Multimed Tools Appl (2014) 69:897–920
DOI 10.1007/s11042-012-1151-8

# HWVP: hierarchical wavelet packet descriptors and their applications in scene categorization and semantic concept retrieval

**Xueming Qian · Danping Guo · Xingsong Hou · Zhi Li · Huan Wang · Guizhong Liu · Zhe Wang**

Published online: 19 June 2012


**Abstract** Wavelet packet transform is an effective texture analysis approach by sub-band filtering. Different texture patterns have distinctive responses to the sub-bands of wavelet packets. The responses are valuable for texture description. Utilizing all the responses of the sub-bands of different resolutions can improve texture pattern discrimination power. In this paper, effective texture descriptors based on hierarchical wavelet packet (HWVP) transform are proposed. The subtle sub-bands of wavelet packet transform improve the discrimination power of HWVP descriptors for the images in different categories. Scene categorization performances of the HWVP descriptors under various decomposition levels and wavelet bases are discussed. Performances of HWVP descriptors of global and local images with different partition patterns are also analyzed. The advantages of HWVP descriptors attribute to the following two aspects. Firstly sub-band filtering is helpful for improving the discrimination power of HWVP descriptors to capture the subtle differences of texture patterns. Secondly hierarchical feature representation makes the HWVP descriptors robust to resolution variations. Comparisons are made with some existing robust descriptors on scene categorization and semantic concept retrieval. Experimental results on the widely used OT, Scene-13, Sport Event, and TRECVID 2007 datasets show the effectiveness of the proposed HWVP descriptors.

**Keywords** Scene categorization · Wavelet packet · TRECVID · Concept retrieval · SVM

## 1 Introduction

Image sharing websites can easily gather huge amount of images by worldwide users. Usually text based image retrieval (TBIR) [30] and content-based image retrieval (CBIR)

This work is supported in part by the National Natural Science Foundation of China (NSFC) Project No.60903121,No.61173109, and Foundations of Microsoft Research Asia

X. Qian (✉) · D. Guo · X. Hou · Z. Li · H. Wang · G. Liu · Z. Wang
School of Electronics and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China
e-mail: qianxm@mail.xjtu.edu.cn

X. Hou
e-mail: houxs@mail.xjtu.edu.cn

[4, 32, 41] can be utilized for users to access the image sharing websites. TBIR is performed by matching the query texts and the auxiliary descriptive texts of the images [30]. Image searching performance is inevitably influenced by the subjectivity, in-completeness and ambiguity of the user annotated texts. CBIR is carried by measuring the similarities of the low level visual features. Due to the semantic gaps, the performances of TBIR and CBIR are far from satisfactory. Textural, visual features, multimodal internet source information [4] and users' relevant feedbacks [32] are often fused to bridge the semantic gaps. There are two important ways to utilized by researchers. The first one is focused on effective feature representation and the other is on learning robust models. Usually, learning a high-performance classifier is very difficult than a set of weak classifiers. Thus, multi-model fusion based approaches often adopted to minimize the semantic gaps in image retrieval [3, 4, 25, 32, 39, 41]. In this paper, we focus on the first problem by proposing effective feature descriptors.

How to extract robust and effective features are very important for image representation. Recently, SIFT feature is proved to be very effective to represent scale and transform invariant features [20, 21]. It is effective to extract local feature by detecting scale and transform invariant feature points. Thus the image feature is represented by the sparse feature points. Each feature point is expressed by a 128-dimensional directional edge histogram of a block with its sizes scale-related. SIFT is very efficient in representing images with salient structures. However, for complex image the sparse characteristics cannot be guaranteed. Thus the computational cost is very high. Most importantly the salient features are contaminated by the large number of un-important points. Thus, in this paper, we represent image feature by the responses of the image to a bank of filters to improve the feature discriminative power. Each filter has certain localized characteristics. Different objects or texture patterns have distinct response to the filters. These are important for object categorization [29]. To improve the discrimination power of feature descriptors, an image is filtered by a set of multi-resolution and multi-direction filters, which aims at decomposing different object into different sub-band of filters.

The main contributions of this paper are as follows: 1) in terms of the fact that different objects or texture patterns have different responses to the multi-resolution and multi-direction filters, we propose to improve feature discrimination power by multi-bands filtering. 2) Hierarchical feature representation approach collects the responses of the objects to the multi-resolution and multi-direction filters. It discloses the object characteristics from multi-resolutions and directions. 3) Systematically analyzed the relationship of transform kernels of wavelet packets and localized texture pattern, which is useful for transform kernel selection. 4) The impacts of wavelet packet bases, wavelet transform levels, and global/local representations of HWVP descriptors are systematically analyzed which providing some guidelines for descriptors selection. Comprehensive comparisons are made for HWVP descriptors with PHOG [2], GIST [34] and SPM [16] in the applications of scene categorization and semantic concept retrieval.

The rest of this paper is organized as follows. In Section 2, related works on scene categorization are briefly reviewed. In Section 3 the proposed HWVP are illustrated in detail. Applications of HWVP descriptors based scene categorization and semantic concept retrieval are given in Section 4. The test datasets are given in Section 5. Experimental results and discussions are given in Section 6. And finally conclusions are drawn in Section 7.

## 2 Related works

Scene categorization is one of the promising ways to bridge the semantic gaps in image retrieval [41]. Bag-of-Words (BOW) based scene categorization and semantic concept

retrieval approaches have been paid much attention by many researches [12, 18, 35, 37]. BOW models such as the probabilistic latent semantic analysis (pLSA) and latent Dirichlet allocation have been widely adopted [1, 12, 32, 41]. These approaches model scenes as geometric-free structures, which are represented by the spatial constraints of local patches. The BOW based approaches are robust to the illumination, occlusion, and scale variations. Probabilities of co-occurrence of visual words are also utilized in the BOW based scene categorization [35, 37].

Discriminative part-based models [13, 26] are effective in representing scenes with rigorous geometric structures by modeling the relationships between different parts. Usually each image is represented by a set of local patches. Each patch is represented by local descriptors which are robust to illumination, scale, orientation and transform variations [20]. In [18] and [8], the local patches of an image are assumed to be independent from each other. This assumption simplifies the computations for the ignorance of the spatial co-occurrences and dependences of local patches. However, one of the shortcomings of BOW models is that objects with different appearances may have similar statistics of visual words which tend to be confused during classification. Thus, some approaches model the co-occurrences, dependences or linkages of the salient parts for improving scene categorization performances [35]. Probabilities of co-occurrence of visual words are also taken into consideration in the training of BOW models [35, 37]. Despite of using the co-occurrences, the spatial relationships of the local patches can be modeled [6, 31, 36, 37]. Hierarchical Dirichlet process (HDP) is a nonparametric Bayesian model that infers latent themes from the training samples under the assumption that a hierarchical structure in different groups shares the same themes [33]. Extensions of the HDP have been proposed by modeling the relative spatial locations of local patches [31] and using dependent hierarchical Dirichlet process (DHDP) [35]. The DHDP is performed by introducing a linkage structure over the latent themes to encode the dependencies of the patches. The linkage enforces the semantic connections among the patches by facilitating better clustering of the themes. A visual language modeling method is utilized to incorporate the spatial context of the local appearance features into statistical language model [36]. The visual language models capture both co-occurrence and spatial proximity of local image features.

Statistical learning based methods are often utilized to improve object categorization performance by discovering the salient structures of objects [19]. Hence, the local appearance, shape and texture information are usually fused by generative and discriminative models to improve object categorization performances [3, 19, 39]. The spatial dependency between neighboring patches is modeled by a two-dimensional multi-resolution hidden Markov model [19]. Markov random fields [15] and conditional random fields [27] are adopted to model the dependencies of local patches. Statistical learning models maximize contextual constraints over the object labels and reduce the ambiguities during object categorization [27]. A generative model is utilized to determine object categories and carry out object segmentation in a unified framework [6]. Zhang et al. utilize support vector machines (SVM) classifiers to integrate BOW features for image classification [39]. Pyramid histogram of oriented gradients (PHOG) is good at representing the shapes and spatial layouts of objects [2]. SVM classifiers with spatial pyramid kernels are utilized to improve the object classification performance [2].

