# Discriminative Spatial Pyramid

Tatsuya Harada[1,2], Yoshitaka Ushiku[1], Yuya Yamashita[1], and Yasuo Kuniyoshi[1]
[1]The University of Tokyo, [2]JST PRESTO
harada@isi.imi.i.u-tokyo.ac.jp

## Abstract

*Spatial Pyramid Representation (SPR) is a widely used method for embedding both global and local spatial information into a feature, and it shows good performance in terms of generic image recognition. In SPR, the image is divided into a sequence of increasingly finer grids on each pyramid level. Features are extracted from all of the grid cells and are concatenated to form one huge feature vector. As a result, expensive computational costs are required for both learning and testing. Moreover, because the strategy for partitioning the image at each pyramid level is designed by hand, there is weak theoretical evidence of the appropriate partitioning strategy for good categorization. In this paper, we propose discriminative SPR, which is a new representation that forms the image feature as a weighted sum of semi-local features over all pyramid levels. The weights are automatically selected to maximize a discriminative power. The resulting feature is compact and preserves high discriminative power, even in low dimension. Furthermore, the discriminative SPR can suggest the distinctive cells and the pyramid levels simultaneously by observing the optimal weights generated from the fine grid cells.*

## 1. Introduction

In recent years, generic image recognition has attracted many researchers and has drastically developed. One inherent challenge in image recognition is the handling of spatial information. The meanings of images are embedded in the spatial distribution of pixels; therefore, the means of representing spatial information in a feature is an important topic in understanding images.

Spatial information is usually embedded in the feature extraction process. The feature can be classified into a local feature (e.g., SIFT [14]) and a global feature (e.g., GIST [19]). The typical process to build one image feature from local features can be broken down into two steps [5]: 1) coding of local features and 2) spatial pooling of semi-local features. For each step, the embedding of spatial information has been well studied.

A common framework for the coding step is Bag of Features (BoF) [7]. However, BoF is built with a histogram of the vector-quantized local features and lacks the spatial distribution of local features in the image space. Therefore, many studies have attempted to embed the spatial orders of the local features into BoF (e.g., [6, 16]). Recently, sparse coding [20] has been reported to outperform BoF in this area [27, 5]. Sparse coding permits a linear combination of a small number of codewords, while in BoF, one local feature corresponds to only one codeword. Sparse coding also lacks the spatial orders of local features. To embed spatial orders into sparse codes, [17] considers a pair of spatially close features as a new local feature followed by sparse coding. BoF and sparse codes are the sparse representations of the distribution of the local descriptors in the feature space. On the other hand, the dense representation of the distribution has been studied. [18] proposed the Global Gaussian (GG) approach that estimates the distribution as one Gaussian distribution and builds the feature by arranging the elements of the mean and covariance of the Gaussian. Similarly, [11], which is a general form of the GG, proposed to embed local spatial information into a feature by calculating the local auto-correlations of any local features.

In spatial pooling, Spatial Pyramid Representation (SPR) [12] is popular for encoding the spatial distribution of local features. Spatial Pyramid Matching (SPM) with BoF has been remarkably successful in terms of both scene and object recognitions. As for sparse codes, the state-of-the-art variants of the spatial pyramid model with linear SVMs work surprisingly well [27, 5]. The variations of sparse codes [29, 28] also utilize SPR. Adopting some normalizations and SPM, [22] improved the Fisher vector [21], and obtained a discriminative feature.

As stated above, SPR contributes as a major component to the state-of-the-art methods. In SPR, the input image is divided into a sequence of increasingly finer grids on each pyramid level. Features are extracted from all of the grid cells, and are concatenated to form one huge image feature. As a result, expensive computational costs are required for both learning and testing. Moreover, SPR has two major parameters: the number of pyramid levels and the strategy for

partitioning the image. Figure 1 (a) shows the spatial pyramid structure used in [12, 27]. Figure 1 (b) was used in the winner of VOC 2007 [15], and recently, many state-of-the-art methods have employed this structure [29, 28, 22]. As shown, the spatial pyramid structures have been designed by hand, and therefore, it is not clear how to design an optimal spatial pyramid structure.


(a) Spatial pyramid representation in [12]


(b) Spatial pyramid representation in [15]

Figure 1. Variations on Spatial Pyramid Representation

In this paper, we propose the discriminative SPR that forms a feature as the weighted sum of the semi-local features over all of the pyramid levels. The weights are automatically selected to maximize a discriminative power. To find the discriminative weights, we also propose the partial least squares SPR (PlsSPR) and the Fisher SPR (FisherSPR). The resulting feature is compact, and preserves high discriminative performance even in low dimension. In addition, the discriminative SPR can suggest distinctive cells and pyramid levels simultaneously by observing the weights generated from the fine grid cells.

Our contributions are summarized as follows: 1) proposal of the discriminative and compact SPRs, and 2) proposal of the PlsSPR and the FisherSPR as the implementations of 1). To the best of our knowledge, there are few methods that optimize SPR by estimating the discriminative weights of cells and pyramid levels simultaneously, and hence, the designs of SPR have been created by hand. Our method can estimate discriminative weights of each cell on each level. By using the weights, we can directly compare discriminative power between cells on the different levels. We believe that our new representation can contribute to the automatic, effective, and proper design of SPR.

## 2. Related Work

SPR [12] has been widely used to encode the spatial information of local features. The original SPR gives fixed, different weights for each pyramid level. [4] proposed a method for learning the spatial pyramid weights. This method finds optimal weights by using cross-validation, and thus, it is computationally inefficient. While the weights for a pyramid level can be learned, the optimization of the weight of each grid cell is not considered. Similarly, [3] used a random forest and ferns for image categorization. A 1 or 0 weight is randomly selected for each pyramid level at the nodes. Therefore, it is difficult to provide the explicit importance for each pyramid level. In addition, it does not consider the weights of the grid cells.

