Pattern Anal Applic (2017) 20:717–732
DOI 10.1007/s10044-016-0529-z

THEORETICAL ADVANCES

Class-specific image representation for image classification using multiple scale-invariant region detectors

Hui-Jin Lee[1] · Ki-Sang Hong[1]

Received: 4 June 2014 / Accepted: 5 January 2016 / Published online: 19 January 2016

Abstract We propose a new class-specific image representation for image classification using multiple region detectors. The new representation is designed to solve the problem of increasing variation in object location and size within images of a class, for which traditional spatial pyramid matching shows limited classification accuracy. We propose a new region-division method that divides the image region into two class-specific regions, called class-specific region-of-interest (C-ROI) and focal region (FR). Using multiple region detectors and appropriate mixing of their responses avoids the problem of selecting a region detector that gives the best classification accuracy for a given image class, and thereby yields better results than using only one region detector. Several scale-invariant region detectors are used to obtain C-ROI and FR by considering their importance over a given image class. In experiments using several well-known datasets, the proposed method improved the accuracy and achieved results that were better than or comparable to those achieved by the related methods.

Keywords Image representation · Class-specific region-of-interest (C-ROI) · Focal region (FR) · Classification accuracy · Bag-of-words

1 Introduction

Image classification is the process of classifying images according to the objects contained in them. One of the main challenges in designing a classification system is to develop an appropriate method of image representation [2, 14, 16, 28, 29, 34, 36]. Bag-of-words (BoW) has been widely used as the image representation method for image classification [14, 34, 36].

In the traditional BoW framework, each image is represented as a histogram of word frequency by assigning all local features to visual words. This model is insensitive to scale and illumination change, but suffers from lack of spatial information. Hence, pyramid structure representation such as spatial pyramid matching (SPM) [14] has been used to extend the global BoW representation by partitioning images into progressively finer sub-regions. The SPM computes a histogram of word frequency within each sub-region, and concatenates all the histograms to form the final image representation. However, SPM suffers from degradation of classification accuracy due to varying locations of objects in images of a same class (Fig. 1a). If the object locations in images are different, the spatial partition based on SPM (Fig. 1b) may mismatch them. This problem can be solved by partitioning the images separately into object and background areas.

In this paper, we propose a class-specific image representation (Fig. 1c) to match objects and background areas more accurately. For the object area, we define two kinds of region: class-specific region-of-interest (C-ROI) and focal region (FR). The C-ROI is defined as a region that can be found in all images of the same class (Fig. 1c, Region 2). A C-ROI contains almost all the region of an object of interest. The FR is defined as the most informative region in the C-ROI, i.e., the most informative part of

✉ Hui-Jin Lee
huijin@postech.ac.kr

[1] Image Information Processing Lab., POSTECH, San 31 Hyojadong, Pohang, South Korea

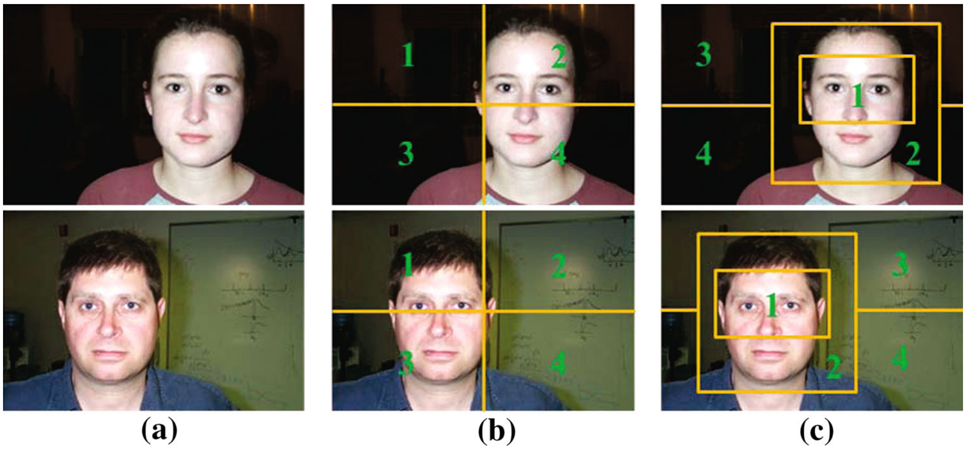

Fig. 1 Spatial layout for image representation. Numbers are indices of regions to be matched in two images. **a** Example images, **b** SPM layout, **c** proposed layout

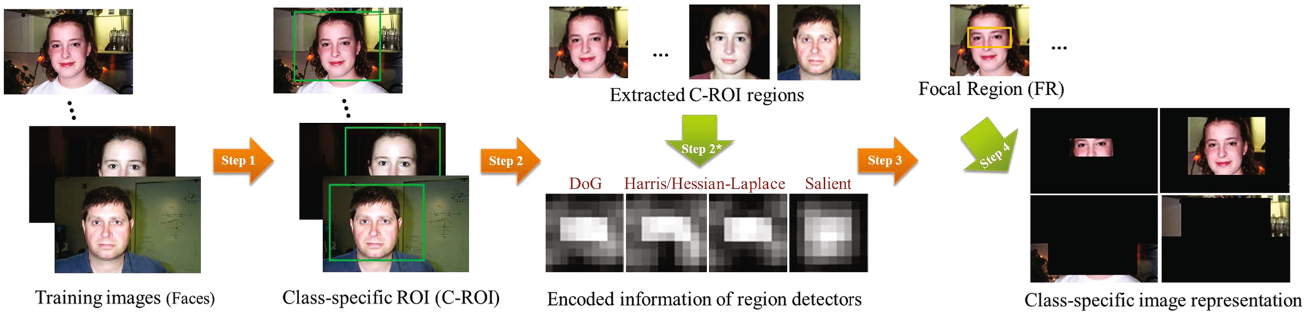

Fig. 2 Framework of proposed method for class-specific image representation. Steps are described in Sect. 3

the object of interest (Fig. 1c, Region 1). For the background area, we define the remaining region excluding the object area as the region in which to match the background scene. The region is divided horizontally into two regions (Fig. 1c, Regions 3 and 4) to model the background scene. By concatenating feature vectors of each region extracted in this way, we can construct an image representation that is more class-specific to objects of interest than is traditional SPM.

To extract C-ROI and FR, we use multiple region detectors. Usually, the classification accuracy depends on the region detector used, because they have different characteristics. The region detector that is most suitable for classification depends on the image class given. In this paper we use four scale-invariant region detectors: DoG [22], Harris-Laplace [24], Hessian-Laplace [25], and salient [12]. These region detectors capture a variety of characteristic information such as blob-like, corner-like, and entropy-based features; hence, combining the detectors' output can be helpful to classify a variety of objects. We use the similarity of information obtained from multiple region detectors to extract C-ROIs in images of a class. We use the spatial distribution and appearance characteristics of region detectors to compute the similarity for extraction of C-ROIs.

The characteristics of images in various image classes can vary widely, which means that different information is required to describe images of different classes. For example, for faces (Fig. 1), the eye regions may be the most informative region to describe the class, so capturing blob-like structures is quite useful. Therefore, in this paper, the most informative region of a class, i.e., the FR, is obtained by considering the class-specific importance of each region detector for the class. The class-specific importance of a region detector presents how strongly the region detector affects extraction of C-ROIs.

The proposed method to construct a specific representation for a class consists of four steps (Fig. 2). Given training images of a class, step 1 extracts C-ROIs in them; these C-ROIs are defined by the similarity of information obtained from multiple region detectors. When the C-ROIs are extracted, the class-specific importance of each region detector is also computed to indicate which region detectors give dominant effects on extraction of C-ROIs. Step 2 extracts the spatial distribution of keypoints extracted by each region detector in the C-ROIs. We use nonnegative matrix factorization (NMF) to obtain the semantic spatial distribution for each region detector. Step 3 finds the FR of the class by summing the semantic spatial distributions of region detectors weighted by the class-specific importance and by thresholding. Step 4 defines spatial pooling regions, and concatenates encodings of each spatial region to form the final image representation. Here, the encoding of each spatial region is the BoW representation for obtained features of the spatial region.

2 Related work

Image representation has been developed as the main objective for various tasks such as image classification and retrieval [14, 16, 17, 34, 36]. For image classification, the pyramid structure representation based on Bag-of-words (BoW) [14, 34, 36] has been widely used as the image representation. This representation extends the global BoW representation and models approximate geometric layout by partitioning the image plane into progressively finer sub-regions; this procedure has become standard in the image classification task. Yang et al. [36] and Wang et al. [34] proposed extensions of the SPM approach [14]; the extensions compute a pyramid image representation based on effective coding schemes, instead of the k-means vector quantization in the SPM. The extensions obtained better classification accuracy than traditional SPM, and attained state-of-the-art accuracy on some benchmarks. However, if corresponding object locations and scene layout differ among images, these methods also suffer from misalignment.

