Color CENTRIST: Embedding Color Information in Scene Categorization

Wei-Ta Chu, Chih-Hao Chen, and Han-Nung Hsu
National Chung Cheng University, Taiwan
wtchu@cs.ccu.edu.tw, w7376ms46@hotmail.com, h3607912@hotmail.com

ABSTRACT

A new color descriptor has now been proposed to embed color information into the framework of CENsus Transform histogram (CENTRIST), so that such a state-of-the-art visual descriptor can be further improved to categorize image scenes. In the proposed color CENTRIST descriptor, global structure characteristics are described by both gradients derived from intensity values and color variations between image pixels. The spatial pyramid scheme has also been adopted to convey information in different scales. Comprehensive studies based on various datasets were conducted to verify the effectiveness of the color CENTRIST from different aspects, including the way to quantize the color space, selection of color space, and categorization performance on various datasets. We demonstrated that the color CENTRIST descriptor was not only easy to implement, but also reliably achieved superior performance over CENTRIST. An application was also proposed to demonstrate the possibility of applying the color CENTRIST in various domains.

Index Terms: Census transform histogram; color index; scene categorization; color descriptor.

I. INTRODUCTION

Scene categorization, or scene recognition, has become a fundamental process for efficient image browsing, retrieval, and organization. For example, if an image's scene category can be recognized, we would reduce the search space of object recognition, or more accurately, detect semantic concepts present in this image. The results of scene categorization may also help a robot to localize itself in a building. Detecting semantic category of an image is undoubtedly important, and devising good visual descriptors plays the core role in this task.

In the literature, many visual descriptors have been proposed for image scene recognition. They can be roughly divided into two groups: 1) part-based representation, with the consideration of multiple scales or spatial distributions, and 2) holistic representation that directly models global configurations. The former approach describes texture information in image patches, and has been proven to be extremely effective when detecting objects under various conditions. By considering the distribution of local descriptors over image patches, sometimes in a multiscale manner, global information is captured. One of the most popular part-based descriptors is Scale-Invariant Feature Transform (SIFT) [7], and one of the most prominent approaches to consider the

global distribution is the spatial pyramid approach [3]. Despite the SIFT descriptors associated with the bag of visual words model [8] which have shown discriminative power on scene categorization, directly modeling global texture information often more reliably describes spatial structure of a scene. The same scene may be taken from various viewpoints, and objects with significantly different appearances may appear in the same type of scene. In contrast with the local texture information, holistic representation, such as GIST [2], captures global structure and achieves high accuracy in natural scene categorization. Recently, CENsus TRansform hISTogram (CENTRIST) [1] was proposed to provide accurate and stable performance on various scene image datasets.

We found that most works were targeted on gray images, and existing visual descriptors mainly relied on oriented gradient calculated based on intensity values. However, we argue that color information also plays an important role, although it would not be as important as intensity, and should not be neglected in scene categorization. Figure 1 shows an example about how color information is used in distinguishing scene categories. Without color information, these two images have similar structure and are hard to recognize. With color information we realize that the open country image has a blue region on the top half, while the coast image has two distinct blue regions at the top half and the bottom half, respectively. It is clear that considering color information benefits scene categorization.

In this work, we devise a visual descriptor called color CENTRIST to embed color information into the framework of CENTRIST, and demonstrate its effectiveness through evaluating various color image datasets. Through comprehensive evaluation, we verify effectiveness of the color CENTRIST. The main contributions of this work are briefly described as follows, which were also shown in our preliminary work [21].

- We devise a color index scheme to embed HSV color information into the framework of CENTRIST. Information of three color channels is encapsulated into an 8-bit representation, so that the framework of CENTRIST can be directly employed, and various performance comparisons can be impartially conducted. We verify that different color channels should be allocated different numbers of bits to more accurately characterize image content.
- Performance of the proposed descriptor is evaluated based on various datasets, including the 8-class scene dataset, the 8-event dataset, the 67-indoor scene dataset, the KTH-IDOL and the KTH-INDECS datasets. Working on various datasets shows robustness and effectiveness of the proposed descriptor.

The unique contributions of this work over our previous work [21] are described as follows.

- We verify the best multilevel representation of the proposed descriptor by carefully evaluating performances obtained by different levels of descriptors. Moreover, statistical analysis is conducted to show that the performance superiority is statistically

significant.

- We verify that extracting the proposed descriptor from the HSV color space gives stable performance.
- We verify that combining the proposed descriptor with CENTRIST further yields better performance.
- We compare performance obtained by the proposed descriptor with that obtained by SIFT, and its color variants, based on the bag of words framework.
- We compare performance obtained by the proposed descriptor with that obtained by several promising color LBPs.
- An application on object detection is proposed to demonstrate the possibility of applying the color CENTRIST in different domains.

The rest of this paper is organized as follows. Section II provides a literature survey. The color CENTRIST descriptor is proposed after briefly reviewing conventional CENTRIST in Section III. Preliminary analysis of different descriptor settings is described in Section IV. We provide comprehensive evaluation on various datasets in Section V, and a novel application based on color CENTRIST in Section VI. Section VII concludes this paper with discussions of the proposed descriptor and future research.

II. Related Works

In recent years, significant advancement had been made for scene recognition by the computer vision and pattern recognition community. Some studies focused on feature/descriptor design to more reliably describe scene characteristics, while some studies focused on distance metric or recognition scheme to achieve a more accurate classification. Because related literature was rich, we just made a brief survey from the perspective of feature/descriptor design in the following.

A. Scene Categorization by Local Descriptors

Currently, SIFT [7] and other local descriptors associated with the bag of words (BOW) model [8] were the dominant scheme in scene categorization. Fei-Fei and Perona [5] described images with a collection of local regions, which were represented by codewords derived from a visual word codebook. They proposed the theme models, modified from the Latent Dirichlet Allocation, to represent the distribution of codewords in each scene category. Lazebnik et al. [3] argued that describing bags of visual words in multiple scales provided an encouraging performance on natural scene recognition. Focusing on codebook design, van Gemert et al. [6] dealt with codeword uncertainty and codeword plausibility. They proposed a kernel codebook method that allowed some degree of ambiguity in assigning a visual descriptor to multiple codewords. Also based on BOW representation, Bosch et al. [11] investigated classification methodologies for scene categorization. They proposed a hybrid approach that first discovered latent

topics in scene images by pLSA (probabilistic latent semantic analysis), and then topic distributions were fed to discriminative classifiers. Rather than directly modeling texture features, Vogel and Schiele [20] first detected semantic concepts for image patches, and then modeled an image by the distribution of concept occurrence.

Recently, Vasconcelos's group proposed a series of works to construct semantic spaces based on bag of local features [27][28]. They demonstrated that, by representing images in the semantic space (manifold), distances between images had been well measured and thus scene categories had been more accurately recognized. In [29], a multiclass problem was treated as a collection of one-versus-one binary problem. For each binary problem, a unified objective function was designed to jointly optimize parameters of SIFT-based codebook construction and classifier training. To encode spatial layout, Krapac et al. [30] employed Gaussian mixture models associated with Fisher kernels to describe spatial information of local features. Fornoni and Caputo [34] argued that saliency information had been used in feature pooling, and thus spatial context was more reliably captured. They adopted SIFT descriptors and demonstrated their proposed scheme was especially useful in indoor scenes. In [38], local context and spatially regularized characteristics were jointly considered to construct codebooks.

One emerging idea was recognizing objects first and then using relationships between objects (rather than local features extracted from image patches) to facilitate scene recognition. Yu et al. [31] proposed a reasoning module that iteratively detected objects in each run and decided the scene class based on the response of object detection results. Zheng et al. [40] modeled relationships between objects based on response of object part filters, which were implemented by deformable part-based models. They also showed that performance was further improved if object part information and global texture like GIST [2] were jointly modeled. Niu et al. [35] developed a context aware topic model that jointly considered global and local contexts between scene elements, e.g., sky and car, in different scene categories. In [39], an image was viewed as a collection of regions, which were represented by region models. Jiang et al. [41] focused on determining optimal spatial layout of images based a randomized spatial partition scheme. The most descriptive pattern for each scene category was discovered to boost scene recognition accuracy.

B. Scene Categorization by Global Descriptors

Oliva and Torralba [2] argued that modeling object information was not necessarily needed for recognizing a scene. They proposed the GIST descriptor to model the structure of a scene, and assumed that images coming from the same scene category had similar configurations. This idea had been proven effective in recognizing outdoor scenes, e.g., mountain and coast, but performance decreased significantly for indoor scenes. Based on census transform, Wu and Rehg [1] proposed a simple yet effective visual descriptor, called CENsus TRansform hISTogram (CENTRIST), to model global configurations of scenes. They demonstrated that structure information can be effectively described by comparing the intensity value of a pixel with its eight

neighboring pixels. By considering the global distribution of visual descriptors in images using spatial pyramids, they constructed a holistic representation for images. Comprehensive studies were provided in [1] to show multilevel CENTRISTs yield superior performance over SIFT and GIST in most cases. The idea of CENTRIST was similar to LBP (local binary pattern) [22], which had also been widely adopted in various computer vision applications [42], such as face detection, facial expression recognition [37], and moving object tracking. Global descriptors had also been incorporated with object information to facilitate scene categorization. Pandey and Lazebnik [32] utilized deformable part models to detect recurring visual elements and salient objects. By integrating object information and global image features (i.e., GIST), promising recognition performance was reported.

Various features had been proposed to associate with effective classification schemes. However, most features employed texture or gradient information, and much fewer studies had been conducted to investigate how color information affected scene categorization. The work in [11] was one of the few studies that investigated color descriptors. From their reported results, color information consistently brought performance increment if it was appropriately incorporated into visual descriptors. Van de Sande et al. [13] evaluated color variants of SIFT descriptors on object and scene recognition. Their results also conformed to the trend, but only SIFT-based descriptors were evaluated. In this paper, we designed a method to incorporate color information into one of the state-of-the-art visual descriptor, i.e., CENTRIST [1], and demonstrated its effectiveness through comprehensive evaluation from various perspectives.