Eigenfaces [25] and Fisherfaces [2] are well-known methods for weighting the regions in an image. Although these methods were originally designed for face recognition, they are considered to be frameworks for finding the distinctive pixels in an image. However, they can be applied only to images consisting of scalar values (e.g., brightness) at the pixel, and thus, cannot be applied to an image where each pixel is described by the vector. To overcome this limitation, Fisher Weight Maps (FWM) was proposed in [24], which maximizes the Fisher criterion of the feature vectors on each pixel. In general, images have different scales and aspect ratios, and the pixel-wise weight maps are not directly utilized in generic images. In order to absorb the varieties of scales and aspect ratios, [11] proposed to divide the image into regular grid cells to calculate the semi-local features in each cell and to apply FWM to those features.

Finding discriminative regions are of special interest in human detection. [8] indicated that SVM provides a measure of importance to each cell in the final discrimination decision, and the obtained weights represent major human contours. [8] used only the HOG, thereby ignoring some other useful features (e.g., skin color and textures of clothing) for human detection. [23] obtained good results by using a richer set of features and partial least squares (PLS), and showed that PLS weight vectors are good indicators of the contribution of each feature for human detection.

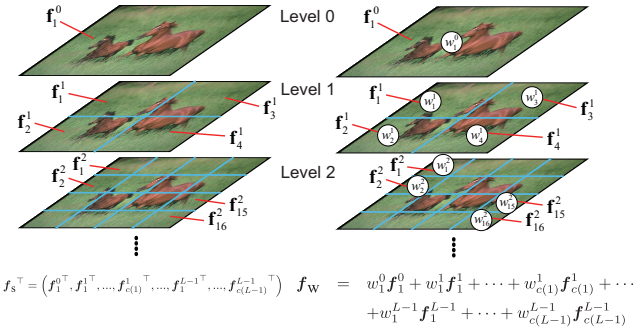
## 3. Weighted Spatial Pyramid

In this section, we first explain SPR, and then propose a new weighted SPR. First, SPR extracts the global feature from the input image (the top level $l = 0$). Next, the image is divided into a sequence of increasingly finer grids on each pyramid level. Features are extracted from each grid cell on each pyramid level $l$, and are concatenated to form one large feature. Let $\boldsymbol{f}_k^l \in \mathbb{R}^d$ denote the feature extracted from the cell $k$ on level $l$, and let $c(l)$ and $L$ denote the number of cells on the level $l$ and the number of levels, respectively. The feature vector $\boldsymbol{f}_\mathrm{S} \in \mathbb{R}^{d_s}$, $d_s = d\sum_{l=0}^{L-1} c(l)$ with SPR is defined by the following equation:

$$\boldsymbol{f}_\mathrm{S}{}^\top = \left(\boldsymbol{f}_1^{0\top}, \boldsymbol{f}_1^{1\top}, ..., \boldsymbol{f}_{c(1)}^{1\top}, ..., \boldsymbol{f}_1^{L-1\top}, ..., \boldsymbol{f}_{c(L-1)}^{L-1\top}\right).$$

The original SPR [12] defined the number of cells at level $l$ as $c(l) = 4^l$, and divided each cell at this level into four cells at the next level $l + 1$. As the number of levels increases, the feature $\boldsymbol{f}_s$ becomes a huge vector. For this reason, SPR

is usually used up to level 2 ($L = 3$). The original SPR is illustrated in Fig. 2 (a).




(a) Spatial Pyramid Representation (b) Weighted Spatial Pyramid Representation

Figure 2. (a) Original Spatial Pyramid Representation. (b) Proposed Weighted Spatial Pyramid Representation

As stated above, because the feature vector with SPR becomes huge as the number of levels increases, an SPR vector with large pyramid levels is inefficient for the following learning and testing phases. However, if we stop feature extraction at a lower pyramid level, the SPR feature fails to capture fine spatial information. Here, we propose the weighted SPR, which is a compact SPR that contains coarse-to-fine spatial information effectively. The standard SPR usually uses at most $4 \times 4$ cells because of a huge vector. Our method theoretically can use more finer cells (e.g., $16 \times 16$, $32 \times 32$), while the resulting vector has the constant dimensionality. In a similar way as SPR, the weighted SPR first extracts the global feature from the input image (the top level $l = 0$). Next, the image is divided into a sequence of increasingly finer grids on each pyramid level. Features are extracted from each grid cell on each pyramid level $l$. We consider that cell $k$ on level $l$ has a weight $w_k^l \in \mathbb{R}$ representing the importance of the cell. We define the weighted SPR feature as the weighted sum of features from each cell on each level. The weighted SPR feature $\boldsymbol{f}_{\mathrm{W}} \in \mathbb{R}^d$ can be written by the following equation:

$$\begin{aligned} \boldsymbol{f}_{\mathrm{W}} &= w_1^0 \boldsymbol{f}_1^0 + w_1^1 \boldsymbol{f}_1^1 + \cdots + w_{c(1)}^1 \boldsymbol{f}_{c(1)}^1 + \cdots \\ &\quad + w_1^{L-1} \boldsymbol{f}_1^{L-1} + \cdots + w_{c(L-1)}^{L-1} \boldsymbol{f}_{c(L-1)}^{L-1}. \end{aligned} \quad (1)$$

Now, we define a weight vector $\boldsymbol{w} \in \mathbb{R}^{d_w}, d_w = \sum_{l=0}^{L-1} c(l)$ and a feature matrix $\boldsymbol{F} \in \mathbb{R}^{d \times d_w}$ as the following equations:

$$\begin{aligned} \boldsymbol{w} &= \left(w_1^0, w_1^1, ..., w_{c(1)}^1, ..., w_1^{L-1}, ..., w_{c(L-1)}^{L-1}\right)^\top, \\ \boldsymbol{F} &= \left(\boldsymbol{f}_1^0, \boldsymbol{f}_1^1, ..., \boldsymbol{f}_{c(1)}^1, ..., \boldsymbol{f}_1^{L-1}, ..., \boldsymbol{f}_{c(L-1)}^{L-1}\right). \end{aligned}$$