The misalignment problem between objects in images can be solved by the object-centered representation. The part-based approach [5–8] and the interest region-based approach [1, 9, 11, 26, 31, 35] are two widely used object-centered representations. The part-based approach represents an object as a spatial layout of multiple parts, where the deformable configuration is characterized by spring-like connections between them [8] or by a joint Gaussian density of the locations of parts obtained from a random constellation [6]. Other methods [5, 7] constructed with spring-like connections introduce many local ambiguities and limited parts. The disadvantage of existing part-based models is that they depend heavily on the representations of each part. The interest region-based approaches represent an image by focusing on the specific interest region considered in their work. Galleguillos et al. [9] focused on the interest region for image classification by incorporating multiple stable segmentations and Bag-of-features (BoF) image representation into a multiple instance learning (MIL) framework. Chai et al. [1] proposed segmenting images into foreground and background within a co-segmentation scenario to improve image classification accuracy. Yakhnenko et al. [35] used a latent-SVM model, which uses all regions to score an image, and associates each region with a latent variable that indicates whether or not the region represents the object of interest. Nguyen [26] used segment-based Support Vector Machines which simultaneously localize the most discriminative set of segments and use them to learn an SVM. However, all of these methods are based on segments, and are sensitive to the segmentation result. Recently, some studies [11, 31] presented methods based on the saliency map for the interest region. Sharma [31] proposed a method to learn the discriminative spatial saliency of images while simultaneously learning a max margin classifier for a given visual classification task, but this work focused mainly on image classes like ‘riding horse’ in which the spatial relation between a person and an object (‘horse’) is important information to obtain the saliency map. Jiang et al. [11] used supervised learning to map the regional feature vector to a saliency score that yielded the saliency map. Because the method is only evaluated in terms of salient object detection, not image classification, the interest region obtained from this method has not been proven to be effective for image classification.

The proposed method also aims to find interest regions for the object-centered representation. Unlike most existing region-based approaches that obtain interest regions from segmentation results or saliency maps, our method exploits scale-invariant region detectors to model the interest regions. Scale-invariant region detectors such as DoG, Harris-Laplace, Hessian-Laplace, and Salient can capture important information (e.g., blob-like, corner-like, entropy-based) in images. Traditionally, region detectors have been used to extract keypoints for image matching and object class recognition. Some previous studies [6, 23] proposed frameworks that used scale-invariant region detectors to classify images. Fergus et al. [6] used the Salient detector to construct a probabilistic representation. Mikolajczyk et al. [23] compared the classification accuracy of local detectors and descriptors in the context of object class recognition. However, the question of which region detectors are most effective for a specific class is seldom discussed. In this paper, we try to define effects of region detectors for a specific class and use them in the class-specific modeling.

Nonnegative matrix factorization (NMF) [15] is an effective factor analysis method. It aims to find two nonnegative matrices whose product provides a good approximation to the original matrix. It is optimal for learning the parts of objects because the nonnegative constraints allow only additive combinations. The methods based on NMF and its variants have been applied to various tasks such as feature selection or data dimension reduction [18–21]. In this paper, we use NMF to obtain the semantic spatial information of keypoints extracted by each region detector; the semantic spatial information is used to construct class-specific representation.

3 Proposed method

In this section, we propose a framework that uses multiple scale-invariant region detectors for class-specific image representation. For class-specific object area, we define two kinds of region, i.e., the C-ROI (Sect. 3.2) and the FR (Sect. 3.3); a class-specific image representation (Sect. 3.4) is obtained by using these two regions.

3.1 Region detectors

In this paper, we use four different scale-invariant region detectors to obtain class-specific spatial layouts: DoG, Harris-Laplace, Hessian-Laplace and salient detectors. The region detectors provide locations and scale of keypoints, and capture different kinds of information: the DoG and Hessian-Laplace detectors are suitable for finding blob-like structure; the Harris-Laplace detector captures corner-like structures; and, the salient detector extracts regions that have high entropy (or information). These region detectors have been successfully used in object classification. In the preprocessing step, we use these detectors to extract keypoints and their local information (locations and scale of the local region) for all training images.

3.2 Class-specific region-of-interest (C-ROI)

The C-ROI is a region which can be commonly found in images of same class. To find this region, we first obtain candidate C-ROIs at different scales and locations. The spatial distribution and appearance characteristics of region detectors are used to select the C-ROI among candidate C-ROIs (Algorithm 1).

Algorithm 1 Extraction of C-ROIs of images in a class

```
Input: $N$ training images $x_i$ in a class, $i = 1, ..., N$.
Output: $N$ C-ROIs, $P_{ik^*}$, $i = 1, ..., N$.
for $i = 1$ to $N$ do
    Generate candidate C-ROIs $P_{ik}$ of $x_i$, $k = 1, ..., M_i$.
    for $k = 1$ to $M_i$ do
        for $f = 1$ to $N_f$ do
            For each region detector $f$,
            1. Compute spatial histogram $S_f^{P_{ik}}$ of $P_{ik}$.
            2. Construct appearance histogram $A_f^{P_{ik}}$ of $P_{ik}$.
        end for
        For each candidate C-ROI $P_{ik}$,
        1. Compute $Z_{ik}^j$ for all $x_j$ $(j \neq i)$ using eq.(2).
        2. Rank $Z_{ik}^j$ for all $x_j$.
        3. Keep $W$ smallest values $B_{ik}$.
        4. Compute score $T_{ik}$ over $B_{ik}$ using eq.(4).
    end for
    Select $P_{ik^*}$ in $x_i$ using eq.(5).
end for
```

3.2.1 Candidate C-ROIs

To extract C-ROIs in images of the same class, we must first identify candidate C-ROIs in images. In many previous studies, candidate regions to be processed were obtained from all possible locations and scales, so tens of thousands of candidates may have been identified. Although accurate target regions could be obtained by considering all possible candidates, this approach is impractical because it entails huge computational cost. Therefore, we try to reduce the number of candidate C-ROIs by choosing a limited number. In experiments, we observed that if extracted keypoints are concentrated in a region, the region is worth considering closely. This means that we must compute the density of the 2-dimensional distribution of keypoint and find its local peaks. To do this, we apply the four region detectors to every training image of same class. For each image, we superimpose the four kinds of detected keypoints on one image, then use the MeanShift algorithm [3] to identify local peaks of the keypoint distribution for the use as locations of candidate C-ROIs. To increase the accuracy of locating C-ROI, we add some extra locations around the local peaks (in our work, these are located in ±10 and ±20 pixels from local peaks). To maintain the classification accuracy even when the C-ROIs vary in size, we use three different sizes of candidate C-ROI at each candidate location (in our work, sizes of the regions are 0.5, 0.7, and 0.9 times of the image's width/height ratio). Given a set of images $I = \{x_1, x_2, \ldots, x_N\}$ in a class, we obtain a collection of M_i candidate C-ROIs in each image x_i; this collection is denoted as $P_i = \{P_{i1}, P_{i2}, \ldots, P_{iM_i}\}$. Even if a small number of candidate regions is considered, the obtained candidate C-ROIs cover most objects of interest in images (Fig. 3).

The next step is to select a C-ROI among the M_i candidate C-ROIs for each image. To do this, we define two features; spatial histogram $S_f^{P_{ik}}$, and appearance histogram $A_f^{P_{ik}}$, for each candidate C-ROI, where index f represents a region detector. Using these two features, we select a C-ROI for each image that gives the best matching score with candidate C-ROIs from other images of the same class. Although finding a C-ROI for each image is nearly exhaustive matching, it is practical because a small number of candidate C-ROIs is considered in this process.

3.2.2 Spatial histogram $S_f^{P_{ik}}$

C-ROIs in images of a class should show similar distributions of keypoints extracted by region detectors. To describe this similarity, we compute spatial histogram $S_f^{P_{ik}}$

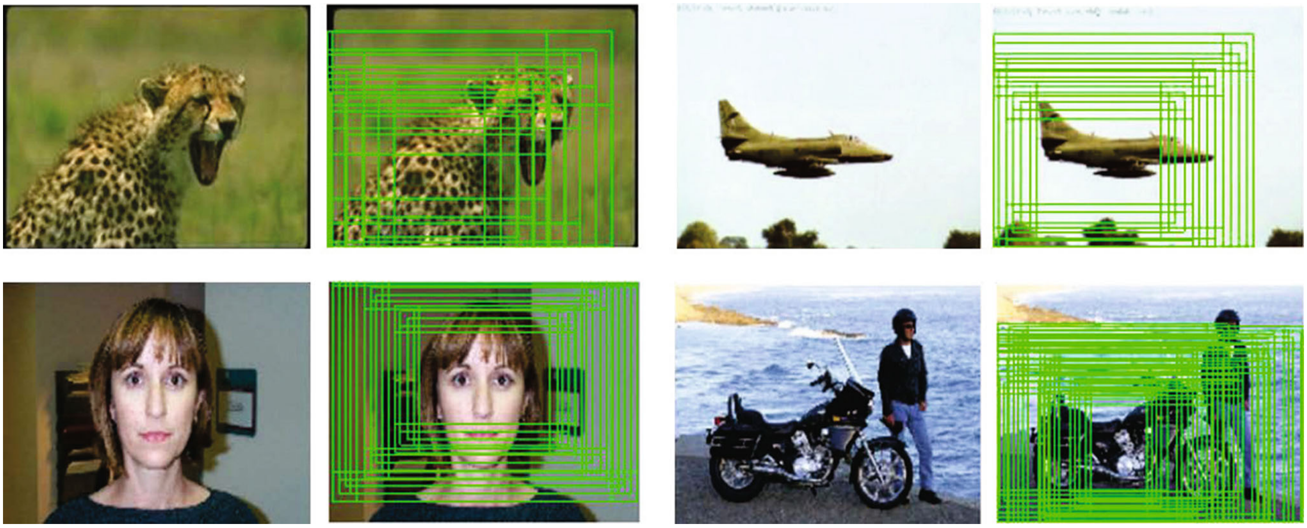

Fig. 3 Obtained candidate C-ROIs (*green rectangles*) in some images of Caltech-4

(Fig. 4) of keypoints detected using region detector f within each candidate C-ROI P_{ik}.