III. DESCRIPTORS

A. CENTRIST

To handle scene categorization, Wu and Rehg described desired properties of appropriate visual descriptors [1], including that holistic representation may be robust, structural properties should be captured, rough geometry of scenes may be helpful, and the descriptor should be general for different scene categories. By considering these, Wu and Rehg proposed a holistic representation modeling distribution of local structures, called CENsus TRansform hISTogram (CENTRIST). Rough geometrical information was captured by CENTRISTs extracted from spatial pyramids in different levels. In the following, we briefly review conventional CENTRIST before we propose its color extension.

To describe the relationship between a pixel and its neighboring pixels, census transform was carried out by comparing characteristics of pixels in local patches [4]. For instance, Figure 2 shows an example of census transform based on comparing intensity value of a pixel with that of its eight spatially neighboring pixels. Namely, replacing a neighboring pixel with bit 1 if its intensity value was less than or equal to the center pixel's intensity value. Otherwise, a bit 0 was set. By concatenating these bits from top-left to bottom right, an 8-bit binary representation was constructed, and the corresponding base-10 number was called

Census Transform value (CT value) of the center pixel. Note that CT values of pixels at the image borders were undefined. Because the CT value described the relative intensity distribution of a pixel in a local patch, it was robust to gamma variations and illumination changes. Basically, the Census Transform was similar to the local binary pattern code $LBP_{8,1}$ [22], except that a bit shifting mechanism was designed to make $LBP_{8,1}$ rotation invariant.

After assessing a CT value for each pixel, the histogram of CT values, i.e., CENTRIST, was constructed to describe an image. Note that CENTRIST was 256-dimensional because CT values were described by eight bits, and CT values may range from 0 to 255. Wu and Rehg discussed detailed properties of CT values and CENTRIST descriptors in [1].

Missing spatial information and lack of multilevel representation were common drawbacks of histogram-based descriptors. To improve the robustness of CENTRIST, Wu and Rehg proposed a spatial pyramid scheme, as illustrated in Figure 3. To construct level k spatial pyramids, an $N \times N$ image was equally divided into 2^k blocks in the horizontal direction and in the vertical direction, respectively. That is, each block was of size $\frac{N}{2^k} \times \frac{N}{2^k}$. To avoid artifacts caused by non-overlapping division, the blocks centered at the common corners of four neighboring blocks were considered as well. Taking level 2 spatial pyramids as an example, an image was split into $2^2 \times 2^2 + 9 = 25$ blocks, as illustrated in Figure 3. With the same splitting scheme, an image was split into five blocks to construct level 1 spatial pyramids, and one block to construct the level 0 spatial pyramid. Note that images were resized to ensure all blocks from different levels were of the same size.

From each block, the CENTRIST descriptor was extracted, and descriptors from all blocks were concatenated to describe the image. Different dimensions of the CENTRIST descriptor were not independent, and thus Wu and Rehg used principal component analysis (PCA) to reduce dimensionality of CENTRIST to 40. This compact representation was called spatial Principal component Analysis of Census Transform (spatial PACT) histogram, or abbreviated to sPACT. In this case, an image with level 2 spatial pyramids was thus described by a $40 \times (25 + 5 + 1) = 1240$-dimensional descriptor. Note that when we say *level 2 spatial pyramids*, descriptors extracted from blocks generated by levels 2, 1, and 0 split were all considered together.

B. Color CENTRIST

In this work, we devise a color index scheme to embed color information into the framework of CENTRIST. Through extensive experimental studies, we will demonstrate that the proposed *color CENTRIST* descriptor effectively enhanced the performance of scene categorization.

We represent color in the hue-saturation-value (HSV) color space, where three channels are normalized to range from 0 to 255[1]. Theoretically, representing color of a pixel needs 24 bits in this setting. To make the proposed representation comparable to CT

[1] Converting RGB to HSV is implemented based on the OpenCV library.

values in CENTRIST, we devise a color index scheme to represent color information of a pixel by 8 bits, with the design of different quantization granularities for different color channels. For example, if we respectively allocate b_1, b_2, and b_3 bits ($b_1 + b_2 + b_3 = 8$) to represent hue, saturation, and value components, three channels were uniformly quantized into 2^{b_1}, 2^{b_2}, and 2^{b_3} levels, respectively. Let us denote the hue, saturation, and value components (in the base 10 numeric system) of a pixel by h, s, and v, respectively. The hue component is transformed into a (base 10) color index $\hat{i}_h = \lfloor \frac{h \cdot 2^{b_1}}{256} \rfloor$, which is then represented in the base 2 system by b_1 bits. Similarly, the color indices for saturation and value components are computed as $\hat{i}_s = \lfloor \frac{s \cdot 2^{b_2}}{256} \rfloor$ and $\hat{i}_v = \lfloor \frac{v \cdot 2^{b_3}}{256} \rfloor$, and are represented in the base 2 system by b_2 and b_3 bits, respectively.

Figure 4 shows the flowchart to extract a color CENTRIST, especially with an example showing how color of a pixel is represented by 8 bits. In this example, color indices corresponding to the hue, saturation, and value components are allocated 1, 2, and 5 bits, respectively. The hue axis is divided into $2^1 = 2$ ranges, i.e., [0, 127] and [128, 255]. The saturation axis is divided into $2^2 = 4$ ranges, i.e., [0, 63], [64, 127], [128, 191], [192, 255]. The value axis is divided into $2^5 = 32$ ranges, i.e., [0, 7], [8, 15], …, [240, 247], and [248, 255]. The pixel's hue component is 183, and thus the corresponding color index is $\hat{i}_h = \lfloor \frac{183 \cdot (2^1)}{256} \rfloor = 1$ in decimal. The value $\hat{i}_h$ is then represented in the base 2 system (by one bit) as $i_h = 1$. The saturation component of the pixel is 220, the corresponding color index is $\hat{i}_s = \lfloor \frac{220 \cdot (2^2)}{256} \rfloor = 3$, and thus in the base 2 system (by two bits) $i_s = 11$. Similarly, the color index corresponding to the value component 91 is $\hat{i}_v = \lfloor \frac{91 \cdot (2^5)}{256} \rfloor = 11$, and thus in the base 2 system (by five bits) $i_v = 01011$. To give the highest priority to the value component, the color indices in binary forms are concatenated in the manner $(i_v \ i_s \ i_h)$. Therefore, the pixel originally represented by 24 bits, i.e., (183, 220, 91), is represented by eight bits $i = (i_v, i_s, i_h) = 01011\ 11\ 1$ in binary, or 95 in decimal.

Through the process mentioned above, we represent each pixel by a color index. Just like an index to a color palette, the number 95 indicates that the color (183, 220, 91) falls into the 95th range of the quantized HSV color space. We then conduct a census transform based on color indices of pixels, rather than on the intensity of pixels. If the color index of the center pixel is larger than or equal to one of its neighbors, a bit 1 is set at the corresponding position. Otherwise, a bit 0 is set. From top-left to bottom right, these bits are concatenated to form a binary representation, which is then evaluated to a base 10 number called the color Census Transform value (cCT value) of the center pixel. The histogram of cCT values over the whole image is finally constructed to form a *color CENTRIST*. Note that a color CENTRIST is also 256-dimensional because there are 256 different types of cCT values. Similar to CENTRIST, we reduce dimensionality of a color CENTRIST by PCA, and model rough global structure of an image based on spatial pyramids.

By considering color information, the cCT value represents whether a pixel's color index is ahead of, or behind, the color indices of its neighboring pixels, and thus describes color distribution around it. Although the physical meaning of cCT values is not as intuitive as that of CT values, we will verify that this representation effectively benefits scene categorization.

C. Properties of Color CENTRIST

We use Figure 5 and Figure 6 to underline the difference between CENTRIST and color CENTRIST. One image from the "open country" category and one image from the "coast" category are compared based on CENTRISTs and color CENTRISTs, respectively. The second column of Figure 5 shows two images in gray, and from which we can see that they look similar when only intensity information is considered. The third columns of Figure 5 and Figure 6 show the corresponding census transformed images, where each pixel is replaced by its CT value and cCT value, respectively. The last columns in two figures show the corresponding CENTRISTs and color CENTRISTs, respectively. Visually, from CENTRISTs, we see that both open country and coast images have two peaked ranges. From color CENTRIST, however, the coast image has two peaked ranges, while the open country image has only one clear peaked range. To quantify this observation, similarity between two CENTRISTs (or two color CENTRISTs) of the open country image and the coast image is measured by histogram intersection, i.e., $s_{ij} = \frac{\sum_{k=1}^{256} \min(H_i(k), H_j(k))}{\sum_{k=1}^{256} \max(H_i(k), H_j(k))}$, where H_i and H_j are CENTRISTs (or color CENTRISTs) of two images, respectively. Based on CENTRIST, the histogram intersection between the open country image and the coast image is 0.525. Based on color CENTRIST, the histogram intersection between the open country image and the coast image reduces to 0.380. This indicates that color CENTRIST more accurately discriminates these two images, because the similarity between these two different images reduces when color CENTRISTs are used as image representation.

We further illustrate the difference between two descriptors based on a designed image as shown in Figure 7. The image of size 33×33 pixels is constituted by three regions in pure blue (R,G,B = 0,0,255), pure red (R,G,B = 255,0,0), and pure green (R,G,B = 0,255,0), from left to right, respectively. The HSV values corresponding to these three regions are (170, 255, 255), (0, 255, 255), and (85, 255, 255), respectively. Because all pixels have the same intensity value, CT values of all pixels are equal to 255, except for the ones at the image borders, which are undefined. The top of Figure 7(b) shows CT values of pixels, and the top of Figure 7(c) shows the corresponding CENTRIST. According to the quantization table illustrated in Figure 4, the color indices corresponding to those three regions are 255, 254, and 254, respectively. The bottom of Figure 7(a) shows a sample patch that is centered by a pixel located in the red region, and can be seen at the right side of the boundary between the blue and the red regions (we call it right boundary pixels in the following, in contrast to the pixels at the left boundary between the blue and the red regions). In contrast to intensity values, the blue region and the red region have different color indices, and the corresponding cCT value is 107. Comparing

CT values with cCT values of these pixels, we clearly see that cCT values provide more cues to discriminate color regions. The bottom of Figure 7(b) shows detailed cCT values of pixels located at different positions, and the bottom of Figure 7(c) shows the corresponding color CENTRIST. The color CENTRIST is basically the same as CENTRIST, except at the 107th bin, where cCT values come from the right boundary pixels between blue and red. From this example, we see that, with color CENTRIST, not only structure characteristics but also color variations can be described.