Using the above $\boldsymbol{w}$ and $\boldsymbol{F}$, Eq. (1) can be rewritten as the following simple equation:

$$\boldsymbol{f}_{\mathrm{W}} = \boldsymbol{F}\boldsymbol{w}. \quad (2)$$

Suppose that we have a set of weight vectors $\mathcal{W} = \{\boldsymbol{w}_i \in \mathbb{R}^{d_w}\}_{i=1}^{N_w}$ obtained under various conditions, the multiple weighted SPR vectors can be calculated by multiplying the feature matrix by each weight vector. In fact, the set of weight vectors corresponds to a set of eigen vectors in the discriminative SPR. This topic is discussed in Sec 4.2 and Sec 4.3. Then, we redefine the weighted SPR $\boldsymbol{f}_{\mathrm{W}}^{(N_w)} \in \mathbb{R}^{N_w d}$ by concatenating those multiple vectors as follows:

$$\boldsymbol{f}_{\mathrm{W}}^{(N_w)} = \left((\boldsymbol{F}\boldsymbol{w}_1)^\top, \cdots, (\boldsymbol{F}\boldsymbol{w}_{N_w})^\top\right)^\top. \quad (3)$$

We expect that the weighted SPR becomes compact and discriminative if a small number of weights is effectively selected for image categorization.

Note that the weighted SPR is equal to the single flat-level SPR [11, 24] with the finest cells when using average spatial pooling. For simplicity, we now consider an SPR with levels 0 and 1. We assume that a set of local descriptors $\mathcal{U} = \{\boldsymbol{u}_m \in \mathbb{R}^d\}_{m=1}^M$ is obtained at level 0. At level 1, the image is divided into $c(1)$ cells.

$$\mathcal{U} = \mathcal{U}_1 \cup \mathcal{U}_2 \cup \cdots \cup \mathcal{U}_{c(1)},\ \mathcal{U}_i \cap \mathcal{U}_j = \emptyset\ (i \neq j).$$

The feature at level 0 can be represented using average spatial pooling as follows:

$$\boldsymbol{f}_1^0 = \frac{1}{M} \sum_{\boldsymbol{u}_i \in \mathcal{U}} \boldsymbol{u}_i \quad (4)$$

A feature at cell $k$ on level 1 can be represented in the same manner: $\boldsymbol{f}_k^1 = \frac{1}{M_k} \sum_{\boldsymbol{u}_i \in \mathcal{U}_k} \boldsymbol{u}_i$. When we plug this equation into Eq. (4), we have:

$$\begin{aligned} \boldsymbol{f}_1^0 &= \frac{1}{M} \left( \sum_{k=1}^{c(1)} \sum_{\boldsymbol{u}_i \in \mathcal{U}_k} \boldsymbol{u}_i \right) \\ &= \sum_{k=1}^{c(1)} \frac{M_k}{M} \left( \frac{1}{M_k} \sum_{\boldsymbol{u}_i \in \mathcal{U}_k} \boldsymbol{u}_i \right) = \sum_{k=1}^{c(1)} \alpha_k \boldsymbol{f}_k^1, \end{aligned} \quad (5)$$

where $\alpha_k = \frac{M_k}{M}$. We let $L = 2$ in Eq. (1), and plug Eq. (5) into Eq. (1), so we obtain the following equation:

$$\begin{aligned} \boldsymbol{f}_{\mathrm{W}} &= w_1^0 \boldsymbol{f}_1^0 + \sum_{i=1}^{c(1)} w_i^1 \boldsymbol{f}_i^1 \\ &= \sum_{i=1}^{c(1)} (w_1^0 \alpha_i + w_i^1) \boldsymbol{f}_i^1 = \sum_{i=1}^{c(1)} \beta_i^1 \boldsymbol{f}_i^1, \end{aligned} \quad (6)$$

where $\beta_i^1 = w_1^0 \alpha_i + w_i^1$. Eq. (6) is equal to the weight maps [11, 24], ignoring the pyramid structure. Average spatial pooling is popular, and is used in BoF. If we apply the weighted SPR to average spatial pooling, the pyramid structure becomes of unknown significance. The weighted SPR becomes of great significance when it is applied to nonlinear spatial pooling (e.g., max spatial pooling [27]).

## 4. Discriminative Spatial Pyramid

In this section, we propose a novel representation called the discriminative SPR that automatically finds the discriminative weights of the weighted SPR. As implementations of the discriminative SPR, we also propose the PlsSPR and the FisherSPR. Here, we first explain the relationship between PLS and canonical correlation analysis (CCA) as the basis of the proposed representations. Then, we describe the details of the discriminative SPRs.

### 4.1. PLS and CCA

PLS is a method to extract common information between sets of observed variables. The origins of PLS are traced to nonlinear iterative partial least squares (NIPALS) [26]. The connection among PLS, CCA, and Fisher Linear Discriminant Analysis (Fisher LDA) is proved in [1]. According to [1], we here explain how the modified PLS reduces to the eigenvalue problem of the between-class covariance matrix.