Except the BOW models, the spatial pyramids of local appearance and shape can capture the salient structures of objects, too [2, 16]. The effectiveness of the spatial pyramids has been shown in image categorization [2, 16, 23]. Pyramid histogram of oriented gradients (PHOG) is good at representing the shape information and spatial layouts of objects [2]. The spatial layout is obtained by partitioning an image into

non-overlapping grids with multiple resolutions. Local shape information of a grid is represented by a histogram of oriented gradients (HOG). PHOG descriptor of the image is a concatenation of all the HOG vectors over the grids at all the resolutions. Spatial pyramid matching descriptor (SPM) is also very effective [16]. In SPM visual vocabulary histogram in spatial pyramid domain is constructed. GIST feature is very effective, because localized Gabor transforms are carried out for each grid [34]. To speed up the computation, firstly each image is resized into 128×128 pixels. Secondly, the image partitioned into 4×4 grids and each grid is decomposed by a bank of multi-scale oriented filters [34].

Different scenes have salient structures, shapes and texture patterns which can be utilized for their categorization [24, 34]. The local appearances [39, 40] and shape information [3] have been shown their effectiveness in scene classification. However, texture information often plays assistant roles [3, 28, 38]. Different scenes usually have distinctive texture patterns. While image in the same category may have different visual appearances and arbitrary shapes, and images in different categories may have similar appearances and shapes. Sometimes texture patterns are the most discriminative. Different scenes have distinctive responses to different filters. Features represented in multi-resolution transform domain can improve their discriminative power. Features extracted in hierarchical wavelet packet transform domain inherit the advantages of multi-resolution transform and spatial pyramid representation [9, 24]. This paper is extended from our previous version [24]. The motivations of the multi-resolution and multi-direction filters valuable for scene categorization and retrieval are analyzed. More experimental results are given which is helpful to show the effectiveness of the HWVP descriptors. More discussions for the proposed HWVP descriptors are provided, which provide a guideline for selecting appropriate descriptor for scene categorization and semantic concept retrieval.

## 3 The proposed hierarchical wavelet packet texture descriptors

Wavelet packet analysis has been successfully used for data compression, texture classification [14] and face recognition [11]. Wavelet packet decomposition is performed by filtering the original signal with a set of sub-band filters. Images of different categories have different properties to the wavelet packet filters.

### 3.1 Wavelet packet transform

Wavelet packets consist of orthonormal and compactly supported wavelets. Wavelet packets represent a generalization of multi-resolution decomposition. Wavelet packets comprise the entire family of sub-band tree decompositions which permit the choice of any decomposition topological structures [14]. Figure 1a–d show a three level wavelet packet transform for an one dimensional (1-D) signal S with four different decomposition trees. Figure 1a is a regular wavelet packet tree. Figure 1b and c are two arbitrary wavelet packet trees. Figure 1d is a wavelet tree. Only the leftmost nodes (low-frequency sub-bands) have two children. In wavelet packet tree, the parent nodes also have two children for the nodes if the corresponding sub-bands are further decomposed. Wavelet packet transform constructs a tree-structured and multi-band extension of the wavelet transform. Wavelet packets are described by the collection of functions $\{W_i(x)|i \in Z^+\}$ as follows:

$$2^{\frac{s-1}{2}} W_{2n}\left(2^{\frac{s-1}{2}}x - t\right) = \sum_m h_{m-2t} 2^{\frac{s}{2}} W_{2n}(2^s x - m) \tag{1}$$

$$2^{\frac{s-1}{2}} W_{2n+1}\left(2^{s-1}x - t\right) = \sum_m g_{m-2t} 2^{\frac{s}{2}} W_{2n}(2^s x - m) \tag{2}$$

where $s$ and $t$ denote the scale and translation indexes, $W_0(x) = \phi(x)$, $W_1(x) = \psi(x)$, $\phi(x)$ is a scaling function and $\psi(x)$ is a basic wavelet. $h_k$ and $g_k$ are the low-pass and high-pass filters. The basis functions are obtained by changing scale and translation. The inverse relationship between wavelet packets of different scales is as follow:

$$2^{\frac{s}{2}} W_{2n}(2^s x - k) = \sum_t h_{k-2t} 2^{\frac{s-1}{2}} W_{2n}\left(2^{s-1}x - t\right) + \sum_t g_{k-2t} 2^{\frac{s-1}{2}} W_{2n+1}\left(2^{s-1}x - t\right) \tag{3}$$

### 3.2 Wavelet packet transform for 2-D image

The extension of wavelet packet transform from 1-D signal to 2-D image is straight forward by using separable 2-D wavelet packets. Generally, a low-pass filter (denoted $H$) and a high-pass filter (denoted $G$) are used. The convolutions of original image with the low pass filter results in an **approximation** and the convolutions with the high-pass filter in specific directions result in **details** [11]. In wavelet packet transform, the **approximation** and the **details** are further split into a second level of **approximation** and **details** respectively. For an $L$-level decomposition, the 2-D wavelet transform is carried out as follows:

$$F_L^{AA} = \left(Hx * (Hy * F_{L-1})_{\downarrow 2}\right)_{\downarrow 2} \tag{4}$$

$$F_L^{AD} = \left(Hx * (Gy * F_{L-1})_{\downarrow 2}\right)_{\downarrow 2} \tag{5}$$

$$F_L^{DA} = \left(Gx * (Hy * F_{L-1})_{\downarrow 2}\right)_{\downarrow 2} \tag{6}$$

$$F_L^{DD} = \left(Gx * (Gy * F_{L-1})_{\downarrow 2}\right)_{\downarrow 2} \tag{7}$$

where* denotes the convolution operator, ↓2denotes sub-sampling along the rows (or columns) and $F_{L\text{-}1}$ represents one of the sub-bands at level ($L$-1), $Hx$ and $Hy$ denote the separated low-pass filters of $H$ in x- and y- directions respectively, $Gx$ and $Gy$ denote the separated high-pass filters of $G$ in x- and y- directions respectively. $F_0$ is the original image. $F_L^{AA}$ is obtained by low pass filtering with sub-band image $F_{L\text{-}1}$. The **details** $F_L^{AD}$ , $F_L^{DA}$ and $F_L^{DD}$ are obtained by band-pass filtering in vertical, horizontal and diagonal sub-bands respectively.

### 3.3 Hierarchical wavelet packet texture descriptors

By hierarchical wavelet packet decomposition, the original image $F_0 = I(x, y)$ is thus represented by a complete set of sub-band images. There are $4^L$ sub-band images at level $L$ of the regular wavelet decomposition tree topology as shown in Fig. 1a. Figure 2 shows the sub-band of the regular wavelet packet transform under decomposition depth $L$=0, 1 and 2 respectively. The energy and its variations of all the sub-bands are collected and used for feature description. Hereinafter, we call the sub-band of the top-left corner of a decomposition level as **approximation** and the others as **details**. There are three and 15 **details** for wavelet packet transform under $L$=1 and $L$=2 as shown in Fig. 2b and c respectively.

In this paper, the mean and standard deviation of each sub-band are used for texture descriptions. Let $\mu_b^l$ and $\sigma_b^l$ denote the mean and standard deviation of a sub-band image $W_b^l$ under a specified decomposition level $l$ respectively, which are calculated as follows.

$$\mu_b^l = \frac{1}{S\times T}\sum_{s=1}^{S}\sum_{t=1}^{T}\left|W_b^l(s,t)\right|; \quad l = 0,\cdots,L\,;\, b = 1,\cdots,4^l \tag{8}$$

$$\sigma_b^l = \sqrt{\frac{1}{S\times T}\sum_{s=1}^{S}\sum_{t=1}^{T}\left(\left|W_b^l(s,t)\right| - \mu_b^l\right)^2}; \quad l = 0,\cdots,L\,;\, b = 1,\cdots,4^l \tag{9}$$

where $W_b^l(s,t)$ is the coefficient of the coordinate $(s,t)$ of the $b$-th sub-band of the $l$-th level, $S$ and $T$ are the height and width of the sub-band image $W_b^l$.

