	75	60	gameroom	35	20	25	25
pantry	75	65	60	60	office	10	24	24	24
tv studio	50	50	50	56	laboratorywet	14	18	23	23
auditorium	33	50	61	56	shoeshop	16	26	26	21
kindergarden	40	30	55	55	deli	5	16	11	21
locker room	38	29	29	52	bakery	26	11	21	21
dentaloffice	48	38	43	52	artstudio	15	10	10	20
nursery	65	55	50	50	warehouse	29	10	14	19
hospitalroom	20	45	45	50	gym	33	6	17	17
bar	33	44	39	44	livingroom	20	5	35	15
winecellar	38	24	29	43	airport inside	10	0	15	15
kitchen	52	38	29	43	waitingroom	33	14	19	14
meeting room	77	41	45	41	jewelleryshop	5	5	18	14
computerroom	44	39	44	39	toystore	18	0	0	9
clothingstore	33	44	39	39	restaurant	10	10	15	0
children room	11	28	28	39					

Table 7. Average classification rate on MIT-indoor dataset obtained by different numbers of LPRs

number and size of LPRs	Accuracy
(1) 100%	37.32
(2) 100%,50%	38.15
(3) 100%,50%,65%	40.27
(4) 100%,50%,65%80%	41.22
(5) 100%,35%,50%65%,80%	40.91

of categories for all the three datasets. In our experiments we have trained the model with three LPRs for the 15-Scenes and UIUC-Sports dataset. Due to the larger number of categories in the MIT-Indoor dataset we trained LPR representation with four regions.

For each category in the datasets, Fig. 1 presents five examples of the test images in which LPR has discovered discriminative regions well (the first five columns) and one example in which inappropriate region is selected by the LPR (last column). In all the images the LPRs are shown with colored bounding boxes and the region with the biggest size(i.e 100% of the image) is not shown. The examples in the last column show that LPR in confused whenever a pattern appears in the image that seldomly occurs in the images of that category. In these images, the smaller region has not been able to find a discriminative pattern, however the larger region has captured the global context of the scene.

Fig. 1. The detected regions found by LPR in (MITMountain, CALsuburb, bedroom) categories of 15-Scenes, (Polo, Sailing) categories of UIUC-Sports and (movietheater, operating room, staircase, gym and livingroom) categories of MIT-indoor dataset. For each category five example of the test images in which LPR has discovered discriminative regions is presented (the first five columns) along with one example in which inappropriate region is selected by the LPR(last column).

For the 15-Scenes dataset, we have provided examples of the discovered regions by LPR in natural outdoor (*MITMountain*), natural man-made (*CALsuburb*) and indoor (*bedroom*) scenes. We observe that for each of the scene categories, the smaller region captures a discriminative recurring pattern of the scene that can be part of an object (e.g. *bed corner* in *bedroom* or *peak* of the *mountain*) or just a recurrent pattern (e.g. *sidewalk* and *lawn* in *suburb*). The larger region also detects the general pattern of the scene and yet has the flexibility to discard noise patterns and select a sub-window with the most information.

Similarly, in the UIUC-Sports dataset, the smaller region captures the *mast* in the *Sailing* category, whereas the larger region captures the *sailing boat, sky* and *water*. Likewise, in the *Polo* category, the smaller region has learned the pattern of *man riding a horse*, while the larger region has most of the important elements of the scene which is *man riding a horse in a field*. These examples show that our latent region detectors can efficiently find discriminative scene structures.

The MIT-indoor dataset, contains images of indoor scenes where usually several different objects occur in the images of one category. In these images, our LPR representation searches for recurring patterns of different objects and their combinations. For example, the smallest detected region contains *screen* in the *movietheater*, *sofa* in the *living room* or *bed* in the *operating room*. The bigger regions capture the combination of *screen* and *seats* in *movietheater*, *steps, wall* and *floor* in *staircase* or *sofa* and *desk* in the *living room*. Based on these examples we understand that an obvious advantage of LPR over DPM is that LPR can find discriminative regions of the scenes which contain a discriminative part of an object, a whole object, or the combination of several objects with their context; whereas DPM only seeks to find a recurring object of the scene. We conclude that our latent region detection approach has the ability to capture both local and global discriminative features and is general enough to be used for recognition tasks in all types of scenes.