Now, we assume a set of data $\mathcal{D} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^N$, $\mathcal{X} = \{\boldsymbol{x}_i \in \mathbb{R}^{d_x}\}_{i=1}^N$, $\mathcal{Y} = \{\boldsymbol{y}_i \in \mathbb{R}^{d_y}\}_{i=1}^N$, and transform $\mathcal{X}$ and $\mathcal{Y}$ into a new coordinate system: $s_i = \boldsymbol{a}^\top(\boldsymbol{x}_i - \bar{\boldsymbol{x}})$, $t_i = \boldsymbol{b}^\top(\boldsymbol{y}_i - \bar{\boldsymbol{y}})$, where $\bar{\boldsymbol{x}} = \frac{1}{N}\sum_i \boldsymbol{x}_i, \bar{\boldsymbol{y}} = \frac{1}{N}\sum_i \boldsymbol{y}_i$ are the mean of $\boldsymbol{x}$, $\boldsymbol{y}$, respectively. We let $C_{xy} = \frac{1}{N}\sum_i (\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{y}_i - \bar{\boldsymbol{y}})^\top$ be the sample estimate of the cross-product covariance matrix between $\mathcal{X}$ and $\mathcal{Y}$. PLS searches for vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ with unit norm that maximize the sample covariance $\mathrm{cov}(s,t) = \frac{1}{N}\sum_i s_i t_i = \boldsymbol{a}^\top C_{xy}\boldsymbol{b}$.

$$\max \frac{[\mathrm{cov}(s,t)]^2}{(\boldsymbol{a}^\top\boldsymbol{a})(\boldsymbol{b}^\top\boldsymbol{b})} \tag{7}$$

Estimates of the weight vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ are given as the solution of the following eigenvalue problem:

$$\begin{pmatrix} 0 & C_{xy} \\ C_{yx} & 0 \end{pmatrix} \begin{pmatrix} \boldsymbol{a} \\ \boldsymbol{b} \end{pmatrix} = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \begin{pmatrix} \boldsymbol{a} \\ \boldsymbol{b} \end{pmatrix},$$

where $\lambda_1$, $\lambda_2$ are eigenvalues, and $C_{yx} = C_{xy}^\top$.

The correlation $\mathrm{corr}(s,t)$ between $s,t$ can be represented using the covariance $\mathrm{cov}(s,t)$ and the variances $\mathrm{var}(s) = \frac{1}{N}\sum_i s_i^2$, $\mathrm{var}(t) = \frac{1}{N}\sum_i t_i^2$ as follows:

$$[\mathrm{cov}(s,t)]^2 = \mathrm{var}(s)[\mathrm{corr}(s,t)]^2\mathrm{var}(t). \tag{8}$$

CCA maximizes $\mathrm{corr}(s,t)$, subject to $\mathrm{var}(s) = 1$, $\mathrm{var}(t) = 1$, and thus, PLS can be seen as a form of penalized CCA with $\mathrm{var}(s)$ and $\mathrm{var}(t)$. Suppose that our goal is discrimination and that $\mathcal{Y}$ space represents a category where the category information is coded in $\boldsymbol{y}$ with a 1-of-$K$ coding scheme, the $\mathcal{Y}$ space penalty is not meaningful. Therefore, we consider a new objective function by removing the $\mathcal{Y}$ space penalty from the original objective function:

$$\mathrm{var}(s)[\mathrm{corr}(s,t)]^2 = \frac{[\mathrm{cov}(s,t)]^2}{\mathrm{var}(t)}. \tag{9}$$

The modified PLS maximizes Eq. (9), subject to the norm that $\boldsymbol{a}$ is equal to 1 and that the variance $\mathrm{var}(t)$ is equal to 1.

$$\max \frac{[\mathrm{cov}(s,t)]^2}{\mathrm{var}(t)(\boldsymbol{a}^\top\boldsymbol{a})} \tag{10}$$

An estimate of the weight vector $\boldsymbol{a}$ is given as the solution of the following eigenvalue problem:

$$\boldsymbol{C}_{xy}\boldsymbol{C}_y^{-1}\boldsymbol{C}_{yx}\boldsymbol{a} = \lambda\boldsymbol{a}, \tag{11}$$

where $\lambda$ is an eigenvalue, and $C_y = \frac{1}{N}\sum_i(\boldsymbol{y}_i - \bar{\boldsymbol{y}})(\boldsymbol{y}_i - \bar{\boldsymbol{y}})^\top$. Here, [1] proved that $\boldsymbol{C}_{xy}\boldsymbol{C}_y^{-1}\boldsymbol{C}_{yx}$ is equal to the between-class covariance matrix $\boldsymbol{S}_b$. Thus, Eq. (11) can be rewritten as follows:

$$\boldsymbol{S}_b\boldsymbol{a} = \lambda\boldsymbol{a}. \tag{12}$$

The modified PLS reduces to the eigenvalue problem of the between-class covariance matrix.

### 4.2. PLS Spatial Pyramid

We propose an efficient calculation of discriminative weights of Eq. (2) based on Eq. (12). Now, we have the labeled training samples $\mathcal{I} = \{(\boldsymbol{F}_i, \boldsymbol{y}_i)\}_{i=1}^N$, $\boldsymbol{f}_{\mathrm{W}i} = \boldsymbol{F}_i\boldsymbol{w}$ with $K$ classes $\{\omega_k\}_{k=1}^K$. The between-class covariance matrix $\boldsymbol{S}_b$ can be written as follows:

$$\boldsymbol{S}_b = \frac{1}{N}\sum_{k=1}^K n_k(\bar{\boldsymbol{f}}_{\mathrm{W}k} - \bar{\boldsymbol{f}}_{\mathrm{W}})(\bar{\boldsymbol{f}}_{\mathrm{W}k} - \bar{\boldsymbol{f}}_{\mathrm{W}})^\top, \tag{13}$$

where $\bar{\boldsymbol{f}}_{\mathrm{W}k} = \frac{1}{n_k}\sum_{f_{\mathrm{W}i}\in\omega_k}\boldsymbol{f}_{\mathrm{W}i}$, $\bar{\boldsymbol{f}}_{\mathrm{W}} = \frac{1}{N}\sum_i \boldsymbol{f}_{\mathrm{W}i}$, and $n_k$ is the number of samples in class $\omega_k$. The trace of $\boldsymbol{S}_b$ is given by:

$$\mathrm{tr}\boldsymbol{S}_b = \boldsymbol{w}^\top\boldsymbol{\Sigma}_b\boldsymbol{w}, \tag{14}$$

where

$$\boldsymbol{\Sigma}_b = \frac{1}{N}\sum_{k=1}^K n_k(\boldsymbol{M}_k - \boldsymbol{M})^\top(\boldsymbol{M}_k - \boldsymbol{M}), \tag{15}$$