We denote the texture feature of each level $l$ as $\mathrm{x}^l = \left(\mu_1^l, \sigma_1^l, \cdots, \mu_{4^l}^l, \sigma_{4^l}^l\right)$ regular wavelet tree topology [24]. Four different HWVP are used and compared in this paper: 1) texture information of the sub-bands of the last level (denoted WVPK), 2) texture information of the details of WVPK (denoted NoDC), 3) texture information of all the details of the decomposition level $l$=1,⋯, $L$ (denoted HIGH), 4) texture information of all the sub-bands of the decomposition level $l$=0, 1,⋯, $L$ (denoted HWVP). Let $X^{\mathrm{WVPK}}(L)$, $X^{\mathrm{NoDC}}(L)$, $X^{\mathrm{HIGH}}(L)$ and $X^{\mathrm{HWVP}}(L)$ denote the hierarchical wavelet texture descriptors of WVPK, NoDC, HIGH and HWVP at decomposition level $L$. The corresponding dimensions and descriptions of NoDC, WVPK, HIGH and HWVP under different decomposition levels are shown in Table 1 respectively. For example, at the decomposition level $L$=4, the dimensions of the texture descriptors WVPK, NoDC, HIGH and HWVP are 512, 510, 672 and 682 respectively.

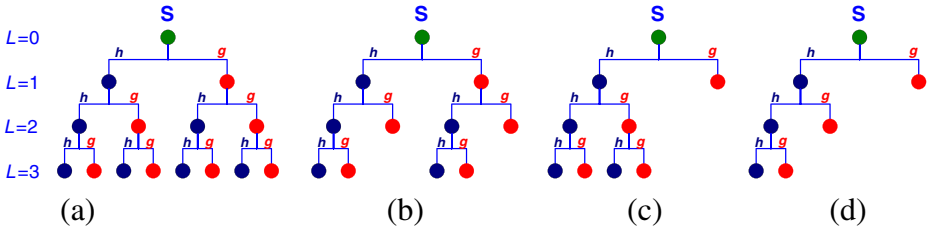
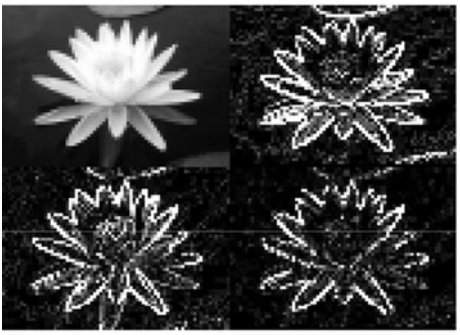
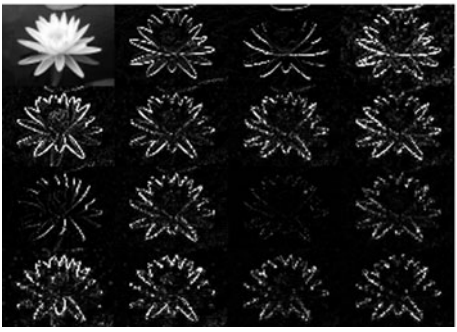



**Fig. 1** Four 1-D wavelet packet trees under decomposition depth $L$. **a** regular wavelet packet tree, **b** and **c** arbitrary wavelet packet tree, **d** wavelet tree. S is a 1-D signal, $h$ and $g$ are low-pass and high-pass filters respectively

(a) Original color image($L$=0)



(b) wavelet packet transform under $L$=1



(c) wavelet packet transform under $L$=2

**Fig. 2** Wavelet packet decomposition for an image under decomposition level $L$=0, $L$=1 and $L$=2. **a** Original color image($L$=0). **b** wavelet packet transform under $L$=1. **c** wavelet packet transform under $L$=2

### 3.4 Local HWVP descriptors

The local texture descriptors can capture the salient texture information and can decrease the influence of complex background. In this paper local and global HWVP are compared in scene categorization. Texture descriptors of four partitioning patterns are compared. The four patterns are Global, Local4, Local5 and Local9 which are shown in Fig. 3a–d respectively. In Local4 each image is equally partitioned into 2×2 grids. Local5 consists of Local4 and a

**Table 1** Descriptions for different hierarchical wavelet packet descriptors WVPK, NoDC, HIGH and HWVP under decomposition level $L$=1, $L$=2, $L$=3, and $L$=4

| | $L$=1 | $L$=2 | $L$=3 | $L$=4 | Descriptions of texture descriptors |
|---|---|---|---|---|---|
| WVPK | 8 | 32 | 128 | 512 | $X^{\mathrm{WVPK}}(L) = \left(\mu_1^L, \sigma_1^L, \cdots, \mu_{4^l}^L, \sigma_{4^l}^L\right)$ |
| NoDC | 6 | 30 | 126 | 510 | $X^{\mathrm{NoDC}}(L) = \left(\mu_2^L, \sigma_2^L, \cdots, \mu_{4^l}^L, \sigma_{4^l}^L\right)$ |
| HIGH | 6 | 36 | 162 | 672 | $X^{\mathrm{HIGH}}(L) = (X^{\mathrm{NoDC}}(1); \cdots; X^{\mathrm{NoDC}}(L))$ |
| HWVP | 10 | 42 | 170 | 682 | $X^{\mathrm{HWVP}}(L) = (X^{\mathrm{WVPK}}(0); \cdots; X^{\mathrm{WVPK}}(L))$ |

centralized sub-image, which is with the same sizes of the four grids in Local4. In the Local9, the original image is partitioned into 3×3 equal sized grids. Figure 4a, b and c show the local hierarchical wavelet packet transform of the texture descriptors of Local4 at decomposition level: $L$=0, 1 and 2 respectively. Similar to the global HWVP descriptors, the mean and standard deviation of each sub-band of the grids are collected for texture representation. Hence, in a specified decomposition level $L$, the dimensions of Local4, Local5 and Local9 are 4, 5, and 9 times of the corresponding global texture descriptors.

## 4 Applications of HWVP descriptors in scene categorization and semantic concept retrieval

In order to show the effectiveness of the proposed HWVP descriptors (including WVPK, NoDC, HIGH and HWVP), two kind of similarity measurement approaches are utilized in this paper. The first approach is based the feature similarity measurement [24]. This method is used for evaluating the impacts of decomposition levels, wavelet bases, and local/global patterns on HWVP descriptors. The second approach is SVM based scene categorization and semantic concept retrieval. This approach aims at making objective comparisons for HWVP descriptors with PHOG, SPM and GIST.

### 4.1 Feature similarity based scene categorization

This method consists of following two steps: 1) feature centroids determination for each category by training samples; 2) the category of a given test image determination according to the distances of the feature to the centroids of the categories.

Let $X_i^k$ denote the corresponding texture feature of the $i$-th image of the $k$-th category and $\overline{X^k}$ denote the feature centroid of a category obtained by $N$ ($N \geq 1$) training images per category as follow

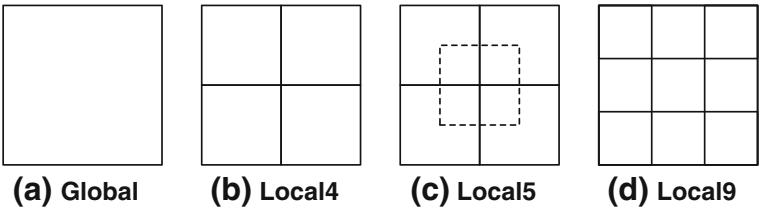



**Fig. 3** HWVP under different partitioning patterns. **a** Global (with no partitioning). **b** Local4 with 2×2 grids. **c** Local5 with four grids of Local4 and a centralized sub-image. **d** Local9 with 3×3 grids

