Learning Important Spatial Pooling Regions for Scene Classification

Di Lin[†] Cewu Lu[§] Renjie Liao[†] Jiaya Jia[†]
[†] The Chinese University of Hong Kong
[§] The Hong Kong University of Science and Technology

Abstract

We address the false response influence problem when learning and applying discriminative parts to construct the mid-level representation in scene classification. It is often caused by the complexity of latent image structure when convolving part filters with input images. This problem makes mid-level representation, even after pooling, not distinct enough to classify input data correctly to categories. Our solution is to learn important spatial pooling regions along with their appearance. The experiments show that this new framework suppresses false response and produces improved results on several datasets, including MIT-Indoor, 15-Scene, and UIUC 8-Sport. When combined with global image features, our method achieves state-of-the-art performance on these datasets.

1. Introduction

Finding discriminative parts [23] to construct mid-level representation is one of the main streams in scene classification. Discriminative parts, such as beds in bedroom, washing machines in laundry, and other distinct components, are important to identify scenes. They are generally more useful than simultaneously considering all pixels in an image. State-of-the-art results are yielded using this strategy, as described in [4].

These advanced methods learn a set of discriminative parts (filters). Given an image, response map are computed by filtering in a convolution way. To consider spatial information, mid-level representation is built upon the part response map, similar to Spatial Pyramid Matching (SPM) [12]. The constructed mid-level representation is the input to discriminative classifiers, e.g., SVM.

In this framework, we observe a common issue. That is, part learning in the first step could produce many false responses, which adversely influence mid-level representation construction.

Take the images in Figure 1 as examples. The "screen" filter is learned from the images of "movie theater" category. When applying the part filter to the images in (a)-(b) selected from categories "movie theater" and (c) from "florist" by convolution, respective region-level responses are obtained. All resulting maps in (d)-(f) contain many high response points in the "screen" region. It is also noticeable that the many false responses are produced on "non-screen" pixels.

To generate mid-level representations based on these response maps, several systems pool and concatenate these responses in spatial pyramids (3×2 spatial pooling in Figure 1(j)). The mid-level representation is shown as $F_a - F_c$ in (j). This routine possibly takes false responses into account and suffers from two drawbacks.

First, because the false responses in (d)-(f) are strong and clutter in image space, the concatenated response histograms F_b and F_c in (j) have high cross-correlation in many dimensions though images (b) and (c) are from different categories. Second, histograms F_a and F_b are different in many dimensions though the images (a) and (b) are from same category. Thus it is not that easy to classify images correctly into corresponding categories due to the negative impact of false responses.

Note these are not special cases. In fact, most discriminative parts are not similarly semantic as the "screen". Filters learned from these parts are more likely to generate false responses than the semantic ones. We examine many results output from this framework, the false responses generally affect discriminative power of image representation. Pre-defined spatial pooling strategies, like SPM, are empirically designed and could leave out these false responses in image representation. When the classification framework such as SVM is performed to classify images, a lot of misclassification errors are caused. At the very root, false response is neglected after learning discriminative parts and constructing mid-level representation in many systems.

We address this issue in this paper by introducing important spatial pooling regions (ISPRs) visualized in Figure 1, which are learned jointly with discriminative part appearance in a unified optimization framework. For the examples in Figure 1, our response maps are processed with the learned ISPRs to form the score maps shown in (g)-(i).

