HSOG: A Novel Local Image Descriptor Based on Histograms of the Second-Order Gradients

Di Huang, *Member, IEEE*, Chao Zhu, Yunhong Wang, *Member, IEEE*, and Liming Chen, *Member, IEEE*

***Abstract*— Recent investigations on human vision discover that the retinal image is a landscape or a geometric surface, consisting of features such as ridges and summits. However, most of existing popular local image descriptors in the literature, e.g., scale invariant feature transform (SIFT), histogram of oriented gradient (HOG), DAISY, local binary Patterns (LBP), and gradient location and orientation histogram, only employ the first-order gradient information related to the slope and the elasticity, i.e., length, area, and so on of a surface, and thereby partially characterize the geometric properties of a landscape. In this paper, we introduce a novel and powerful local image descriptor that extracts the histograms of second-order gradients (HSOGs) to capture the curvature related geometric properties of the neural landscape, i.e., cliffs, ridges, summits, valleys, basins, and so on. We conduct comprehensive experiments on three different applications, including the problem of local image matching, visual object categorization, and scene classification. The experimental results clearly evidence the discriminative power of HSOG as compared with its first-order gradient-based counterparts, e.g., SIFT, HOG, DAISY, and center-symmetric LBP, and the complementarity in terms of image representation, demonstrating the effectiveness of the proposed local descriptor.**

***Index Terms*— Local image descriptor, feature extraction, second order gradients, image matching, object categorization, scene classification.**

I. Introduction

LOCAL image descriptors, *e.g.*, SIFT [1], computed densely or from interest regions, have many applications in computer vision, ranging from traditional vision tasks, *e.g.*, panoramic stitching [2], wide-baseline matching [3], to advanced visual recognition problems, *i.e.*, visual object classification [1], [4], scene categorization [5], [6], and image retrieval [7]. As such, they have been the focus of great attentions from the research community in recent years and a number of effective local image descriptors have been proposed for various purposes, *e.g.*, object recognition or image indexing.

Manuscript received February 14, 2014; revised June 16, 2014 and August 21, 2014; accepted August 22, 2014. Date of publication September 4, 2014; date of current version September 23, 2014. This work was supported in part by the National Basic Research Program of China under Grant 2010CB327902, in part by the National Natural Science Foundation of China under Grant 61202237, in part by the Beijing Municipal Natural Science Foundation under Grant 4142032, in part by the French Research Agency through the Videosense Project 2009 CORD 026 02, in part by the Visen Project under Grant ANR-12-CHRI-0002-04 within the CHIST-ERA Program, in part by the Specialized Research Fund for the Doctoral Program of Higher Education under Grant 20121102120016, in part by the Research Program of State Key Laboratory of Software Development Environment under Grant SKLSDE-2013ZX-31, in part by the Joint Project through the LIA 2MCSI Laboratory between the Group of Ecoles Centrales and Beihang University, Beijing, China, and in part by the Fundamental Research Funds for the Central Universities. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Chun-Shien Lu.

D. Huang and Y. Wang are with the Laboartory of Intelligent Recognition and Image Processing, School of Computer Science and Engineering, Beihang University, Beijing 100191, China (e-mail: dhuang@buaa.edu.cn; yhwang@buaa.edu.cn).

C. Zhu is with the Institute of Computer Science and Technology, Peking University, Beijing 100871, China (e-mail: zhuchao@pku.edu.cn).

L. Chen is with the LIRIS Laboratory, MI Department, École Centrale de Lyon, Lyon 69134, France (e-mail: liming.chen@ec-lyon.fr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2014.2353814

A. Related Work

A landmark representative of these local image descriptors is Scale Invariant Feature Transform (SIFT) proposed by Lowe [1]. SIFT has been widely studied and has played a dominant role in object recognition. Its descriptor is represented by a 3D histogram of the gradient locations and orientations whose contributions are weighted by their gradient magnitudes. The quantization of gradient locations and orientations makes the descriptor robust to small geometric distortions and errors in the previous step of region detection.

Two years before Lowe's work in [1], Belongie *et al.* [11] introduced a descriptor named Shape Context, similar to SIFT but is based on edges. It is represented by the 2D histogram of these edge point locations, where the log-polar location grid is utilized. It aims at describing the distribution of edge points on a shape with respect to the reference point.

Following the two descriptors, *i.e.* SIFT and Shape Context, great strides have been achieved to ameliorate the performance of local image descriptors in the literature, in reinforcing the discriminative power [9], [13], [16], improving the efficiency while decreasing the memory requirements [8], [12], [15], [17], [18], increasing the invariance to lighting changes [10], [14], [19], [20] or scale variations [21], [22], and making them robust to local distortions [23] or even occlusions [24], [25].

Ke and Sukthankar [8] proposed the PCA-SIFT descriptor, which directly applies Principal Component Analysis (PCA) to the normalized gradient patches to enhance the distinctiveness and reduce the dimensionality of the SIFT features.

Mikolajczyk and Schmid [9] extended SIFT to the Gradient Location and Orientation Histogram (GLOH) descriptor to increase both robustness and distinctiveness. It replaces the rectangular location grid utilized in SIFT with a log-polar one, and applies PCA to reduce the size of the descriptor.

In [19] and [10], Van de Sande *et al.* extracted SIFT features in different color spaces and compared their accuracies, including HSV-SIFT [26], HueSIFT [27], OpponentSIFT, C-SIFT, rgSIFT, RGB-SIFT, and Transformed color SIFT, showing that

TABLE I
ATTRIBUTE SUMMARY OF MAIN LOCAL IMAGE DESCRIPTORS

Descriptor	Type	Information	Neighborhood	Computation	Dimension
SIFT [1]	Sparse	1st-order gradients	Rectangular	Distribution	128
PCA-SIFT [8]	Sparse	1st-order gradients	Rectangular	Distribution	36
GLOH [9]	Sparse	1st-order gradients	Polar	Distribution	128
Color SIFT [10]	Sparse	1st-order gradients	Rectangular	Distribution	384
Shape Context [11]	Sparse	Edge points	Polar	Distribution	60
SURF [12]	Sparse	1st-order wavelets	Rectangular	Filter	64
HOG [13]	Dense	1st-order gradients	Rectangular & Polar	Distribution	36
CS-LBP [14]	Sparse	1st-order binary patterns	Rectangular	Distribution	256
DAISY [15]	Dense	1st-order gradients	Polar	Filter	200
HSOG	Sparse	**2nd-order gradients**	Polar	Distribution	128

combining the SIFT descriptor with color clues is a promising way to improve the performance in object recognition.

Inspired by SIFT, Bay *et al.* [17], [12] introduced Speeded-Up Robust Features (SURF). Instead of the gradient information used in the SIFT descriptor, SURF computes Haar wavelet responses, and exploits integral images to save computational cost. As a result, it runs times faster than SIFT.

Dalal and Triggs [13] presented the Histogram of Oriented Gradient (HOG) descriptor. HOG combines both the properties of SIFT and GLOH, because it is also represented by the 3D histogram of gradient locations and orientations, and employs both rectangular and log-polar location grids. The main difference between HOG and SIFT is that HOG is computed on a dense grid of uniformly spaced cells, with overlapping local contrast normalization.

Ojala *et al.* [28] proposed Local Binary Patterns (LBP) for texture classification, and such a descriptor encodes the sign information between the central pixel and its surrounding ones within a given neighborhood. It was soon successfully applied to face recognition [29] and object categorization [20].

Heikkila *et al.* [14] combined both the strengths of SIFT and LBP to build the Center-Symmetric LBP (CS-LBP) descriptor. It adopts the SIFT-like approach for descriptor construction, but replaces its gradient information with CS-LBP features. Instead of thresholding each pixel with the central one within the neighborhood, CS-LBP only compares center-symmetric pairs of pixels, which reduces the size of the LBP histogram.

Similar to [19], Zhu *et al.* [20] attempted to embed the color information to the original LBP operator for object recognition, pointing out that the six proposed color LBP descriptors increase the photometric invariance and discriminative power of the original LBP and their joint use achieves the comparable performance as SIFT does. They further proposed an extension of this color LBP, namely orthogonal color LBP or OC-LBP [6], which drastically decreases the feature vector dimension while keeping the same discriminative power.

In order to improve the efficiency of local descriptors, Tola *et al.* [18], [15] proposed the DAISY descriptor which replaces the weighted sum rule used in SIFT by sum of convolutions. Recently, Zhu *et al.* [30] introduced DAISY into the domain of object recognition, and proved that when displaying a similar recognition accuracy to SIFT, DAISY operates 12 times faster.

The performance of local image descriptors, especially for object recognition and image categorization, is discussed and compared in several studies, see [9], [10], [31], [32]. The attributes of some popular ones so far proposed within the domain are summarized in Table I, including representation type (sparse or dense), encoded information, spatial pooling scheme (neighborhood), computation method, and dimensionality. It should be noted that the items in the column of category and dimension can be changed according to different tasks, and the ones listed are directly cited from the original papers.

As it can be seen from Table I, most local image descriptors are based on the first order gradient information along with a given pooling scheme to simulate the features of the human visual cortex following the findings of Hubel and Wiesel [33]. However, recent psychophysical and physiological studies on human vision, see [34], [35], have shown that the first order gradient information is far from being sufficient accurate in capturing the perceived visual features by human beings.

B. Motivation and Contribution

In this paper, we concentrate on the discriminative power of local image descriptors and investigate a novel one based on second order gradient clues, namely Histograms of the Second Order Gradients (HSOG), which is able to simulate the human perceived visual features. Indeed, some latest studies on human vision [34] suggest that the neural image is a landscape or a surface, consisting of features such as cliffs, ridges, summits, valleys, or basins, whose geometric properties can be uniquely and accurately charactered by local curvatures of differential geometry through second order gradient related information as depicted in Fig.1, whereas first order gradients only measure the slope of the luminance profile at each point [35] and thus give quantities of the elasticity of a surface, *e.g.*, length, area.

In the theory of differential geometry, slope and curvature are different geometric clues which one can measure at each point on a 1D curve. As we can observe from Fig.1 (a), the first order gradient computed on a point delivers the slope or the velocity of the curve at that point which encodes the metrics, *e.g.*, the length of that curve, whereas the second order gradient at that point is the quantity related to its local curvature or how much the curve bends. Now the retinal image is a landscape or

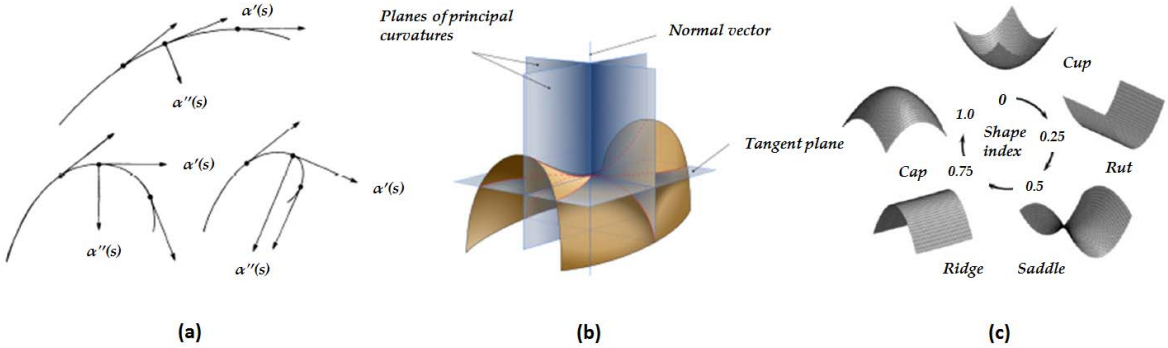


Fig. 1. Slope, curvature, and shape in differential geometry. (a) The first order gradient computes the slope at a given point on a 1D curve whereas the second order gradient delivers the curvature at that point; (b) On a smooth 2D surface embedded in a 3D space, one can compute the two principle curvatures which help to characterize the local shape; (c) The shape index is computed from the principle curvatures and its different values correspond to different shapes.