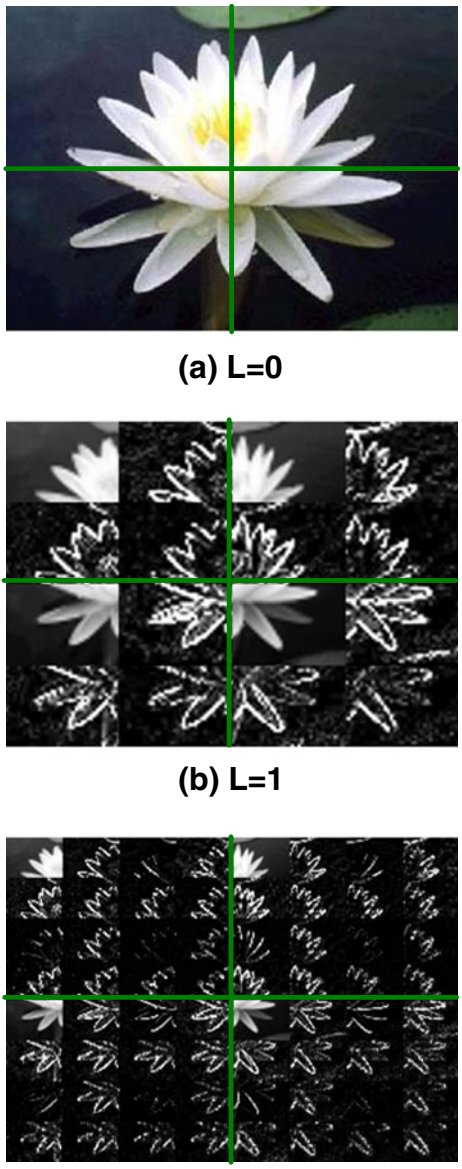



**Fig. 4** Local hierarchical wavelet packet transform for the sub-images of Local4 under decomposition level $L$=0, 1, and 2

$$\overline{X^k} = \frac{1}{N}\sum_{i=1}^{N} X_i^k; \quad k = 1, \cdots, K \tag{10}$$

where $K$ is category number. Each element $X_i^k(j)(j = 1, \cdots, d)$ is normalized as follows

$$X_i^k(j) = \frac{X_i^k(j) - MinX(j)}{MaxX(j) - MinX(j)}; \quad i \in \mathbf{I} \tag{11}$$

$$MaxX(j) = \max_{k=1,\cdots,K; i \in \mathbf{I}} \left\{ X_i^k(j) \right\} \tag{12}$$

$$MinX(j) = \min_{k=1,\cdots,K; i \in \mathbf{I}} \left\{ X_i^k(j) \right\} \tag{13}$$

where $\mathbf{I}$ is the set of all images, $d$ is the dimension of $X_i^k$ . For a given testing image with its feature $X$, we classify it into the $k_0$-th category according to the minimum distance as follow

$$k_0 = \arg\min_k Dist\left(X, \overline{X^k}\right) \tag{14}$$

where $Dist\left(X, \overline{X^k}\right)$ is the distance of $X$ and $\overline{X^k}$ . In this paper, the Euclidean distance is utilized which is calculated as follow

$$Dist\left(X, \overline{X^k}\right) = \frac{1}{d} \sum_{j=1}^{d} \left( X(j) - \overline{X^k}(j) \right)^2 \tag{15}$$

### 4.2 SVM based scene categorization and semantic concept retrieval

Scene categorization is to assign each test image to one of the pre-defined categories. In the SVM based scene categorization and semantic concept retrieval, we systematically compare the performances of SPM [16], PHOG [2], GIST [34] and HWVP descriptors. A brief description for the four features is shown in Table 2. The four features are all with high dimensions and all of them using the global features while catching some local information. For an $M$ class scene classification problem accurate classification is done by using $M$ one-versus-all SVM classifiers [7]. The kernel of the SVM is radical basis function (RBF). The

**Table 2** Descriptions of PHOG, SPM, GIST and HWVP descriptors utilized in scene categorization and semantic concept retrieval

| Features | Dimension | Description |
|---|---|---|
| PHOG | 850 | In extraction of **PHOG**, local shape is represented by a histogram of edge orientations which are quantized into K bins. The final PHOG descriptor for the image is a concatenation of all the HOG vectors over each grid. The PHOG descriptor of the entire image at level S is a vector with its dimension $K \times \left(\sum_{s=0}^{S} 4^s\right)$ . In this paper we set the level $S$=3 and $K$=10, then the histogram is a vector with 850 bins [2]. |
| SPM | 6300 | In extraction of **SPM** feature, local appearance features are quantized into V visual vocabularies. Then visual vocabulary histogram in spatial pyramid domain is constructed. In this paper, we set the spatial pyramid level $P$=2 and $V$=300, thus the dimension of SPM is 6300 [16]. |
| GIST | 640 | In extraction of **GIST** feature, 8 orientations and 5 scales are utilized. Finally, the magnitude of each filter is utilized for feature representation. Thus the dimension of GIST [34] of a gray-level image is 5×8×16=640. For a color image, the dimension 640×3=1920. |
| HWVP | 850 | The proposed HWVP under decomposition level $l$=3 and the partition pattern is Local5. The wavelet packet bases are db5. |

parameters of SVM are learned adaptively by using $R$-fold cross-validation. Firstly, randomly selected ($R$-1) folds are used for model training and the left one fold is used for testing. Then the optimal parameters obtained by cross-validation are used as the parameters of RBF kernel for the training of the SVM classifiers. In this paper, $R$ is set to be 5.

In scene categorization, the label of a test image is assigned as the label $k_0$ under the input feature $X$ as follow.

$$k_0 = \arg \max_{k=1,\cdots,M} f_k(X) \tag{16}$$

where $f_k(X)$ is the response of the $k$-th one-versus-all SVM which is calculated as follow

$$f_k(X) = \sum_{i=1}^{N_k} \alpha_k^i y_k^i K\left(X, S_k^i, \sigma_k\right) + b_k; k = 1, \cdots, M \tag{17}$$

where $a_k^i$ , $y_k^i$ , $S_k^i$ , and $b_k$ are the parameters of the $k$-th one-versus-all SVM. $N_k$ is the support vector number of the $k$-th SVM. $S_k^i$ is the $i$-th ($i$=1, ⋯, $N_k$) support vector of the $k$-th SVM ($k$=1, ⋯, $M$). $y_k^i$ , $a_k^i$ , and $b_k$ are the label index, weight and bias of the $i$-th support vector of the $k$-th SVM. The parameters of the SVM classifiers are trained using the images of the $k$-th category as positive samples and the images from the other $M$-1 categories as negative samples. The kernel function $K(X, Y, \sigma_k)$ of the $k$-th SVM is as follows

$$K(X, Y, \sigma_k) = \exp\left(-\|X - Y\|^2 / \left(2\sigma_k^2\right)\right); k = 1, \cdots, M \tag{18}$$

The SVM classifiers are with optimal parameters $\sigma_k$ which are learned during cross validation, Y is support vector, X is input feature. In the tasks of semantic concept retrieval, we sort the images in descending order according to the responses to the SVM classifier of the corresponding concepts. Then we can determine whether an image is belonging to the concept or not. Different from the scene categorization, in scene retrieval an image can be classified into many concepts.

## 5 Datasets

The proposed hierarchical wavelet packet descriptors are evaluated on four widely used datasets:

1) Oliva and Torralba dataset (denoted OT) [22]. This dataset has 2,688 images with eight categories: 360 **coast**, 328 **forest**, 374 **mountain**, 260 **highway**, 308 **insidecity**, 410 **open country**, 292 **street**, and 356 **tallbuilding**. Each image in this dataset are with the same sizes 256×256.
2) Scene-13 dataset [18]. This dataset consists of the 2,688 images of the eight categories of the OT dataset and another five categories with 1,071 images: 241 **suburb**, 174 **bedroom**, 151 **kitchen**, 289 **living room**, and 216 **office**. Totally there are 3,759 images in this dataset.
3) Sport event dataset [17]. This dataset contains 1,579 images of eight Sport event classes: 200 **badminton**, 137 **bocce**, 236 **croquet**, 182 **polo**, 194 **rock climbing**, 250 **rowing**, 190 **sailing**, and 190 **snowboarding**.
4) TRECVID 2007 sound and vision dataset for high level feature extraction task (we also call it semantic concept retrieval in this paper) in 2009. Totally about 50 h development

dataset is used for models training, and about 50 h are used for performance evaluation. This test data is obtained from news magazine, science news, news reports, documentaries, educational programs and archival videos. There are 20 concepts (high level features) needed to be submitted for high level feature task, in 2009. The 20 concepts are **Classroom**, **Chair**, **Infant**, **Traffic intersection**, **Doorway**, **Airplane_flying**, **Person-playing-a-musical-instrument**, **Bus**, **Person-playing-soccer**, **Cityscape**, **Person-riding-a-bicycle**, **Telephone**, **Person-eating**, **Demonstration_Or_Protest**, **Hand**, **People-dancing**, **Nighttime**, **Boat_Ship**, **Female-human-face-closeup**, and **Singing** respectively. More detailed definitions and descriptions of the concepts can be found from the website of TRECVID 2009[1].