IV. Analysis of Descriptor Settings

This section presents how different experiment settings influence scene categorization accuracy. Experiments in this section (and also in Section V) were conducted in five runs, and the recognition accuracies of five runs are averaged to show the overall performance. At each run, parts of a dataset in each scene category were randomly selected for training, and the remaining is for testing. We call it the *five-random-run scheme* in the following.

In the following experiments, we remove the two bins with CT or cCT values equal to 0 and 255 in both CENTRISTs and color CENTRISTs, and normalize them into unit vectors. Similar to sPACT in [1], to reduce dimensionality of color CENTRIST, 40 eigenvectors corresponding to 40 largest eigenvalues are found, and 254-dimensional color CENTRIST descriptors are projected into the eigenspace to form a 40-dimensional sPAcCT (spatial Principal component Analysis of color Census Transform histogram). To include more image statistics, mean and standard deviation of intensity values in a block are concatenated at the end of a sPACT in [1]. We analogize this setting and concatenate mean and standard deviation of color indices at the end of a sPAcCT as well. Therefore, the feature vectors of both level 2 sPACT and level 2 sPAcCT have $(40+2)\times(25+5+1)=1302$ dimensions. Based on these visual descriptors, SVM classifiers with RBF kernels are constructed to conduct scene categorization, where kernel parameters are chosen based on the cross-validation scheme provided by the LIBSVM package [25].

A. Color Quantization

To represent color information, we quantize the HSV color space into a number of color ranges, and transform color components of a pixel into color indices. To determine the number of bits to describe color indices, we examine scene recognition accuracy for the 8-class scene dataset [2], by using 4 bits, 8 bits, 12 bits, or 16 bits to describe quantization levels, respectively. Namely, the HSV color space is quantized into 2^4, 2^8, 2^{12}, or 2^{16} ranges. For the 8-bit settings, for example, we test different allocation schemes that present the hue, saturation, and value components by different numbers of bits. Table 1 shows detailed scene recognition rates for the 8-class scene dataset, based on the 8-bit setting. The result in the sixth row H-S-V (1-1-6), for example, means that the hue channel and saturation channel are respectively quantized into $2^1=2$ levels, and the value channel is quantized into $2^6=64$ levels.

The first row H-S-V (8-0-0) means that saturation and value channels are discarded, and the hue channels are quantized into 256 levels. Note that the setting (H-S-V 0-0-8) in the third row is similar to CENTRIST (but not exactly the same) because only (quantized) intensity is considered to do census transform.

By comparing the first three rows in Table 1, we clearly see that intensity values still play the most important role in scene description. However, by jointly considering hue and saturation, and appropriately quantizing different color channels, better performance can be further achieved. We see that in the 8-bit setting the best performance is obtained by the (H-S-V 1-2-5) scheme, the worst performance is obtained by the (H-S-V 8-0-0) scheme, and the average accuracy is 86.00%. To show that performance superiority of the (H-S-V 1-2-5) scheme is statistically significant, recognition accuracies obtained by this scheme and another in five runs are compared pairwise. Table 1 shows statistical significance in terms of the p-values of the paired two-sample t-tests [17] between the best scheme (H-S-V 1-2-5) and others, and demonstrates that, in most cases, performance superiority of the best scheme over others is significant. Note that in Table 1, we do not exhaustively list performances of all possible schemes because of space limitation. Although we did exhaustively evaluate all combinations, only a few samples are shown in this table to reveal the performance variation.

We evaluate different allocation schemes based on the 4-bit, 12-bit, and 16-bit settings as well. Table 2 shows detailed schemes evaluated based on different settings. From each setting, we respectively find the best, the worst, and the average performance, and illustrate them in Figure 8. From this figure we can see that the 8-bits setting averagely achieves the best recognition performance. Note that in this figure we only show the worst cases where hue, saturation, and value components are all used. According to Table 1 and Figure 8, the (H-S-V 1-2-5) scheme in the 8-bit setting is used in the following experiments.

Based on the experimental results and the color index setting mentioned above, we need to emphasize that we are not claiming color information is more important than intensity. In color CENTRIST, we still largely employ intensity (by setting the value component in most significant bits). What we claim, with appropriate arrangement, the performance of scene categorization can be improved if color information is considered together.

B. Color Space Selection

The designed color CENTRISTs are extracted based on the HSV color space. To verify the influence of different color spaces on scene categorization and that motivates our choice, we compare the performance of level 2 sPAcCT extracted based on the HSV, RGB, and Lab color spaces, respectively for the 8-class scene dataset [2] and the 67-scene indoor scene dataset [10]. Similar to the procedure described in Sec. IV.A, we try different allocation schemes to find the best quantization setting for each color space, and achieve the best accuracy for two datasets.

Table 3 shows that color CENTRISTs from the RGB color space achieve the best performance for the 8-class scene dataset. The p-values of RGB vs. HSV and RGB vs. Lab are 0.068 and 0.016, respectively. Color CENTRISTs from the Lab color space achieve the best performance for the 67-class indoor scene dataset. The p-values of Lab vs. HSV and Lab vs. RGB are 0.213 and 0.005, respectively. Although not being the one achieving the best performance, the difference between HSV and the best methods in both datasets is not statistically significant. Color CENTRISTs extracted from the HSV color space stably achieve in-between results, and this is why the HSV color space is chosen to extract color CENTRIST for large-scale and various scene categorization tasks in this study.

C. Comparing Color CENTRIST with CENTRIST Extracted from Multiple Channels

An index scheme is designed to embed color information from three channels into a compact color index able to be represented by 8 bits. To verify effectiveness of the proposed index scheme is effective, here we evaluate scene categorization performance obtained by color CENTRISTs with this index scheme, and that obtained by concatenating CENTRISTs respectively extracted from H, S, and V color channels rather than the intensity channel solely.

From the hue channel, for example, the census transform for a pixel is conducted by comparing its hue component with that of its eight neighbors. The CT values are then collected to construct the CENTRIST in the hue channel, denoted as h-CENTRIST. With the spatial pyramid scheme and principal component analysis, a 1302-dimensional level 2 h-sPACT can then be constructed from the hue channel. In this experiment, a 1302-dimensional level 2 sPAcCT is compared with the 3906-dimensional ($1302 \times 3 = 2906$) concatenation of h-sPACT, s-sPACT, and v-sPACT, which is denoted by hsv-sPACT. Based on the 8-class scene dataset, the recognition rate obtained by the level 2 sPAcCT is 86.92%, while the recognition rate obtained by the level 2 hsv-sPACT is 86.27%. The p-value of the paired two-sample t-tests between two sets of experimental results is 0.037, which shows that performance superiority of the proposed color index scheme is statistically significant, though the dimension of level 2 sPAcCT is much smaller than that of level 2 hsv-sPACT. The reason for evaluating based on level 2 representation is that it consistently gives the best performance for both cCENTRIST and CENTRIST, which will be shown in the following section.

V. PERFORMANCE EVALUATION

With the experiment settings discovered above, the color CENTRIST descriptor is tested based on four data sets: 8-class scene category [2], 8-class sports event [9], 67-class indoor scene recognition [10], and KTH-IDOL/KTH-INDECS [14][15]. These datasets include a variety of images with various visual characteristics.

A. *The 8-Class Scene Category Dataset*

The 8-class scene recognition data set was built by Oliva and Torralba [2]. Although this dataset was gradually extended to 13 classes and 15 classes by Fei-Fei and Perona [5], and Lazebnik et al. [3], respectively, only the original 8 classes of images are colorful. We thus evaluate CENTRIST and color CENTRIST (abbreviated as cCENTRIST in the following) based on this smaller dataset. This data set contains a wide range of scene categories in outdoor environments, such as coast, forest, mountain, and so on. Figure 9 shows some sample images in this dataset. Resolutions of all these color images are 256×256 pixels, and there are 260 to 410 images in each category.

Experiments of image scene categorization are conducted in five runs, and the recognition accuracies of five runs are averaged to show the overall performance. At each run, 100 images in each category are randomly selected for training, and the remaining images are for testing. A multiclass SVM classifier with RBF kernel is constructed for recognition. Note that the multiclass SVM classifier implemented in the LIBSVM package [25] is constituted by a collection of binary SVM classifiers with a one against one approach. We compare CENTRIST with cCENTRIST based on level 0 representation, and based on representations of levels 1, 2, and 3, with and without PCA. For cCENTRIST with spatial pyramid scheme but without dimension reduction, we call it spatial color Census Transform (spatial cCT) histogram, or abbreviated to scCT. The dimensions of scCT with level 1 pyramids, level 2 pyramids, and level 3 pyramids are $(254+2) \times 6 = 1536$, $(254+2) \times 31 = 7936$, and $(254+2) \times 144 = 36864$, respectively. You may recall that the term $(254+2)$ comes from removing the two bins with cCT values equal to 0 and 255 from the 256-dimensional color CENTRIST, and concatenating mean and standard deviation of color indices in a block.

Table 4 shows the experimental results. At each level the best result is shown in boldface, and p-values of comparing the best results with others at the same level are calculated to show statistical significance. This table demonstrates the following trends. First, the proposed cCENTRIST stably has superior performance over CENTRIST at all levels. These results verify that, if color information is appropriately embedded, scene images are better categorized. Second, the level 2 representation with PCA provides the best performance over all other settings. This conforms to the trend reported in [1] and [3]. For descriptors modeling global structure, if we appropriately extract descriptors in local patches and consider information in different levels as well, better performance can be obtained. Third, by comparing the performance obtained with PCA and without PCA at levels 1, 2, and 3, we found that applying PCA can effectively eliminate noisy features and improve performance.

The confusion matrices of scene recognition based on level 2 sPAcCT and level 2 sPACT are shown in Figure 10, where rows are true labels and columns are predicted labels. The level 2 sPAcCT yields the best performance in forest and tall building categories. The level 2 sPACT also works best for forest, but does not work that well for tall building. There is a clear structure and color

difference between tall buildings and the sky, and thus cCENTRIST brings more clues for recognizing tall buildings. The most confused case comes from open country versus coast, which also conforms to the trend reported in [1] and [3].