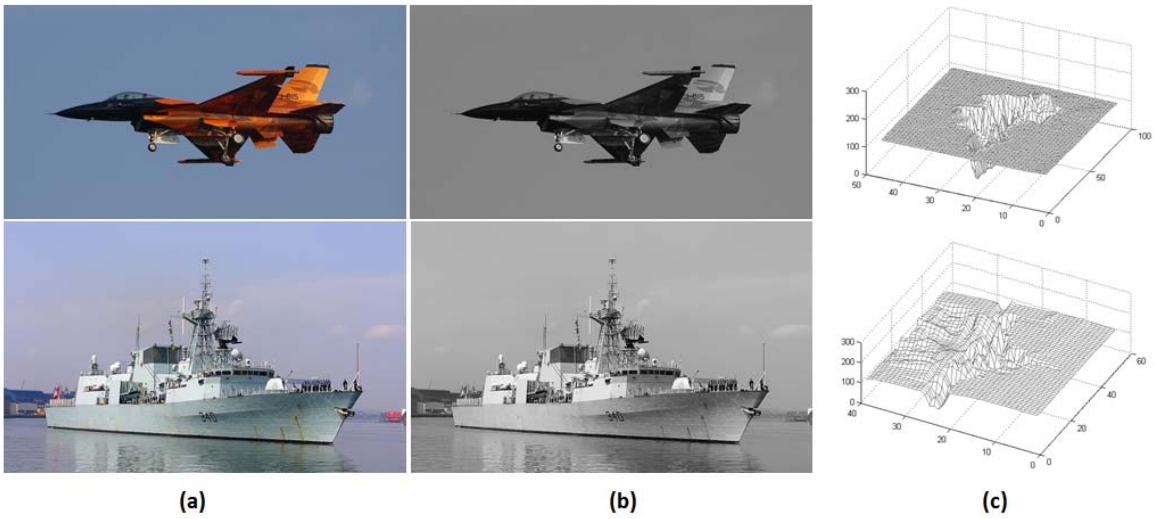


Fig. 2. Retinal images are perceived as landscapes or surfaces. (a) Two images showing an aircraft and a warship, respectively; (b) Their gray level images; (c) These gray level images plotted as landscapes where one can see different geometric properties: cliffs, ridges, summits, valleys, basins, *etc.*

a surface embedded in a 3D space, and the local shape around a point, as illustrated in Fig.1 (b), is characterized by the two principle curvatures, *i.e.* maximum and minimum curvatures, that one can compute by the second fundamental form [36] which is closely related to the second order gradient cue. Their joint variations, *e.g.*, through the value of shape index, define various local shapes as in Fig.1 (c).

Following the insights conveyed by recent investigations on human vision as well as existing tools of differential geometry, we conjecture that local image descriptors calculated over a point should exploit the second order gradient information to account for its local shape attributes of a retinal image in terms of curvature and thus provide additional discriminative power as compared with their first order gradient-based counterparts. Nevertheless, since first order gradients are quantities related to the metrics of a surface, *e.g.*, length, angle, and area, while second order gradients correspond to the curvatures, these two categories of quantities should present some complementarity in the description of a local surface shape. Fig.2 illustrates two images and their corresponding landscapes plotted as surfaces. The first one shows an aircraft and the second one a warship. As we can see from this figure, each of these two landscapes displays various local shapes, including cliffs, ridges, summits, valleys, or basins.

This paper proposes a novel local image descriptor, namely *Histograms of Second Order Gradients* abbreviated as *HSOG*, to characterize local shape changes for images represented as landscapes. Following the findings of Hubel and Wiesel [37], we also apply a pooling strategy, as most state of the art local descriptors do, to enable small displacements of second order gradients in the neighborhood of a certain point. Specifically, for a given image region, HSOG begins by computing a set of its first order Oriented Gradient Maps (OGMs), each of which is for a quantized direction. The histograms of the second order gradients are then extracted from all these OGMs, and finally concatenated to achieve the HSOG descriptor. Additionally, we embed the multi-scale strategy to further reinforce the descriptive completeness of local shape changes and thereby increase discriminative power and performance. Extensive experiments are carried out in three different applications, and the results

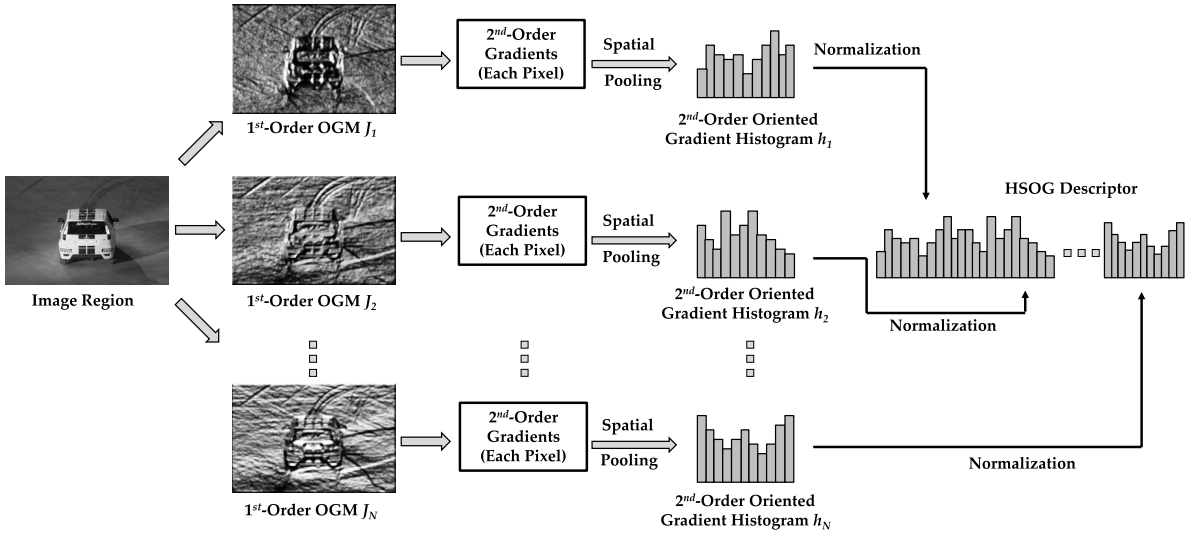


Fig. 3. Construction process of the proposed HSOG descriptor.

clearly demonstrate the effectiveness of the proposed HSOG descriptor and its complementarity with respect to these first order gradient based ones.

The main contribution of this work is three-fold:

- We highlight recent findings on human vision that retinal images are landscapes and their local geometric properties, *e.g.*, cliffs, ridges, summits, can be characterized by quantities of differential geometry, *e.g.*, curvatures.
- We propose a local descriptor, HSOG, to encode second order gradients, and these quantities are closely related to principle curvatures, accounting for local shape variations and thereby offering good discriminative power as its first order gradient-based counterparts do.
- We test HSOG in three applications, namely local image matching, visual object categorization (VOC), and scene classification, and prove its effectiveness with respect to first order gradient-based ones and their complementarity in terms of descriptive power, especially for the application where the major challenges occur in viewpoint and illumination.

A preliminary version of this work appeared in [38], which compares HSOG with state of the art local image descriptors in VOC using sparse sampling. This paper includes that work but significantly extends it in the following ways. Firstly, we motivate HSOG using the recent findings of psychophysicists on human vision and explain why local shape characterization should rely on second order gradients in the viewpoint of differential geometry. Secondly, using the Notre Dame-Yosemite-Liberty (NYL) database [16], we highlight the discriminative power of HSOG compared with the state of the art by comprehensive experiments in local descriptor matching whose major challenges are viewpoint and illumination changes. Thirdly, we evaluate HSOG and its first order gradient related counterparts on two other applications, *i.e.* VOC and scene classification. These two applications possess increased challenges including in particular background clutter, scale variations, occlusions, intra-class dissimilarities, and inter-class similarities. However, scene images are generally wide views of a large depth-of-field and therefore scale changes are not so important as in VOC. We improve their performance by the dense sampling strategy and the experimental results on these two applications provide insights into the properties of HSOG contrasted with the first order gradient based ones.

C. The Organization of the Paper

The rest of this paper is organized as follows. In Section II, the HSOG descriptor is introduced in detail, including the way of construction and utilization. Section III presents and discusses the experimental results achieved in three applications. Section IV concludes the paper.

II. HSOG Descriptor Construction

In this section, we introduce the Histograms of the Second Order Gradient (HSOG) descriptor in detail. The construction of HSOG is composed of three steps: (1) computation of the first order Oriented Gradient Maps (OGMs); (2) computation of the second order gradients based on these computed OGMs; and (3) spatial pooling.

The entire process is illustrated in Fig. 3.

A. Computation of 1st-Order Oriented Gradient Maps (OGMs)

The input of the proposed HSOG descriptor is a local image region around a given keypoint, which is either detected by an interest point detector, *e.g.*, Harris-Laplace, or located on a dense sampling grid. For each pixel (x, y) within the given region I, a certain number of gradient maps G_1, G_2,..., G_N, one for each quantized direction o, are first computed. They are formally defined as:

$$G_o = \left(\frac{\partial I}{\partial o}\right)^+; \quad o = 1, 2, \ldots, N. \tag{1}$$

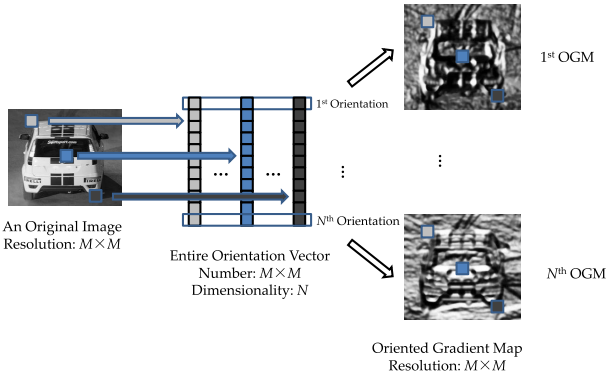


Fig. 4. An illustration of the oriented gradient maps for each of the quantized orientations o.

where the '+' sign means that only positive values are kept to preserve the polarity of intensity changes, while negative ones are set to zero.