## 6 Experimental results and discussions

In this Section, scene categorization performances of SPM, PHOG, GIST and HWVP are evaluated on the OT, Scene-13, and Sport Event datasets. The performances of HWVP are also comparisons with some of the algorithms on these datasets are given. The object categorization performances based on SVM classifiers are carried out [7]. We compare it with the authors' approaches using their own datasets [17, 18, 22]. Comparisons of the SPM, PHOG, GIST and HWVP descriptors based semantic concept retrieval are evaluated on TRECVID 2007 test dataset for the detection of 20 concepts in the high level feature extraction task of TRECVID 2009.

### 6.1 Scene categorization performance evaluation for SPM, PHOG, GIST and HWVP

Scene categorization performances of SVM based method with $N$ ($N$=5, 10, 20, 30, 50) training images per category on the OT, Scene-13, and Sport event datasets are compared for the four features: SPM [16], PHOG [2], GIST [34] and HWVP which are shown in Table 3 respectively. The average recognition rates and their standard deviations of 10 runs are provided. From Table 3, HWVP and GIST are effective than PHOG and SPM in scene categorization. When the training samples per category is set to be $N$=50, HWVP outperforms GIST, PHOG and SPM by 3.8 %, 15.6 % and 17.9 % respectively on Scene-13 dataset. For OT and Sport Event datasets, the performances of GIST and HWVP are very close. They are better than the performances of SPM and PHOG.

Table 4 shows the performances of HWVP and the approaches in [17, 18, 22]. The performance of HWVP is 83.1 % which is as good as that of the authors' method in [22] (with recognition rate 83.7 %) when the training image number per category is set to be 100. The confusion matrix of HWVP is shown in Fig. 5. The categories **forest, opencountry, coast** and **mountain** are very confusing during classification. We find that 10 %, 10 % and 5 % images of the **opencountry** are falsely classified into **coast, forest and mountain**. Moreover 10 % and 4 % images of the category **coast** are missed classified into the categories **opencountry** and **mountain**.

For the Sport Event dataset the performances of the authors' is 73.4 % using 70 training images per category. Our method achieves 73.6 % under the same testing conditions. The corresponding confusion matrix of ours is shown in Fig. 6a. We find that **bocce** is with lowest recognition rate 56 %. Some of the images in this class are falsely classified into the other seven categories. The [52] outperforms ours for the three categories: **badminton**,

[1] TRECVID 2009 Website: http://www-nlpir.nist.gov/projects/tv2009/tv9.hlf.for.eval.txt

**Table 3** Evaluation for SPM, PHOG, GIST and HWVP descriptors in scene categorization on OT, scene-13, and sport event datasets. The training image number N is set to be 5, 10, 20, 30, and 50 per category. The wavelet packet bases of the texture descriptors are db5. The texture descriptor is HWVP under Local5. The decomposition level of HWVP is set to be 4

| | | $N$=5 | $N$=10 | $N$=20 | $N$=30 | $N$=50 |
|---|---|---|---|---|---|---|
| OT dataset | SPM | 53.6±4.1 | 56.4±3.5 | 62.3±3.9 | 70.2±1 | 72.9±1.8 |
| | PHOG | 54.9±3.1 | 59.4±3.1 | 65.7±2.4 | 69.7±1.4 | 72.7±0.9 |
| | GIST | 56.3±3.1 | 64.5±2.5 | 72.1±2.3 | 75.8±1.6 | 77.0±0.8 |
| | HWVP | 55.8±3.5 | 64.7±3.1 | 71.8±2.0 | 75.3±1.2 | 77.2±0.7 |
| Scene-13 dataset | SPM | 40.7±3.1 | 46.8±2.8 | 50.9±2.6 | 53.2±2.3 | 55.3±1.7 |
| | PHOG | 41.1±2.1 | 48.6±1.5 | 53.2±1.2 | 55.3±1.2 | 57.6±1.1 |
| | GIST | 47.6±2.7 | 57.4±1.9 | 62.3±2.2 | 67.9±1.9 | 69.4±1.3 |
| | HWVP | 48.2±2.5 | 59.8±2.0 | 64.6±1.8 | 69.7±1.6 | 73.2±1.1 |
| Sport event dataset | SPM | 42.6±3.1 | 48.0±3.3 | 52.3±2.4 | 55.1±2.3 | 57.6±2.1 |
| | PHOG | 44.1±3.6 | 54.7±4.4 | 55.6±2.5 | 57.8±2.4 | 59.4±2.2 |
| | GIST | 45.2±3.1 | 56.8±4.0 | 64.1±2.1 | 67.4±2.2 | 68.9±2.3 |
| | HWVP | 45.6±2.8 | 56.2±3.3 | 64.3±2.1 | 67.6±1.9 | 69.3±1.8 |

**rockclimbing** and **snowboarding**. Compared to [17], our method improves the recognition rates of the most confusing three categories: **bocce**, **croquet** and **polo** by 4 %, 7 % and 16 % respectively.

For the Scene-13 dataset, the average recognition rate of the authors' approach is 65.2 % when the training image number is set to be 100 per category [18]. Under the same conditions our method achieves the mean recognition rate 78.9 %. The corresponding confusion matrix of our approach is shown in Fig. 6b. The most confusing two categories in this dataset are **bedroom** and **livingroom.** Their recognition rates of the proposed HWVP are 55 % and 57 % respectively. About 22 % missing classified images in the category **bedroom** are falsely classified into **livingroom**. And about 16 % missing classified images in the category **livingroom** are falsely classified into **bedroom**. From the datasets, we find that the texture patterns of **bedroom** and **livingroom** are very close. So they are very confusing during classification.

## 6.2 Semantic concept retrieval performances evaluation for SPM, PHOG, GIST and HWVP

In this section the semantic concept retrieval performances of SPM, PHOG, GIST and HWVP are evaluated on TRECVID 2007 dataset [5]. For this dataset we used the most representative key-frame of each video shot (43,616 key-frames). This data is also split into three disjoint sets: 26,170 key-frames for SVM parameters' training, 8,723 key-frames for classifiers' weights learning during decision fusion (this part is not utilized in this paper) and 8,723 for testing.

**Table 4** Comparisons for HWVP and the approaches [17, 18, 22] for scene categorization

| Dataset | $N$ | Ref. | HWVP |
|---|---|---|---|
| OT | 100 | 83.7 [22] | 83.1 |
| Scene-13 | 100 | 65.2 [18] | 78.9 |
| Sport event | 70 | 73.4 [17] | 73.6 |

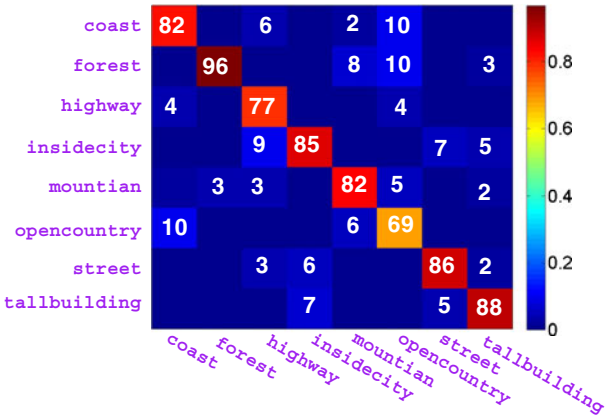



**Fig. 5** Confusion matrix of the HIGH under decomposition level $L$=4 with 30 training images per category for OT dataset

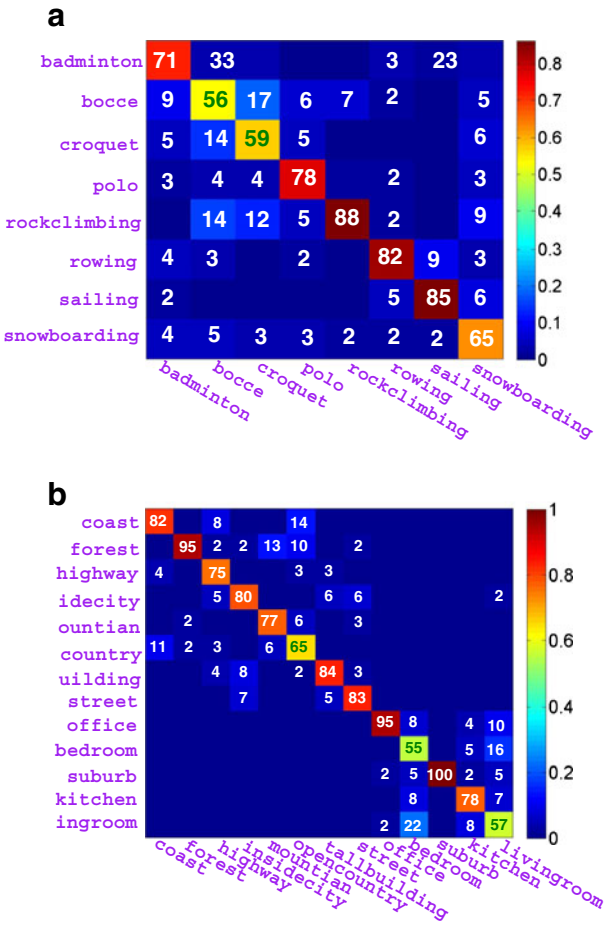



**Fig. 6** **a** Confusion matrixes of the descriptor HIGH on Sport event dataset. **b** Confusion matrixes on Scene-13 dataset