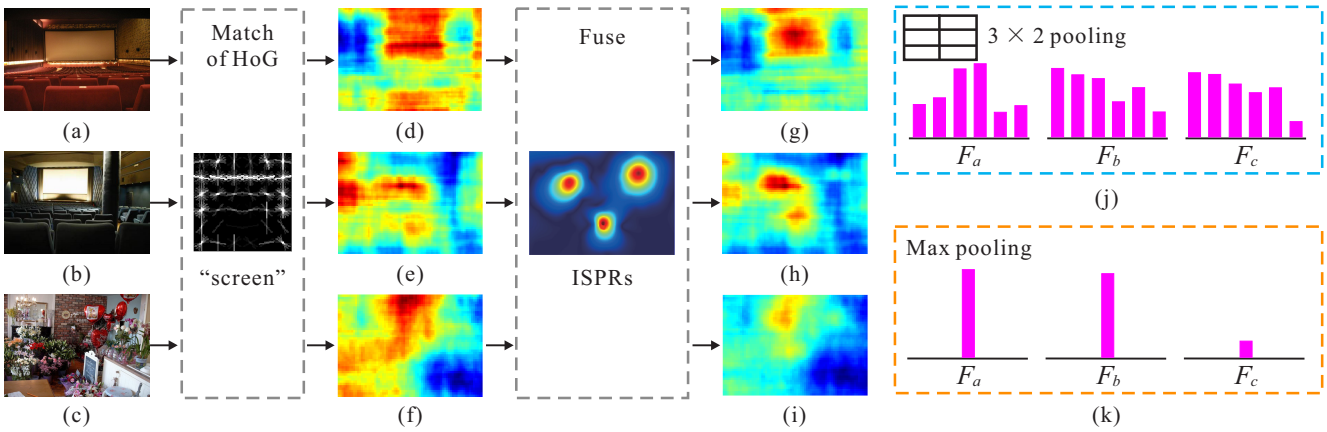


Figure 1. False response influence in scene classification. (a)-(b) Images in categories "movie theater". (c) Image in categories "florist". (d)-(f) Part response maps resulted from convolution of part filter "screen" and corresponding HoG maps. (g)-(i) Score maps after our ISPR process. (j) Mid-level representation by applying 3×2 pooling to (d)-(f). (k) Representation by applying max pooling to (g)-(i). In (j) and (k), F_a-F_c are representations corresponding to images (a)-(c) respectively.

False responses in (d)-(f) are suppressed in the results (g)-(i). Now, even a simple max pooling performed on them can make the final scores sufficiently different as shown in (k), good for categorizing novel images into correct scene classes.

Our contribution is twofold. First, we provide a new insight of spatial pooling to achieve discriminative mid-level representation. Second, a joint model to learn part appearance and ISPR is proposed. In the learning framework, ISPR can tap the potential of classifiers by suppressing false response in training samples.

We apply our method to scene classification on MIT-Indoor [23], 15-Scene [12] and UIUC 8-Sport [14] datasets. Experimental results show that our new mid-level representation enhances classification accuracy in general. Moreover, state-of-the-art performance is yielded by combining our mid-level representation with improved Fisher vector (IFV) [22], which is a global image feature.

2. Related Work

Discovering discriminative parts is an effective technique for scene classification. The term "discriminative part" was originally introduced in object recognition [5]. As explained in [23], scene can also be regarded as a combination of parts, which are called regions of interest (ROI). Because discriminative parts provide powerful representation of scene, exploiting them drew much attention recently.

This type of methods can be understood in three ways. First, distinct power of learned parts is used to alleviate visual ambiguity. Recent work [8, 24, 15, 16, 27, 17] discovered parts with specific visual concepts – that is, the learned part is expected to represent a cluster of visual objects. Second, unsupervised discovery of discriminative parts is dominating. Though handcrafted part filters are easier to comprehend, unsupervised frameworks [17, 28, 9, 10, 13, 20, 21, 25, 37] are more practical and efficient especially for large volume data.

Third, mid-level representation is employed to enhance the discriminative power in classification [2, 15, 16, 37]. State-of-the-art methods [27, 17, 8] discover discriminative parts and part responses obtained from convolution. They are concatenated as mid-level representation applied to discriminative classifiers. As described in [35], mid-level representation constructed with part responses maintains fine-grained power to discriminate a large set of inter and intra categories. Decent results manifest that this type of representation is an advantageous substitution or complementarity of traditional low-level ones [12, 22, 3, 18, 19, 31, 34].

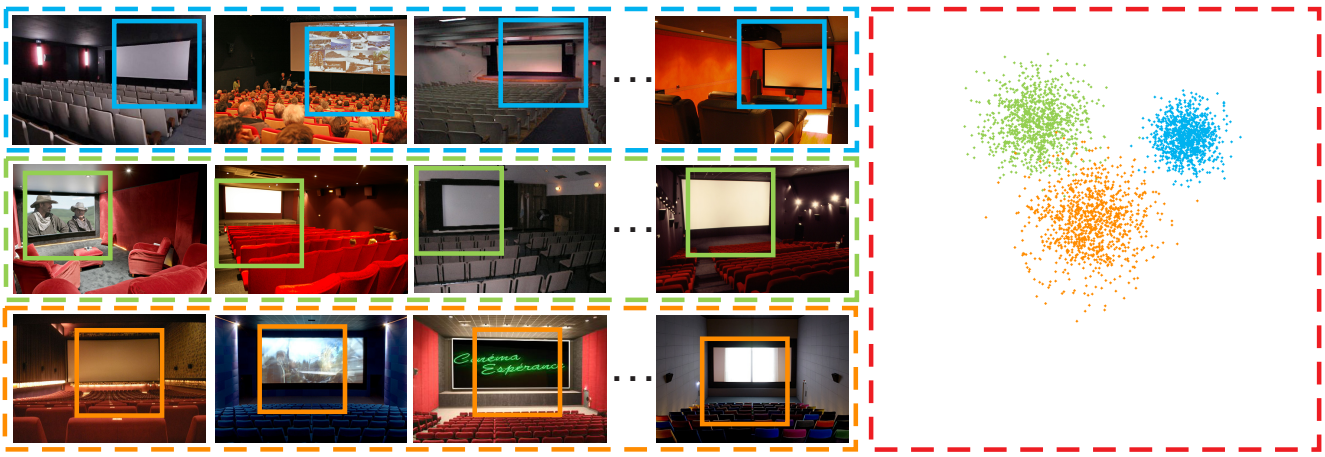
Noticing that previous methods still ignore the adverse impact of false response when constructing image representation, we develop a new scheme with better suppression of false response in order to generate more discriminative mid-level representation.