Figure 11 and Figure 12 show sample images that are correctly and incorrectly recognized based on the level 2 sPAcCT, respectively. The caption "highway(coast)", for example, means the corresponding image is detected as highway, while the true label is coast. From Figure 11 we can see that cCENTRIST achieves reliable performance even there is significant intra-class variation. On the other hand, in Figure 12, some cases that may also confuse humans still annoy the proposed descriptor.

B. The 8-Class Event Dataset

The 8-class event dataset [9] includes images of eight sports: badminton, bocce, croquet, polo, rowing, rock climbing, sailing, and snowboarding (see Figure 13). Although this dataset was designed for event recognition, in this experiment we classify events by classifying scenes, and do not attempt to recognize individual objects or persons. Images in this dataset are in high resolutions (from 800×600 to thousands of pixels per dimension). There are 137 to 250 images in each category. With the five-random-run scheme, 70 images per class are randomly selected for training, and the remaining images are for testing. Based on the LIBSVM package, we respectively construct multiclass SVM classifiers with the RBF kernel, based on CENTRIST or cCENTRIST in the representation of level 0, the representations of levels 1, 2, and 3, with or without PCA.

Table 5 shows the experimental results, where at each level the best result is shown in boldface. Similar to the results of the 8-class scene dataset, cCENTRIST achieves better performance over CENTRIST in all levels of representations. The benefits brought by PCA are also confirmed. However, comparing Table 5 with Table 4, performance superiority of cCENTRIST over CENTRIST for the 8-class event dataset is slightly less than that for the 8-class scene dataset. The 8-class event dataset was not specifically collected to represent image scenes. Because a variety of players and sports alliances largely occupy the image space, there are fewer regular-texture regions inside images. Recognition performance obtained based on cCENTRISTs and CENTRISTs is thus getting close.

Figure 14 shows the confusion matrices of scene recognition based on level 2 sPAcCT and level 2 sPACT, respectively. In both matrices, the most confused case is bocce versus croquet. Based on sPAcCT, 22% of bocce images are misrecognized as croquet, and 25% of croquet images are misrecognized as bocce. Based on sPACT, 16% of bocce images are misrecognized as croquet, and 23% of croquet images are misrecognized as bocce. This trend is expectable because bocce images and croquet images share very similar backgrounds. When comparing these two matrices, sPAcCT especially works better for recognizing rock climbing (91%), rowing (91%), and sailing (88%), which are over 85%, 87%, and 82% obtained by sPACT, respectively. We conjecture that considering color information in sPAcCT helps to describe large-area colorful regions, such as water, sky, and rock.

C. *The 67-Class Indoor Scene Dataset*

The 67-class indoor scene dataset was proposed in [10]. The indoor scenes range from specific categories (e.g., dental office) to generic concepts (e.g., mall), and contain a total of 15,620 images. It was argued in [10] that both local and global information are needed to recognize complex indoor scenes. In [10], the global GIST feature averagely achieved 21% recognition accuracy for this challenging data set. By jointly considering local information, the accuracy was improved to 25%.

Following the experiment settings in [10] and [1], 80 images were randomly selected from each category for training, and 20 images were selected for testing. The five-random-run scheme is also used. Multiclass SVM classifiers with the RBF kernel were constructed, respectively based on CENTRIST and cCENTRIST in the representation of level 0, the representation of level 1 with PCA, and the representation of level 2 with PCA. Table 6 shows the experimental results, where at each level the best result is shown in boldface. The average recognition accuracy based on level 2 sPAcCT is 36.09%, which shows the performance of sPAcCT derived from cCENTRIST is significantly better than GIST and sPACT.

D. *The KTH-IDOL and The KTH-INDECS Dataset*

The KTH Image Database for rObot Localization (IDOL) dataset [14] includes pictures captured by two mobile robot platforms, Minnie and Dumbo, that move in an indoor environment consisting of five rooms of different functionalities, including a one-person office, a two-person office, a kitchen, a corridor, and a printer area. A complete image sequence was captured when a robot moved around all the five rooms, under one of three weather conditions (cloudy, night, and sunny). For each robot and each weather condition, four image sequences were captured on different days, and thus there are $2 \times 3 \times 4 = 24$ sequences. Resolution of these images is 320×240. In different image sequences, various objects like a walking person or furniture may be added or removed. The first two rows of Figure 15 show sample images captured by Minnie and Dumbo in a one-person office, under different weather conditions. In the same environment as the IDOL dataset, images in the KTH-INDECS dataset [15] were captured by cameras mounted in several fixed locations inside each room. The third row of Figure 15 shows three sample images in this dataset.

We use the first two runs of image sequences captured by each robot in each weather condition. The following four experimental settings were evaluated:

- Setting 1: Train and test using the data captured by the same robot, and under the same weather conditions. Run 1 is used for training and run 2 is used for testing, and vice versa.
- Setting 2: Train and test using the data captured by the same robot, but under different weather conditions. This experiment tests generality over variations of object locations and illumination.

- Setting 3: Train and test using the data captured under the same weather conditions, but captured by different robots. Cameras mounted at different heights on the robots, and this experiment tests generality over scene layout variations.
- Setting 4: The KTH-INDECS dataset is used for training, and images from INDECS under different weather conditions are used for testing.

Following the settings mentioned above, multiclass SVM classifiers with the RBF kernel were constructed, respectively based on level 2 sPACT and level 2 sPAcCT. Table 7 shows the average recognition accuracies based on Setting 1. In this experiment, sPACT and sPAcCT have similar performances for cloudy and sunny conditions. However, sPAcCT achieves nearly 1% accuracy behind that of sPACT for the night conditions. In the images captured at night, light from fluorescent lamps may have caused color shift and influenced the robustness of color CENTRIST.

Table 8 shows average recognition accuracies when training and testing data are in different weather conditions (Setting 2). The training and test conditions in the first row, for example, mean that the sunny sequence captured in the first run was used for training, and the night sequence captured in the second run was used for testing. We found that sPAcCT has inferior performance when night images were used to train or test. On the other hand, when the cloudy or sunny images were used to train or test, sPAcCT has promising performance.

Table 9 shows the average recognition accuracies when images captured by different robots were separately used for training or testing. From this table we can see that sPAcCT has slightly weaker robustness for this experimental setting. Table 10 shows the average recognition accuracies for the KTH-INDECS dataset. sPACT and sPAcCT generally have similar performances.

Overall, sPAcCT achieves slightly worse performance than sPACT in the KTH-IDOL and KTH-INDECS datasets. We conjecture that fewer color variations exist in those datasets, and thus embedding color information, with the cost of allocating a few bits to represent hue and saturation, gives no benefit to the original CENTRIST framework. To verify this conjecture, we randomly selected ten image pairs from each scene category of the 8-class scene dataset and calculated the relative entropy of each pair based on HSV histograms of images. All relative entropy values from sampled image pairs are then averaged to show the intra-class color variation. The same procedure was also conducted based on the KTH-IDOL/KTH-INDECS datasets. Quantitatively, the average relative entropy within classes of the 8-class scene dataset is four times that of the KTH-IDOL/KTH-INDECS datasets. Therefore, we conclude that color CENTRIST benefits scene recognition more when images in the same scene convey more color variations. The relationship between intra-class color variations and categorization performance is thus worth future study.

E. Combining Color CENTRIST with CENTRIST

We have evaluated cCENTRIST on various datasets and verified that color information is helpful in describing images. In this section, based on the 8-class scene dataset, we investigate whether it is more helpful if we combine CENTRIST and cCENTRIST.

The same procedure is used to extract sPAcCT and sPACT in levels 0, 1, and 2 pyramids. Then sPAcCT and sPACT of an image on each level are concatenated, respectively. The combined descriptor describes level 0 images by $256 \times 2 = 512$ dimensions (without PCA), level 1 images by $252 \times 2 = 504$ dimensions (with PCA), and level 2 images by $1302 \times 2 = 2604$ dimensions (with PCA). The five-random-run scheme is used for training and testing.

Table 11 shows the experimental results, where at each level the best result is shown in boldface. The result shows that cCENTRIST combining with CENTRIST achieves better performance than only using CENTRIST or only using cCENTRIST. We also notice that the combined descriptor can only marginally improve cCENTRIST (recognition rates improve by 0.31%~1.57%), but can moderately improve CENTRIST (recognition rates improve by 2.09%~3.06%).

F. Bag of Words Framework

In this section, we compare the proposed holistic color descriptor with the local descriptors mentioned in [13], based on the bag of words (BOW) framework. We follow the dense sampling scheme and pyramid construction used in [12], [1], and [13] to extract descriptors. From an image we extract color CENTRIST and other local descriptors from patches of size 16 by 16, which are sampled over a grid with a spacing of 8 pixels. One fourth of the image patches sampled from the training set are used to generate two codebooks (by the k-means algorithm), which contains 200 code words and 400 code words, respectively. For a level 2 pyramid representation, the final descriptor that uses the 200 code words codebook has a dimension of $200 \times (16 + 4 + 1) = 4200$, and the dimension changes to 8400 if the 400 code words codebook is used. SVM classifiers with the histogram intersection kernel are constructed to conduct scene categorization. Based on the 8-class scene dataset, we compare color CENTRIST with other color descriptors, including c-SIFT, opponent-SIFT, and RGB-SIFT [13], and other gray-level descriptors, including CENTRIST [1] and SIFT [7]. We also compare the total time needed to extract these descriptors, based on the *coast* category in the 8-class scene dataset, which contains 360 images.

The average recognition rates are reported in Table 12. From this table, we can see that the performance of cCENTRIST is similar to that of other color descriptors. Comparing Table 12 with Table 4, performance of cCENTRIST in the BOW framework is worse than that of level 1 and level 2 sPAcCT, which is expectable because cCENTRIST is inherently a global descriptor. Although performance of cCENTRIST in the BOW framework is not superior to other color descriptors, cCENTRIST can be extracted much faster. Figure 16 shows the total time needed to extract descriptors in all of the 360 images in the *coast* category of the 8-class scene

dataset. Extracting CENTRIST and cCENTRIST is simple and fast. By jointly considering the categorization performance and extraction time, color CENTRIST is especially suitable for high-resolution images or large-scale image collections, which inspires us to develop an application described in Section VI.