The average precision is utilized to evaluate the scene retrieval performances. In this paper the precision is the number of relevant documents retrieved divided by the total number retrieved. Average precision (AP) is defined as follow

$$AP = \frac{\sum_{r=1}^{C} (P(r) \times R(r))}{RV} \tag{19}$$

where $r$ is the rank, $C$ is the number of retrieved video shots. $RV$ is total number of retrieved video shots. $R(r)$ is a binary function on the relevance of a document at the given rank $r$, and $P(r)$ is precision at the given rank $r$. The mean average precision (MAP) is used for the evaluation of the $M$ ($M$=20) concepts on the TRECVID 2007 dataset, which is expressed as follow

$$MAP = \frac{1}{M} \sum_{i=1}^{M} AP_i; i = 1, \cdots, M \tag{20}$$

where $AP_i$ is the average precision for the $i$-th categories.

The corresponding definitions of the four features are shown in Table 2. For each concept, 2,000 shots are returned according to the rank of the responses. The MAP values of SPM, PHOG, GIST and HWVP for the 20 concepts are 7.28 %, 7.54 %, 8.20 % and 8.87 % respectively as listed in Table 5. HWVP achieves best MAP values for the eight classes: **Classroom (#1)**, **Infant (#3)**, **Doorway (#5)**, **Airplane_flying(#6)**, **Bus(#8)**, **Cityscape (#10)**, **Telephone(#12)**, and **Nighttime (#17)**. Its performances for the other 12 concepts are also satisfactory. We find that the four descriptors have high compensations for semantic

**Table 5** AP and MAP values (%) of the 20 concepts of TRECVID 2009 high level feature extraction task for the features PHOG, SPM, GIST and HWVP on TRECVID 2007 test dataset

| Concept # | | SPM | PHOG | GIST | HWVP |
|---|---|---|---|---|---|
| Classroom | #1 | 0.73 | 0.38 | 1.14 | 1.17 |
| Chair | #2 | 7.52 | 12.65 | 8.84 | 9.32 |
| Infant | #3 | 1.87 | 0.40 | 2.02 | 4.49 |
| Traffic-intersection | #4 | 18.93 | 28.07 | 9.87 | 22.03 |
| Doorway | #5 | 2.01 | 1.84 | 0.79 | 2.69 |
| Airplane_flying | #6 | 1.66 | 1.32 | 1.55 | 2.52 |
| musical-instrument | #7 | 22.52 | 19.83 | 28.46 | 27.63 |
| Bus | #8 | 0.28 | 0.06 | 0.06 | 0.31 |
| Person-playing-soccer | #9 | 4.26 | 0.9 | 3.61 | 1.9 |
| Cityscape | #10 | 3.06 | 3.4 | 2.92 | 4.62 |
| Person-riding-a-bicycle | #11 | 12.62 | 18.84 | 16.35 | 14.01 |
| Telephone | #12 | 0.80 | 0.49 | 2.26 | 2.65 |
| Person-eating | #13 | 21.59 | 23.83 | 23.68 | 22.89 |
| Demonstration_or_protest | #14 | 5.94 | 1.7 | 6.16 | 3.84 |
| Hand | #15 | 5.36 | 4.86 | 7.70 | 5.23 |
| People-dancing | #16 | 2.97 | 1.38 | 2.85 | 2.23 |
| Nighttime | #17 | 1.56 | 1.91 | 1.74 | 9.47 |
| Boat_ship | #18 | 12.33 | 15.02 | 11.45 | 12.44 |
| Female-human-face-closeup | #19 | 8.29 | 6.32 | 11.79 | 11.59 |
| Singing | #20 | 11.23 | 7.53 | 20.85 | 16.28 |
| MAP of the 20 concepts | | 7.28 | 7.54 | 8.20 | 8.87 |

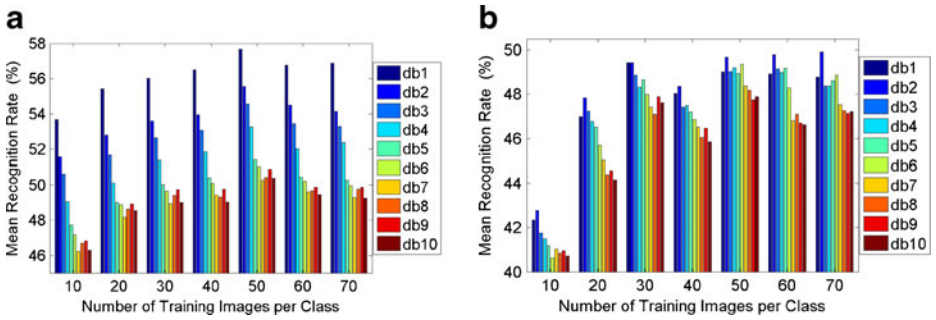



**Fig. 7** Recognition rates of the texture descriptor HIGH under the wavelet packet bases db1-db10 with decomposition level $L=4$ on (**a**) Scene-13 and (**b**) Sport event datasets

concept retrieval. Thus the final concept retrieval performances can be improved by fusing the descriptors using Adaboost algorithms [10, 25].

### 6.3 Impacts of wavelet packet bases on HWVP descriptors

Different wavelet packet bases may influence the object category performances. In this Section, we discuss the impacts of different wavelet bases to the performances of the HWVP based scene categorization performances on Scene-13 and Sport event datasets. Performances of HWVP descriptors under the Daubechies wavelet bases db1-db10 are compared as shown in Fig. 7. The average recognition rates of HIGH under 10, 20, 30, 40, 50, 60, and 70 training images per category are given. The performances of HWVP descriptors under wavelet bases db1 and db2 are comparatively better than the others. This is due to the fact that the wavelet packet transform with short length filters is good at keeping local texture patterns. As we know that wavelet packet transform with long length filters take more spatial information, thus the local structures are weakened in feature extraction. Performances of WVPK, NoDC, HIGH and HWVP under decomposition level $L=4$ with 50 training images per category on Scene-13 and Sport event datasets are shown in Fig. 8a and b respectively. From Figs. 7 and 8, it is obvious that better performances are achieved by the HWVP descriptors under wavelet packet bases db1 and db2.

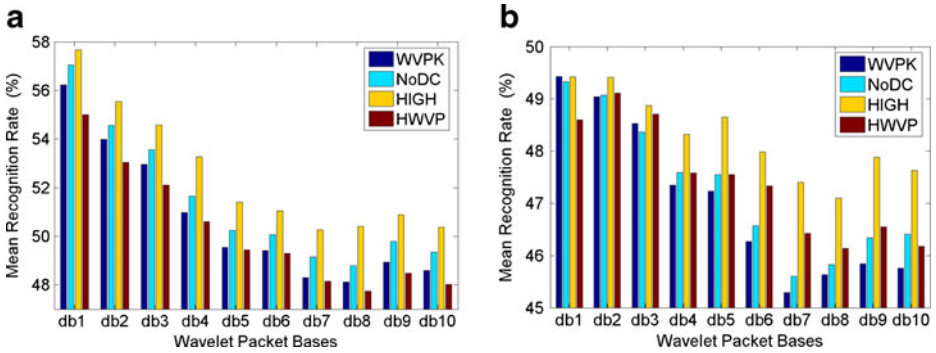



**Fig. 8** Object categorization performance of WVPK, NoDC, HIGH and HWVP under Global with decomposition level $L=4$ with 50 training images per category on (**a**) Scene-13 dataset; (**b**) Sport event dataset

### 6.4 Impacts of decomposition level $L$ on HWVP descriptors

In order to give some guidance to select the appropriate texture descriptors, we compare WVPK, NoDC, HIGH and HWVP under different training images and decomposition level $L$. The average recognition rates of WVPK, NoDC, HIGH and HWVP on OT dataset with $N$ ($N$=1, 5, 10, 15, 20, 25 and 30) training images per category are shown in Fig. 9a.