3. Our Model

We describe the joint model of part appearance and ISPR in this section.

3.1. Importance of Spatial Pooling Regions

The false response problem described in Section 1 can be understood in another way. For the images shown in Figure 2, useful visual cues include the projection screen and stage in order to label images as belonging to movie theaters. The chairs are repetitive texture and light may vary wildly in different images. They are not that distinct even for human to identify a theater scene.



(a) (b)

Figure 2. Discriminative part location modeling. (a) Screen structure learned from category "movie theater". These specific discriminative parts are mostly located within several clusters illustrated in (b).

Therefore, even if the discriminative part, e.g., the screen, has been correctly and statistically learned from data, when applying it to new images, those unrelated but structurally diversified environmental objects could mistakenly yield high response. In our experiments, this problem is ubiquitous.

Our solution is to eliminate the influencing structure by knowing statistically where false response could be produced. An observation experiment conducted by us is to plot centers of all discriminative parts in training data in a normalized image space, as shown in Figure 2(a). It is intriguing that most centers are located within a few clusters, as shown in Figure 2(b). Discriminative parts in other scene categories, such as beds in bedroom and the stage in concert hall, form similar patterns.

These patterns provide us with spatial clusters of part's occurrence. It implies that parts have higher chances to occur within clusters. In other words, the locations far away from the clusters primarily correspond to false response. By modeling and learning spatial clusters, we develop an adaptive method to infer responses in image space and make pooling more robust against incorrect input.

3.2. Model of Important Spatial Pooling Region (ISPR)

According to the above mentioned properties, our ISPRs should cover several clusters where discriminative parts frequently appear. Given $\Theta^i(p_i^c)$ denoting the i^{th} cluster centered at p_i^c, the overall important pooling region is

$$\bigcup_{i=1,\ldots,k} \Theta^i(p_i^c), \tag{1}$$

where k is the number of clusters and $\bigcup$ is the union operator. We model $\Theta^i(p_i^c)$ in the pixel level as

$$\phi(p, p^c, \sigma) = \exp(-\frac{||p - p^c||^2}{\sigma}), \tag{2}$$

where p and p^c are 2D coordinates of the part and its cluster center respectively. σ is the parameter controlling the coverage of ISPR. $\phi(p, p^c, \sigma)$ denotes the potential of part appearing in location p.

With Eq. (2), we use a mixture model to update the union process in Eq. (1) as

$$\Phi(p, p^c, \sigma) = \sum_{i=1}^{k} d_i \cdot \phi_d(p, p_i^c, \sigma_i), \tag{3}$$

where d_i is the weight of each ISPR.

3.3. Joint Model of Appearance and ISPR

Given a discriminative part, the occurrence of this part is affected by the appearance at certain locations besides the spatial clusters. The appearance is evaluated by the convolutional response of part filter [8, 15, 16, 27, 17, 37, 35]. Together with the above ISPR term, our joint model of part's occurrence at location p in image I is formulated as

$$f(I, p) = \underbrace{F \cdot H(I, p)}_{\text{Appearance Term}} + \underbrace{\Phi(p, p^c, \sigma)}_{\text{ISPR Term}}. \tag{4}$$

$F \cdot H(I, p)$ is the appearance term in convolution to obtain the response map. F is the part filter vector and $H(I, p)$ is the feature vector extracted from location p in image I.

Because structure of scenes varies spatially, we involve root filter similar to that of [5] to capture the global structure. The root filter is convolved with each input image to