Given a candidate C-ROI P_{ik}, $k \in M_i$ in an image, P_{ik} is decomposed into N_R regular small sub-regions (in our work, N_R is set to 100, i.e., 10×10 grids). For a region detector f in P_{ik}, a spatial histogram $S_f^{P_{ik}}$ with N_R bins is constructed by counting the number of keypoints detected using f in each bin. The spatial histogram of bin $r \in N_R$ is computed by

$$S_f^{P_{ik}}(r) = \sum_{e \in P_{ik}} \delta(e \in R_r), \tag{1}$$

where e is the keypoint extracted using f, R_r is a sub-region that corresponds to bin r, and $\delta(P)$ returns 1 if P is true and 0 otherwise. The value of each bin is normalized to a proportion of the maximum value of the spatial histogram.

3.2.3 *Appearance histogram $A_f^{P_{ik}}$*

To obtain the appearance histogram from each candidate C-ROI P_{ik}, we use the Bag-of-words (BoW) representation, for which denseSIFT [14] features are extracted from subsampled training images and a codebook is constructed from them using k-means clustering as a preprocessing step. Only one codebook is constructed for all classes. Once the codebook is prepared, all that is required to compute the appearance histogram is to collect denseSIFT features in P_{ik} and to use the codebook to generate a histogram. To extract the individual characteristics of each of the four region detectors used in this paper, we modify P_{ik} in two aspects. First, instead of using the whole region of P_{ik} to obtain the histogram, we exclude some sub-regions N_r in which the number of keypoints detected using f is very small. Here, the sub-region is the same as the sub-grid region used for spatial histogram computation; i.e., we use only the sub-regions that are informative enough for each f. Second, we modify the size s and location l of each remaining sub-region that is used for histogram computation. Instead of using a fixed-sized sub-region at the regular location, we relocate each remaining sub-region to the average location of all the keypoints detected in it. We also set its size to $s = k_s\overline{\sigma_s}$, where $\overline{\sigma_s}$ is the average scale of all detected keypoints in the sub-region. Here, k_s is set to 8.

We obtain four $A_f^{P_{ik}}$ for each P_{ik} (Fig. 4). Notice that we use denseSIFT features for histogram computation; we use the four region detectors only to define the newly changed sub-regions in each P_{ik}.

3.2.4 *C-ROI selection*

C-ROI is a region that can be commonly found in images of the same class. To select the region $k \in M_i$ among candidate C-ROIs P_{ik} in an image x_i, the similarity of P_{ik} over the class must be measured. For a candidate C-ROI P_{ik}, the closest distance to candidate C-ROIs in another image x_j is defined as

$$Z_{ik}^j = \min{}_{l \in M_j} Z(P_{ik}, P_{jl}), \tag{2}$$

The distance $Z(P_{ik}, P_{jl})$ between two regions is

$$\begin{aligned} Z(P_{ik}, P_{jl}) = \min_{f \in F}[&\alpha D(S_f^{P_{ik}}, S_f^{P_{jl}}) \\ &+ (1-\alpha) D(A_f^{P_{ik}}, A_f^{P_{jl}})], \end{aligned} \tag{3}$$

where F is the set of region detectors, α is the trade-off parameter between the two types of information (in our case, $\alpha = 0.5$), and D computes the χ^2 distance between two histograms.

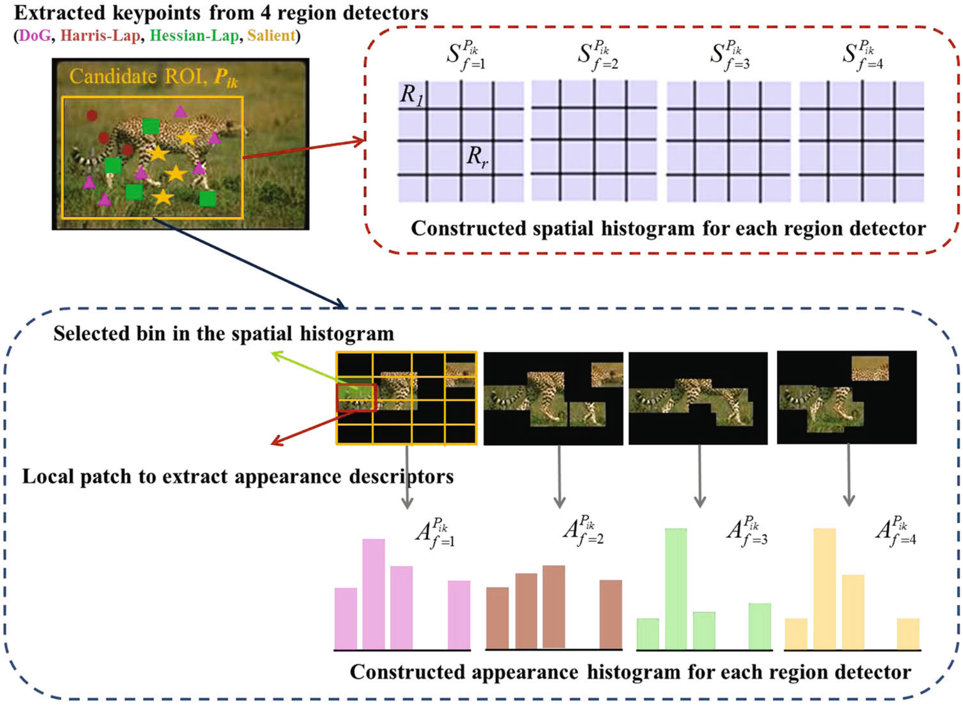
Extracted keypoints from 4 region detectors
(DoG, Harris-Lap, Hessian-Lap, Salient)
Candidate ROI, P_{ik}
$S_{f=1}^{P_{ik}}$
$S_{f=2}^{P_{ik}}$
$S_{f=3}^{P_{ik}}$
$S_{f=4}^{P_{ik}}$
R_1
R_r
Constructed spatial histogram for each region detector
Selected bin in the spatial histogram
Local patch to extract appearance descriptors
$A_{f=1}^{P_{ik}}$
$A_{f=2}^{P_{ik}}$
$A_{f=3}^{P_{ik}}$
$A_{f=4}^{P_{ik}}$
Constructed appearance histogram for each region detector

Fig. 4 Conceptual illustration of spatial histogram and appearance histogram for a given candidate C-ROI

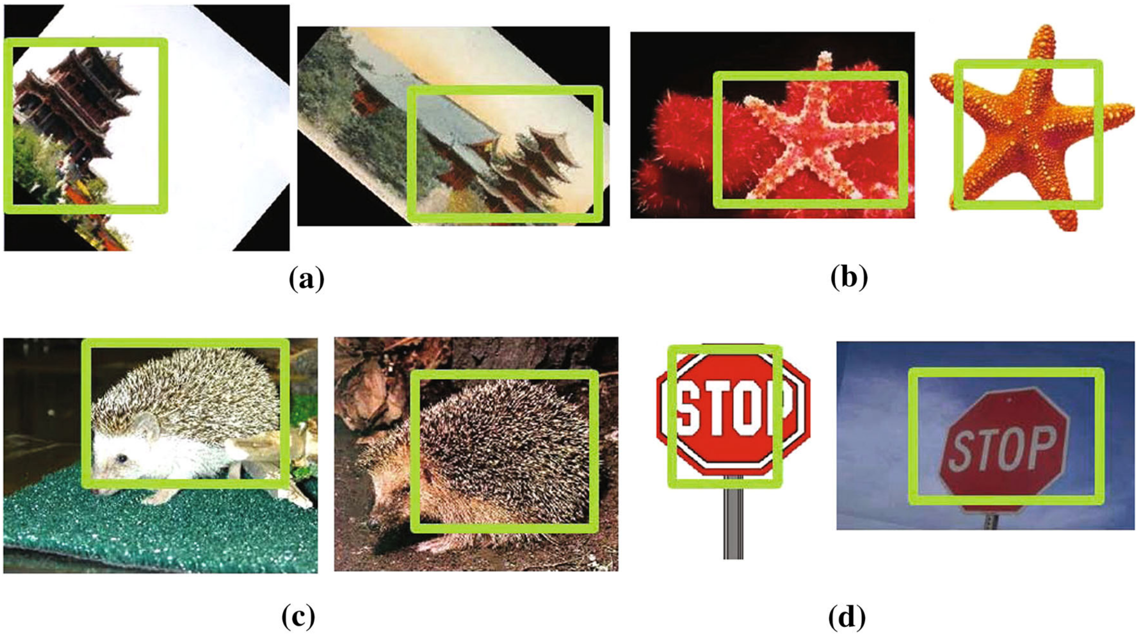

Fig. 5 Examples of extracted C-ROIs for some classes of Caltech-101: **a** pagoda, **b** starfish, **c** hedgehog and **d** stop-sign classes

To measure the similarity of the candidate C-ROI over a class, we use Multi-ranking Amalgamation Strategy [10]. For region k in x_i, we re-rank Z_{ik}^{j} with all j and keep the W smallest values, which are defined as $B_{ik} = \{B_{ik}^{1}, B_{ik}^{2}, \ldots, B_{ik}^{W}\}$, for which $B_{ik}^{m} < B_{ik}^{n}$ and $m < n$. The score of the region k as a C-ROI increases as the values in B_{ik} decrease. Therefore, the score of P_{ik} is defined as

$$T_{ik} = \frac{1}{W}\sum_{w=1}^{W}\frac{1}{log(1+B_{ik}^{w})}. \tag{4}$$

The C-ROI P_{ik^*} in x_i is selected by

$$k^* = \arg\max_{k} T_{ik}. \tag{5}$$

As examples we present extracted C-ROIs (Fig. 5) for several classes of the Caltech-101 dataset.