Each gradient map describes the gradient norms of the input image region in a direction o at every pixel location. We then convolve these gradient maps with a Gaussian kernel G. The standard deviation of the Gaussian kernel G is proportional to the radius of the given neighborhood, R, as in (2).

$$\rho_o^R = G_R * G_o \tag{2}$$

The purpose of the convolution with a Gaussian kernel is to allow the gradients to shift within a neighborhood without abrupt changes and thereby simulates the operation of simple cells in the human visual processing.

At a given pixel location (x, y), we collect all the values of these convolved gradient maps at that location and build the vector $\rho^R(x, y)$.

$$\rho^R(x, y) = \left[\rho_1^R(x, y), \cdots, \rho_N^R(x, y)\right]^T \tag{3}$$

This vector, $\rho^R(x, y)$, is further normalized to a unit norm vector, which is called entire orientation vector and denoted by $\underline{\rho^R}$ in the subsequent, and the image region can hence be represented by entire orientation vectors. Specifically, given a local image region I, we generate an Oriented Gradient Map (OGM) J_o for each orientation o defined as:

$$J_o(x, y) = \underline{\rho_o^R}(x, y) \tag{4}$$

Fig. 4 illustrates such a process. As it can be seen from that figure, OGMs highlight the details of local texture changes and potentially offer high distinctiveness. Due to the computation of gradient maps and the following normalization step, OGMs possess the property of being invariant to affine lighting transformations. Indeed, an OGM J_o is simply the normalized convolved gradient map at orientation o according to (4), while a brightness change often adds a constant intensity value, so it does not affect the gradient computation. Moreover, a change in image contrast in which the intensities of all the pixels are multiplied by a constant will thus result in the multiplication of gradient calculation; however, this change of contrast will be cancelled by the normalization of the response vector. As such, it has been successfully applied to face recognition [39]. All these properties will be inherited by the HSOG descriptor.

B. Computation of 2nd-Order Gradients

Once these first order OGMs of all quantized directions are generated, they are exploited as the inputs for computing the second order gradients over the same image region. Precisely, for each of OGMs, $J_o(x, y)$, $o = 1, 2, \ldots, N$, we consider it as a regular image, and calculate the gradient magnitude mag_o and orientation θ_o at every pixel location as (5) and (6).

$$mag_o(x, y) = \sqrt{\left(\frac{\partial J_o(x, y)}{\partial x}\right)^2 + \left(\frac{\partial J_o(x, y)}{\partial y}\right)^2} \tag{5}$$

$$\theta_o(x, y) = \arctan\left(\frac{\partial J_o(x, y)}{\partial y} \Big/ \frac{\partial J_o(x, y)}{\partial x}\right) \tag{6}$$

where $o = 1, 2, \ldots, N$;

$$\frac{\partial J_o(x, y)}{\partial x} = J_o(x + 1, y) - J_o(x - 1, y) \tag{7}$$

$$\frac{\partial J_o(x, y)}{\partial y} = J_o(x, y + 1) - J_o(x, y - 1) \tag{8}$$

Each orientation (denoted as θ_o) is then mapped from the range of $[-\pi/2, \pi/2]$ to that of $[0, 2\pi]$, and quantized into N dominant orientations, which keeps consistent with the number of the first order OGMs. After quantization, the entry n_o of each direction θ_o is calculated as (9).

$$n_o(x, y) = \mathrm{mod}\left(\left\lfloor \frac{\theta_o(x, y)}{2\pi/N} + \frac{1}{2} \right\rfloor, N\right), \quad o = 1, 2, \ldots, N \tag{9}$$

A crucial issue to be dealt with when computing the second order gradients is the sensitivity of the resultant local image descriptor with respect to noise. As it can be seen in the latter section of experimental results, the fact of using the Gaussian kernel to simulate human simple cells and smooth first order gradients by (2) gives HSOG a desirable robustness to noise.

C. Spatial Pooling

The local shape descriptor associated with a given keypoint is computed to simulate the operation of human complex cells in the visual cortex [37] so that they are robust to small transformations, *e.g.*, spatial shifting, non rigid deformations. This is achieved as most state of the art local image descriptors, *e.g.*, SIFT, through a spatial pooling strategy. It consists of dividing the input local image region into sub-regions and accumulating a histogram of certain property (gradients, edge points, binary patterns, etc.) within each sub-region. All these histograms are then concatenated to construct the final descriptor. Brown *et al.* [16] analyzed different spatial pooling schemes and compared their performance, suggesting that the best performance can be achieved by the DAISY-style arrangement, as illustrated in Fig. 5. As a result, this spatial pooling strategy is adopted to build the HSOG descriptor.

As illustrated in Fig. 5, the input image region is separated into a number of circles of different sizes located on a series of concentric rings (3 in Fig. 5), whose radius values form an arithmetic sequence and are hence controlled by the radius of the region. In each ring, the centers of these circles are evenly distributed, and the radius value of each circle is proportional to its distance (the radius value of that ring) from the central pixel. Therefore, there are four parameters that determine the

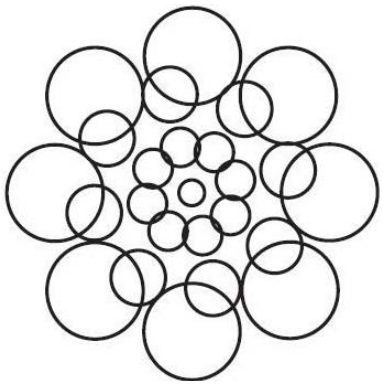


Fig. 5. Spatial pooling arrangement (DAISY-style in [16]) of the proposed HSOG descriptor.

spatial arrangement of the HSOG descriptor, *i.e.* the radius of the region (R); the number of quantized orientations (N); the number of concentric rings (CR); the number of circles on each ring (C). The influence of different parameters will be discussed experimentally in Section III.

Recall that the standard deviation of the Gaussian kernel is proportional to the size of the region in the HSOG descriptor. Specifically, it is defined as:

$$\sigma_i = \frac{R(i+1)}{2CR} \tag{10}$$

and circle locations are formulated in polar coordinates as:

$$r_i = \frac{R(i+1)}{CR}; \quad \theta_j = \frac{2\pi j}{C} \tag{11}$$

where i is the ith layer in the circular neighborhood; and j is the jth circle in each ring.

The total number of the divided circles can be calculated as $T = CR \times C + 1$. Within each circle CIR_j, $j = 1, 2, \ldots, T$, and for each first order OGM J_o, $o = 1, 2, \ldots, N$, a second order gradient histogram, h_{oj}, is constructed as (12) by accumulating the gradient magnitudes mag_o of all the pixels with the same quantized orientation entry n_o.

$$h_{oj}(i) = \sum_{(x,y)\in CIR_j} f(n_o(x,y) == i) * mag_o(x,y) \tag{12}$$

where $i = 0, 1, \ldots, N-1$; $o = 1, 2, \ldots, N$, $j = 1, 2, \ldots, T$,

$$f(x) = \begin{cases} 1, & \text{if } x \text{ is true} \\ 0, & \text{otherwise} \end{cases} \tag{13}$$

Then, for each first order OGM J_o, its second order gradient histogram h_o is generated by concatenating all the histograms from T circles:

$$h_o = [h_{o1}, h_{o2}, h_{o3}, \cdots, h_{oT}]^{\mathrm{T}} \tag{14}$$

where $o = 1, 2, \ldots, N$. The HSOG descriptor is obtained by concatenating all N histograms of the second order gradients as (15). Each histogram h_o is normalized to a unit norm vector $\hat{h}_o$ before the concatenation.

$$\mathrm{HSOG} = [\hat{h}_1, \hat{h}_2, \hat{h}_3, \cdots, \hat{h}_N]^{\mathrm{T}} \tag{15}$$

Some descriptors perform a weighting scheme during pooling, to highlight different contributions of pixels. For example, SIFT adopts the Gaussian-weighted gradient magnitude, and CS-LBP exploits the uniform weighting (i.e. without weighting). In order to control the computational cost of HSOG, we do not employ any weighting scheme as CS-LBP does.

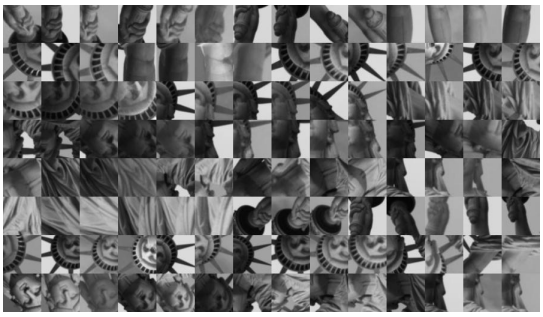


Fig. 6. Example image patches of the descriptor matching dataset.

III. Experimental Evaluation

We evaluate the proposed HSOG descriptor in three different applications: (1) *local image matching*; (2) *visual object categorization* (VOC); as well as (3) *scene classification*, and compare its performance with that of several state-of-the-art ones including SIFT [1], HOG [13], DAISY [18], and CS-LBP [14]. T3-S4 (steerable filters with DAISY-style spatial pooling) is only discussed in local matching [16]. These applications represent different levels of challenges. The major problems in local image matching are viewpoint and illumination variations whereas VOC implies to handle a number of additional ones, *i.e.* large scale changes, background clutter, occlusions, intra-class dissimilarities and inter-class similarities. Finally, scene classification requires to deal with all the challenges in VOC except scale change, since it is not so important as compared to VOC. Indeed, scene images are generally wide views that are captured adopting cameras with short focal lengths displaying large depth-of-fields. These experiments shed different light on the proposed HSOG descriptor and therefore provide insights into its discriminative power and complementarity to these first order related counterparts.

A. Experiments on Descriptor Matching

Local matching aims to match local features extracted from image patches of a scene and/or an object captured in different viewpoints and illumination conditions. It has a broad range of applications, including in particular wide-baseline matching [3] and image stitching [2]. In this experiment, we evaluate the discriminative power of HSOG and its complementarity in terms of descriptive completeness in comparison with several state of the art first order gradient-based descriptors, *e.g.*, SIFT [1], HOG [13], DAISY [18], as well as CS-LBP [14]. The best performance achieved in this work are further compared with the leading ones reported in literature on this database.