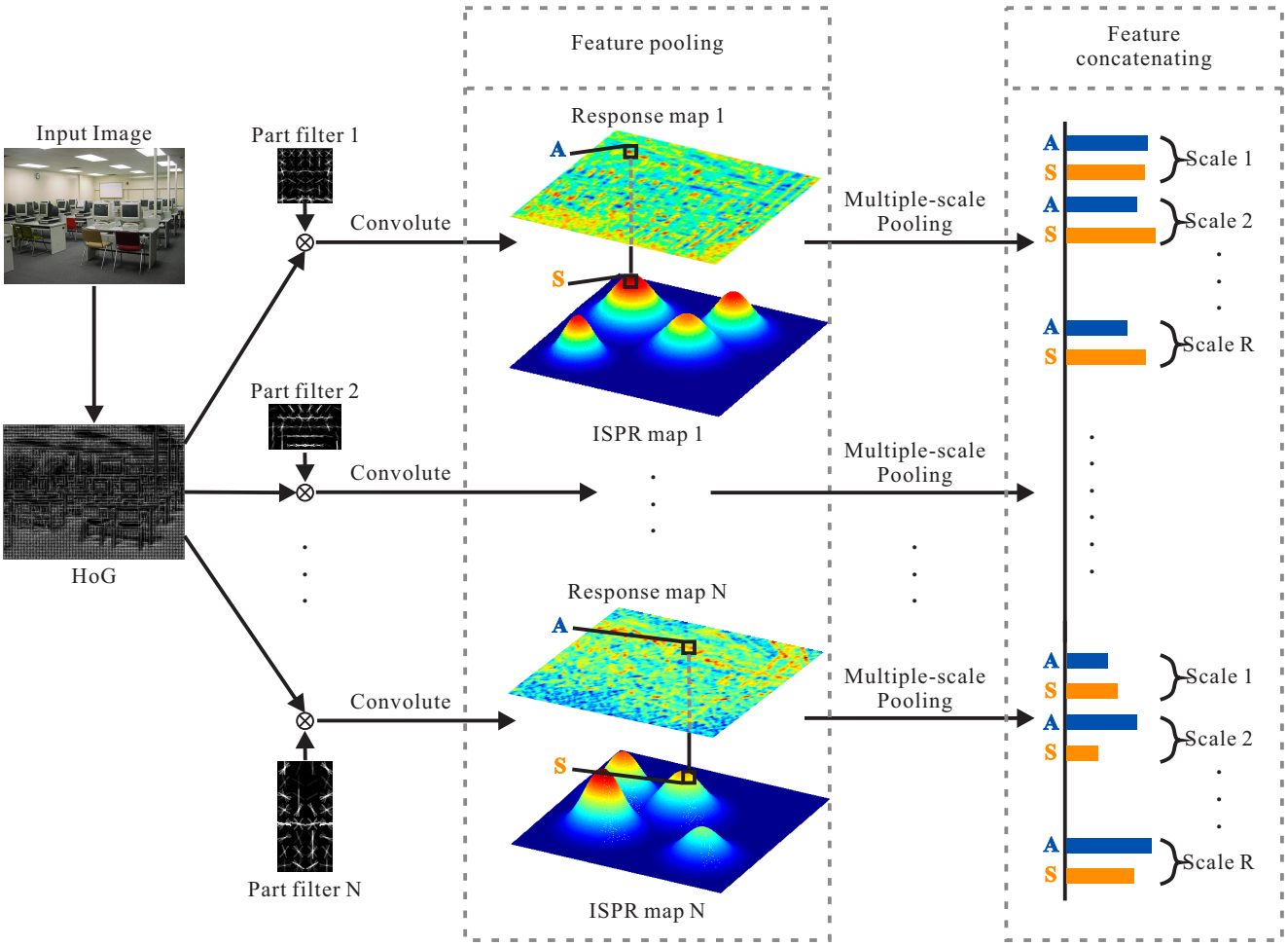


Figure 3. Construction of mid-level representations for scene classification.

localize global scene appearance. The locations of ISPRs are fixed with respect to the root filter. In addition, we note that scene can be recognized by several discriminative parts. With the above settings, we further adapt Eq. (4) to the multi-part version as

$$f(I,P) = \sum_{i=0}^{n} F_i \cdot H(I,p_i) + \sum_{i=1}^{n}\sum_{j=1}^{k} d_{ij} \cdot \phi(p_i, p_{ij}^c, \sigma_{ij}), \tag{5}$$

where

$$\phi(p_i, p_{ij}^c, \sigma_{ij}) = \exp(-\frac{||p_i - p_{ij}^c - p_0||^2}{\sigma}),$$

$P = [p_0, ..., p_n]$, and p_0 and F_0 are the 2D coordinates and vector of the root filter. The locations of ISPRs, which are denoted as p_{ij}^c in Eq. (5), take p_0 as origin. n and k are the numbers of parts and the center corresponding to the i^{th} part. d_{ij} and σ_{ij} are weight and coverage modeling the j^{th} cluster corresponding to the i^{th} part. p_{ij}^c is the central location of the j^{th} cluster corresponding to the i^{th} part. The component $\phi(p_i, p_{ij}^c, \sigma_{ij})$ defined in Eq. (2) is modified to the one in Eq. (5) by involving the reference location p_0.

3.4. Construct Mid-level Representation

In this paper, we use HoG [3] as an operator. Figure 3 visualizes two main processes of applying our model to mid-level representation construction.

The first step is feature pooling. Note the i^{th} part filter in Figure 3 is denoted as F_i in Eq. (5). The locations of root filter p_0 and part filter p_i are computed by maximizing Eq. (5) (described in Section 4) over P. "Feature pooling" extracts values A and S based on their definition: $A = F_i \cdot H(I, p_i)$ and $S = \sum_{j=1}^{k} d_{ij} \cdot \phi(p_i, p_{ij}^c, \sigma_{ij})$. This step is equivalently visualized as extracting A and S from p_i in the response and ISPR maps respectively. To deal with parts in diverse scales, we apply the multiple-scale scheme by resizing the image into different scales and repeating the pooling on the resized images. For each part filter, there are several pairs of A and S extracted in this manner.