$\boldsymbol{M}_k = \frac{1}{n_k}\sum_{F_i\in\omega_k}\boldsymbol{F}_i$ is the mean of $\boldsymbol{F}_i$ belonging to the class $\omega_k$, and $\boldsymbol{M} = \frac{1}{N}\sum_i \boldsymbol{F}_i$ is the mean of the total data set. By maximizing Eq. (14) under the condition $\boldsymbol{w}^\top\boldsymbol{w} = 1$, we obtain the weight $\boldsymbol{w}$ as the eigen vector of the following eigenvalue problem:

$$\boldsymbol{\Sigma}_b\boldsymbol{w} = \lambda\boldsymbol{w}, \tag{16}$$

where $\lambda$ is the eigenvalue corresponding to the eigen vector $\boldsymbol{w}$. We select the $N_w$ largest eigen values $\lambda_1, \cdots, \lambda_{N_w}$, and the corresponding eigen vectors $\boldsymbol{w}_1, \cdots, \boldsymbol{w}_{N_w}$.

We call the obtained representation the PlsSPR. The PlsSPR is concerned with finding discriminative weights of

each cell on each pyramid level. It is not uncommon for the feature to be several tens of thousands of dimensions (e.g., 170,000 in [23]). Solving the eigenvalue problem with a raw feature is ineffective. On the other hand, our method in Eq. (16) is the eigenvalue problem of $\mathbf{\Sigma}_b \in \mathbb{R}^{d_w \times d_w}$. The total number of cells $d_w$ in the SPR is usually at most 100. Thus, because $d_w \ll d$, our method can find discriminative spatial pyramid weights very fast. Moreover, the computation of the PlsSPW is more stable than that of the Fisher Spatial Pyramid Weight, which is discussed in the next section. In addition, Eq. (16) that uses not the vector, but the matrix consisting of mid-local features, is considered to be the generalized PLS.

### 4.3. Fisher Spatial Pyramid

We propose the FisherSPR to obtain the discriminative SPR. Let $\boldsymbol{S}_w$ be the within-class covariance matrix of mid-local features. The Fisher criterion is given by $J(\boldsymbol{w}) = \frac{\mathrm{tr}\boldsymbol{S}_b}{\mathrm{tr}\boldsymbol{S}_w}$. The trace of $\boldsymbol{S}_w$ is given by:

$$\mathrm{tr}\boldsymbol{S}_w = \boldsymbol{w}^\top \mathbf{\Sigma}_w \boldsymbol{w}, \tag{17}$$

where

$$\mathbf{\Sigma}_w = \frac{1}{N}\sum_{k=1}^{K}\sum_{i\in\omega_k}(\boldsymbol{F}_i - \boldsymbol{M}_k)^\top(\boldsymbol{F}_i - \boldsymbol{M}_k). \tag{18}$$

By maximizing the Fisher criterion under the condition $\boldsymbol{w}^\top \mathbf{\Sigma}_w \boldsymbol{w} = 1$, we obtain the weight vector $\boldsymbol{w}$ as the eigen vector of the following generalized eigenvalue problem:

$$\mathbf{\Sigma}_b \boldsymbol{w} = \lambda \mathbf{\Sigma}_w \boldsymbol{w}, \tag{19}$$

where $\lambda$ is the eigenvalue corresponding to the eigen vector $\boldsymbol{w}$. We select the $N_w$ largest eigen values $\lambda_1, \cdots, \lambda_{N_w}$, and the corresponding eigen vectors $\boldsymbol{w}_1, \cdots, \boldsymbol{w}_{N_w}$. We call the representation with those weight vectors the FisherSPR.

If we apply the FisherSPR to a single-level SPR, the FisherSPR is equal to that in [11, 24]; therefore, the FisherSPR is the generalized form that includes those methods. As stated in Sec. 4.1, CCA maximizes $\mathrm{corr}(s,t)$, subject to $\mathrm{var}(s) = 1$, $\mathrm{var}(t) = 1$. Assuming that $\mathcal{Y}$ space represents the category where the category information is coded in $\boldsymbol{y}$ with a 1-of-$K$ coding scheme, CCA reduces to Fisher LDA. Thus, the difference between PLS and Fisher LDA is the penalty $\mathrm{var}(s)$, and if $\mathrm{var}(s)$ is critical for the discrimination, the FisherSPR is more effective than the PlsSPR. Similarly in the PlsSPR, Eq. (19) is the generalized eigenvalue problem of $\mathbf{\Sigma}_b \in \mathbb{R}^{d_w \times d_w}$, $\mathbf{\Sigma}_w \in \mathbb{R}^{d_w \times d_w}$. Because $d_w \ll d$, Eq. (19) can be quickly solved.

However, if $\mathbf{\Sigma}_w$ is a singular matrix, there is a problem in terms of numerical stability. To avoid singularity, we can use Eigen Weight Maps [24] or use the regularization term; however, determining these parameters is also difficult. In addition, if we use one-vs-the-rest classifiers for a large number of classes, the variance of the "rest class" is likely to be large. The Fisher criterion minimizes the within-class variance while maximizes the between-class variance, and those optimizations are mutually related. When variance of the rest class is large, minimizing the within-class variance becomes dominant, while maximizing the between-class variance is relatively ignored.