1) The NYL Dataset: We make use of the ground truth data provided by Brown *et al.* [16], which were generated for the purpose of local descriptor matching through multi-view stereo

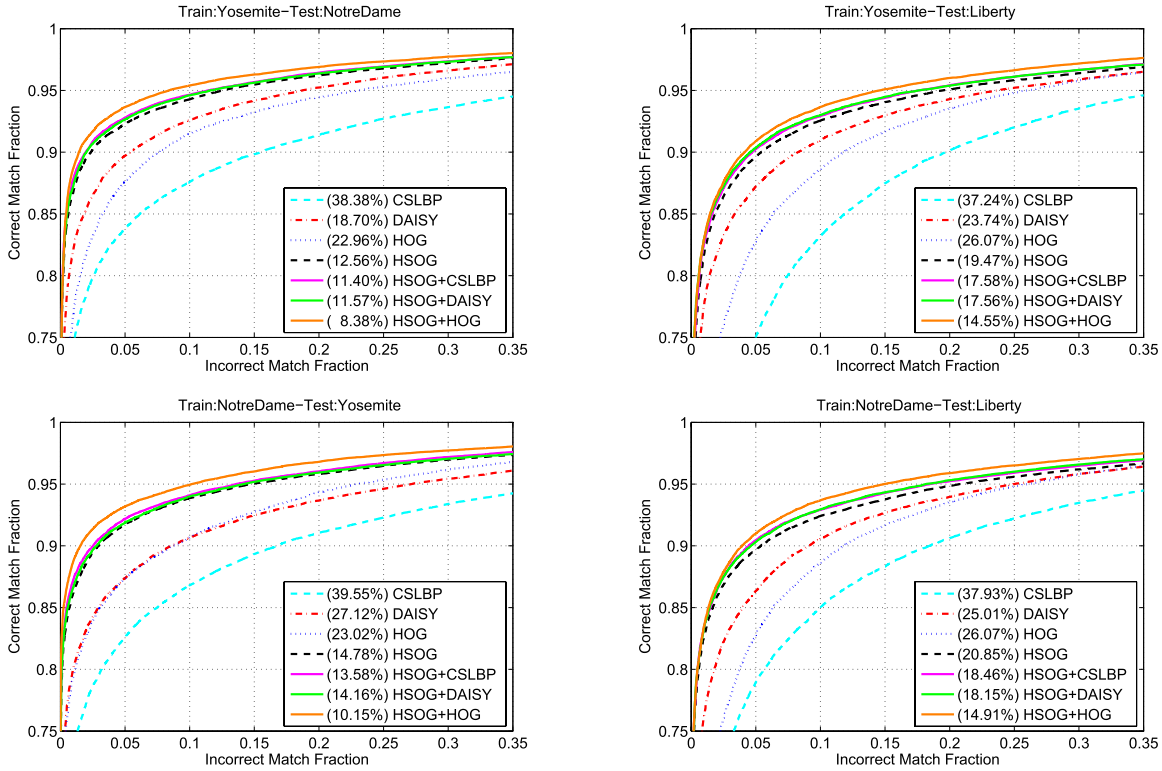


Fig. 7. ROC curves of matching results for HSOG and other state-of-the-art descriptors (with false positive rate at 95% recall in parenthesis).

correspondences from large 3D reconstructions. This database consists of approximately 2.5×10^6 labeled patches of Notre Dame (Paris), Half Dome (Yosemite), and the Statue of Liberty (New York). It is called in the subsequent as NYL. All patches are sampled to the size of 64×64 around each Difference of Gaussian (DoG) interest point with associated position, scale and orientation. Two patches are considered to be "matches" if their corresponding interest points are detected within 5 pixels of position, 0.25 scale octaves as well as $\pi/8$ radians in angle, while those outside 10 pixels of position, 0.5 octaves of scale, and $\pi/4$ radians in angle are defined to be "non-matches." The interest points detected between these ranges are considered to be ambiguous and not used. This dataset is publicly available online,[1] and some example image patches are shown in Fig. 6.

2) Experimental Setup: To conduct the experiments in local descriptor matching, we follow the settings as adopted in [16]. More precisely, four combinations of training and test set are used (the former one of each pair is the training set): Yosemite-Notre Dame, Yosemite-Liberty, Notre Dame-Yosemite, as well as Notre Dame-Liberty. The training sets contain from 10,000 to 500,000 patch pairs depending on various applications while the test sets always contain 100,000 patch pairs. The training and test sets have 50% "match" and 50% "non-match" pairs.

For each image patch, the proposed HSOG and other state-of-the-art descriptors are extracted respectively to represent its visual content. Then, for each patch pair in the training and test sets, we compute the Euclidean distance between their feature vectors and decide whether they are "matches" according to a threshold. When investigating the complementation of the 1st- and 2nd-order gradient based descriptors through combining HSOG with DAISY, HOG, CS-LBP, *etc.*, respectively, their similarity scores measured by normalized Euclidean distances are combined at a late stage using the simple sum rule in order not to increase the dimensionality of the feature space, and the final measurement is compared with a threshold. Therefore, by sweeping this threshold on the values of descriptor distances, we can obtain an ROC curve which plots the correctly detected matches as a fraction of all true matches against the incorrectly detected matches as a fraction of all true non-matches. In addition, we also compare the accuracies of different descriptors or their combinations in terms of the percentage of false matches present as 95% of all true matches are detected, *i.e.* the false positive rate at 95% recall.

For the state-of-the-art descriptors including DAISY, HOG, and CS-LBP, we optimize the parameters on the training set, and report their best performance for comparison. The detailed parameter values are as follows. CS-LBP: $\text{CS-LBP}_{1,8,0.01}$ with the 4×4 grid; DAISY: $R = 20$, $N = 8$, $CR = 3$, $C = 8$; HOG: 9 orientation bins in 0°-180° using the cell size of 8. For SIFT, we directly cite its performance from [16]. We tune the parameters of HSOG on the training set as well, and the detailed values are $R = 24$, $N = 8$, $CR = 3$, $C = 8$.

3) Experimental Results: The matching results at 95% recall and the corresponding ROC curves in comparison with the state of the art are shown in Table II and Fig. 7, respectively. In addition to these popular first order gradient-based descriptors, we also compare our results with the state of the art ones achieved on this dataset in the literature in Table III. It can be seen from the results that:

[1] http://www.cs.ubc.ca/~mbrown/patchdata/patchdata.html

TABLE II
COMPARISON OF DESCRIPTORS IN TERMS OF FALSE POSITIVE RATE (%) AT 95% RECALL
(YOS: YOSEMITE, ND: NOTRE DAME, LIB: LIBERTY)

Train	Test	SIFT [16]	CS-LBP	DAISY	HOG	T3h-S4-25 [16]	HSOG	HSOG+CS-LBP	HSOG+DAISY	HSOG+HOG
Yos	ND	26.10	38.38	18.70	22.96	14.43	12.56	11.40	11.57	**8.38**
Yos	Lib	35.09	37.24	23.74	26.07	20.48	19.47	17.58	17.56	**14.55**
ND	Yos	28.50	39.55	27.12	23.02	16.35	14.78	13.58	14.16	**10.15**
ND	Lib	35.09	37.93	25.01	26.07	21.85	20.85	18.46	18.15	**14.91**

TABLE III
COMPARISON OF BEST PERFORMANCES IN TERMS OF FALSE POSITIVE RATE (%) AT 95% RECALL
(YOS: YOSEMITE, ND: NOTRE DAME, LIB: LIBERTY)

Train	Test	Brown et al. [16]	Simonyan et al. [40]	Trzcinski et al. [41]	Boix et al. [42]	Trzcinski et al. [43]	This Work
Yos	ND	11.98	**7.11**	13.73	8.52	14.54	8.38
Yos	Lib	18.27	16.27	21.03	15.52	21.67	**14.55**
ND	Yos	13.55	10.36	15.86	**8.81**	18.97	10.15
ND	Lib	16.85	**13.63**	18.05	15.60	20.49	14.91

- In local shape representation through curvature related quantities, HSOG shows its effectiveness and outperforms the existing popular descriptors based on the first order gradients, *i.e.*, HOG, CS-LBP, DAISY, and SIFT by a large margin, hence clearly demonstrating its discriminative power.
- HSOG also surpasses T3h-S4-25, namely the parametric descriptor in [16] achieving the best accuracy based on the 4th-order steerable filters, whose parameters were optimized using the Powell's multidimensional direction set method. It suggests that the proposed HSOG descriptor better captures local shape information and thereby provides better discriminative power than the 4th order steerable filters.
- As first and second order gradient-based descriptors are related to different geometric properties of an image landscape, one can expect more comprehensive representation of a local shape when they are jointly used. Indeed, the fusion of HSOG with DAISY, HOG, or CS-LBP further improves the matching performance, indicating that HSOG provides complementary information to that conveyed by the first order gradient based descriptors, and their combination is thus a promising manner for visual content description.
- The fusion of HSOG with HOG produces better performance than that when HSOG is combined with other 1st-order gradient based descriptors, and its scores are even comparable to the state of the art ones reported on this dataset as shown in Table III that are always reached through comprehensive learning where a set of building blocks whose parameters or combinations are jointly optimized as in [16] and [40], etc.

4) Robustness to Noise: An important concern of the image descriptor when encoding the local shape information through curvature related second order gradient cues is its robustness or sensitivity to noise as images captured in real-life applications are more or less disrupted by noises for various reasons, *e.g.*, acquisition. In the design of HSOG, the first step is to generate first order Oriented Gradient Maps (OGMs) which are further smoothed by adopting a Gaussian kernel to allow the gradients to shift within the neighborhood without abrupt changes. This smoothing process should provide some robustness to noises. However, how does it behave facing different noise intensities? We aim to answer this problem in comparison with several other state of the art local image descriptors, including HOG, SIFT, DAISY, CS-LBP, and 2nd order Steerable Filter.

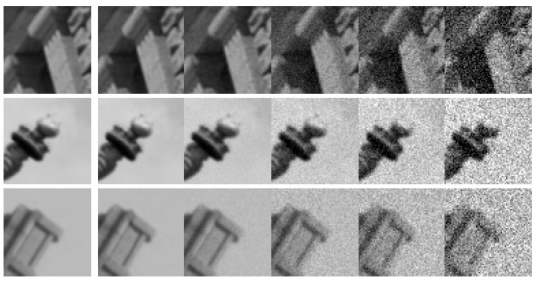


Fig. 8. Illustration of samples corrupted by white Gaussian noise in different intensities. First column presents the original image patches; the other columns from left to right show the same samples with noise added in 0db, −10db, −20db, −25db, and −30db SNR, respectively.

We follow the protocol as previously defined in Subsection III-A.2, and conduct local descriptor matching in four scenarios according to different training and test subsets. For each pair of image patches, we add random white Gaussian noises with varying intensities measured by the Signal Noise Ratio (SNR). Fig. 8 depicts several image patches corrupted by noises in different intensities.

Fig. 9 (a) to (d) illustrate the curves of the performance (at 95% recall) of these local descriptors with regard to increasing white Gaussian noise in different local matching scenarios.

To detailedly analyze the curves in these 4 local matching scenarios, we divide the variation of SNR into three different ranges: [8db, −8db], [−8db, −20db], and [−20db, −30db].