Algorithm 1 Inference and Learning

Input: positive sample images $\{I_1^+, ..., I_u^+\}$; negative sample images $\{I_1^-, ..., I_v^-\}$; threshold δ; maximum buffer size $MSIZE$.
Initialize β and ω; $D^- = \emptyset$.
repeat
  $D^+ = \emptyset$.
  for t $= 1$ to u **do**
    $P = \arg\max_P f(I_t^+, P)$.
    $D^+ = D^+ \cup (1, I_t^+, P)$.
  repeat
    for t $= 1$ to v **do**
      while $\exists(-1, I_t^-, P) \notin D^-$, $f(I_t^-, P) \geq \delta$ **do**
        $D^- = D^- \cup (-1, I_t^-, P)$.
        if $|D^-| > MSIZE$ **then**
          break
    $D = D^+ \cup D^-$.
    $\beta = \arg\min_\beta L_D(\beta)$.
    for t $= 1$ to v **do**
      while $\exists(-1, I_t^-, P) \in D^-$, $f(I_t^-, P) < \delta$ **do**
        $D^- = D^- \setminus (-1, I_t^-, P)$.
  until β converges
  $\omega = \arg\max_\omega \sum_{h=1}^m \sum_{i=1}^n \sum_{j=1}^k d_j \cdot \phi(p_{hi}, p_{ij}^c, \sigma_{ij})$.
until ω and β converge
Output: β and ω.

The second step is feature concatenation. The first step is repeated by enumerating all the part filters. As for each part filter, there are A and S, we concatenate all these sets of values into a feature vector, which is the mid-level representation of the image as visualized in Figure 3.

The scene image representation is finally fed into SVM. Assume the number of part filters is N. The step of feature pooling is performed on R scales of one image. Then the dimensionality of the image representation is $2NR$.

4. Inference and Learning

Parameters in Eq. (5) include positions of part and root filters P, root and part filters F_i, weights d_{ij}, positions of centers p_{ij}^c, and σ_{ij}. Two parameters, i.e., β and ω, group the above parameters except P into two sets. β and ω are the model parameters to be learned. In this section, we describe inference and learning of these unknown parameters. As shown in Algorithm 3.3, inference of P and learning of β are performed in an iterative manner. With fixed P and β, we estimate ω.

Inference of P P groups the locations of root and part filters in Eq. (5). To infer it, we follow Relabel Positive and Mine Hard Negative Samples described in [5]. Obtaining P^+ of each positive sample image I^+ is equivalent to optimizing

$$P^+ = \arg\max_P f(I^+, P) \tag{6}$$

over P using dynamic programming. On the other hand, to obtain parameters P^- of each negative sample image I^-, we set a threshold δ and extract any P^- if $f(I^-, P^-)$ exceeds δ.

Learning β β groups parameters including filters F_i and weights d_{ij}. Eq. (5) can be expressed as a dot product:

$$F_\beta(I, P) = \beta \cdot \psi(I, P), \tag{7}$$

where β and $\psi(I, P)$ are vectors defined as

$$\beta = [F_0, \cdots, F_n, d_{11}, \cdots, d_{1k}, \cdots, d_{n1}, \cdots, d_{nk}], \tag{8}$$

$$\begin{aligned}\psi(I, P) = [&H(I, p_0), \cdots, H(I, p_n), \\ &\phi(p_1, p_{11}^c, \sigma_{11}), \cdots, \phi(p_1, p_{1k}^c, \sigma_{1k}), \cdots, \\ &\phi(p_n, p_{n1}^c, \sigma_{n1}), \cdots, \phi(p_n, p_{nk}^c, \sigma_{nk})].\end{aligned} \tag{9}$$

Since parameters P and ω are fixed in this step, vector $\psi(I, P)$ does not change. Learning parameters β becomes minimizing the objective function

$$\begin{aligned}L_D(\beta) = \frac{1}{2}||\beta||^2 + C\sum_{h=1}^m \max(0, 1 - y_h \cdot F_\beta(I_h, P_h)), \\ D = \{(y_1, I_1, P_1), ..., (y_m, I_m, P_m)\}\end{aligned} \tag{10}$$

over β using gradient descent. In Eq. (10), m is the number of samples, y_h is the binary label of image I_h, and constant C controls the relative weight in regularization.

Learning ω ω groups parameters, spatial coordinates of centers p_{ij}^c, and σ_{ij} as

$$\omega = \{\omega_{ij} | i = 1, 2, ..., n, j = 1, 2, ..., k\}, \tag{11}$$

where $\omega_{ij} = [p_{ij}^c, \sigma_{ij}]$. Eq. (5) can be treated as function $G_\omega(I, P)$ with respect to ω. The solution of ω is obtained by maximizing the objective function in Eq. (5), which reduces to

$$L_E(\omega) = \sum_{h=1}^m G_\omega(I_h, P_h) \tag{12}$$

over ω, where $E = \{(I_1, P_1), \cdots, (I_m, P_m)\}$. Note that ω reflects ISPRs in scene categories of positive samples. We denote by m the number of positive samples, I_h the h^{th} positive sample image, and by P_h the position parameters of the h^{th} image. Not related to ω, all appearance terms

described in Eq. (12) are safely ignored. With these algebraic operations, we update optimization to

$$\omega^* = \arg\max_{\omega} \sum_{h=1}^{m} \sum_{i=1}^{n} \sum_{j=1}^{k} d_j \cdot \phi(p_{hi}, p_{ij}^c, \sigma_{ij}). \quad (13)$$

Since each pair of ω_{ij} is independent, the solution boils down to estimation of parameters for each ω_{ij} by coordinate ascent, which alternatively optimizes p_{ij}^c and σ_{ij}.