## 5. Experiment

In this experiment, we compare the original SPR with our discriminative SPRs (PlsSPR and FisherSPR). Although the discriminative SPR is a general framework and can be applied to any combinations of local descriptors, spatial pooling, and classifiers, we apply it to ScSPM and linear SVMs [27]. The reasons are as follows:

- The combination of ScSPM and linear SVMs is one of the state-of-the-art methods [27, 5, 17]. Thus, this method is appropriate for the baseline.
- The discriminative SPR can find the optimal weight for each cell, but cannot find the weights for each element of the feature vector. On the other hand, sparse coding can be seen to discover the importance for each element of the feature vector, and thus, the discriminative SPR and sparse coding are mutually beneficial.
- As stated in Sec. 3, the weighted SPR becomes of great significance when it is applied to nonlinear spatial pooling, such as max spatial pooling [27].

In the training step, the feature vector in each cell is built with SIFT extraction followed by sparse coding and max spatial pooling. The weights of the discriminative SPR are learned with those features using Eq. (16) or Eq. (19). Then, we plug the obtained weights into Eq. (3) and obtain the discriminative SPR. Finally, the classifiers are learned with the discriminative SPR features and linear SVMs. In the testing step, the features are built by SIFT + sparse coding + max spatial pooling, and the discriminative SPR can be obtained with those features and discriminative weights. The classification results are obtained by inputting the discriminative SPR features into learned linear SVMs. We used the parameters of ScSPM as defaults, which are used in the matlab codes[1] distributed by [27].

We tested the 15 Scenes, Caltech101, and Caltech256 datasets. We investigated the effects of the dimension of the discriminative SPRs for the classification performance. Note that the dimensions of discriminative SPR $N_w d$ is proportional to the number of eigenvectors in Eqs (16) and (19). Moreover, following [4], we investigated two methods for learning the weights:

[1] http://www.ifp.illinois.edu/ jyang29/ScSPM.htm

**GSPW** Global Spatial Pyramid Weights. Instead of giving a fixed weight to each cell at each pyramid level, we learn the weights that give the best classification performance over all the categories.

**CSPW** Class-Specific Spatial Pyramid Weights. Instead of learning the weights that are common across all the classes, the weights are learned for each class separately by optimizing the classification performance for that class using a one-vs-the-rest classification.

### 5.1. 15 Scenes

We experimented with a commonly used scene classification benchmark dataset by Lazebnik et al., [12] (15 Scenes dataset). The 15 Scenes dataset consists of the gray images of OT8 [19], plus seven additional classes. OT8 consists of 2,688 color images of eight classes. Each class has sample images from 260 to 410. In total, it contains 4,492 gray images. We randomly chose 100 training images for each class. We used the remaining samples as test data, and calculated the mean of the classification rate for each class. This score was averaged over 10 trials, randomly replacing the training and test samples. This is the same methodology as was used in the previous studies.

Figure 3 shows the performance of a variety of the discriminative SPRs. The baseline method is ScSPM + linear SVMs with level 0 ($1 \times 1$, $d_s = 1$), level 0-1 ($1 \times 1$, $2 \times 2$, $d_s = 5$), and level 0-2 ($1 \times 1$, $2 \times 2$, $4 \times 4$, $d_s = 21$) without spatial weights. We used two pyramid structures for our representations: level 0-2 ($d_w = 21$), and level 0-3 ($1 \times 1$, $2 \times 2$, $4 \times 4$, $8 \times 8$, $d_w = 85$). All discriminative SPRs outperform ScSPM in low dimension. In high dimension, the discriminative SPRs are comparable to or slightly better than ScSPM. PlsSPR + CSPW or PlsSPR + GSPW with level 0-3 obtain the highest performance. Their classification scores saturate at $N_w = 7$. The direct comparison between ScSPM and PlsSPR + CSPW is shown in Table 1. In the 15 Scene dataset, there is no difference between GSPWs and CSPWs. The PlsSPRs are slightly better than the FisherSPRs.

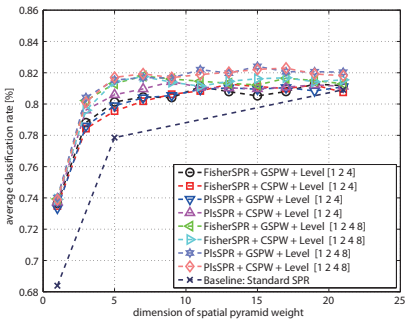



Figure 3. Performance of discriminative SPR on 15 Scenes.

Table 1. Classification rate [%] on 15 Scenes. PlsSPR was applied to level 0-3 ($1 \times 1,\ 2 \times 2,\ 4 \times 4,\ 8 \times 8$) with CSPW.

| Algorithms | Level 0 $N_w = d_s = 1$ | Level 0-1 $N_w = d_s = 5$ | Level 0-2 $N_w = d_s = 21$ |
|---|---|---|---|
| ScSPM | $68.39 \pm 0.56$ | $77.86 \pm 0.96$ | $80.91 \pm 0.51$ |
| PlsSPR | $73.89 \pm 0.89$ | $81.69 \pm 0.54$ | $81.81 \pm 0.54$ |

Here, we discuss the extra computational costs of the discriminative SPRs. In the training step, the PlsSPR and the FisherSPR require the eigenvalue problem (Eq. (16)) and the generalized eigenvalue problem (Eq. (19)) once, respectively. The CSPWs must calculate Eq. (16) or Eq. (19) $K$ times. In both the training and testing steps, the features are transformed with the weights. To obtain the final feature, GSPWs calculate the transformation of Eq. (2) $N_w$ times, and the CSPWs $K \times N_w$ times. The calculation costs of Eq. (16), Eq. (19), and Eq. (2) are shown in Table 2. These computational costs on 15 Scenes were evaluated on a dual Xeon5160 with 14 GB RAM. As shown, the extra computational costs for both training and testing are low. Considering that the discriminative SPRs obtain good performance in low dimension, the discriminative SPRs are highly effective on the 15 Scenes dataset.