G. Comparing Color CENTRIST with Color LBP

As color CENTRIST is essentially similar to local binary patterns (LBP) considering color information, a straightforward question arises: how about comparing color CENTRIST with color LBPs? Recently, Zhu et al. [26] extended original LBPs to construct six color LBP descriptors. Based on the 8-class scene dataset, we compare color CENTRIST with the best three color LBPs reported in [26], which are Hue-LBP, Opponent-LBP, and nOpponent-LBP. The Hue-LBP is obtained by computing LBP from the hue channel of the HSV color space. Each pixel is represented by eight bits. The Opponent-LBP is obtained by computing LBP over all three channels of the opponent color space, and the nOpponent-LBP is obtained by computing LBP over two channels of the normalized opponent color space. Note that each pixel is represented by 8, 24, and 16 bits when Hue-LBP, Opponent-LBP, and nOpponent-LBP are calculated, respectively. To make fair comparison, we employ the spatial pyramid scheme (with PCA-based dimensionality reduction) to all color LBPs and color CENTRIST, though a multi-scale scheme was proposed in [26] as well.

Table 13 shows performance comparison between color CENTRIST and the three color LBPs. It can be seen that color CENTRIST works the best over all levels. Compared with the Hue-LBP, we jointly consider hue, saturation, and value channels with the proposed color index scheme. When comparing with the Opponent-LBP and nOpponent-LBP that represent color information by 24 bits and 16 bits, color information in color CENTRIST is embedded into 8 bits. Evaluating results in different levels of representation verify that the proposed color CENTRIST provides robust and superior performance over existing color LBPs.

VI. APPLICATIONS

We have shown that cCENTRIST can be extracted faster than most of the other descriptors. This property is suitable to be applied in studies which need a great number of operations. In this section, we apply cCENTRIST on object detection in a high-resolution image, which is an important step in many research topics. Object detection is conducted based on the panorama of a grocery store produced in [23]. The resolution of the panorama is 14569×711 pixels. For convenience, we only use the left part of this panorama, which is the food region. The size of the left part of panorama is 7220×711, and the panorama is as shown in Figure 17. In the object detection experiment, we select 20 patches from this panorama as queries, extract cCENTRIST from queries, and aim to find the positions of these query patches in the panorama.

To efficiently detect where a query patch comes from, we adopt a search method similar to the logarithmic search method in fast motion estimation. The panorama is first divided into grids without overlapping, which have the same size as the query patch. Based on level 0 cCENTRIST, the histogram intersection between a grid and the query patch is calculated. The three grids that have the largest histogram intersection to the query patch are then selected as candidate patches. We then apply a sliding window pixel by pixel from left to right, and from top to bottom through each candidate patch. Each sliding window has the same size as the query patch, and is centered by a pixel in a candidate patch. The histogram intersection between a sliding window and the query patch is again calculated. Finally, the window with the largest histogram intersection to the query patch is where the query patch was located. Other than the histogram intersection, we also tried to evaluate the Euclidean distance and the chi square distance between query and targeted patches, but found that the histogram intersection gives the best detection performance.

The twenty query patches are shown in Figure 18, from which we see the queries includes patches of varied sizes, structures, and appearances. For each query, the spatial Euclidean distance between centers of the detected location and the ground truth is calculated to show the spatial error. As can be seen in Figure 19, two descriptors have similar performance in most cases, except for the query 4, where cCENTRIST significantly improves performance. The average spatial error over 20 queries based on cCENTRIST is 54 pixels, and that based on CENTRIST is 68 pixels. By carefully comparing detection performance query by query, cCENTRIST achieves better or equal performance to CENTRIST in 13 of 20 queries. From this experiment, we can see that incorporating color information also benefits object detection.

Note that the developed descriptors can be utilized with more efficient object detection approaches, such as efficient subimage search [24]. But we simply focus our attention on evaluating the effectiveness of color CENTRIST in this paper.

VII. CONCLUSION

In our presentation we have shown that embedding color information into the CENTRIST framework consistently provides better performance on scene categorization, through comprehensive evaluation studies from various perspectives, including color quantization, color index, color space selection, multilevel representation, and dimension reduction. By appropriately quantizing the HSV color space and with the designed color index scheme, color information is elaborately represented and incorporated to construct the color CENTRIST descriptor. With spatial pyramids, structure information in multiple levels can be described, and thus robust performance can be obtained for various datasets, including the 8-class scene dataset, the 8-class event dataset, and 67-indoor dataset. Based on the evaluation results for the KTH IDOL/INDECS datasets and the intra-class color variations measured by average relative entropy, we now suggest that the proposed color CENTRIST is especially suitable for images with higher color variations. To demonstrate the possibility of applying color CENTRIST in different domains, an application on object detection is

proposed. We believe this would be one of the most comprehensive experimental studies on scene categorization based on color holistic descriptors.

In the future, using color CENTRIST in more applications will also be investigated. Because color CENTRIST shares the same limitations as CENTRIST, i.e., not invariant to rotation and scale, we need to enhance the descriptor by designing a rotation-invariant extraction scheme. How intra-class color variations relate to categorization performance is also worth further study. Whereas with the current classification methodology, we simply utilize the standard SVM classifiers. By considering manifold assumption, more advanced SVM like Hessian regularized SVM [36], or dimension reduction techniques considering geometry preservation [18], can be adopted to improve scene categorization. Combining color CENTRIST with other features, scene categorization based on multimodal or multiview features [19] can be accomplished by multiview SVMs [33] or sparse coding schemes [16].

REFERENCES

[1] Wu, J. and Rehg, J.M. 2011. CENTRIST: a visual descriptor for scene categorization. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 33, no. 8, pp. 1489-1501.

[2] Oliva, A. and Torralba, A. 2001. Modeling the shape of the scene: a holistic representation of the spatial envelope. International Journal of Computer Vision, vol. 42, no. 3, pp. 145-175.

[3] Lazebnik, S., Schmid, C., and Ponce, J. 2006. Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 2169-2178.

[4] Zabih, R. and Woodfill, J. 1994. Non-parametric local transforms for computing visual correspondence. Proceedings of European Conference on Computer Vision, vol. 2, pp. 151-158.

[5] Fei-Fei, L. and Perona, L. 2005. A Bayesian hierarchical model for learning natural scene categories. Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 524-531.

[6] Van Gemert, J.C., Geusebroek, J.-M., Veenman, C.J., and Smeulders, A.W.M. 2008. Kernel codebooks for scene categorization. Proceedings of European Conference on Computer Vision, vol. 3, pp. 696-709.

[7] Lowe, D. 2004. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, vol. 60, no. 2, pp. 91-110.

[8] Sivic, J. and Zisserman, A. 2003. Video Google: a text retrieval approach to object matching in videos. Proceedings of IEEE International Conference on Computer Vision, pp. 1470-1477.

[9] Li, L.-J. and Fei-Fei, L. 2007. What, Where and Who? Classifying events by scene and object recognition. Proceedings of IEEE International Conference on Computer Vision.

[10] Quattoni, A. and Torralba, A. 2009. Recognizing indoor scenes. Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition.

[11] Bosch, A., Zisserman, A., and Munoz, X. 2008. Scene classification using a hybrid generative/discriminative approach. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 30, no. 4, pp. 712-727.

[12] Pronobis, A., Caputo, B., Jensfelt, P., and Christensen, H.I. 2006. A discriminative approach to robust visual place recognition. Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems.

[13] Van de Sande, K.E.A, Gevers, T., and Snoek, C.G.M. 2010. Evaluating color descriptors for object and scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, no. 9, 1582-1596.

[14] Luo, J., Pronobis, A., Caputo, B., and Jensfelt, P. 2006. The KTH-IDOL2 database. Technical Report CVAP304, Kungliga Tekniska Hoegskolan, CVAP/CAS, Oct. 2006.

[15] Pronobis, A. and Caputo B. 2005. The KTH-INDECS database. Technical Report CVAP297, Kungliga Tekniska Hoegskolan, CVAP, Sep. 2005.

[16] Liu, W., Tao, D., Cheng, J., and Tang, Y. 2013. Multiview Hessian discriminative sparse coding for image annotation. To appear in Computer Vision and Image Understanding.

[17] Devore, J. 2011. Probability and Statistics for Engineering and the Sciences. Cengage Learning.

[18] Song, D., and Tao, D. 2010. Biologically inspired feature manifold for scene classification. IEEE Transactions on Image Processing, vol. 19, no. 1, pp. 174-184.

[19] Yu, J., Tao, D., Rui, Y., and Cheng, J. 2013. Pairwise constraints based multiview features fusion for scene classification. Pattern Recognition, vol. 46, pp. 483-496.

[20] Vogel, J. and Schiele, B. 2007. Semantic modeling of natural scenes for content-based image retrieval. International Journal of Computer Vision, vol. 72, no. 2, pp. 133-157.

[21] Chu, W.-T., and Chen, C.-H. 2012. Color CENTRIST: A color descriptor for scene categorization. Proceedings of ACM International Conference on Multimedia Retrieval.

[22] Ojala, T., Pietikainen, M., and Maenpaa, T. 2002. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 7, pp. 971-987.

[23] Agarwala, A., Agrawala, M., Cohen, M., Salesin, D., and Szeliski, R. 2006. Photographing long scenes with multi-viewpoint panoramas. Proceedings of ACM SIGGRAPH, pp. 853-861.

[24] Lampert, C.H. 2009. Detecting objects in large image collections and videos by efficient subimage retrieval. Proceedings of IEEE International Conference on Computer Vision, pp. 987-994.

[25] Chang, C.-C. and Lin, C.-J. 2011. LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology, vol. 2, no. 3, Article No.: 27.

[26] Zhu, C., Bichot, C.-E., and Chen, L. 2010. Multi-scale color local binary patterns for visual object classes recognition. Proc. of International Conference on Pattern Recognition, pp. 3065-3068.

[27] Rasiwasia, N., and Vasconcelos, N. 2008. Scene classification with low-dimensional semantic spaces and weak supervision. Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition.

[28] Kwitt, R., Vasconcelos, N., and Rasiwasia, N. 2012. Scene recognition on the semantic manifold. Proceedings of European Conference on Computer Vision, pp. 359-372.