7 Conclusion

In this paper, we addressed the problem of scene classification by using a latent SVM framework in which a set of region detectors are learned to capture the key characteristics of the scenes. Each region is called Latent Pyramidal Region and it is represented by a spatial pyramid and using nonlinear locality constraint coding. By learning the pattern of these regions. Our model can learn the local and global characteristics of the scene images efficiently. We conducted our experiments on the 15-Scenes dataset, UIUC-Sports dataset, and MIT-indoor scene dataset with 67 categories. The results show that our proposed method can obtain state-of-the art performance on a variety of scene datasets without needing to be combined with known global features like GIST and KSPM.

Acknowledgements. This work was supported by NSF grants IIS-0905387 and IIS-0916868.

References

1. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. IJCV 60(2) (2004)
2. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR, pp. 886–893 (2005)
3. Lee, H., Battle, A., Raina, R., Ng, A.Y.: Efficient sparse coding algorithms. In: NIPS, pp. 801–808 (2007)
4. Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., Gong, Y.: Locality-constrained linear coding for image classification. In: CVPR, pp. 3360–3367 (2010)
5. Yang, J., Yu, K., Gong, Y., Huang, T.: Linear spatial pyramid matching using sparse coding for image classification. In: CVPR, pp. 1794–1801 (2009)
6. Gemert, J.C., Geusebroek, J.M., Veenman, C.J., Smeulders, A.W.: Kernel Codebooks for Scene Categorization. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part III. LNCS, vol. 5304, pp. 696–709. Springer, Heidelberg (2008)
7. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: CVPR, pp. 2169–2178 (2006)
8. Pandey, M., Lazebnik, S.: Scene recognition and weakly supervised object localization with deformable part-based models. In: ICCV (2011)
9. Felzenszwalb, P., Girshick, R., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE Trans. PAMI 32(9), 1627–1645 (2010)
10. Yu, C.N.J., Joachims, T.: Learning structural svms with latent variables. In: ICML, pp. 1169–1176 (2009)
11. Blaschko, M.B., Lampert, C.H.: Learning to Localize Objects with Structured Output Regression. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part I. LNCS, vol. 5302, pp. 2–15. Springer, Heidelberg (2008)
12. Do, T.M.T., Artières, T.: Large margin training for hidden markov models with partially observed states. In: ICML, pp. 265–272 (2009)
13. Teo, C.H., Smola, A., Vishwanathan, S.V., Le, Q.V.: A scalable modular convex solver for regularized risk minimization. In: ACM SIGKDD, pp. 727–736 (2007)
14. Teo, C.H., Vishwanthan, S., Smola, A.J., Le, Q.V.: Bundle methods for regularized risk minimization. JMLR, 311–365 (2010)
15. Quattoni, A., Torralba, A.: Recognizing indoor scenes. In: CVPR, pp. 413–420 (2009)
16. Li, L.-J., HaoSu, E.P.X., Fei-Fei, L.: Object bank: A high-level image representation for scene classification & semantic feature sparsification. In: NIPS (2010)
17. Li, L.-J., Hao Su, Y.L., Fei-Fei, L.: Objects as attributes for scene classification. In: ECCV (2010)
18. Li, L.J., Fei-Fei, L.: What, where and who? classifying event by scene and object recognition. In: ICCV (2007)
19. Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A.: Sun database: Large-scale scene recognition from abbey to zoo. In: CVPR, pp. 3485–3492 (2010)
20. Han, Y., Liu, G.: Efficient learning of sample-specific discriminative features for scene classification. SPLetters 18(11), 683–686 (2011)
21. Oliva, A., Torralba, A.: Modeling the shape of the scene: A holistic representation of the spatial envelope. IJCV 42, 145–175 (2001)
22. Zhu, J., Li, L.J., Fei-Fei, L., Xing, E.P.: Large margin learning of upstream scene understanding models. In: NIPS (2010)
23. Wu, J., Rehg, J.: Centrist: A visual descriptor for scene categorization. IEEE Trans. PAMI 33(8), 1489–1501 (2011)