3.2.5 *Class-specific weights of region detectors for C-ROIs*

We can measure the class-specific weights of region detectors that represent the relative importance of a region detector for detecting C-ROIs from a given class of images (Algorithm 2).

```
Algorithm 2 Class-specific weights of region detectors
  Input: $N$ C-ROIs, $P_{ik^*}$, $i = 1, ..., N$, in a class $c$
  Output: Weights $w_f^c$ of region detectors over $c$
  for $i = 1$ to $N$ do
      For each C-ROI $P_{ik^*}$,
      1. Obtain the set of $W$ regions, $Q_{ik^*}$ having the smallest distance with $P_{ik^*}$.
      2. Find the index set of region detectors, $F_i$ using eq.(6).
  end for
  Obtain a collection of index sets of the class $c$, $F^c$.
  for $f = 1$ to $N_f$ do
      For each region detector $f$,
      Compute a weight, $w_f^c$.
  end for
```

Let P_{ik^*} be a C-ROI in an image x_i and $Q_{ik^*} = \{Q_{ik^*}^1, Q_{ik^*}^2, \ldots, Q_{ik^*}^W\}$ be the set of W candidate C-ROIs in other images having the smallest distance from P_{ik^*}. Then we can find the region detector that gives the minimum distance between P_{ik^*} and $Q_{ik^*}^w$, which we denote as $f_{ik^*}^w$, using the equation:

$$f_{ik^*}^w = \arg\min_{f \in F} [\alpha D(S_f^{P_{ik^*}}, S_f^{Q_{ik^*}^w}) + (1-\alpha) D(A_f^{P_{ik^*}}, A_f^{Q_{ik^*}^w})]. \quad (6)$$

For the set Q_{ik^*}, we can get a set F_i; $F_i = \{f_{ik^*}^1, f_{ik^*}^2, \ldots, f_{ik^*}^W\}$. For a given image class c, we can get a collection of index sets defined as: $F^c = \{F_1, F_2, \ldots, F_N\}$, where N is the number of images in c. Then, $|F^c| = WN$.

The relative importance, or weight w_f^c of the region detector f for c can be measured by counting the number N_f^c of each region detector in F^c:

$$N_f^c = \sum_{a \in F^c} \delta(a == f),$$
$$w_f^c = \frac{N_f^c}{\max_{f \in F} N_f^c}, \quad (7)$$

where $\delta(P)$ returns 1 if P is true, and 0 otherwise. The relative importance of region detectors shows relatively large variation depending on image classes (Table 1).

3.3 Focal region (FR)

The extracted C-ROI in Sect. 3.2 generally contains the whole structure of the object of interest. In this section, we

Table 1 Relative importance of detectors for image classes in Fig. 5

Classes	DoG	Harris-lap	Hessian-lap	Salient
Pagoda	0.74	1.00	0.82	0.94
Hedgehog	1.00	0.72	0.93	0.61
Starfish	0.99	0.97	1.00	0.77
Stop-sign	0.74	0.92	0.65	1.00

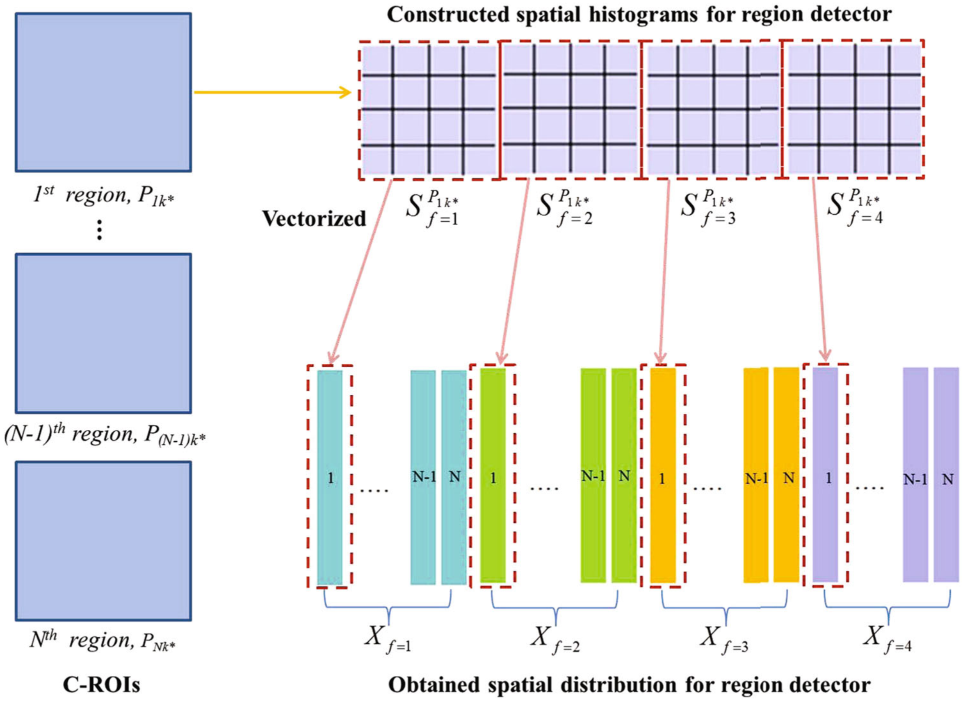

Fig. 6 Illustration of constructing spatial distributions to apply NMF for FR

aim to find the most informative region called FR in the C-ROI (Algorithm 3).

Algorithm 3 Extraction of FRs in a class

```
Input: $N$ C-ROIs, $P_{ik^*}$, $i = 1, ..., N$, in a class $c$
Output: $N$ FRs, $Y_i$ on the $P_{ik^*}$, $i = 1, ..., N$
for $f = 1$ to $N_f$ do
    For each region detector $f$,
    1. Construct spatial distribution $X_f^c$.
    2. Apply NMF to the $X_f^c$ using eq.(8).
    3. Obtain semantic spatial histogram $U_f^c$.
end for
Obtain an activation map $A^c$ using eq.(9) and thresholding.
For each $P_{ik^*}$, activate a FR $Y_i$.
```

Toward this goal, the algorithm follows four steps. Given N training images of a class c, the first step is to get the N_R-dimensional spatial histograms $S_f^{P_{ik^*}}$, $i = 1 \ldots N$, for region detector f from C-ROIs in N images (Sect. 3.2.2), and obtain a $N_R \times N$ matrix X_f^c by putting them in a matrix form (Fig. 6), that is, $X_f^c = [S_f^{P_{1k^*}} S_f^{P_{2k^*}} \ldots S_f^{P_{Nk^*}}]$.

The second step is to obtain the semantic spatial distribution by applying nonnegative matrix factorization (NMF) [15] to X_f^c. NMF is known to be able to learn parts or semantic information of some content. NMF determines a 2-factor decomposition:

$$X_f^c \approx U_f^c V_f^c, \tag{8}$$

where U_f^c is an $N_R \times K$ matrix that contains K bases (in our case, $K = 1$[1]), and V_f^c is a $K \times N$ matrix that contains K weights for each basis.

The third step is to get the activation map A^c by summing U_f^c weighted with the class-specific weights w_f^c:

$$A^c = \sum_{f=1}^{N_f} w_f^c U_f^c. \tag{9}$$

The final step is to get the FR from activation map A^c by thresholding. The FR is fixed relative to C-ROI for all images of same class. As examples, we present the FRs detected in C-ROIs for two classes of image (Fig. 7). The detected FRs are almost the same regions that can be designated by human intuition.

3.4 Class-specific image representation

The C-ROI and the FR described so far can be used to construct the class-specific image representation. To compute an image-level descriptor for an image, we define spatial pooling regions, then concatenate encodings of each spatial region. Here, the encoding of each spatial region is the BoW representation with denseSIFT features of the spatial region. As in the traditional SPM where the spatial pooling is done in a spatial pyramid fashion (1×1, 2×2 and 4×4 grids), we also construct 3-level spatial pooling structure as:

- Level 1: The whole image is used as a spatial pooling region (1×1 grid).
- Level 2: Class-specific spatial pooling is designed. Unlike SPM pooling, the proposed method uses four regions divided differently (Fig. 1), i.e., Region 1 of FR, Region 2 of C-ROI, Regions 3 and 4 on the remaining area that correspond to background. Here, Regions 3 and 4 are partitioned horizontally as shown in Fig. 1c; the properties of background are generally changed to the horizontal direction.
- Level 3: Regular 4×4 grids are constructed on the C-ROI of the class so that we can obtain more detailed information on the object of interest.

3.4.1 *Classifier learning*

For learning a classifier, we use the PEGASOS SVM [30] as a linear SVM solver. To use non-linear additive kernels instead of the linear kernel, we use the χ^2 explicit feature map [33]. The regularization-loss trade-off parameter C of the SVM is set to 10. For a specific class, training images are divided into two groups, positive (training images of the specific class) and negative (all training images of the other classes). Then, a 1-vs-rest classifier is trained with the training data for the specific class. For multi-class image classification, a 1-vs-rest classifier has to be prepared for each class so that the number of classifiers is the same as the number of classes.

3.4.2 *Testing*

To test an image i, candidate C-ROIs are extracted first. To select a C-ROI for a class c among them, equation (2) is modified to

$$Z_{ik}^c = Z(P_{ik}, P_{jl^*}^c), \tag{10}$$

where $P_{jl^*}^c$ denotes the C-ROI (not candidate C-ROI) of training image j of class c, which were obtained in the training stage. Equations (3)–(5) can be used without modification. Notice that C-ROI is extracted differently depending on the class to test. FR in the extracted C-ROI can be obtained using the activation map A^c of the class to test, which is fixed for each class. Using these two regions yields a class-specific image representation of the class, and which we use to evaluate over the c classifier. Finally, if the evaluation of all classes is finished, the test image is

[1] We used only one basis because in experiments we observed that one basis is sufficient to represent the particular information of the given data.