Figure 9b shows the comparisons of HWVP under Global and Local5 with various decomposition levels $L$ ($L$=1, 2, 3, and 4) and training image numbers $N$ ($N$=1, 5, 10, 15, 20, 25, and 30) per category on Scene-13 dataset. We find that with the increment of decomposition level $L$, the scene categorization performances of the descriptors are improved. From $L$=1 to $L$=2, the average recognition rate improves by about 5 %. However, the improvement from $L$=2 to $L$=3 is about 3 % in average. The improvements of $L$=3 to $L$=4 are very small for Global. In our opinion, this is caused by the fact that the responses of the image to the subtle sub-band filters are very weak when the decomposition level is more than three. In this case the average energy and standard deviations are too small for the details to provide discriminative power. Thus, the contributions of those sub-bands are comparatively limited. From this point of view, $L$=4 is enough for both global and local HWVP descriptors.

From Figs. 8 and 9, we find that HIGH outperforms WVPK, NoDC and HWVP. This is the reason that only the texture information of **details** is utilized in HIGH. It is robust to luminance variations. This can be revealed by the fact that he performances of NoDC are better than these of WVPK. This also shows the fact that **details** of any other decomposition level have positive contributions to object categorization. HIGH and NoDC are both using the **details** of wavelet packets. Performances of HIGH are better than those of NoDC. This further shows the effectiveness of feature representation in hierarchical multi-resolution domain. Due to the fact that the descriptor HWVP uses the **approximations** of all the resolutions during hierarchical wavelet packet transform, the impact of luminance variations to HWVP is larger than that of WVPK. This makes the scene categorization performances of WVPK better than these of HWVP under the same conditions.

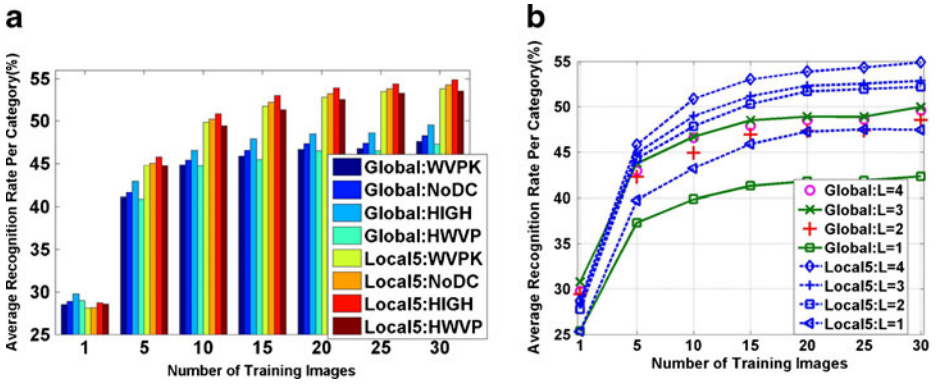



**Fig. 9** **a** Comparison for HWVP, NoDC, WVPK and HIGH on OT dataset under Global and Local5 patterns with wavelet packet bases db5. **b** Comparisons of HWVP under Global and Local5 with wavelet packet bases db5 under various decomposition levels on Scene-13

**Table 6** Scene categorization performances of HWVP descriptors under Global, Local4, Local5, and Local9 on Scene-13 dataset. The training image number is set to be 15 per category. The wavelet packet bases of the texture descriptors are db5

| | $L$=1 | | | | $L$=2 | | | | $L$=3 | | | | $L$=4 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | WVPK | NoDC | HIGH | HWVP | WVPK | NoDC | HIGH | HWVP | WVPK | NoDC | HIGH | HWVP | WVPK | NoDC | HIGH | HWVP |
| Global | 27.65 | 41.92 | 41.92 | 25.01 | 40.84 | 46.87 | 47.20 | 34.62 | 48.28 | 48.60 | 48.94 | 43.45 | 46.42 | 47.06 | 48.47 | 46.21 |
| Local4 | 35.31 | 42.07 | 42.07 | 32.37 | 47.05 | 47.71 | 47.68 | 41.28 | 50.34 | 48.27 | 48.60 | 48.30 | 49.90 | 50.10 | 50.68 | 51.68 |
| Local5 | 35.91 | 45.72 | 45.74 | 32.94 | 47.86 | 50.28 | 50.33 | 41.87 | 52.18 | 50.81 | 51.18 | 49.44 | 52.04 | 52.53 | 53.20 | 51.68 |
| Local9 | 40.04 | 45.04 | 45.04 | 37.04 | 49.05 | 49.33 | 49.26 | 44.40 | 52.04 | 50.33 | 50.50 | 49.96 | 52.02 | 50.52 | 51.33 | 51.88 |

### 6.5 Impacts of local and global patterns on HWVP descriptors

In this section, the impacts of local and global patterns to HWVP descriptors on scene categorization performance are discussed. Table 6 shows the detailed comparison on Scene-13 dataset with 15 training images per category where the performances of Global, Local4, Local5, and Local9 are compared. The **details** of hierarchical wavelet packet give more positive contribution than **approximations** for object categorization. Thus, performances of NoDC and HIGH are better than those of HWVP and WVPK. In general the performances of proposed texture descriptors under Local4, Local5 and Local9 are better than that of Global. This shows the fact that the roles of local patterns are enhanced by using localized hierarchical wavelet packet transform. For the texture descriptors NoDC and HIGH, the performances of Local9 are higher than those of Local4 and lower than those of Local5. This is due to the fact that salient texture characters are broken with subtle partitioning.

## 7 Conclusions

In this paper, effective texture descriptors based on hierarchical wavelet packet transform are proposed. Systematical comparisons are made with PHOG, SPM and GIST features in the applications of scene categorization and semantic concept retrieval on widely used OT, Scene-13, Sport event, and TRECVID 2007 datasets. The HWVP descriptors achieve better performances, and they have good compensations to the other descriptors for semantic concept retrieval.

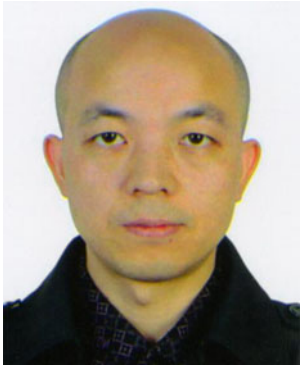
Scene categorization can be benefited from the texture information of the sub-bands of hierarchical wavelet packet transform. Four hierarchical wavelet packet descriptors are evaluated. Details of hierarchical wavelet packet sub-bands provide positive contribution for feature representation. The wavelet packet bases have significant impacts of the HWVP descriptors, with the increase of decomposition level, HWVP based scene categorization performance improved. HWVP descriptors with four decomposition levels are enough for both their computational costs and discrimination powers. HWVP descriptors under wavelet packet bases with short lengths achieve better performances than the descriptors under filters with long length. Local texture patterns are blurred when using long length wavelet packet bases. The discriminative powers of HWVP descriptors with long length filters are larger than those of the descriptors with short length. HWVP descriptors extracted from the local images can improve the feature discrimination power. However, it is important that partitioning images into many grids may break the salient texture patterns.