Table 2. Extra calculation costs on 15 Scenes.

| Training [ms] PlsSPR Eq. (16) | Training [ms] FisherSPR Eq. (19) | Compressing [$\mu s$] Eq. (2) |
|---|---|---|
| $189.2 \pm 82.5$ | $404.7 \pm 39.8$ | $55.9 \pm 9.4$ |

### 5.2. Caltech101

Caltech101 [9] is the de-facto, standard, object-recognition dataset. This dataset consists of images from 101 object categories and one background class, and contains images from 31 to 800 per category. This dataset has large intra-class variety. The spatial information is essential for image recognition, because in this dataset, the images in the same category are well centered. To evaluate the classification performance, we followed the most standard methodology. 15 images were randomly selected from all 102 categories for training purposes, and the remaining images were used for testing. The classification score was averaged over 10 trials.

Figure 4 shows the performance of a variety of the discriminative SPRs on Caltech101. The baseline method is ScSPM + linear SVMs with level 0, level 0-1, and level 0-2 without spatial weights. We used two pyramid structures for the discriminative SPRs: level 0-2 and level 0-3. All of the discriminative SPRs outperform ScSPM in low dimension. In high dimension, our methods are comparable to or slightly better than ScSPM, except for FisherSPR + CSPW + Level 0-2, because as discussed in Sec. 4.3, CSPW uses one-vs-the-rest classifiers for a large number of classes (102 classes), and hence, the variance of the "rest

class" becomes large. This condition inhibits the effect of the variance of the between-class relatively, and lowers the discriminative power. PlsSPR + CSPW with level 0-3 outputs the highest performance. Their classification scores saturate at $N_w = 9$. The direct comparison between ScSPM and PlsSPR + CSPW is shown in Table 3. The computation costs are the same as for the 15 Scenes dataset, if the same $N_w$ is used.

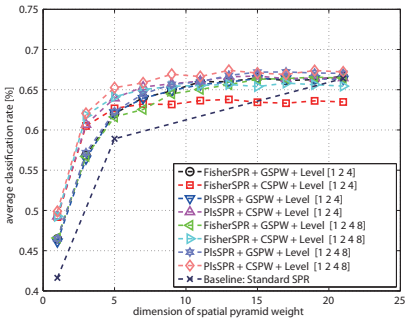



Figure 4. Performance of discriminative SPR on Caltech101.

Table 3. Classification rate [%] on Caltech101. PlsSPR was applied to level 0-3 ($1 \times 1,\ 2 \times 2,\ 4 \times 4,\ 8 \times 8$) with CSPW.

| Algorithms | Level 0 $N_w = d_s = 1$ | Level 0-1 $N_w = d_s = 5$ | Level 0-2 $N_w = d_s = 21$ |
|---|---|---|---|
| ScSPM | $41.62 \pm 0.57$ | $58.90 \pm 0.61$ | $66.40 \pm 0.55$ |
| PlsSPR | $49.88 \pm 0.36$ | $65.26 \pm 0.83$ | $67.21 \pm 0.67$ |

Figure 5 shows the absolute value of the spatial pyramid weights with PlsSPR + CSPW + Level 0-3 on Caltech101. This figure represents that the optimal discriminative weights can be obtained for each cell on each pyramid level with our method. For example, the most distinctive cell is the cell at level 0 for "windsor chair," the upper right cell at level 1 for "revolver," and the both side cells at level 2 for "watch." Furthermore, the most important cells appear by level 2, therefore, the fact that many studies have used SPM with level 0-2 is supported by this experiment.

### 5.3. Caltech256

Caltech256 [10] consists of images from 256 object categories. This dataset contains images from 80 to 827 per category. The significance of this database is its large inter-class variability, as well as an intra-class variability larger than that found in Caltech101. Moreover, there is no alignment among the object categories. To evaluate the classification performance, we followed the common methodology. Fifteen images were randomly selected from all 256 categories for training purposes, and the remaining images were used for testing. The classification score was averaged over 10 trials.

Figure 6 shows the performance of a variety of discriminative SPRs on Caltech256. The baseline method is Sc-

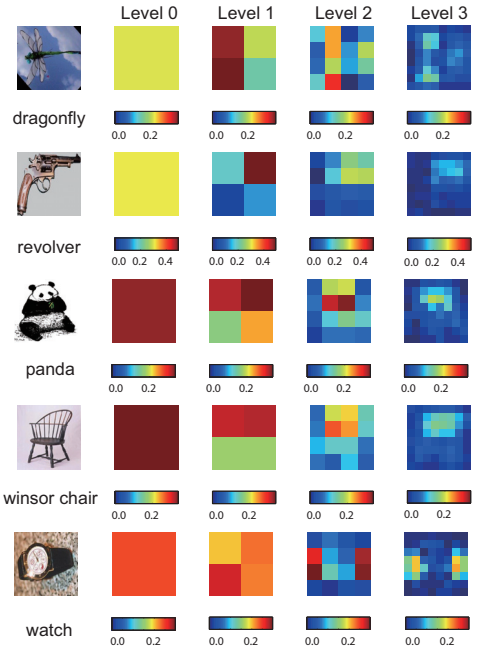



Figure 5. Some examples of discriminative spatial pyramid weights on Caltech101 dataset with PlsSPR+CSPW.

SPM + linear SVMs, which is the same as that used in the previous experiments. We used only one pyramid structure on level 0-2, because of a shortage of the main memory. Our methods outperform ScSPM in low dimension, except for FisherSPR + CSPW, whose score is worse than that of Caltech101. This is because as the number of classes increases from 102 to 256, the variance of the "rest class" becomes larger than that of Caltech101. Among our methods, PlsSPR + CSPW produces the highest performance. Their classification scores saturate at $N_w = 9$. The direct comparison between ScSPM and PlsSPR + CSPW is shown in Table 4. For the computation costs, see the discussion in the 15 Scenes experiment.