Alternating Algorithm The training framework is sketched in Algorithm 3.3. Lines 4-7 implement relabel positive samples. Lines 9-13 implement mine hard negative samples. We exploit a fixed buffer D^- to store negative samples. Lines 14-15 are to compute parameter β. After updating β, lines 16-18 remove negative samples from buffer D^- if their scores are smaller than the threshold δ. Line 20 describes estimation of parameter ω.

5. Experiments

We evaluate our method on three public datasets, i.e., MIT-Indoor [23], 15-Scene [12] and UIUC 8-Sport dataset [14]. Average classification accuracy and per-category accuracies are reported.

We set the number of part filters to 8 for each scene category. The number of ISPRs is set to 4 for each part filter. Though flexibly setting the numbers of parts and ISPRs can further enhance the classification accuracy, these numbers are enough to achieve satisfactory performance in our experiments. In addition, ISPRs are applied to 15 scales for each image in different resolutions.

5.1. MIT-Indoor Dataset

MIT-Indoor dataset contains 15,620 indoor scene images in 67 scene categories. Each category in this dataset has about 80 training and 20 testing images. 67 part models (8 part filters in each part model) are trained for the construction of mid-level representation. It means that the total number of part filters is 67×8=536. Following the calculus of feature dimensionality described in Section 3.4, we construct 2×536×15=16,080 dimension mid-level representation for each image.

We compare single-feature approaches. Table 1 lists the performance of previous single-feature approaches [23, 4, 8, 15, 27, 17, 20, 21, 25, 37, 38, 33] and ours, together with the accuracies. Our method yields the accuracy of 50.10%, which outperforms other approaches except "mode seeking" [4].

Though the accuracy achieved by "mode seeking" [4] is higher than ours, our mid-level representation proves to be an excellent complementarity to existing image representation, such as IFV [22]. We use the implementation of IFV in toolkit [30]. By combining IFV and our mid-level representation, we compare this scheme with other methods [20, 4, 37, 27, 17, 8], which also apply multiple features to scene classification. Table 2 shows the difference – our solution achieves 68.50% accuracy.

Method	Accuracy(%)
ROI [23]	26.05
MM-scene [38]	28.00
DPM [20]	30.40
CENTRIST [33]	36.90
Object Bank [15]	37.60
RBoW [21]	37.93
Patches [27]	38.10
Hybrid-Parts [37]	39.80
LPR [25]	44.84
BoP [8]	46.10
VC [17]	46.40
VQ [17]	47.60
Mode Seeking [4]	**64.03**
ISPR (our approach)	50.10

Table 1. Average classification accuracy of single-feature approaches on the MIT-indoor dataset.

Method	Accuracy(%)
DPM + GIST-color + SP [20]	43.10
Hybrid-Parts + GIST-color + SP [37]	47.20
Patches + GIST + SP + DPM [27]	49.40
VC + VQ [17]	52.30
BoP + IFV [8]	63.10
Mode Seeking + IFV [4]	66.87
ISPR(our approach) + IFV	**68.50**

Table 2. Average classification accuracy of state-of-the-art approaches fusing multiple features and our mid-level representation combining IFV on MIT-indoor dataset.

5.2. 15-Scene Dataset

15-Scene dataset [12] contains 4485 images of 15 scene categories. Since these images describe different indoor and outdoor scenes, the variety is large enough to evaluate the generality of our approach. Following the training/testing split as [12], 10 random splits of data are taken. In each split, 100 random images per category are used for training and the rest are for testing. Totally there are 15×8=120 part filters in this experiment. The dimensionality of the mid-level representation is 2×120×15=3,600.

We compare our approach with state-of-the-arts in Table 3. Our framework reaches classification accuracy of 85.08%. The combination of our mid-level representation and IFV achieves accuracy as high as 91.06%. In addition, Figure 4 shows per-category accuracies in the confusion matrix.