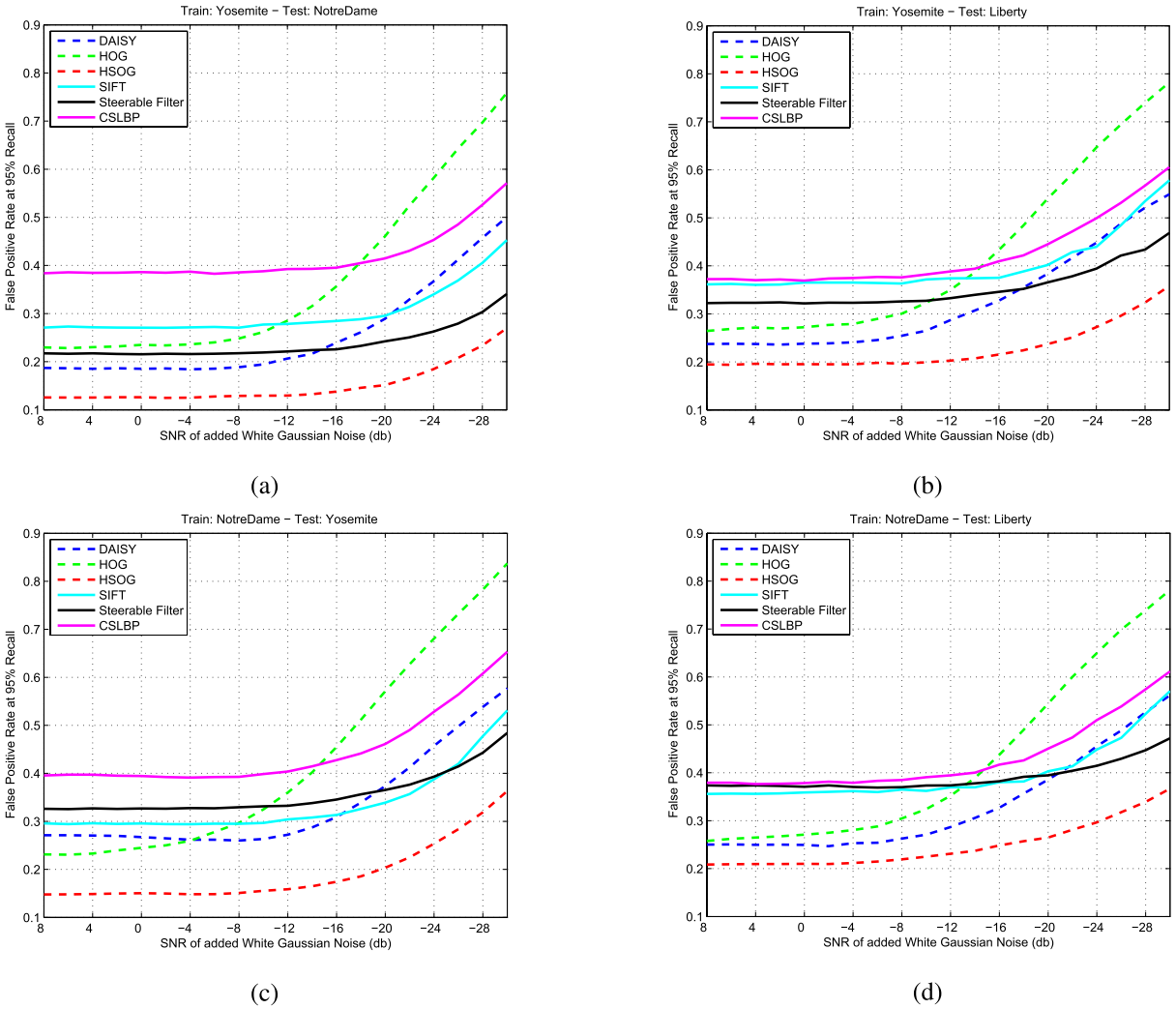


Fig. 9. Curves of matching performance (false positive rate at 95% recall) of different local image descriptors with regard to increasing white Gaussian noise (a) Yosemite-NotreDame, (b) Yosemite-Liberty, (c) NotreDame-Yosemite, and (d) NotreDame-Liberty.

- SNR in [8db, −8db]

The performance of HSOG, Steerable Filter (SF), SIFT, CS-LBP, and DAISY generally remains stable and that of HOG slightly drops, which indicates that these local descriptors are not seriously impacted by weak noise.

- SNR in [−8db, −20db]

The accuracies of all these local descriptors are obviously degraded, and they change in different styles, indicating that the robustness to noise of each local descriptor is different. The sensitivity of the local descriptor is illustrated by the slope of the curve. We can see that SIFT possesses good robustness to noise thanks to its histogram statistics of gradient distribution, while HSOG achieves a comparable robustness as SIFT does for its map-based manner of gradient calculation as well as its overlapped arrangement of spatial pooling. SF also presents good robustness, since in this experiment it is computed using the map-based gradient generation as in HSOG, showing the necessity of such a process to decrease the sensitivity to noise. Furthermore, in CS-LBP, this robustness on flat image regions is obtained by thresholding the gray level differences with a small value T (set at an experimentally optimized value 0.01). The robustness of these four descriptors is superior to that of DAISY and HOG. Since DAISY adopts Gaussian smoothing and a similar pooling strategy as HSOG, its robustness is not far behind. The robustness of HOG is the worst and its slope is quite large as the noise increases. We think that this sensitivity is mainly caused by its simple sampling grid.

- SNR in [−20db, −30db]

In this range, the slopes of these curves are similar, which are all quite large, illustrating that when the additive noise is strong enough, the robustness of local image descriptors does not make sense any more. Their errors dramatically increase.

To sum up, HSOG not only achieves the best performance in local matching, but also owns good robustness to noise, in comparison with these first order gradient based descriptors.

B. Experiments on Object Categorization

The previous part shows that HSOG depicts high discriminative power in local image matching, and proves complementary to state of the art first order gradient-based descriptors in terms of descriptive completeness. But how does HSOG behave when it is applied to visual object categorization

Fig. 10. Illustration of intra-class dissimilarities and inter-class similarities. (a) Examples are all from the class of speed boat in the Caltech 256 database, but possess very different appearances. (b) Images in the first row are from the bike class of the Caltech 256 database, while the ones in the second row are from the class of motorbike in the same dataset, and they are very similar in appearance.

(VOC)? This is a much harder problem since it implies to deal with, in addition to viewpoint and lighting variations, those challenges such as scale changes, background clutter, occlusions, intra-class dissimilarities and inter-class similarities as illustrated in Fig.10. We make use of the bag-of-visual words approach [44] which requires more stages as compared to local matching. It additionally consists of building a dictionary of visual words using training data, encoding of sparsely or densely sampled local descriptors, and classification.

1) The Caltech 101 and Caltech 256 Datasets: We evaluate the proposed HSOG descriptor in the context of visual object categorization (VOC) on two standard databases: Caltech 101 [45] and Caltech 256 [46]. Caltech 101 contains a total number of 9146 images split into 101 different object classes including chairs, faces, airplanes, animals, vehicles, flowers, etc. and an additional background category. The number of images in each class varies from 31 to 800, and most categories have about 50 images. As an extension of Caltech 101, Caltech 256 consists of 30607 images from 256 object categories and an additional clutter category. Each of categories contains at least 80 images. Compared with Caltech 101, Caltech 256 is more challenging because it contains more categories and presents higher inter-class similarities and larger intra-class dissimilarities in object scale, location, viewpoint, *etc.*

2) Experimental Setup: We follow the method whose general flowchart is illustrated in Fig. 11 for object categorization.

For each image in the database, these HSOG based features are extracted from a dense grid with a 6-pixel spacing.

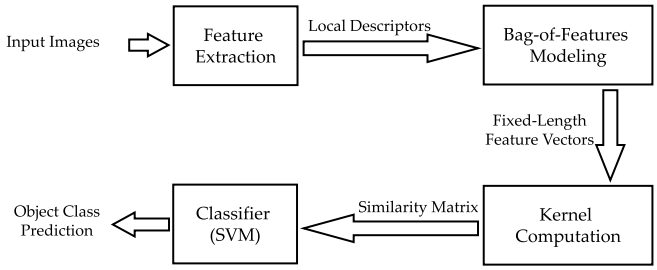


Fig. 11. Flowchart of our approach for visual object categorization.

It is in contrast to our preliminary work [38] where sparse sampling was performed by using the Harris-Laplace keypoint detector [47]. Nevertheless, the dense sampling technique consistently displays better performance than sparse sampling in VOC. As a result, we apply it and make comparison of these descriptors. In the configuration of a dense sampling, SIFT is quite similar to HOG, and we therefore do not provide the performance of HOG in this experiment. We implement the CS-LBP descriptor according to [14], and use the source code available online for computing SIFT[2] and DAISY.[3]

The visual content of images is modeled using the popular Bag-of-Features (BoF) framework [44] which achieves great success in the VOC task. The main idea of BoF is to represent an image as an orderless collection of local image descriptors. More precisely, a visual vocabulary is initially constructed by introducing a clustering algorithm on training data, and each cluster center is considered as a visual word in the vocabulary. Instead of hard assignment that all these descriptors extracted from a given image are quantized to their closest visual word in an appropriate metric space, we make use of a more effective soft assignment strategy, namely Locality-constrained Linear Coding (LLC) [48], which employs the locality constraints to project each of the descriptors into its local-coordinate system, and the resulted coordinates are integrated by max pooling to generate the final representation. The number of the descriptors assigned to each visual word is accounted into a histogram as the final BoF based representation.

The Support Vector Machine (SVM) algorithm is applied in classification. When all the local descriptors are transformed to fixed length feature vectors by the BoF modeling, the linear kernel function is utilized for the SVM training and prediction, as non-linear information has been already included in LLC.

Finally, each of test images is classified into the object class with the maximum SVM output decision value. We tune these parameters of the classifier by using the training set via 5-fold cross-validation, and obtain the accuracy on the test set.

To conduct the experiments on the Caltech 101 and Caltech 256 datasets, we follow the common training and test settings as in [49], [50], and 2007Caltech256Griffin. For Caltech 101, 15 and 30 images per category are randomly selected for training respectively, while another 15 random images for testing (except for the categories containing less than 45 images). For Caltech 256, 30 images are randomly chosen for training while the other 25 random images for testing

[2]http://www.vlfeat.org
[3]http://cvlab.epfl.ch/software/daisy

respectively from each category. We report the recognition accuracy on all the 102 classes of Caltech 101 averaged over three splits, and on 256 classes of Caltech 256 (excluding the clutter category) for a single split. In our case, a vocabulary of 1024 and 4000 visual words is constructed for each kind of local descriptors on Caltech 101 and Caltech 256 respectively by applying the k-means clustering algorithm on a subset of the descriptors randomly selected from the training data as in [10].

3) Parameter Selection: Recall that HSOG has 4 parameters: the radius of the region area (R); the number of quantized orientations (N); the number of concentric rings (CR); and the number of circles on each ring (C). To evaluate their impacts on the performance of the descriptor, we draw a series of line graphs of the recognition accuracy on Caltech 101 (15 training images per class) for different R by alternately changing one parameter while fixing the others for N, CR, and C. These results are shown in Fig. 12.