[29] Lian, X.-C., Li, Z., Lu, B.-L., and Zhang, L. 2010. Max-margin dictionary learning for multiclass image categorization. Proceedings of European Conference on Computer Vision, pp. 157-170.

[30] Krapac, J., Verbeek, J., and Jurie, F. 2011. Modeling spatial layout with fisher vectors for image categorization. Proceedings of IEEE International Conference on Computer Vision, pp. 1487-1494.

[31] Yu, X., Fermuller, C., Teo, C.L., Yang, Yezhou, and Aloimonos, Y. 2011. Active scene recognition with vision and language. Proceedings of IEEE International Conference on Computer Vision, pp. 810-817.

[32] Pandey, M., and Lazebnik, S. 2011. Scene recognition and weakly supervised object localization with deformable part-based models. Proceedings of IEEE International Conference on Computer Vision, pp. 1307-1314.

[33] Liu, W., and Tao, D. 2013. Multiview Hessian regularization for image annotation. IEEE Transactions on Image Processing, vol. 22, no. 7, pp. 2676-2687.

[34] Fornoni, M., and Caputo, B. 2012. Indoor scene recognition using task and saliency-driven feature pooling. Proceedings of British Machine Vision Conference.

[35] Niu, Z., Hua, G., Gao, X., and Tian, Q. 2012. Context aware topic model for scene recognition. Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 2743-2750.

[36] Tao, D., Jin, L., Liu, W., and Li, X. 2013. Hessian regularized support vector machines for mobile image annotation on the cloud. IEEE Transactions on Multimedia, vol. 15, no. 4, pp. 833-844.

[37] Liu, W., Wang, Y., and Li, S. 2011. LBP feature extraction for facial expression recognition. Journal of Information & Computational Science, vol. 8, no. 3, pp. 412-421.

[38] Shabou, A., and Le Borgne, H. 2012. Locality-constrained and spatially regularized coding for scene categorization. Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 3618-3625.

[39] Parizi, S.N., Oberlin, J., and Felzenszwalb, P.F. 2012. Reconfigurable models for scene recognition. Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 2775-2782.

[40] Zheng, Y., Jiang, Y.-G., and Xue, X. 2012. Learning hybrid part filters for scene recognition. Proceedings of European Conference on Computer Vision, pp. 172-185.

[41] Jiang, Y., Yuan, J., and Yu, G. 2012. Randomized spatial partition for scene recognition. Proceedings of European Conference on Computer Vision, pp. 730-743.

[42] Pietikäinen, M., Hadid, A., Zhao, G., and Ahonen, T. 2011. Computer Vision Using Local Binary Patterns. Springer.

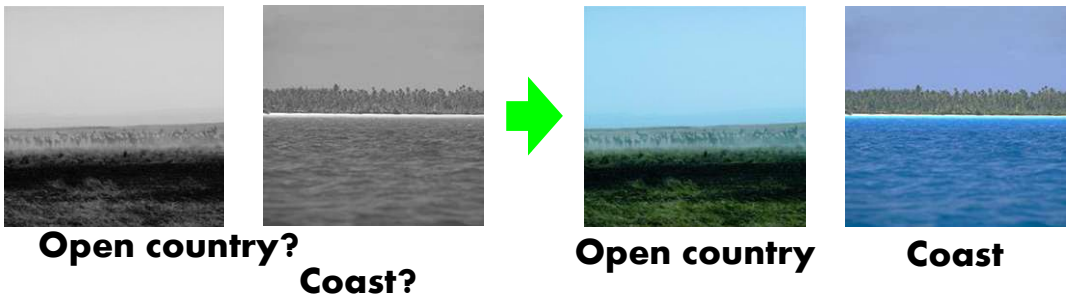

Figure 1. An example showing importance of color information in scene categorization. Left: gray-level images; Right: colorful images (better seen in color).

$$\begin{matrix} 32 & 64 & 128 \\ 32 & \mathbf{100} & 128 \\ 100 & 128 & 256 \end{matrix} \xrightarrow{\text{Census transform}} \begin{matrix} 1 & 1 & 0 \\ 1 & & 0 \\ 1 & 0 & 0 \end{matrix} \Rightarrow CT = (11010100)_2 = 212$$

Figure 2. An example of census transform.

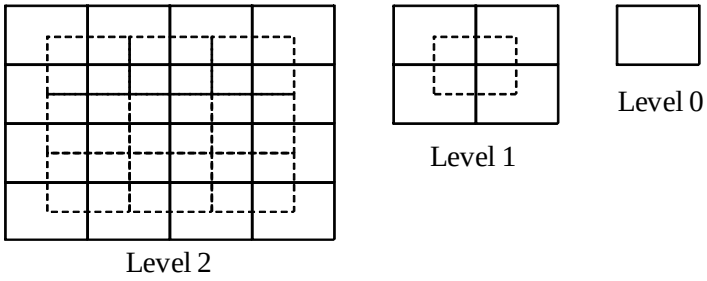

Figure 3. Illustration of levels 2, 1, and 0 split of an image [1].

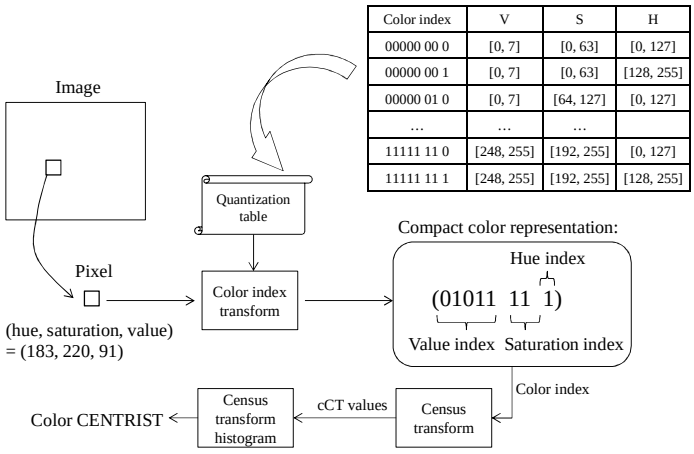

Figure 4. Flowchart for extracting color CENTRIST.

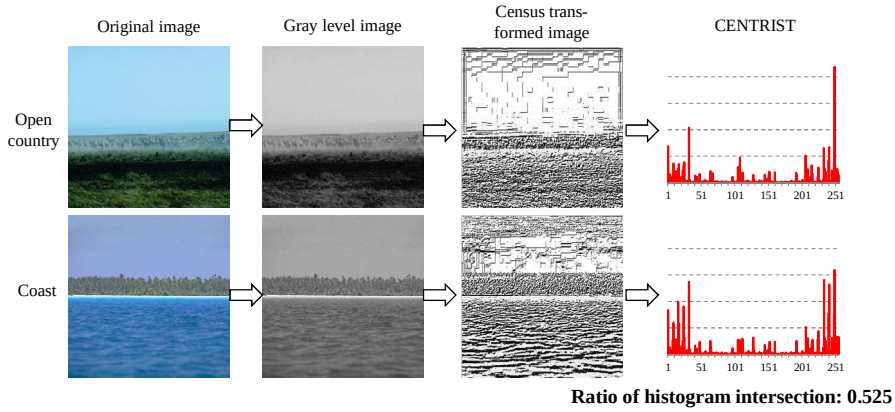

Figure 5. CENTRISTs of two images in different scene categories (better seen in color).

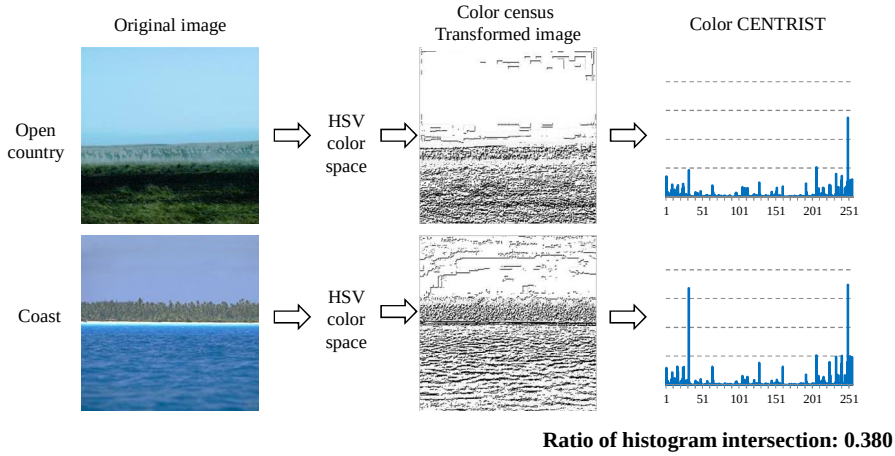

Figure 6. Color CENTRISTs of two images in different scene categories (better seen in color).

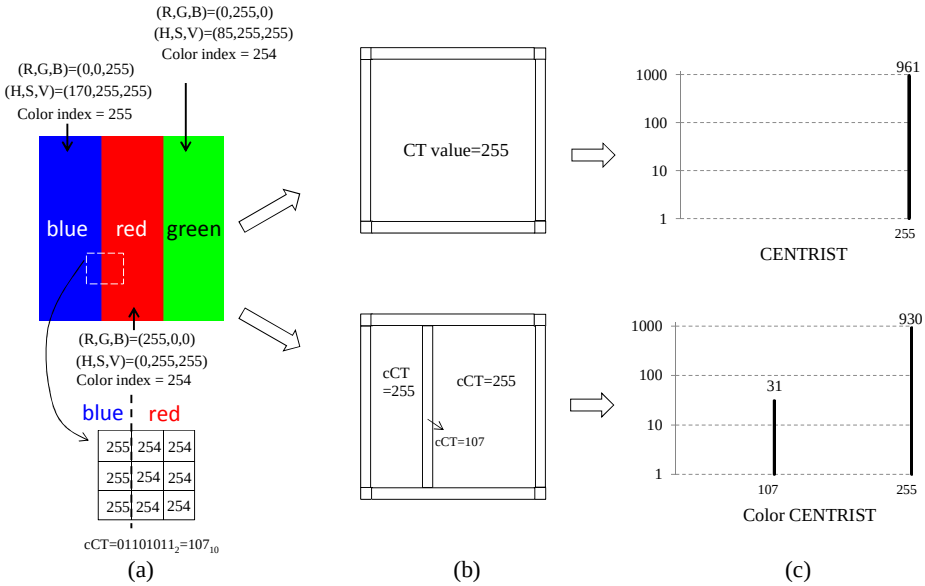

Figure 7. A designed example to illustrate the difference between CENTRIST and color CENTRIST.