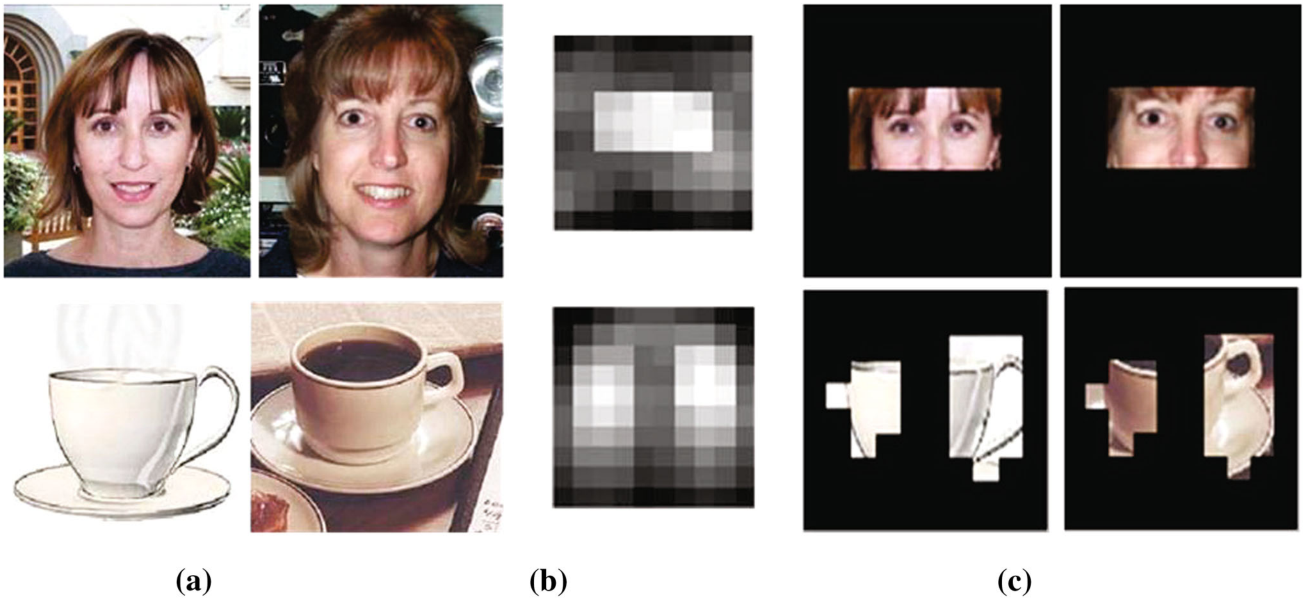

Fig. 7 Results of activation map and FRs obtained from C-ROIs of *face* and *cup* classes. **a** C-ROIs, **b** activation map, **c** FRs

Table 2 Descriptions of the four datasets

Datasets	Caltech-4	Caltech-101	CMU faces	Scene 15
Total # of images	2236	9144	624	4485
# of classes	4	101 (selected 36)	2	15
# of images / class	100–400	31–800	311 and 313	200–400
# of training images / class	Half size [6, 9]	15 or 30	254 [26]	100
# of testing images / class	Half size [6, 9]	~50	370 [26]	The rest

classified into the class c^* that has the maximum score using the equation:

$$c^* = \arg\max_{c \in C} S_i^c, \tag{11}$$

where S_i^c is a score obtained from the c classifier for the test image i and C is the set of classes to test.

4 Experiments

4.1 Datasets

We evaluated our proposed method on the Caltech-4[2], the Caltech-101[3], the CMU Faces[4], and the Scene 15[5] benchmark datasets (Table 2). The Caltech-4 contains four classes of images with large variation in object size and location; the Caltech-101 contains 101 classes of images with even larger variation in object size and location than in the Caltech-4, and with additional large variation in object pose. Among 101 classes in the Caltech-101 dataset, we selected only 36 classes that do not have large variation in object pose, because our approach is based on the assumption that the spatial distribution of region detector responses would be similar over images of a same class. The CMU Faces dataset was of special interest. The goal was to classify images according to whether or not the faces wore sunglasses; this task seemed suitable to demonstrate the power of the FR proposed in this paper for classification. The Scene 15 dataset was included to show the ability of our algorithm to capture similar parts in the scenes as C-ROIs even though unlike other datasets these scenes do not contain objects apparent for classification.

4.2 Implementation details

We used a single descriptor, denseSIFT [14]. The SIFT descriptors extracted from 16 × 16 pixel patches were densely sampled from each image on a grid with step size of 8 pixels. The images were all processed on gray scale. We used k-means to learn a codebook of size 1024, and assigned the SIFT features to the nearest codebook vector

[2] http://www.robots.ox.ac.uk/vgg/data3.html

[3] http://www.vision.caltech.edu/Image_Datasets/Caltech101

[4] http://archive.ics.uci.edu/ml/datasets/CMU+Face+Images

[5] http://www-cvr.ai.uiuc.edu/ponce_grp/data

Fig. 8 Example images for some classes of four datasets

Table 3 Relative importance of detectors for some classes (in Fig. 8) in the datasets of (a) Caltech-4, (b) Caltech-101, (c) CMU Faces and (d) Scene 15

Classes	Dataset	DoG	Harris-lap	Hessian-lap	Salient
Leopard	(a)	0.82	0.84	1.00	0.72
Airplane	(a)	0.93	1.00	0.99	0.53
Stop-sign	(b)	0.74	0.92	0.65	1.00
Sunflower	(b)	0.82	0.99	0.98	1.00
Faces (sunglasses)	(c)	0.62	0.99	1.00	0.47
MITallbuilding	(d)	0.84	0.87	0.99	1.00
Kitchen	(d)	0.81	1.00	0.86	0.98
Livingroom	(d)	0.76	0.97	0.94	1.00

(hard assignment). We used the VLFeat library [32] for SIFT and k-means computation. For all datasets, we set N_R to 10 × 10 grids for spatial histogram computation[6]. For the Caltech-4, Caltech-101, CMU Faces datasets, only two levels (1 and 2) were used for the spatial pooling. For the Scene 15 dataset, all three levels were used for this task. For comparison of classification accuracy, we used SPM as baseline with linear SVM, which was obtained from the Liblinear [4] library.

4.3 Effects of class-specific combination of region detectors

The relative importance (or weights) w_f^c of region detectors computed for some classes of four datasets (Fig. 8; Table 3) shows which region detectors contribute most to extracting C-ROIs from the class. For example, the Hessian-Laplace detector is known to be good at extracting blob-like structure, and showed the highest weights for the classes of leopards and faces with sunglasses which have such blob-like structures. The high weights for these patterns seemed to lead to extraction of FRs that include them.

Table 4 Classification rate (%) on (a) Caltech-4, (b) Caltech-101, (c) CMU faces and (d) Scene 15 datasets

Methods (Ours)	(a)	(b)	(c)	(d)
Only DoG	98.71	78.42	84.59	82.87
Only Harris-Laplace	99.12	77.83	89.19	83.05
Only Hessian-Laplace	99.17	76.71	90.54	83.67
Only salient	98.71	79.56	82.71	82.16
w/o weight	99.42	80.59	90.81	83.81
With weight	99.64	81.22	91.89	84.21

In this section, we show the effects of class-specific combination of region detectors for image classification.

[6] In our experiments, the number of sub-regions has little effect on the classification performance.