It can be observed in Fig. 12(a) that the HSOG descriptors with 8 orientations perform clearly better than those with 4 and 6; whilst the ones with 10 orientations present no superiority to those with 8, demonstrating that 8 orientations are sufficient to describe local image variations. From Fig. 12(b), we also see that the performance keeps improving when the number of concentric rings increases, illustrating that the descriptor based on more rings is better, since more neighboring information is included. When we keep increasing the number of concentric rings to 4 or 5, the performance improvement is more and more limited, but the dimensionality of the HSOG feature increases dramatically. To control the computational cost of HSOG, we have to make a trade-off and hence set this number at 3. Fig. 12(c) shows that raising the number of the circles on each ring does not improve the performance, implying that large number of circles on each ring are unnecessary, due to overlapping of adjacent regions.

Another phenomenon from the three figures lies in that the best performance is achieved when R is set at 15. Therefore, we choose the best parameter setting for the proposed HSOG descriptor as follows: $R = 15$, $N = 8$, $CR = 3$, and $C = 4$.

4) Influence of PCA-Based Dimensionality Reduction: The dimension of HSOG is $(CR \times C + 1) \times N^2$, which is relatively high for the following classification processes (*e.g.* 832 using the best parameter setting). In order to reduce its dimensionality, we apply the well known Principal Component Analysis (PCA) technique, since it has been successfully applied in the PCA-SIFT and GLOH cases for the same objective.

We observe the change in performance with varying dimension using the same protocol as in parameter selection, *i.e.* on Caltech 101 with 15 random training images per category. To build the eigenspace, we localize 76,000 local image patches on a diverse collection of images that belong to Caltech 256 for validation. Each of these patches is exploited to compute its HSOG descriptor, and PCA is then applied on the covariance matrix of these descriptors. The matrix consisting of the top n eigenvectors is stored and utilized as the projection matrix.

For a certain local image region, its HSOG descriptor is first computed and then projected into a low-dimensional feature space by multiplying the pre-generated projection matrix. The dimensionality of the final HSOG descriptor is hence reduced to n. We experimentally evaluate the

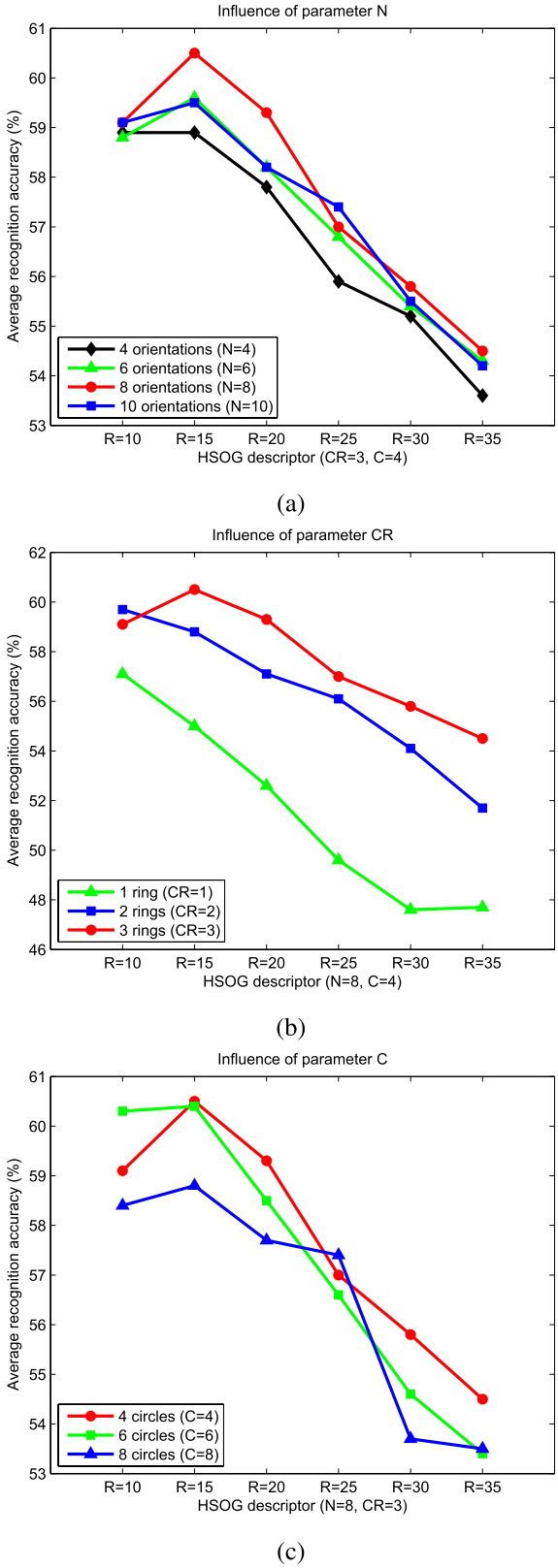


Fig. 12. Influence of different parameters in HSOG. (a) number of quantized orientations N; (b) number of concentric rings CR; and (c) number of circles on each ring C.

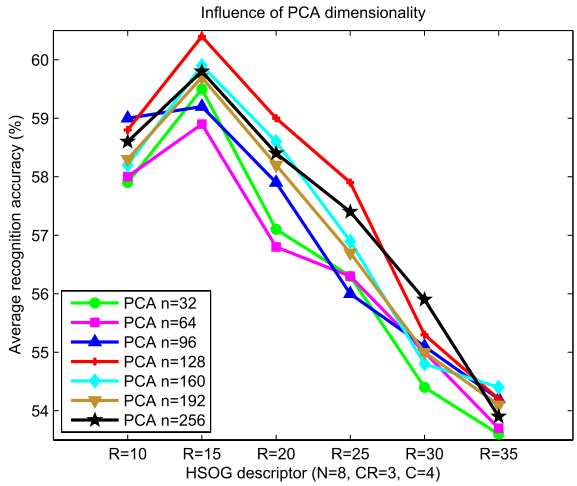


Fig. 13. Influence of the PCA-based dimensionality reduction of the proposed HSOG descriptor.

impact of different values of n for the HSOG performance on Caltech 101. A series of curves of the recognition accuracy based on different region sizes are produced by varying the dimensionality n calculated by PCA from 32 to 256, as shown in Fig. 13.

We calculate the mean and deviation value of these HSOG descriptors over all radii for each given dimension n, and find out that the performance of 128-dimensional HSOG features (57.60 ± 2.37) is better than those of the other choices, such as 32 dimension (56.47 ± 2.20); 64 dimension (56.45 ± 1.91); 96 dimension (56.90 ± 2.10); 160 dimension (57.13 ± 2.19); 192 dimension (57.00 ± 2.14); and 256 dimension (57.33 ± 2.13). Therefore, 128 is a good alternative for the dimensionality of HSOG, and we set $n = 128$ in the following experiments.

The performance difference before and after dimensionality reduction can be measured by these red curves in Fig. 12 (*i.e.* before PCA; $N = 8$, $CR = 3$, $C = 4$, R varies from 10 to 35 with an interval of 5, the performance is 59.1%, 60.5% 59.3%, 57.0%, 55.8%, and 54.5%, respectively) and the red curve in Fig. 13 (*i.e.* after PCA with the same parameter configuration, the classification rate is 58.8%, 60.4%, 59.0%, 57.9%, 55.3%, and 54.2%, respectively). While the performances before and after PCA are rather comparable, the length of the feature is reduced by a factor of 6.5, from 832 to 128, which saves time in the following classification.

5) Performance Evaluation and Comparison: We evaluate the proposed HSOG descriptor with the best parameter setting on the Caltech 101 and Caltech 256 databases, and compare its performance with that of several state-of-the-art ones including SIFT, DAISY, and CS-LBP. It should be noted that in this experiment the parameters of all other descriptors are best tuned except SIFT (its standard configuration is usually regarded as the best in literature and its parameters are seldom changed). The parameter setting of HSOG is $R = 15$, $N = 8$, $CR = 3$, and $C = 4$, with the dimensionality of 128. Additionally, we build the PCA subspace across databases of Caltech 101 and Caltech 256. SIFT utilizes the standard configuration as in [1], thus with 128-dimension.

TABLE IV
PERFORMANCE AND CONSUMED TIME COMPARISON BETWEEN HSOG AND STATE-OF-THE-ART DESCRIPTORS ON CALTECH 101 USING 15 AND 30 TRAINING IMAGES PER CLASS AND ON CALTECH 256 USING 30 TRAINING IMAGES PER CLASS

Descriptor	Cal-101 (%)	Cal-256 (%)	Time (s)
SIFT	**62.48 / 69.89**	**34.58**	0.316
DAISY	58.63 / 67.01	31.02	0.108
CS-LBP	58.50 / 66.86	34.34	0.087
HSOG	60.46 / 67.97	31.14	0.985
HSOG+SIFT	64.05 / **72.99**	36.55	-
HSOG+DAISY	63.46 / 71.73	34.70	-
HSOG+CS-LBP	**64.71** / 72.47	**37.36**	-
All 4 Descriptors	67.06 / 75.13	41.20	-
Yang et al. [51]	- / 73.20	34.02	-
Gao et al. [52]	- / -	35.74	-
Wang et al. [48]	- / 73.44	41.19	-
Liu et al. [53]	- / 74.21	-	-
Shabou & Borgne [54]	- / 73.23	-	-
Kumar et al. [55]	69.10 / 77.20	44.80	-
Gehler & Nowozin [56]	- / 77.70	45.80	-

DAISY applies the parameter setting as $R = 15$, $N = 8$, $CR = 3$, and $C = 4$, and its dimension is 104. The parameters of CS-LBP are set as CS-LBP$_{2,8,0.01}$ with the 4×4 grid, resulting in a 256-dimensional descriptor.

We can see from Table IV that:

- The HSOG descriptor achieves the second best accuracy on Caltech 101, inferior to SIFT but superior to DAISY and CS-LBP, and it reaches the third best result on Caltech 256, inferior to SIFT and CS-LBP but superior to DAISY. These results indicate that HSOG contributes to solving the problem of VOC. Meanwhile, to capture the geometric properties of an image interpreted as a landscape or surface through curvature related quantities (*i.e.*, second order gradients), HSOG is theoretically sensitive to scale variations, which can be observed from the performance difference between HSOG and some of its first order gradient based counterparts.
- Because HSOG captures different local geometric information as compared to first order gradient related descriptors, one could expect the recognition accuracy be improved when they are jointly used and fused through a late fusion strategy. This is indeed the case as we can see in this table. As HSOG is combined with individual first order gradient based descriptors (*i.e*, SIFT, DAISY, and CS-LBP), categorization performance consistently increases.

  At the same time, it is worth noting that, even though SIFT, CS-LBP, and DAISY all convey first order gradient clues, they are still complementary to each other due to their differences in gradient computation, pooling, *etc.*; and thus could improve the final performance when they are combined. However, when HSOG is further added, one can expect extra performance gain again for its additional geometric information.