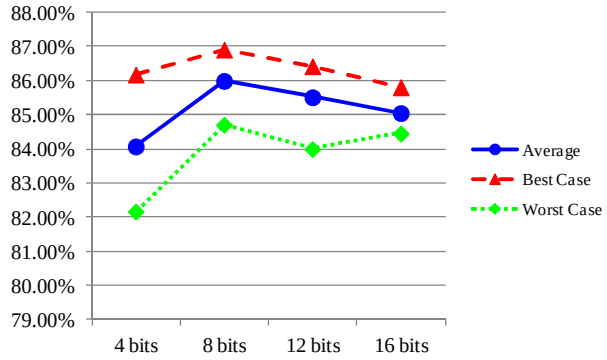

Figure 8. Based on the 8-class scene dataset, the best, the worst, and the average recognition accuracy obtained based on the 4-bit, 8-bit, 12-bit, and 16-bit settings, respectively.

Figure 9. Sample images from eight scene categories. These categories are coast, forest, highway, inside city, mountain, open country, street, and tall building, respectively (from left to right, top to down).

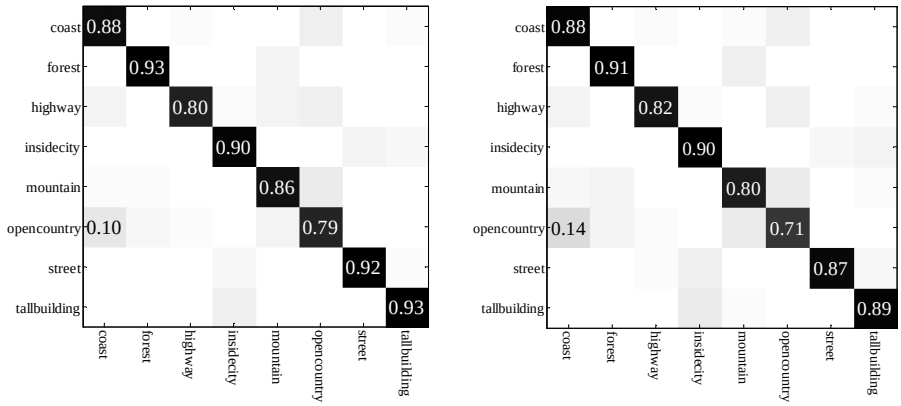

Figure 10. Confusion matrices of the 8-class scene dataset. Only rates higher than 0.1 are shown in the figure. Left: level 2 sPAcCT; right: level 2 sPACT.

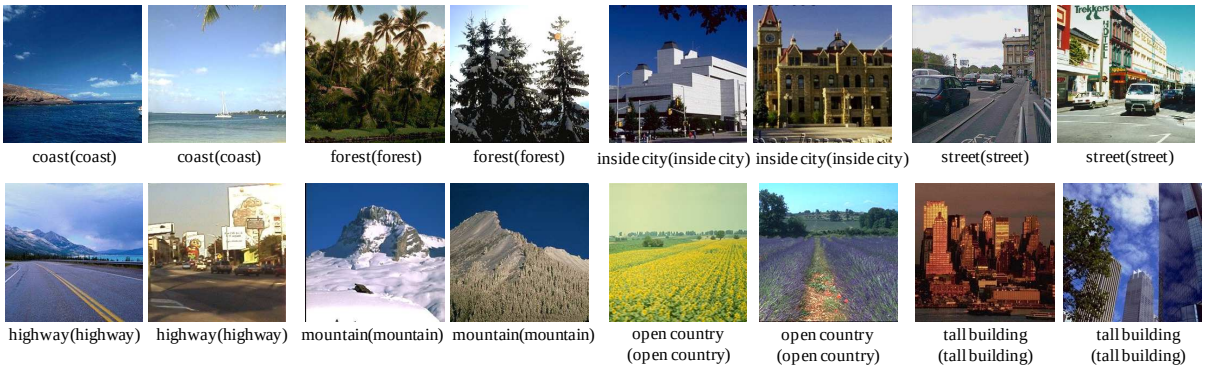

Figure 11. Examples of correctly recognized images.

Figure 12. Examples of incorrectly recognized images.

Figure 13. Sample images from the 8-class event dataset. The categories are badminton, bocce, croquet, polo, rowing, rock climbing, sailing, and snowboarding, respectively (from left to right, top to bottom).

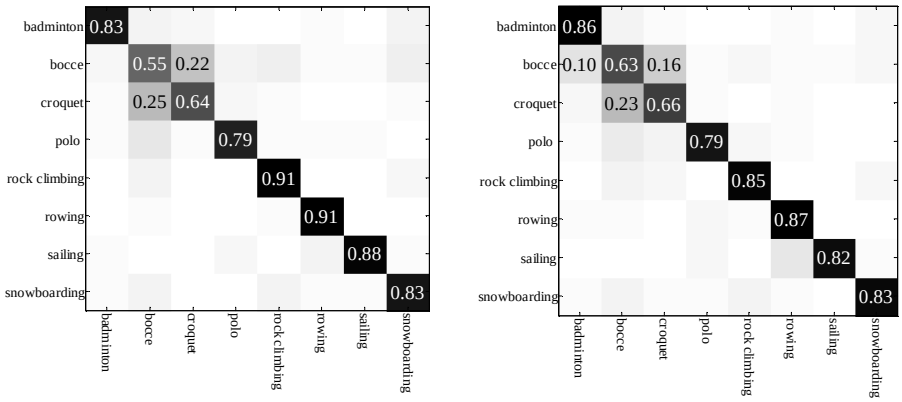

Figure 14. Confusion matrices of the 8-class event dataset. Only rates higher than 0.1 are shown in the figures. Left: sPAcCT; right: sPACT.

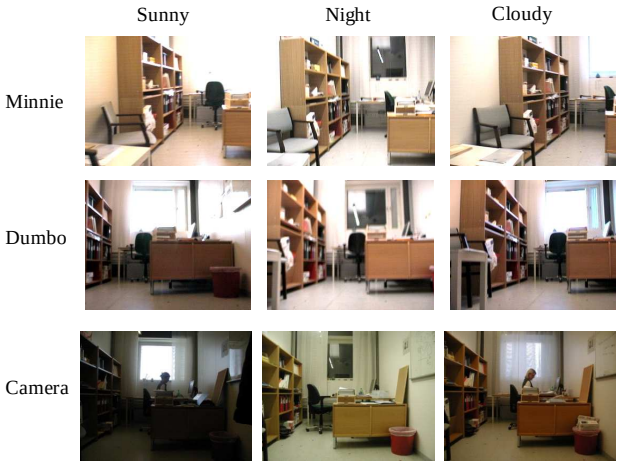

Figure 15. Sample images from the KTH-IDOL dataset (the first and the second rows) and the KTH-INDECS dataset (the third row), in different weather conditions. These examples show nearly the same angle of a one-person office.

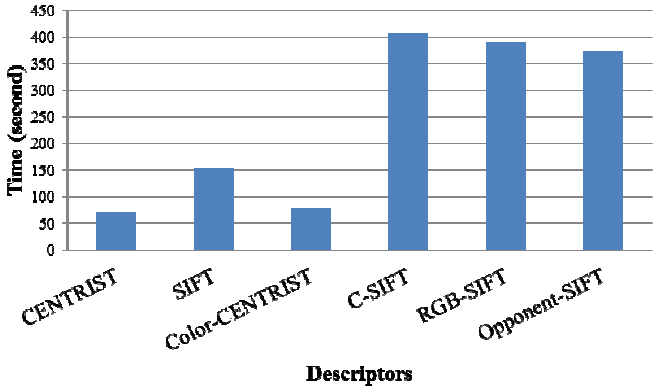

Figure 16. The total time needed to extract color CENTRIST and other descriptors from all images in the *coast* category of the 8-class scene dataset.

Figure 17. The panorama of the food region of a grocery store. Three query patches are also shown with enlarged views.

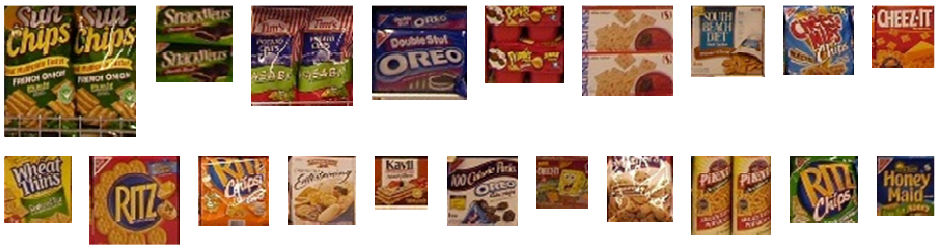

Figure 18. Twenty query patches extracted from the panorama.

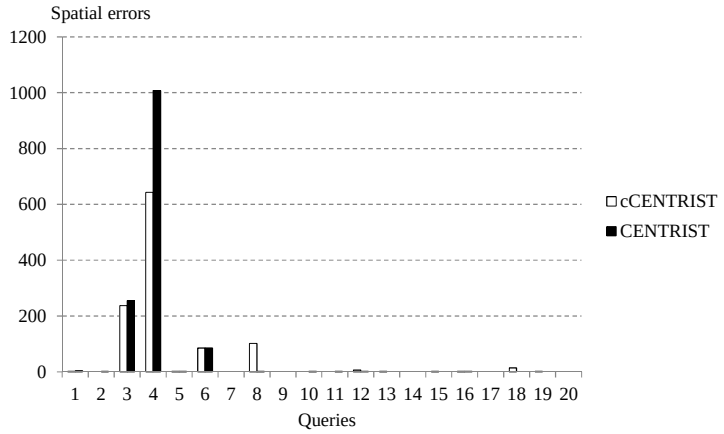

Figure 19. Spatial errors for different queries.