## References

1. Blei D, Ng A, Jordan M (2003) “Latent dirichlet allocation.” J Mach Learn Res (3): 993–1022
2. Bosch A, Zisserman A, Munoz X (2007) “Representing shape with a spatial pyramid kernel.” In: Proc. CIVR
3. Bosch A, Zisserman A, Munoz X (2008) Scene classification using a hybrid generative/discriminative approach. IEEE Trans Pattern Anal Mach Intell 30(4):712–727
4. Cai D, He X, Li Z, Ma W, Wen J (2004) “Hierarchical clustering of WWW image search results using visual, textual and link information.” In: Proc. ACM Multimedia, pp. 952–959
5. Campbell M, Haubold A, Liu M, Natsev A, Smith JR, Tesic J, Xie L, Yan R, Yang J (2007) “IBM research TRECVID-2007 video retrieval system.” In: NIST TRECVID Workshop

6. Cao L, Li F (2007) "Spatially coherent latent topic model for concurrent object segmentation and classification." In: Proc. ICCV
7. Chang C, Lin C (2008) "LIBSVM: a library for support vector machines". http://www.csie.ntu.edu.tw/~cjlin/libsvm/
8. Csurka G, Dance C, Fan L, Willamowski J, Bray C (2004) "Visual categorization with bags of keypoints." In: Proc. ECCV
9. Fidler S, Boben M, Leonardis A (2008) "Similarity-based cross-layered hierarchical representation for object categorization." In Proc. CVPR
10. Freud Y, Schapire R (1996) "Experiments with a new boosting algorithms." Machine Learning: Proceedings of the 13th International Conference
11. Garcia C, Zikos G, Tziritas G (2000) Wavelet packet analysis for face recognition. Image Vision Comput 18:289–297
12. Hofmann T (2001) Unsupervised learning by probabilistic latent semantic analysis. Mach Learn 42 (1):177–196
13. Holub A, Perona P (2005) "A discriminative framework for modeling object classes." In: Proc. ICCV
14. Laine A, Fan J (1993) Texture classification by wavelet packet signatures. IEEE Trans Pattern Anal Mach Intell 15(11):1186–1193
15. Larlus D, Jurie F (2008) "Combining appearance models and markov random fields for category level object segmentation." In: Proc. CVPR
16. Lazebnik S, Schmid C, Ponce J (2006) "Beyond bags of features: spatial pyramid matching for recognizing natural scene categories." In: Proc. CVPR
17. Li L, Li F (2007) "What, where and who? classifying events by scene and object recognition." In: Proc. ICCV
18. Li F, Perona P (2005) "A Bayesian hierarchy model for learning natural scene categories." In: Proc. CVPR
19. Li J, Wang J (2003) Automatic linguistic indexing of pictures by a statistical modeling approach. IEEE Trans Pattern Anal Mach Intell 25(9):1075–1088
20. Lowe D (2004) Distinctive image features from scale-invariant keypoints. ICCV 60(2):91–110
21. Mutch J, Lowe D (2006) "Multiclass object recognition using sparse, localized features." In: Proc. CVPR
22. Oliva A, Torralba A (2001) Modeling the shape of the scene: a holistic representation of the spatial envelope. Int J Comput Vis 42(3):145–175
23. Qian X, Hua X, Chen P, Ke L (2011) PLBP: an effective local binary patterns texture descriptor with pyramid representation. Pattern Recogn 44:2502–2515
24. Qian X, Liu G, Guo D, Li Z, Wang Z, Wang H (2009) "Object categorization using hierarchical wavelet packet texture descriptors." In: Proc. ISM, pp. 44–51
25. Qian X, Yan Z, Hang K (2011) "Boosted scene categorization approach by adjusting inner structures and outer weights of weak classifiers". In: Proc. MMM, pp. 413–423
26. Quattoni A, Collins M, Darrell T (2004) "Conditional random fields for object recognition." In: NIPS
27. Rabinovich A, Vedaldi A, Galleguillos C, Wiewiora E, Belongie S (2007) "Object in context." In: Proc. ICCV
28. Ro Y, Kim M, Kang H, Manjunath B, Kim J (2001) MPEG-7 homogeneous texture descriptor. ETRI J 23 (2):41–51
29. Serre T, Wolf L, Poggio T (2005) "Object recognition with features inspired by visual cortex." In: Proc. CVPR
30. Smeulders A, Worring M, Santini S, Gupta A, Jain R (2000) Content based image retrieval at the end of the early years. IEEE Trans Pattern Anal Mach Intell 22(12):1349–1380
31. Sudderth E, Torralba A, Freeman W, Willsky A (2005) "Describing visual scenes using transformed dirichlet processes." In: NIPS
32. Tao D, Tang X, Li X, Wu X (2006) Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval. IEEE Trans Pattern Anal Mach Intell 28(7):1088–1099
33. Teh Y, Jordan M, Beal M, Blei D (2006) "Hierarchical Dirichlet processes." J Am Stat Assoc
34. Torralba A, William K, Freeman T, Rubin M (2003) "Context-based vision system for place and object recognition." In: Proc. ICCV
35. Wang G, Zhang Y, Li F (2006) "Using dependent regions for object categorization in a generative framework." In: Proc. CVPR
36. Wu L, Hu Y, Li M, Yu N, Hua X (2009) Scale-invariant visual language modeling for object categorization. IEEE Trans Multimedia 11(2):286–294
37. Yuan J, Wu Y, Yang M (2007) "Discovery of collocation patterns: from visual words to visual phrases." In: Proc. CVPR

38. Zhang H, Berg A, Maire M, Malik J (2006) “Svm-knn: discriminative nearest neighbor classification for visual category recognition.” In: Proc. CVPR
39. Zhang J, MarszaÃlek M, Lazebnik S, Schmid C (2007) “Local features and kernels for classification of texture and object categories: a comprehensive study.” Int J Comput Vis
40. Zheng Y, Zhao M, Neo S, Chua T, Tian Q (2008) “Visual synset: towards a higher-level visual representation.” In: Proc. CVPR
41. Zhou X, Wang M, Zhang Q, Zhang J, Shi B (2007) “Automatic image annotation by an iterative approach incorporating keyword correlations and region matching.” In: Proc. CIVR, pp. 25–32



**Xueming Qian** (M’10) received the B.S. and M.S. degrees in Xi’an University of Technology, Xi’an, China, in 1999 and 2004, respectively, and the Ph.D. degree in the School of Electronics and Information Engineering, Xi’an Jiaotong University, Xi’an, China, in 2008. He was awarded Microsoft fellowship in 2006. From 1999 to 2001, he was an Assistant Engineer at Shannxi Daily. From 2008 till now, he is a faculty member of the School of Electronics and Information Engineering, Xi’an Jiaotong University. He was a visit scholar at Microsoft research Asia from Aug. 2010 to March 2011. His research interests include video/image analysis, indexing, and retrieval.

**Danping Guo** received the B.S. degree in ChangAn University, Xi’an, China, in 2006. She received the MS. degree at the School of Electronics and Information Engineering, Xi’an Jiaotong University, Xi’an, China. Now she is a teacher at Hefei, China. Her research interests include video analysis, and processing.


**Xingsong Hou** received the B.S. degree in electronic engineering from North China Institute of Technology, Taiyuan, China, in 1995, and the M.S. degree and Ph.D degree in the School of Electronics and Information Engineering, Xi'an Jiaotong University, Xi'an, China in 2000 and 2005, respectively. From 1995 to 1997, he was an Engineer with the Xi'an Electronic Engineering Institute in the field of radar signal processing. Now he is an associate professor of the School of Electronics and Information Engineering, Xi'an Jiaotong University. His research interests include video/image coding, wavelet analysis, sparse representation, sparse representation and compressive sensing, and radar signal processing.

**Zhi Li** received the B.S. degree in Xi'an Shiyou University, Xi'an, China, in 2005. She is currently working toward the Ph.D. degree at the School of Electronics and Information Engineering, Xi'an Jiaotong University, Xi'an, China. Her research interests include video analysis, processing and compression, semantic-based video analysis, indexing, and retrieval and video/image compression and transmission.

**Huan Wang** received the B.S. in Xi'an University of Technology, in 2004, and M.S. degrees in Xi'an Jiaotong University in 2010 respectively. Her research interests include video/image coding, communication and transmission.

**Guizhong Liu** received the B.S. and M.S. degrees in computational mathematics from Xi'an Jiaotong university, Xi'an, China, in 1982 and 1985, respectively, and the Ph.D. degree in mathematics and computing science from Eindhoven University of Technology, Eindhoven, The Netherlands, in 1989. He is currently a Full Professor with the School of Electronic and Information Engineering, Xi'an Jiaotong University. His research interests include nonstationary signal analysis and processing, image processing, audio and video compression, and inversion problems.

**Zhe Wang** received the B.S. degree from Chongqing University of Posts and Telecommunications, Chongqing, China, in 2003. He is currently pursuing his Ph.D in Xi'an Jiaotong University, Xi'an, China. degree. His research interests are in image processing, image retrieval, and computer vision.