Method	Accuracy(%)
GIST-color [19]	69.50
RBoW [21]	78.60 ± 0.70
Classemes [29]	80.60
Object Bank [15]	80.90
SP [12]	81.40
SPMSM [11]	82.30
LCSR [26]	82.67 ± 0.51
SP-pLSA [1]	83.70
CENTRIST [33]	83.88 ± 0.76
HIK [32]	84.12 ± 0.52
VC + VQ [17]	85.40
LMLF [2]	85.60 ± 0.20
LPR [25]	85.81
Hybrid-Parts + GIST-color + SP [37]	86.30
CENTRIST + LLC + Boosting [36]	87.80
RSP [7]	88.10
LScSPM [6]	89.75 ± 0.50
ISPR(our approach)	85.08 ± 0.01
IFV [30]	89.20 ± 0.09
ISPR(our approach) + IFV	**91.06** $\pm$ **0.05**

Table 3. Average classification accuracies on 15-Scene dataset.

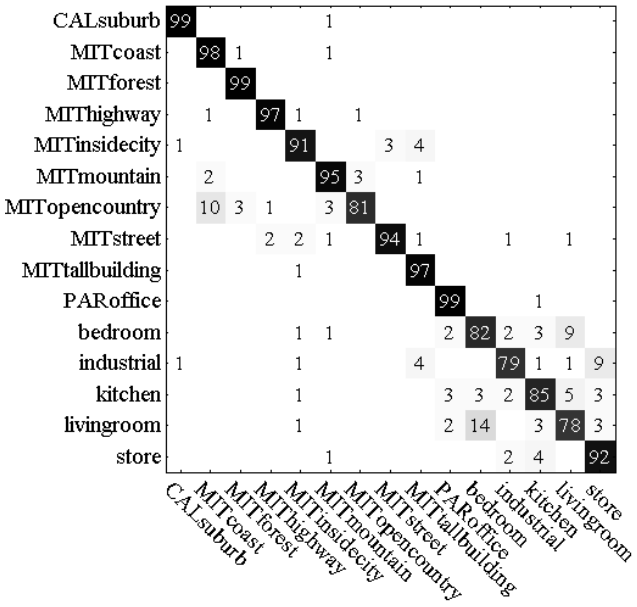


Figure 4. Confusion matrix (in %) of our mid-level representation combining IFV on 15-Scene dataset. Only the rounded rates not lower than 1 % are shown.

5.3. UIUC 8-Sport dataset

UIUC Sport dataset [14] contains 8 sport categories. Following the training/testing split in [14], we take 10 random splits of data. For each split, we select 70 training images and 60 testing images in each category. There are $8 \times 8 = 64$ part filters learned in our system. The dimensionality of the mid-level representation is $2 \times 64 \times 15 = 1{,}920$.

Method	Accuracy(%)
GIST-color [19]	70.70
MM-Scene [38]	71.70
Graphical Model [14]	73.40
Object Bank [15]	76.30
Object Attributes [16]	77.88
CENTRIST [33]	78.25 ± 1.27
RSP [7]	79.60
SP [12]	81.80
SPMSM [11]	83.00
Classemes [29]	84.20
HIK [32]	84.21 ± 0.99
LScSPM [6]	85.30
LPR [25]	86.25
Hybrid-Parts + GIST-color + SP [37]	87.20
LCSR [26]	87.23 ± 1.14
VC + VQ [17]	88.40
ISPR(our approach)	89.50 ± 0.59
IFV [30]	90.80 ± 0.12
ISPR(our approach) + IFV	**92.08** $\pm$ **0.23**

Table 4. Average classification accuracies on UIUC 8-Sport dataset.

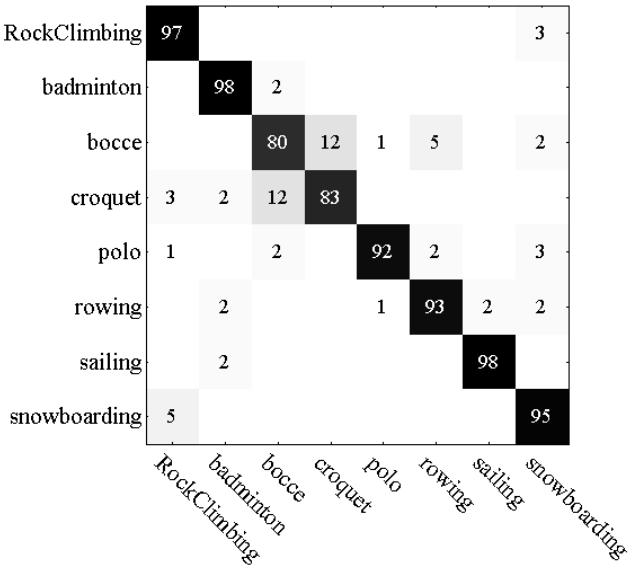


Figure 5. Confusion matrix (in %) of our mid-level representation combining IFV on UIUC 8-Sport dataset. Only the rounded rates not lower than 1 % are shown.

Table 4 shows we get the competitive accuracy 89.50%. When combining our mid-level representation with IFV, it further goes to 92.08%. Figure 5 shows per-category performance in the form of confusion matrix.

6. Conclusion

We have presented a useful model that utilizes part appearance and spatial configuration for improving scene classification. ISPR proposed in this framework encourages

spatial pooling to be performed more adaptively to resist false response. Spatial information extracted from ISPR also enhances the discriminative power of mid-level representation in classification. We have evaluated our method on several representative datasets. Jointly using low-level features and our new model results in high classification accuracy.