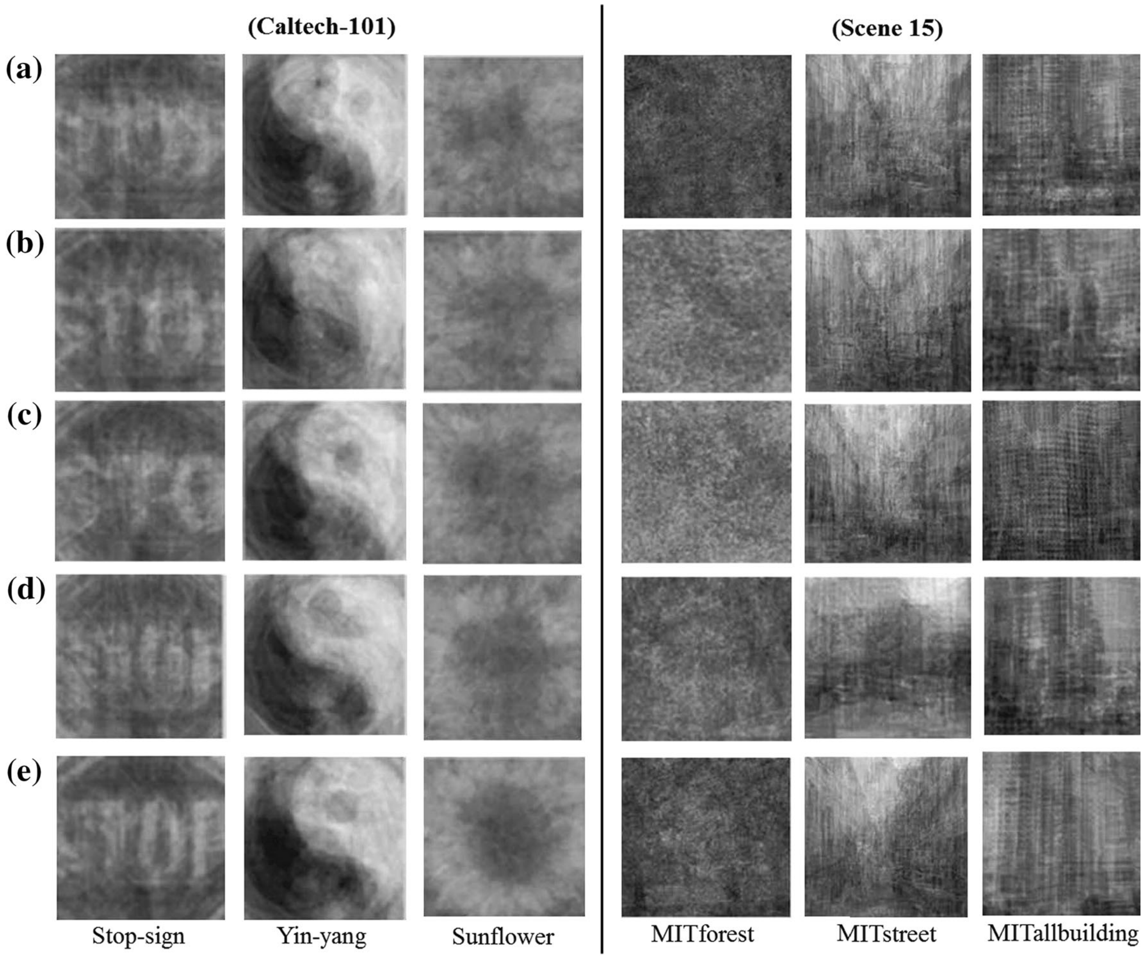

Fig. 9 Average images of selected C-ROIs for some classes from Caltech-101 and Scene 15 datasets; *each row* represents average images of obtained C-ROIs using **a** only DoG region detector, **b** only Harris-Laplace region detector, **c** only Hessian-Laplace region detector, **d** only salient region detector, and **e** multiple region detectors (in proposed method)

To do this, we applied only one region detector to our work for image classification for the four datasets (Table 4). Combining the four detectors always gave results better than any single region detector. This result is very important in that if only one region detector is allowed, we must choose the best one; this choice depends on the data to classify, and is neither easy nor intuitive.

We extracted average images of C-ROIs extracted from some classes of the Caltech-101 and the Scene 15 datasets (Fig. 9). The average images obtained using the combination of region detectors showed clearer boundary of objects than the images obtained using any single region detector. This means that the combination of region detectors localizes the class-specific region-of-interest in images better than does a single detector.

4.4 C-ROI and FR

For qualitative results, we implemented extraction of C-ROIs (Fig. 10) and FRs (Fig. 11) on four datasets in Sect. 4.1. If objects in a given class had similar shapes, the C-ROIs were well extracted regardless of size and location of the objects. The activated FRs in C-ROIs of some classes contained key components of them, i.e., the spines on the hedgehog, the spot pattern on the leopard, the highway sign, and the sunglasses on the face.

To check the effect of two proposed class-specific regions for image classification, we evaluated the classification accuracy using only these two regions over the CMU Faces and the Scene 15 datasets; these two datasets are specially designed to classify specific conditions or scenes without objects apparent for classification, and are therefore suitable to evaluate the power of the C-ROI and the FR for classification. For feature extraction, we defined four different spatial pooling regions: the whole image as baseline; only C-ROI; only FR; C-ROI and FR, instead of 3-level spatial pooling structure (in Sect. 3.4). Classification accuracy for the four different spatial pooling regions is listed in Table 5. For the CMU Faces, the proposed C-ROI or FR (Figs. 10, 11c) was significantly better than the whole

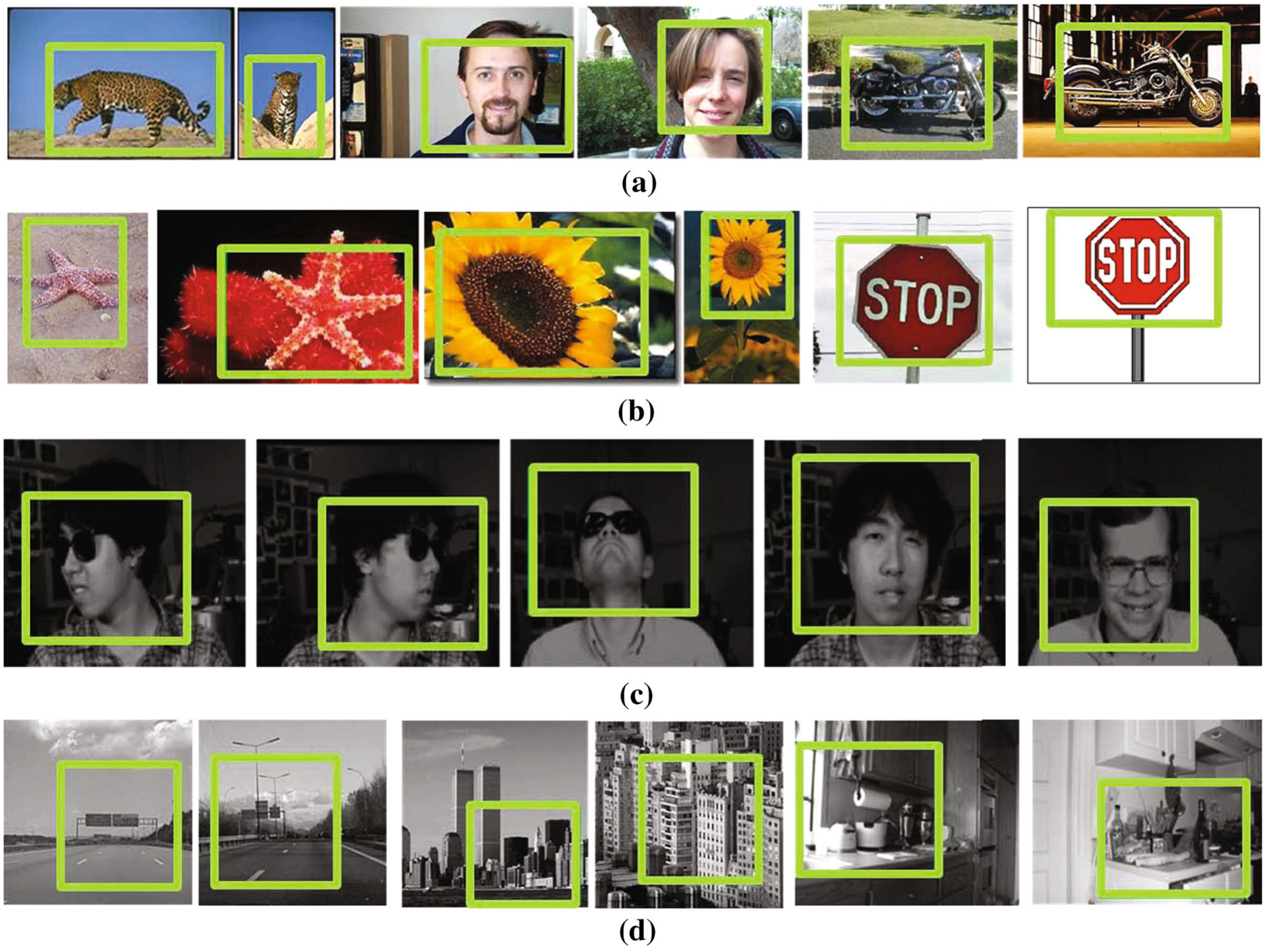

Fig. 10 Examples of extracted C-ROIs for some classes of **a** Caltech-4, **b** Caltech-101, **c** CMU Faces and **d** Scene 15 datasets

image region. This result seems reasonable because the C-ROI or FR captures the region that gives the information about the presence of sunglasses. Similarly, for the Scene 15, the proposed C-ROI or FR showed much better results than the whole image region. To check classes that showed quite good classification accuracy with the proposed C-ROI or FR, we separately evaluated the classification accuracies for each class in the Scene 15, and listed classes with largest improvements over the accuracy of the baseline that considers the whole image region (Fig. 12). Characteristics of some scenes were better described using C-ROI or FR than using the baseline. For both datasets, using both C-ROI and FR for classification gave better results than using only one C-ROI or FR; this result means that both proposed regions play an important role in classification.

C-ROI and FR extraction sometimes failed (Fig. 13). Failures usually occurred in classes that do not have a common structure for the same class (e.g., the beds from various viewpoints in bedroom class). Because we assumed that images in the same class should have similar structures or objects, classification accuracy would be degraded when the assumption was not satisfied. Relaxing this assumption for the condition of classes will be our future work.

4.5 Comparison with existing methods

Image classification results were obtained from the four datasets (Tables 6, 7, 8, 9). In these experiments, we compared our result with SPM as baseline and with other existing methods for both cases of without and with weights of region detectors over each category.

For the Caltech-4 dataset (Table 6), our method showed almost 100 % classification accuracy. Simple SPM showed better accuracy than other methods [6, 9, 26] which were developed to solve problems involved in classifying images in which object location and size vary.

For the Caltech-101 dataset (Table 7), our method showed classification accuracy comparable to the extended versions of SPM [34, 36][7]which adopt efficient encoding

[7] To compare classification results for selected 36 classes, we used codes provided from [34, 36].