Both facts indicate that the information that HSOG provides is complementary to that of first order gradients. A consequent question lies in the statistical significance of the performance gain displayed in Table IV as HSOG is used in addition to other first order gradient related descriptors. For this purpose, we carry out an additional experiment on Caltech 101 by using 10 randomly selected splits instead of 3 splits as in the standard protocol. The mean recognition accuracies and the standard deviations computed over the 10 splits show that the results accord with the ones in Table IV. Additionally, we also perform the StudentâĽ™s t-test to check the statistical significance of the performance improvement for each pair of classifiers, *e.g.*, "CS-LBP+DAISY" vs. "CS-LBP+DAISY+HSOG"; "SIFT+CS-LBP" vs. "SIFT+CS-LBP+HSOG" *etc.*, and validate that the gain is indeed statistically significant.

- The combination of all the four descriptors, *i.e*, HSOG, SIFT, DAISY, and CS-LBP, achieves the best accuracies both on Caltech 101 and Caltech 256. Such scores are among the leading results reported by state of the art systems in the literature. It is worth noting that further performance improvement does not only depend on the features, but also requires more advanced and complex kernel combination techniques as did in [56] and [55].

At the same time, we calculate the average computational time required on the Caltech 101 database for each input image (about the size of 300×250) through different local descriptors using an Intel Core 2 Duo CPU @ 3.16 GHz with 3GB RAM, and it can be seen that the current version of HSOG is 3 times slower than SIFT. Nevertheless, it should be noted that since each first order OGM and its second order gradients can be computed individually, the current implementation of HSOG can be accelerated by GPU programming, which should make the computation of HSOG approximately N times faster (N is the number of OGMs, and 8 in our case), thereby displaying a runtime comparable to the existing ones.

C. Experiments on Scene Classification

The previous experiment in the task of VOC benchmarked HSOG in unconstrained conditions, with these challenging factors including in particular large scale changes in addition to viewpoint and lighting variations, background clutter, *etc.*. In order to gain additional insights into the geometric properties captured by HSOG, we also experiment it using the OT scene dataset consisting of scene images which are generally wide views captured by cameras with short focal lengths displaying large depth-of-fields. The categorization of an given image into a scene class is thus less impacted by scale changes. How does HSOG behave in such a task of an intermediate difficulty in comparison with other state of the art local descriptors? We answer this question in this subsection.

1) The OT Scene Dataset: The proposed HSOG descriptor is evaluated on a dataset from Oliva and Torralba [5], namely the OT database, for the last application of scene classification. It totally consists of 2,688 images from 8 scenery categories: including coast (360 samples), forest (328 samples), mountain (374 samples), highway (260 samples), inside city (308 samples), tall building (356 samples), open country (410 samples), and street (292 samples). Fig. 14 shows some sample images of each category.

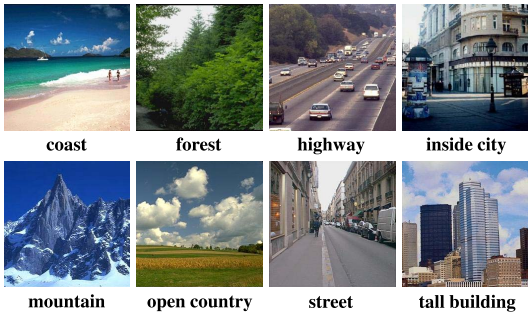


Fig. 14. Example images of the OT scene dataset.

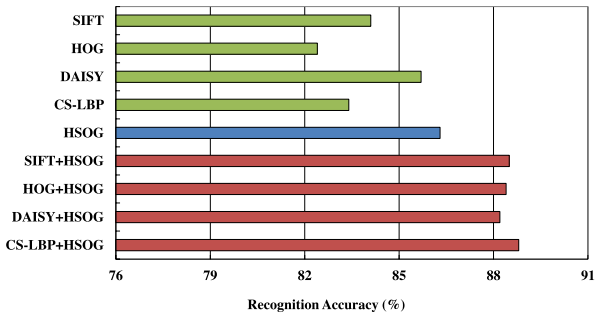


Fig. 15. Classification results on the OT scene dataset.

2) Experimental Setup: For the task of scene classification, our approach is the same as the one used in VOC, as described in Section III-B.2, Once again, the dense sampling strategy is applied to locate keypoints for local feature computation. Such a sampling strategy is expected, because the classification of an image into a scene class mostly requires the whole content of an image, rather than on the "object" region only. Specifically, the sampling spacing is set to 6 pixels, resulting in about 1,700 interest points per image. A visual vocabulary of 2,000 "visual words" is constructed for each kind of local descriptor to build the corresponding Bag-of-Features (BoF) representation.

The parameters of these descriptors in comparison are tuned so that they perform in their best conditions, HSOG: $R = 15$, $N = 8$, $CR = 3$, $C = 4$; CS-LBP: CS-LBP$_{2,8,0.01}$ with the 4×4 grid; DAISY: $R = 15$, $N = 8$, $CR = 3$, $C = 4$; HOG: 9 orientation bins in 0°-180° with the cell size of 8. For SIFT, we also use its standard setting.

We randomly choose half of these images from each scenery category for training, while the other half for test. The recognition accuracy is adopted as the evaluation criterion. We tune these parameters of the classifier on the training set via 5-fold cross-validation, and obtain the classification results on the test set.

3) Experimental Results: The classification results achieved on the OT scene dataset are displayed in Fig. 15. As it can be seen from this figure, in capturing curvature related local shape information, HSOG outperforms all these first order gradient-based local descriptors, with the accuracy gain reaching 3.9% over HOG; 2.9% over CS-LBP; 2.2% over SIFT; and 0.6%

TABLE V
COMPARISON OF THE HSOG DESCRIPTOR (MULTI-SCALE VS. SINGLE SCALE) ON CALTECH 101 USING 15 AND 30 TRAINING IMAGES PER CLASS; CALTECH 256 USING 30 TRAINING IMAGES PER CLASS; AND OT USING HALF OF IMAGES PER CLASS FOR TRAINING

Type	Radius R	Cal-101 (%)	Cal-256 (%)	OT (%)
Single-scale	**10**	58.82 / 67.82	30.56	86.1
	15	**60.46 / 67.97**	**31.14**	**86.3**
	20	59.02 / 67.13	30.22	85.9
	25	57.86 / 66.20	28.86	85.9
Multi-scale	**10 to 25**	**61.90 / 70.18**	**33.81**	**87.4**

over DAISY. This accuracy holds an improvement of 2.6% to that of GIST (83.7%) [5] as well. Furthermore, HSOG proves once more to provide complementary descriptive information with respect to its first order gradient related counterparts, *e.g.*, SIFT, CS-LBP, DAISY, and HOG. Their joint use improves the classification precision, and the performance improvement by combining HSOG and the first order gradient-based descriptors is more than 2 points over HSOG itself and is 4.5 points on average over these first order gradient-based counterparts. These accuracies are in line with those findings in the previous two applications, regarding its discriminative power as well as complementarity to the first order gradient based descriptors.

The best classification accuracy attains 88.8% when HSOG is combined with CS-LBP. This result outperforms the state of the art ones on this dataset, such as 86.65% [57] and 87.8% [26].

D. Summary and Discussion

HSOG captures curvature related local geometric properties which are different from those of the first order gradient related local descriptors. The experimental accuracies in the previous three applications show that HSOG conveys very useful clues which are complementary to that of first order gradient based descriptors in image representation. Furthermore, HSOG tends to extract more discriminative information than its first order gradient-based counterparts and thereby outperforms the latter ones in both the applications of descriptor matching and scene classification where scale changes are limited. While, HSOG indeed presents some sensitivity to scale variations, losing its lead on VOC. The reason lies in that HSOG encodes geometric properties of an image that is interpreted as a surface through curvature related quantities, and it is theoretically sensitive to scale changing. We can image that the earth with mountains, cliffs, valleys, etc. becomes a simple sphere when observed at a large distance.

To improve the robustness of HSOG to severe scale variations, a straightforward way is to consider support regions of various sizes of a given interest point to embed more geometric information. In such a case, the multi-scale strategy is a direct alternative. We introduce the late fusion strategy to combine different HSOG features of multi-scale regions, since it does not increase the dimensionality of feature space, and the similarity scores achieved by different parameters can be calculated individually, leading to a high feasibility in implementation of parallel computing without largely increasing time cost. The kernel matrices of different HSOG descriptors are combined using the Multiple Kernel Learning (MKL) algorithm [58] for decision.

From the preliminary results listed in Table V, we can see that the performance of these single scale HSOG descriptors is improved roughly by 2 points when more scales are combined on both Caltech 101 (15 and 30 training images each class) and Caltech 256 (30 training images each class). The accuracy improvement on the OT scene dataset is about 1 point. These results demonstrate that, facing large scale changes of visual objects, the multi-scale HSOG is an effective solution. On the other hand, the improvement in scene classification is not as significant as in VOC, showing its limited scale variations.

IV. CONCLUSION

In this paper, we present a novel and effective local image descriptor, namely HSOG, by making use of the Histograms of the Second Order Gradients to capture curvature related local geometric properties. These experimental results achieved on three applications with different levels of challenges (descriptor matching, object categorization, scene image classification) indicate that the HSOG descriptor owns a good discriminative power to distinguish different visual contents, especially embedded with more spatial information provided by the multi-scale strategy. Moreover, the information conveyed by HSOG proves complementary to that captured by state of the art first order gradient based local image descriptors, *e.g.*, HOG, SIFT, CS-LBP, and DAISY.

In future work, we continue to go a step in capturing local geometric properties of images interpreted as landscapes and employ true differential geometry quantities which are intrinsically densely computable and rotation invariant, *e.g.*, Gaussian and mean curvatures, and make the resultant descriptor scale invariant. Meanwhile, we will investigate how to make full use of HSOG to improve the accuracy of scene classification on a more comprehensive database, *e.g.* the SUN database [59].

REFERENCES

[1] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.
[2] M. Brown and D. G. Lowe, "Automatic panoramic image stitching using invariant features," *Int. J. Comput. Vis.*, vol. 74, no. 1, pp. 59–73, 2007.
[3] T. Tuytelaars and L. Van Gool, "Matching widely separated views based on affine invariant regions," *Int. J. Comput. Vis.*, vol. 59, no. 1, pp. 61–85, 2004.
[4] N. Liu *et al.*, "Multimodal recognition of visual concepts using histograms of textual concepts and selective weighted late fusion scheme," *Comput. Vis. Image Understand.*, vol. 117, no. 5, pp. 493–512, 2013.
[5] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vis.*, vol. 42, no. 3, pp. 145–175, 2001.
[6] C. Zhu, C.-E. Bichot, and L. Chen, "Image region description using orthogonal combination of local binary patterns enhanced with color information," *Pattern Recognit.*, vol. 46, no. 7, pp. 1949–1963, 2013.
[7] J. Philbin, M. Isard, J. Sivic, and A. Zisserman, "Descriptor learning for efficient retrieval," in *Proc. ECCV*, 2010, pp. 677–691.
[8] Y. Ke and R. Sukthankar, "PCA-SIFT: A more distinctive representation for local image descriptors," in *Proc. IEEE CVPR*, Jun./Jul. 2004, pp. II-506–II-513.