Acknowledgements

This work is supported by a grant from the Research Grants Council of the Hong Kong SAR (project No. 413113) and by NSF of China (key project No. 61133009).

References

[1] A. Bosch, A. Zisserman, and X. Muoz. Scene classification using a hybrid generative/discriminative approach. *PAMI*, 2008.

[2] Y.-L. Boureau, F. Bach, Y. LeCun, and J. Ponce. Learning mid-level features for recognition. In *CVPR*, 2010.

[3] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[4] C. Doersch, A. Gupta, and A. A. Efros. Mid-level visual element discovery as discriminative mode seeking. In *NIPS*, 2013.

[5] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *PAMI*, 2010.

[6] S. Gao, I. W. Tsang, L.-T. Chia, and P. Zhao. Local features are not lonely–laplacian sparse coding for image classification. In *CVPR*, 2010.

[7] Y. Jiang, J. Yuan, and G. Yu. Randomized spatial partition for scene recognition. In *ECCV*, 2012.

[8] M. Juneja, A. Vedaldi, C. Jawahar, and A. Zisserman. Blocks that shout: Distinctive parts for scene classification. In *CVPR*, 2013.

[9] H. Kang, M. Hebert, and T. Kanade. Discovering object instances from scenes of daily living. In *ICCV*, 2011.

[10] G. Kim and A. Torralba. Unsupervised detection of regions of interest using iterative link analysis. In *NIPS*, 2009.

[11] R. Kwitt, N. Vasconcelos, and N. Rasiwasia. Scene recognition on the semantic manifold. In *ECCV*, 2012.

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

[13] Y. J. Lee and K. Grauman. Object-graphs for context-aware category discovery. In *CVPR*, 2010.

[14] L.-J. Li and L. Fei-Fei. What, where and who? classifying events by scene and object recognition. In *ICCV*, 2007.

[15] L.-J. Li, H. Su, L. Fei-Fei, and E. P. Xing. Object bank: A high-level image representation for scene classification & semantic feature sparsification. In *NIPS*, 2010.

[16] L.-J. Li, H. Su, Y. Lim, and L. Fei-Fei. Objects as attributes for scene classification. In *ECCV*, 2012.

[17] Q. Li, J. Wu, and Z. Tu. Harvesting mid-level visual concepts from large-scale internet images. In *CVPR*, 2013.

[18] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2004.

[19] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, 2001.

[20] M. Pandey and S. Lazebnik. Scene recognition and weakly supervised object localization with deformable part-based models. In *ICCV*, 2011.

[21] S. N. Parizi, J. G. Oberlin, and P. F. Felzenszwalb. Reconfigurable models for scene recognition. In *CVPR*, 2012.

[22] F. Perronnin, Y. Liu, J. Sánchez, and H. Poirier. Large-scale image retrieval with compressed fisher vectors. In *CVPR*, 2010.

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

[24] B. C. Russell, W. T. Freeman, A. A. Efros, J. Sivic, and A. Zisserman. Using multiple segmentations to discover objects and their extent in image collections. In *CVPR*, 2006.

[25] F. Sadeghi and M. F. Tappen. Latent pyramidal regions for recognizing scenes. In *ECCV*, 2012.

[26] A. Shabou and H. LeBorgne. Locality-constrained and spatially regularized coding for scene categorization. In *CVPR*, 2012.

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

[28] S. Todorovic and N. Ahuja. Unsupervised category modeling, recognition, and segmentation in images. *PAMI*, 2008.

[29] L. Torresani, M. Szummer, and A. Fitzgibbon. Efficient object category recognition using classemes. In *ECCV*, 2010.

[30] A. Vedaldi and B. Fulkerson. VLFeat: An open and portable library of computer vision algorithms. `http://www.vlfeat.org/`, 2008.

[31] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. Locality-constrained linear coding for image classification. In *CVPR*, 2010.

[32] J. Wu and J. M. Rehg. Beyond the euclidean distance: Creating effective visual codebooks using the histogram intersection kernel. In *ICCV*, 2009.

[33] J. Wu and J. M. Rehg. Centrist: A visual descriptor for scene categorization. *PAMI*, 2011.

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

[35] B. Yao, G. Bradski, and L. Fei-Fei. A codebook-free and annotation-free approach for fine-grained image categorization. In *CVPR*, 2012.

[36] J. Yuan, M. Yang, and Y. Wu. Mining discriminative co-occurrence patterns for visual recognition. In *CVPR*, 2011.

[37] Y. Zheng, Y.-G. Jiang, and X. Xue. Learning hybrid part filters for scene recognition. In *ECCV*, 2012.

[38] J. Zhu, L.-J. Li, L. Fei-Fei, and E. P. Xing. Large margin learning of upstream scene understanding models. In *NIPS*, 2010.