Table 1. Recognition rates under different bit allocation schemes, using the 8-bit setting, based on the 8-class scene dataset.

Different schemes	Setting	Recognition rates	p-values
1	H-S-V (8-0-0)	74.45±1.16	$\ll 0.001$
2	H-S-V (0-8-0)	83.11±0.42	$\ll 0.001$
3	H-S-V (0-0-8)	85.32±0.80	0.004
4	H-S-V (0-1-7)	85.76±0.42	0.004
5	H-S-V (0-2-6)	86.69±0.49	0.266
6	H-S-V (1-1-6)	86.74±0.44	0.298
7	H-S-V (0-3-5)	86.85±0.76	0.443
8	H-S-V (1-2-5)	**86.92±0.58**	
9	H-S-V (0-4-4)	85.39±1.05	0.014
10	H-S-V (1-3-4)	85.74±1.19	0.048
11	H-S-V (2-2-4)	85.87±1.10	0.054
12	H-S-V (2-3-3)	84.71±0.98	0.002
	Average	86.00	

Table 2. Detailed bit allocation schemes evaluated in the 4-bit, 8-bit, 12-bit, and 16-bit settings.

4-bit setting	8-bit setting	12-bit setting	16-bit setting
H-S-V (0-0-4)	H-S-V (8-0-0)	H-S-V (0-4-8)	H-S-V (0-8-8)
H-S-V (0-1-3)	H-S-V (0-8-0)	H-S-V (1-3-8)	H-S-V (1-7-8)
H-S-V (0-2-2)	H-S-V (0-0-8)	H-S-V (2-2-8)	H-S-V (2-6-8)
H-S-V (1-1-2)	H-S-V (0-1-7)	H-S-V (0-5-7)	H-S-V (3-5-8)
	H-S-V (0-2-6)	H-S-V (1-4-7)	H-S-V (4-4-8)
	H-S-V (1-1-6)	H-S-V (2-3-7)	H-S-V (2-7-7)
	H-S-V (0-3-5)	H-S-V (0-6-6)	H-S-V (3-6-7)
	H-S-V (1-2-5)	H-S-V (1-5-6)	H-S-V (4-5-7)
	H-S-V (0-4-4)	H-S-V (2-4-6)	H-S-V (4-6-6)
	H-S-V (1-3-4)	H-S-V (3-3-6)	H-S-V (5-5-6)
	H-S-V (2-2-4)	H-S-V (2-5-5)	
	H-S-V (2-3-3)	H-S-V (3-4-5)	
		H-S-V (4-4-4)	

Table 3. Best average recognition accuracies based on the HSV, RGB, and Lab color spaces.

	HSV	RGB	Lab
8 scene	86.92±0.58	**87.47**±0.46	86.58±0.60
67 scene	36.09±0.70	34.07±1.28	**36.51±0.87**

Table 4. Recognition rates on the 8-class scene dataset, based on CENTRISTs or cCENTRISTs with different settings.

Level	Method	Feature type	Rates	p-values
0	CENTRIST	CENTRIST, not using PCA	77.70±1.04	0.030
0	cCENTRIST	cCENTRIST, not using PCA	**79.19±1.12**	
1	sPACT	CENTRIST, 40 eigenvectors	83.75±0.66	0.002
1	sPAcCT	cCENTRIST, 40 eigenvectors	**85.53±0.77**	
1	scCT	cCENTRIST, not using PCA	83.26±0.72	$\ll 0.001$
2	sPACT	CENTRIST, 40 eigenvectors	84.63±1.08	0.003
2	sPAcCT	cCENTRIST, 40 eigenvectors	**86.92±0.58**	
2	scCT	cCENTRIST, not using PCA	83.32±1.14	$\ll 0.001$
3	sPACT	CENTRIST, 40 eigenvectors	83.92±0.74	0.11
3	sPAcCT	cCENTRIST, 40 eigenvectors	**84.62±0.92**	
3	scCT	cCENTRIST, not using PCA	81.36±0.72	$\ll 0.001$

Table 5. Recognition rates on the 8-class event dataset.

L	Method	Feature type	Rates	p-value
0	CENTRIST	CENTRIST, not using PCA	65.24±1.78	0.04
0	cCENTRIST	cCENTRIST, not using PCA	**67.12±1.06**	
1	sPACT	CENTRIST, 40 eigenvectors	77.37±1.37	0.14
1	sPAcCT	cCENTRIST, 40 eigenvectors	**78.16±0.53**	
1	scCT	cCENTRIST, not using PCA	74.07±0.89	$\ll 0.001$
2	sPACT	CENTRIST, 40 eigenvectors	79.82±0.75	0.45
2	sPAcCT	cCENTRIST, 40 eigenvectors	**79.88±0.59**	
2	scCT	cCENTRIST, not using PCA	74.15±0.54	$\ll 0.001$
3	sPACT	CENTRIST, 40 eigenvectors	78.04±0.42	0.13
3	sPAcCT	cCENTRIST, 40 eigenvectors	**78.39±0.50**	
3	scCT	cCENTRIST, not using PCA	70.93±1.12	$\ll 0.001$

Table 6. Recognition rates on the 67-class indoor scene dataset.

L	Method	Feature type	Rates	p-value
0	CENTRIST	CENTRIST, not using PCA	22.09±1.71	0.084
0	cCENTRIST	cCENTRIST, not using PCA	**23.67±1.57**	
1	sPACT	CENTRIST, 40 eigenvectors	30.84±1.61	0.057
1	sPAcCT	cCENTRIST, 40 eigenvectors	**32.40±1.10**	
2	sPACT	CENTRIST, 40 eigenvectors	34.48±0.98	0.010
2	sPAcCT	cCENTRIST, 40 eigenvectors	**36.09±0.70**	

Table 7. Average recognition accuracies on the KTH-IDOL dataset (Setting 1).

Exp	Train	Test	Condition	sPACT	sPAcCT
1	Minnie	Minnie	Cloudy	94.85%	**95.15%**
2	Minnie	Minnie	Sunny	**97.24%**	97.18%
3	Minnie	Minnie	Night	**93.10%**	92.27%

Table 8. Average recognition accuracies on the KTH-IDOL dataset (Setting 2).

Exp	Train	Test	Train condition	Test Condition	sPACT	sPAcCT
1	Minnie	Minnie	Sunny1	Night2	**80.69%**	79.76%
2	Minnie	Minnie	Night1	Sunny2	**86.10%**	83.04%
3	Minnie	Minnie	Cloudy1	Sunny2	92.93%	**93.40%**
4	Minnie	Minnie	Sunny1	Cloudy2	91.01%	**91.12%**
5	Minnie	Minnie	Night1	Cloudy2	**90.39%**	87.81%
6	Minnie	Minnie	Cloudy1	Night2	**92.72%**	90.09%

Table 9. Average recognition accuracies on the KTH-IDOL dataset (Setting 3).

Exp	Train	Test	Condition	sPACT	sPAcCT
1	Minnie	Dumbo	Cloudy	**74.96%**	73.28%
2	Minnie	Dumbo	Sunny	**78.81%**	76.86%
3	Minnie	Dumbo	Night	**74.19%**	72.20%

Table 10. Average recognition accuracies on the KTH-INDECS dataset (Setting 4).

Exp	Train	Test	Train condition	Test condition	sPACT	sPAcCT
1	Camera	Camera	Sunny	Night	84.52%	**86.54%**
2	Camera	Camera	Night	Sunny	87.04%	**89.26%**
3	Camera	Camera	Cloudy	Sunny	**95.28%**	92.96%
4	Camera	Camera	Sunny	Cloudy	**93.70%**	92.78%
5	Camera	Camera	Night	Cloudy	**92.31%**	91.39%
6	Camera	Camera	Cloudy	Night	89.10%	**91.30%**

Table 11. Recognition rates on the 8-class scene dataset, based on CENTRIST, cCENTRIST, and the combined descriptor.

L	Method	Feature type	Rates	p-value
0	CENTRIST	CENTRIST, not using PCA	77.70±1.04	$\ll$0.001
0	cCENTRIST	cCENTRIST, not using PCA	79.19±1.12	0.016
0	combined		**80.76±0.59**	
1	sPACT	CENTRIST, 40 eigenvectors	83.75±0.66	$\ll$0.001
1	sPAcCT	cCENTRIST, 40 eigenvectors	85.53±0.77	0.236
1	combined		**85.84±0.42**	
2	sPACT	CENTRIST, 40 eigenvectors	84.63±1.08	$\ll$0.001
2	sPAcCT	cCENTRIST, 40 eigenvectors	86.92±0.58	0.136
2	combined		**87.46±0.75**	

Table 12. Average recognition rates on the 8-class scene data set, using different descriptors with the bag of word framework.

Color descriptors					
Pyramid level	Codebook size	Color CENTRIST	C-SIFT	RGB-SIFT	Opponent-SIFT
0	200	78.47±0.72	79.47±0.92	80.40±0.51	80.72±0.19
0	400	81.46±0.59	82.18±0.87	83.16±0.31	83.54±0.54
2	200	82.90±0.43	82.82±0.27	83.42±0.71	83.02±0.82
2	400	84.24±0.68	83.53±0.46	84.40±0.99	83.92±0.48
Grey level descriptors					
Pyramid level	Codebook size	CENTRIST	SIFT		
0	200	78.08±0.59	79.12±0.54		
0	400	80.58±0.57	81.08±0.69		
2	200	83.12±0.67	82.14±0.77		
2	400	84.99±0.39	83.21±0.99		

Table 13. Performance comparison between color CENTRIST and three color LBPs, at various levels, based on the 8-class scene dataset. The p-values of all comparisons between the cCENTRIST and other descriptors are much less than 0.001.

Descriptors	Level 0	Level 1	Level 2	Level 3
Hue-LBP	64.99±1.06	72.18±1.02	75.12±0.45	74.15±0.48
Opponent-LBP	62.24±0.57	72.58±1.56	75.22±1.33	72.82±0.78
nOpponent-LBP	70.41±1.19	77.16±0.83	79.80±0.94	78.15±0.31
cCENTRIST	**79.19±1.12**	**85.53±0.77**	**86.92±0.58**	**84.62±0.92**