[9] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 10, pp. 1615–1630, Oct. 2005.

[10] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek, "Evaluating color descriptors for object and scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1582–1596, Sep. 2010.

[11] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 4, pp. 509–522, Apr. 2002.

[12] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (SURF)," *Comput. Vis. Image Understand.*, vol. 110, no. 3, pp. 346–359, 2008.

[13] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE CVPR*, Jun. 2005, pp. 886–893.

[14] M. Heikkila, M. Pietikainen, and C. Schmid, "Description of interest regions with local binary patterns," *Pattern Recognit.*, vol. 42, no. 3, pp. 425–436, 2009.

[15] E. Tola, V. Lepetit, and P. Fua, "DAISY: An efficient dense descriptor applied to wide-baseline stereo," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 5, pp. 815–830, May 2010.

[16] M. Brown, H. Gang, and S. Winder, "Discriminative learning of local image descriptors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 1, pp. 43–57, Jan. 2011.

[17] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded up robust features," in *Proc. ECCV*, 2006, pp. 404–417.

[18] E. Tola, V. Lepetit, and P. Fua, "A fast local descriptor for dense matching," in *Proc. IEEE CVPR*, Jun. 2008, pp. 1–8.

[19] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek, "Evaluating color descriptors for object and scene recognition," in *Proc. CVPR*, Jun. 2008, pp. 1–8.

[20] C. Zhu, C.-E. Bichot, and L. Chen, "Multi-scale color local binary patterns for visual object classes recognition," in *Proc. 20th ICPR*, Aug. 2010, pp. 3065–3068.

[21] I. Kokkinos and A. Yuille, "Scale invariance without scale selection," in *Proc. IEEE CVPR*, Jun. 2008, pp. 1–8.

[22] T. Hassner, V. Mayzels, and L. Zelnik-Manor, "On SIFTs and their scales," in *Proc. IEEE CVPR*, Jun. 2012, pp. 1522–1528.

[23] F. Moreno-Noguer, "Deformation and illumination invariant feature point descriptor," in *Proc. IEEE CVPR*, Jun. 2011, pp. 1593–1600.

[24] P. Ott and M. Everingham, "Implicit color segmentation features for pedestrian and object detection," in *Proc. IEEE ICCV*, Sep./Oct. 2009, pp. 723–730.

[25] E. Trulls, I. Kokkinos, A. Sanfeliu, and F. Moreno-Noguer, "Dense segmentation-aware descriptors," in *Proc. IEEE CVPR*, Jun. 2013, pp. 2890–2897.

[26] A. Bosch, A. Zisserman, and X. Muoz, "Scene classification using a hybrid generative/discriminative approach," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 4, pp. 712–727, Apr. 2008.

[27] J. van de Weijer, T. Gevers, and A. D. Bagdanov, "Boosting color saliency in image feature detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 1, pp. 150–156, Jan. 2006.

[28] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.

[29] T. Ahonen, A. Hadid, and M. Pietikainen, "Face recognition with local binary patterns," in *Proc. ECCV*, 2004, pp. 469–481.

[30] C. Zhu, C.-E. Bichot, and L. Chen, "Visual object recognition using DAISY descriptor," in *Proc. IEEE ICME*, Jul. 2011, pp. 1–6.

[31] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid, "Local features and kernels for classification of texture and object categories: A comprehensive study," *Int. J. Comput. Vis.*, vol. 73, no. 2, pp. 213–238, 2007.

[32] J. Li and N. M. Allinson, "A comprehensive review of current local features for computer vision," *Neurocomputing*, vol. 71, nos. 10–12, pp. 1771–1787, 2008.

[33] D. H. Hubel and T. N. Wiesel, "Brain mechanisms of vision," *Sci. Amer.*, vol. 241, no. 3, pp. 150–162, 1979.

[34] M. J. Morgan, "Features and the 'primal sketch'," *Vis. Res.*, vol. 51, no. 7, pp. 738–753, 2011.

[35] S. A. Wallis and M. A. Georgeson, "Mach bands and multiscale models of spatial vision: The role of first, second, and third derivative operators in encoding bars and edges," *J. Vis.*, vol. 12, no. 13, pp. 1–25, 2012.

[36] W. Kuhnel, *Differential Geometry: Curves, Surfaces, Manifolds*. Providence, RI, USA: AMS, 2005.

[37] D. H. Hubel and T. N. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," *J. Physiol.*, vol. 160, no. 1, pp. 106–154, 1962.

[38] D. Huang, C. Zhu, C.-E. Bichot, Y. Wang, and L. Chen, "HSOG: A novel local descriptor based on histograms of second order gradients for object categorization," in *Proc. ICMR*, 2013, pp. 199–206.

[39] D. Huang, M. Ardabilian, Y. Wang, and L. Chen, "Oriented gradient maps based automatic asymmetric 3D-2D face recognition," in *Proc. 5th ICB*, Mar./Apr. 2012, pp. 125–131.

[40] K. Simonyan, A. Vedaldi, and A. Zisserman, "Descriptor learning using convex optimisation," in *Proc. ECCV*, 2012, pp. 243–256.

[41] T. Trzcinski, M. Christoudias, V. Lepetit, and P. Fua, "Learning image descriptors with the boosting-trick," in *Proc. NIPS*, 2012, pp. 278–286.

[42] X. Boix, M. Gygli, G. Roig, and L. Van Gool, "Sparse quantization for patch description," in *Proc. IEEE CVPR*, Jun. 2013, pp. 2842–2849.

[43] T. Trzcinski, M. Christoudias, P. Fua, and V. Lepetit, "Boosting binary keypoint descriptors," in *Proc. IEEE CVPR*, Jun. 2013, pp. 2874–2881.

[44] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *Proc. ECCV Workshop*, 2004, pp. 1–2.

[45] F.-F. Li, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories," in *Proc. IEEE CVPRW*, Jun. 2004, p. 178.

[46] G. Griffin, A. Holub, and P. Perona, "Caltech-256 object category dataset," California Inst. Technol., Pasadena, CA, USA, Tech. Rep. CNS-TR-2007-001, 2007.

[47] K. Mikolajczyk and C. Schmid, "Scale and affine invariant interest point detectors," *Int. J. Comput. Vis.*, vol. 60, no. 1, pp. 63–86, 2004.

[48] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *Proc. IEEE CVPR*, Jun. 2010, pp. 3360–3367.

[49] H. Zhang, A. C. Berg, M. Maire, and J. Malik, "SVM-KNN: Discriminative nearest neighbor classification for visual category recognition," in *Proc. IEEE CVPR*, Jun. 2006, pp. 2126–2136.

[50] M. Varma and D. Ray, "Learning the discriminative power-invariance trade-off," in *Proc. IEEE 11th ICCV*, Oct. 2007, pp. 1–8.

[51] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *Proc. IEEE CVPR*, Jun. 2009, pp. 1794–1801.

[52] S. Gao, I. W. Tsang, and L. Chia, "Laplacian sparse coding, hypergraph Laplacian sparse coding, and applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 92–104, Jan. 2013.

[53] L. Liu, L. Wang, and X. Liu, "In defense of soft-assignment coding," in *Proc. IEEE ICCV*, Nov. 2011, pp. 2486–2493.

[54] A. Shabou and H. LeBorgne, "Locality-constrained and spatially regularized coding for scene categorization," in *Proc. IEEE CVPR*, Jun. 2012, pp. 3618–3625.

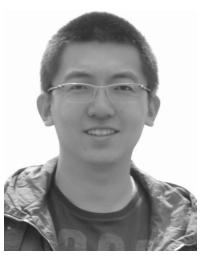
[55] A. Kumar, A. Niculescu-Mizil, K. Kavukcuoglu, and H. Daumé, III, "A binary classification framework for two-stage multiple kernel learning," in *Proc. ICML*, 2012, pp. 1–8.

[56] P. Gehler and S. Nowozin, "On feature combination for multiclass object classification," in *Proc. IEEE 12th ICCV*, Sep./Oct. 2009, pp. 221–228.

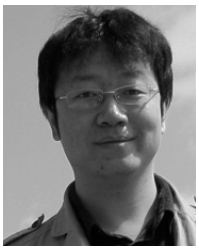
[57] A. Bosch, A. Zisserman, and X. Munoz, "Scene classification via pLSA," in *Proc. ECCV*, 2006, pp. 517–530.

[58] A. Rakotomamonjy, F. Bach, S. Canu, and Y. Grandvalet, "Simplemkl," *J. Mach. Learn. Res.*, vol. 9, pp. 2491–2521, Nov. 2008.

[59] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "SUN database: Large-scale scene recognition from abbey to zoo," in *Proc. IEEE CVPR*, Jun. 2010, pp. 3485–3492.



Di Huang (S'10–M'11) received the B.S. and M.S. degrees in computer science from Beihang University, Beijing, China, and the Ph.D. degree in computer science from the École centrale de Lyon, Lyon, France, in 2005, 2008, and 2011, respectively. He joined the Laboratory of Intelligent Recognition and Image Processing with the Beijing Key Laboratory of Digital Media, School of Computer Science and Engineering, Beihang University, as a Faculty Member. His current research interests include biometrics, in particular, on 2D/3D face analysis, image/video processing, and pattern recognition.



Chao Zhu received the bachelor's degree in automation from the Xi'an University of Electronic and Technology, Xi'an, China, in 2005, the master's degree in system engineering from Xi'an Jiaotong University, Xi'an, China, in 2008, and the Ph.D. degree in computer science from the École centrale de Lyon, Lyon, France, in 2012. He is currently a Post-Doctoral Fellow with Peking University, Beijing, China. His research interests include object detection and recognition, image classification, feature extraction, and image/video processing.

Yunhong Wang (M'98) received the B.S. degree in electronic engineering from Northwestern Polytechnical University, Xi'an, China, in 1989, and the M.S. and Ph.D. degrees in electronic engineering from the Nanjing University of Science and Technology, Nanjing, China, in 1995 and 1998, respectively. She was with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China, from 1998 to 2004. Since 2004, she has been a Professor with the School of Computer Science and Engineering, Beihang University, Beijing, where she is also the Director of the Laboratory of Intelligent Recognition and Image Processing with the Beijing Key Laboratory of Digital Media. Her research interests include biometrics, pattern recognition, computer vision, data fusion, and image processing.

Liming Chen (M'05) received the joint B.Sc. degree in mathematics and computer science from the University of Nantes, Nantes, France, in 1984, and the M.S. and Ph.D. degrees in computer science from the University of Paris VI, Paris, France, in 1986 and 1989, respectively. He served as an Associate Professor with the Université de Technologie de Compiègne, Compiègne, France, and then joined the École centrale de Lyon (ECL), Lyon, France, in 1998, as a Professor, where he leads an Advanced Research Team in multimedia computing and pattern recognition. He has been the Head of the Department of Mathematics and Computer Science at ECL since 2007. His current research interests include computer vision and multimedia, in particular, 2D/3D face analysis, image and video categorization, and affective computing.