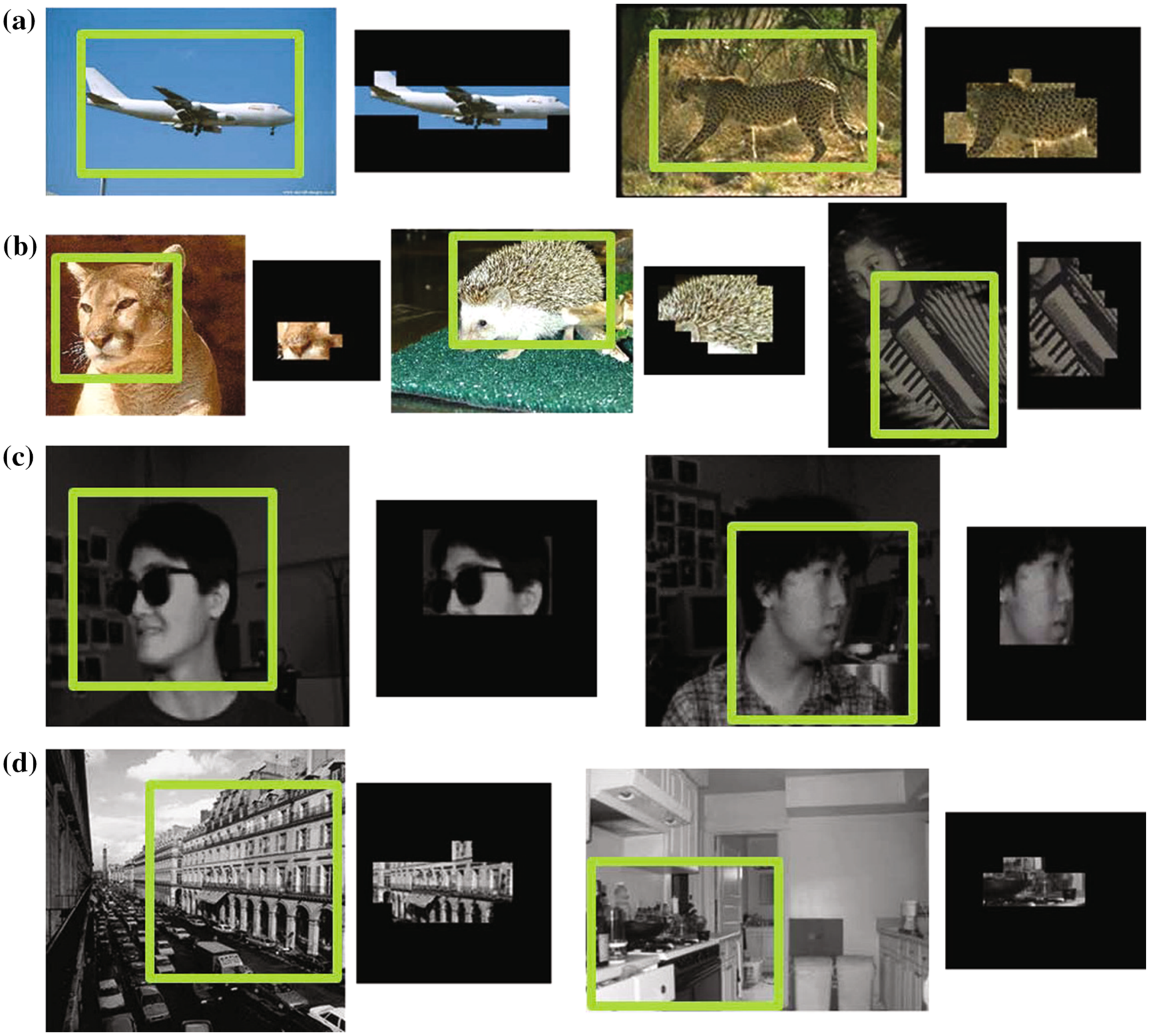

Fig. 11 Activated FRs on C-ROIs for some classes of **a** Caltech-4, **b** Caltech-101, **c** CMU Faces and **d** Scene 15 datasets

Table 5 Classification rate (%) on (a) CMU faces and (b) scene 15 datasets

Methods	(a)	(b)
The whole image (baseline)	81.08	77.94
Only C-ROI	90.77	80.27
Only FR	90.54	82.52
C-ROI and FR	92.97	83.03

techniques although our method uses hard assignment for encoding. Our method showed much more improvement over SPM in this dataset than in the Caltech-4 dataset because variation in object size and location in images is greater in the Caltech-101 dataset than in the Caltech-4 dataset.

For the CMU Faces dataset (Table 8), our result achieved as much as 8.7 % higher classification accuracy than the SPM result; this improvement over the SPM result was larger than achieved in the Caltech-4 dataset and the Caltech-101 dataset, and seemed to be achieved mainly due to our method's ability to detect the most discriminative region (i.e., FR) in the C-ROI. Actually, we obtained the best accuracy (92.97 %, Table 5) when we used only C-ROI and FR without background information. This means that the background information of this dataset disturbs rather than assists image classification. Our method gave results better than Nguyen's method [26]

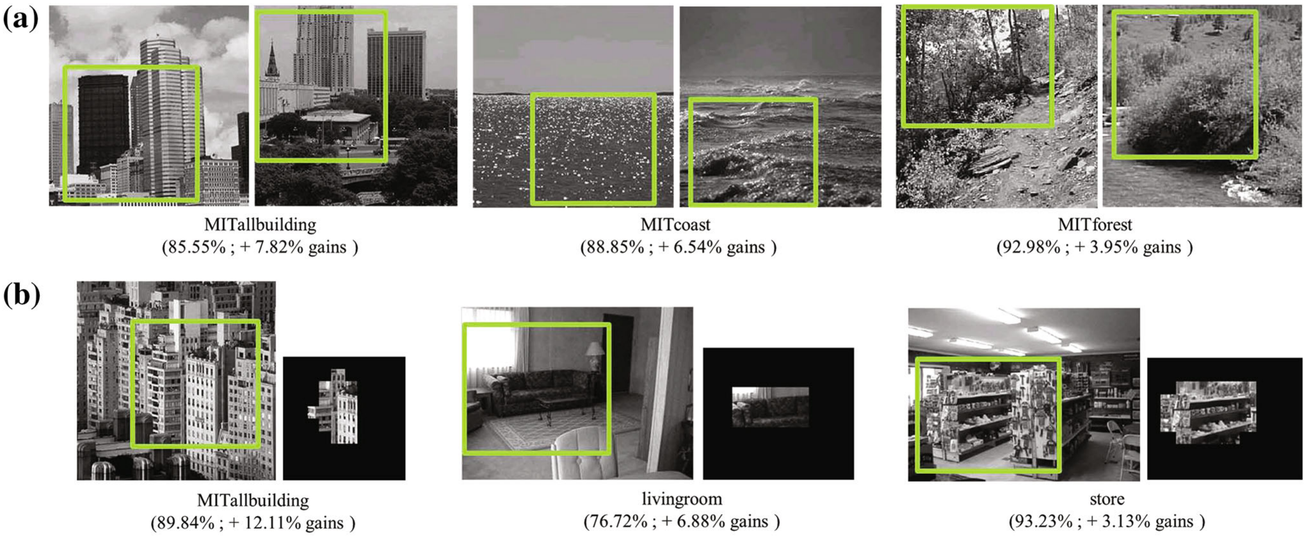
(a)
MITallbuilding
(85.55% ; + 7.82% gains)
MITcoast
(88.85% ; + 6.54% gains)
MITforest
(92.98% ; + 3.95% gains)
(b)
MITallbuilding
(89.84% ; + 12.11% gains)
livingroom
(76.72% ; + 6.88% gains)
store
(93.23% ; + 3.13% gains)

Fig. 12 Example of classes with largest gains obtained by **a** C-ROI and **b** activated FR on C-ROI for the Scene 15 dataset; the classification accuracy are listed below sample images with gains obtained from baseline

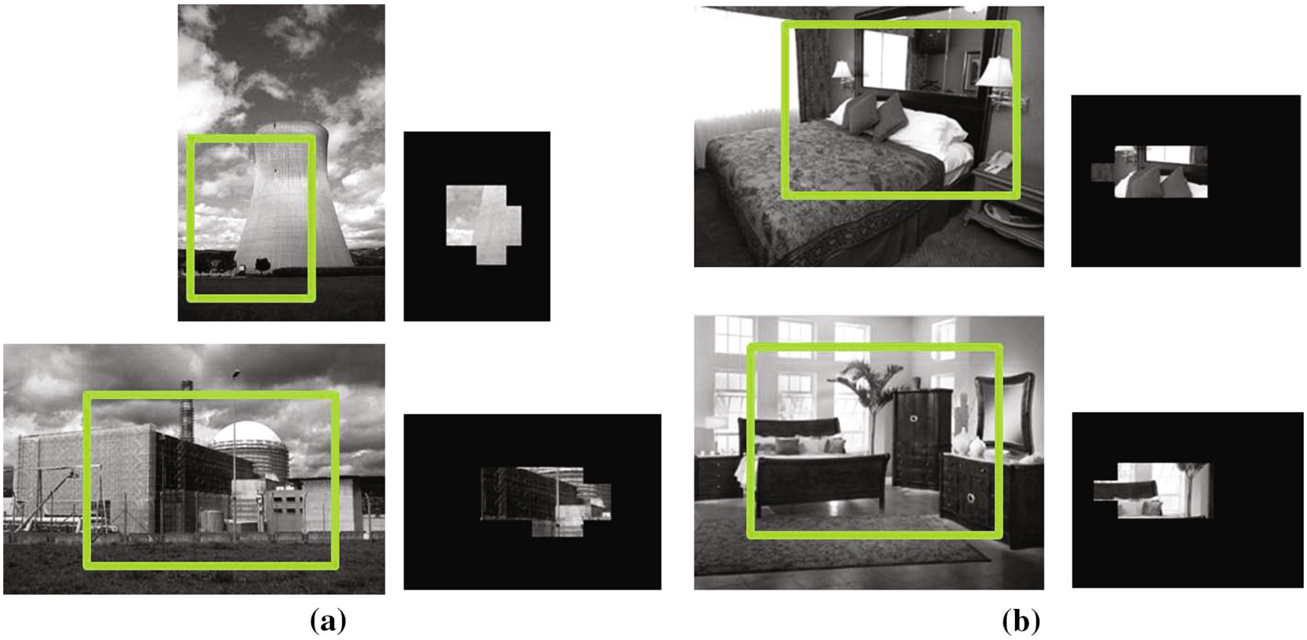

Fig. 13 Examples of failure cases for C-ROI (*left*) and FR (*right*) extraction; **a** failure case 1—*industrial* class and **b** failure case 2—*bedroom* class

which is similar to ours in that automatically localize the subwindows that are most discriminative for classification.