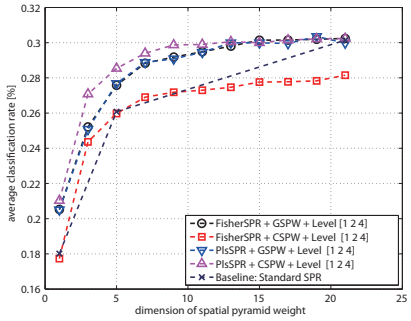



Figure 6. Performance of discriminative SPR on Caltech256.

Table 4. Classification rate [%] on Caltech256. PlsSPR was applied to level 0-2 ($1 \times 1,\ 2 \times 2,\ 4 \times 4$) with CSPW.

| Algorithms | Level 0 $N_w = d_s = 1$ | Level 0-1 $N_w = d_s = 5$ | Level 0-2 $N_w = d_s = 21$ |
|---|---|---|---|
| ScSPM | $18.02 \pm 0.30$ | $26.08 \pm 0.19$ | $30.14 \pm 0.46$ |
| PlsSPR | $21.03 \pm 0.30$ | $28.54 \pm 0.20$ | $30.24 \pm 0.34$ |

## 6. Conclusion

In this paper, we propose the discriminative SPR. As implementations of the discriminative SPR, we also propose the PlsSPR and the FisherSPR. Our methods form the feature as a weighted sum of the mid-local features over all of the pyramid levels. The weights are automatically selected to maximize the discriminative power. By using datasets, our methods showed high performance, especially in low dimension. The discriminative SPR can suggest distinctive cells and pyramid levels simultaneously. In future, we will apply the discriminative SPR to other local descriptors, spatial pooling, and classifiers. The weights of the discriminative SPR are permitted to have negative values, and the weight vector is dense. Non-negative decomposition is related to the extraction of the relevant parts from images [13]. Thus, we will extend this study to incorporate both non-negativity and sparseness into the discriminative SPR.

## Acknowledgement

This work was partially supported by JST PRESTO, and Info-plosion. The authors would like to thank Takio Kurita for useful discussions.

## References

[1] M. Barker and W. Rayens. Partial least squares for discrimination. *J. of Chemometrics*, 17(3):166–173, 2003. 4

[2] P. Belhumeur, J. Hespanha, and D. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *PAMI*, 19(7):711–720, 1997. 2

[3] A. Bosch, A. Zisserman, and X. Munoz. Image classification using random forests and ferns. *ICCV*, 2007. 2

[4] A. Bosch, A. Zisserman, and X. Munoz. Representing shape with a spatial pyramid kernel. *CIVR*, 2007. 2, 5

[5] Y.-L. Boureau, F. Bach, Y. LeCun, and J. Ponce. Learning mid-level features for recognition. *CVPR*, 2010. 1, 5

[6] Y. Cao, C. Wang, Z. Li, L. Zhang, and L. Zhang. Spatial-bag-of-features. *CVPR*, 2010. 1

[7] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. *Int. Workshop on Stat. Learning in Comput. Vision*, 2004. 1

[8] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. *CVPR*, 2005. 2

[9] 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. *CVPR Workshop on Generative-Model Based Vision*, 2004. 6

[10] G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. Tech. Rep. 7694, California Institute of Technology, 2007. 7

[11] T. Harada, H. Nakayama, and Y. Kuniyoshi. Improving local descriptors by embedding global and local spatial information. *ECCV*, 2010. 1, 2, 3, 5

[12] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. *CVPR*, 2006. 1, 2, 6

[13] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, 1999. 8

[14] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004. 1

[15] M. Marszalek, C. Schmid, H. Harzallah, and J. van de Weijer. Learning object representations for visual object class recognition. *Visual Recog. Challange workshop*, 2007. 2

[16] N. Morioka and S. Satoh. Building compact local pairwise codebook with joint feature space clustering. *ECCV*, 2010. 1

[17] N. Morioka and S. Satoh. Learning directional local pairwise bases with sparse coding. *BMVC*, 2010. 1, 5

[18] H. Nakayama, T. Harada, and Y. Kuniyoshi. Global gaussian approach for scene categorization using information geometry. *CVPR*, 2010. 1

[19] A. Oliva and A. Torralba. Modeling the shape of the scene: a holistic representation of the spatial envelope. *IJCV*, 42(3):145–175, 2001. 1, 6

[20] B. A. Olshausen and D. J. Fieldt. Sparse coding with an overcomplete basis set: a strategy employed by v1. *Vision Research*, 37:3311–3325, 1997. 1

[21] F. Perronnin and C. Dance. Fisher kernels on visual vocabularies for image categorization. *CVPR*, 2007. 1

[22] F. Perronnin, J. Sánchez, and T. Mensink. Improving the fisher kernel for large-scale image classification. *ECCV*, 2010. 1, 2

[23] W. R. Schwartz, A. Kembhavi, D. Harwood, and L. S. Davis. Human detection using partial least squares analysis. *ICCV*, 2009. 2, 5

[24] Y. Shinohara and N. Otsu. Facial expression recognition using fisher weight maps. *IEEE FG*, 2004. 2, 3, 5

[25] M. Turk and A. Pentland. Face recognition using eigenfaces. *CVPR*, 1991. 2

[26] H. Wold. Estimation of principal components and related models by iterative least squares. In P. Krishnaiaah, editor, *Multivariate Analysis*, pages 391–420, 1996. Academic Press. 4

[27] J. Yang, K. Yu, Y. Gong, and T. Huang. Linear spatial pyramid matching using sparse coding for image classification. *CVPR*, 2009. 1, 2, 3, 5

[28] J. Yang, K. Yu, and T. Huang. Efficient highly over-complete sparse coding using a mixture model. *ECCV*, 2010. 1, 2

[29] X. Zhou, K. Yu, T. Zhang, and T. S. Huang. Image classification using super-vector coding of local image descriptors. *ECCV*, 2010. 1, 2