For the Scene 15 dataset (Table 9), our method achieved higher classification accuracy than SPM and the extended versions of SPM [34], and achieved comparable accuracy with state-of-the-art method [31] which uses the discriminative spatial saliency as the interest region for classification task.

For all datasets, the use of only one region detector (Table 4) usually gave results better than did SPM; this observation implies that the proposed representation using C-ROI and FR contributed to solve the problems caused by varying size and location of objects in images.

5 Conclusion

We proposed a new method to construct a class-specific representation that is better than SPM for classification of images with large variation of object size and location in

Table 6 Classification rate (%) on Caltech-4 dataset

Methods	Classification rate (%)
Fergus et al. [6]	92.28
Galleguillos et al. [9]	96.75
Nguyen [26]	92.93
SPM	98.37
Ours (w/o weight)	**99.42**
Ours (with weight)	**99.64**

Bold values represent the best classification performance for given experiments

Table 7 Classification rate (%) on 36 classes from Caltech-101

Methods	15 Training images	30 Training images
SPM	71.02	77.59
Yang et al. [36]	75.90	**81.66**
Wang et al. [34]	74.24	**81.91**
Ours (w/o weight)	75.10	80.59
Ours (with weight)	76.17	81.22

Bold values represent the best classification performance for given experiments

Table 8 Classification rate (%) on CMU faces dataset

Methods	Classification rate (%)
Nister and stewenius [27]	80.11
Lampert et al. [13]	86.79
Nguyen [26]	90.00
SPM	83.24
Ours (w/o weight)	**90.81**
Ours (with weight)	**91.89**

Bold values represent the best classification performance for given experiments

Table 9 Classification rate (%) on scene 15 dataset

Methods	Classification rate (%)
SPM	81.40
Wang et al. [34]	82.34
Sharma et al. [31]	**84.60**
Ours (w/o weight)	83.81
Ours (with weight)	**84.21**

Bold values represent the best classification performance for given experiments

images. To obtain good classification accuracy despite these variations, we proposed two kinds of region, called class-specific region-of-interest (C-ROI) and focal region (FR). The C-ROI is the region that is common in images of same class; the FR is the region that is most discriminative in the C-ROI. To extract those two regions, we used the DoG, Harris-Laplace, Hessian-Laplace, and salient multiple scale-invariant region detectors. Image representation using these two regions gave better classification results for several well-known datasets than did SPM. In future, this concept could be extended to find the best combination of macro-features to describe object classes.

References

1. Chai Y, Lempitsky V, Zisserman A (2011) Bicos: a bi-level co-segmentation method for image classification. In: Proceedings of the international conference on computer vision, pp 2579–2586
2. Cheng H, Wang R (2010) Semantic modeling of natural scenes based on contextual bayesian networks. Pattern Recognit 43(12):4042–4054
3. Cheng Y (1995) Mean shift, mode seeking, and clustering. IEEE Trans Pattern Anal Mach Intell 17(8):790–799
4. Fan RE, Chang KW, Hsieh CJ, Wang XR, Lin CJ (2008) Liblinear: a library for large linear classification. J Mach Learn Res 9(4):1871–1874
5. Felzenszwalb PF, Huttenlocher DP (2005) Pictorial structures for object recognition. Int J Comput Vis 61(1):55–79
6. Fergus R, Perona P, Zisserman A (2003) Object class recognition by unsupervised scale-invariant learning. In: Proceedings of IEEE conference on computer vision and pattern recognition, pp 264–271
7. Fergus R, Perona P, Zisserman A (2005) A sparse object category model for efficient learning and exhaustive recognition. In: Proceedings of IEEE conference on computer vision and pattern recognition, pp 380–387
8. Fischler MA, Elschlager RA (1973) The representation and matching of pictorial structures. IEEE Trans Comput 22(1):67–92
9. Galleguillos C, Babenko B, Rabinovich A, Belongie SJ (2008) Weakly supervised object localization with stable segmentations. In: Proceedings of the European conference on computer vision, pp 193–207
10. Gao S, Cheng X, Chia LT (2010) Discovering class-specific informative patches and its application in landmark characterization. Proc Int Multimed Model Conf 5916:218–228
11. Jiang H, Wang J, Yuan Z, Wu Y, Zheng N, Li S (2013) Salient object detection: a discriminative regional feature integration approach. In: Proceedings of IEEE conference on computer vision and pattern recognition (CVPR), pp 2083–2090
12. Kadir T, Brady M (2001) Scale, saliency and image description. Int J Comput Vis 45(2):83–105
13. Lampert CH, Blaschko MB, Hofmann T (2008) Beyond sliding windows: object localization by efficient subwindow search. In: Proceedings of IEEE conference on computer vision and pattern recognition, pp 1–8
14. Lazebnik S, Schmid C, Ponce J (2006) Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In: Proceedings of IEEE conference on computer vision and pattern recognition, pp 2169–2178
15. Lee DD, Seung HS (1999) Learning the parts of objects using non-negative matrix factorization. Nature 401(6755):788–791
16. Li FF, Perona P (2005) A bayesian hierarchical model for learning natural scene categories. In: Proceedings of IEEE conference on computer vision and pattern recognition, pp 524–531
17. Li Z, Liu J, Lu H (2010) Sparse constraint nearest neighbour selection in cross-media retrieval. In: ICIP, pp 1465–1468

18. Li Z, Liu J, Lu H (2013) Structure preserving non-negative matrix factorization for dimensionality reduction. Comput Vis Image Underst 117(9):1175–1189
19. Li Z, Liu J, Tang J, Lu H (2014) Projective matrix factorization with unified embedding for social image tagging. Comput Vis Image Underst 124:71–78
20. Li Z, Liu J, Yang Y, Zhou X, Lu H (2014) Clustering-guided sparse structural learning for unsupervised feature selection. IEEE Trans Knowl Data Eng 26(9):2138–2150
21. Li Z, Yang Y, Liu J, Zhou X, Lu H (2012) Unsupervised feature selection using nonnegative spectral analysis. In: Proceedings of the twenty-sixth AAAI conference on artificial intelligence, July 22–26, 2012, Toronto, Ontario, Canada
22. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vis 60(2):91–110
23. Mikolajczyk K, Leibe B, Schiele B (2005) Local features for object class recognition. In: Proceedings of the international conference on computer vision, pp 1792–1799
24. Mikolajczyk K, Schmid C (2004) Scale and affine invariant interest point detectors. Int J Comput Vis 60(1):63–86
25. Mikolajczyk K, Tuytelaars T, Schmid C, Zisserman A, Matas J, Schaffalitzky F, Kadir T, Gool LV (2005) A comparison of affine region detectors. Int J Comput Vis 65(1–2):43–72
26. Nguyen M (2012) Segment-based svms for time series analysis. Ph.D. thesis, The Robotics Institute, Carnegie Mellon University, Pittsburgh, PA
27. Nister D, Stewenius H (2006) Scalable recognition with a vocabulary tree. In: Proceedings of IEEE conference on computer vision and pattern recognition, pp 2161–2168
28. Perronnin F, Sánchez J, Mensink T (2010) Improving the fisher kernel for large-scale image classification. Proc Eur Conf Comput Vis 6314:143–156
29. Serrano N, Savakis A, Luo J (2004) Improved scene classification using efficient low-level features and semantic cues. Pattern Recognit 37(9):1773–1784
30. Shalev-Shwartz, S., Singer, Y., Srebro, N.: Pegasos: Primal estimated sub-gradient solver for svm. In: Proceedings of the international conference on machine learning, pp 807–814
31. Sharma G (2012) Discriminative spatial saliency for image classification. In: Proceedings of IEEE conference on computer vision and pattern recognition (CVPR), pp 3506–3513
32. Vedaldi A, Fulkerson B (2010) Vlfeat: an open and portable library of computer vision algorithms. In: Proceedings of the international conference on Multimedia, pp 1469–1472
33. Vedaldi A, Zisserman A (2012) Efficient additive kernels via explicit feature maps. IEEE Trans Pattern Anal Mach Intell 34(3):480–492
34. Wang J, Yang J, Yu K, Lv F, Huang T, Gong Y (2010) Locality-constrained linear coding for image classification. In: Proceedings of IEEE conference on computer vision and pattern recognition, pp 3360–3367
35. Yakhnenko O, Verbeek J, Schmid C (2011) Region-based image classification with a latent svm model. Tech Rep RR-7665, INRIA. http://hal.inria.fr/inria-00605344
36. Yang J, Yu K, Gong Y, Huang T (2009) Linear spatial pyramid matching using sparse coding for image classification. In: Proceedings of IEEE conference on computer vision and pattern recognition, pp 1794–1801