Color CENTRIST: A Color Descriptor for Scene Categorization

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

ABSTRACT
We design a method to incorporate color information into the framework of CENsus Transform histogram (CENTRIST), a state-of-the-art visual descriptor for scene categorization. The newly proposed color CENTRIST descriptor describes global shape information by not only gradient derived from intensity values but also color variations between pixels in local image patches. Through extensive evaluations on various datasets, we demonstrate that the color CENTRIST descriptor is not only easily to be implemented, but also reliably achieves performance over that of CENTRIST.

Categories and Subject Descriptors
I.4.8 [**Image Processing and Computer Vision**]: Scene Analysis – color, shape. I.4.7 [**Image Processing and Computer Vision**]: Feature Measurement – *feature representation.*

General Terms
Algorithms, Performance, Experimentation.

Keywords
Census transform histogram, scene categorization, color descriptor.

1. INTRODUCTION
Scene categorization, or scene classification, has become a fundamental process for efficient image browsing, retrieval, and organization. For example, if an image's scene category can be recognized, such as office and street, we would reduce the search space of object recognition, or more accurately detect semantic concepts present in this image. Place recognition, a subproblem of scene recognition, may 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 such task.

In the literature, many visual descriptors have been proposed for image scene recognition. The existing descriptors can be roughly divided into two groups: 1) part-based representation, with some considerations of multiple scales or spatial distribution, and 2) holistic representation that directly models global configurations. The former approach describes texture/shape information in local image patches, which has been proven effective to detect objects under various conditions. By considering distributions of local descriptors over all 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 global distribution is spatial pyramids [3]. Despite the SIFT descriptors plus the bag of visual words model [8] 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 appearance would present in the same type of scene. Without considering detailed 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 target on gray images, and the proposed visual descriptors mainly rely on oriented gradient calculated based on intensity values. In this paper, we would like to study scene categorization for *color images*. We devise a new visual descriptor, i.e. *color CENTRIST* that incorporates color information into the framework of CENTRIST, and demonstrate its effectiveness through evaluating various color image datasets. Based on comprehensive evaluation, we conclude that considering color information indeed benefits scene categorization.

The rest of this paper is organized as follows. Section 2 provides brief literature survey. The color CENTRIST descriptor is proposed after briefly reviewing conventional CENTRIST in Section 3. We provide comprehensive evaluation on various datasets in Section 4. Section 5 concludes this paper with discussions of the proposed method and future research.

2. RELATED WORK
Scene change detection has been a critical issue in video analysis for many years. Many studies evaluate visual coherence between video keyframes, mainly based on color and motion information, and accordingly detect scene boundaries by identifying the timestamps at which visual information changes significantly [16][17].

Based on the experience of video scene detection, color and texture information was widely used in image scene recognition [18][19]. However, as the variations of scene categories increase, more elegant features are needed to provide more reliable performance. Currently, SIFT descriptors [7] associated with the bag of word model [8] have dominated the descriptor choice in

ICMR'12, June 5-8, Hong Kong, China

scene categorization. Fei-Fei and Perona [5] describe images by a collection of local regions, which are represented by codewords derived from a visual word codebook. They propose the theme models modified from the Latent Dirichlet Allocation to represent the distribution of codewords in each scene category. Lazebnik et al. [3] argue that describing bags of visual words in multiple scales provide encouraging performance on recognizing natural scene categories. Focusing on codebook design for scene categorization, van Gemert et al. [6] deal with the issues of codeword uncertainty and codeword plausibility. They propose a kernel codebook method to allow some degree of ambiguity in assigning a visual descriptor to codewords. Also based on bag of word representation, Bosch et al. [11] investigate classification methodologies for scene categorization. They proposed a hybrid approach that first discovers latent topics in scene images by pLSA (probabilistic latent semantic analysis), and then topic distributions are fed to discriminative classifiers based on KNN or SVM. Rather than directly modeling an image by a collection of shape and texture features in local image patches, Vogel and Schiele [20] first detect semantic concepts for each image patch, and then model an image by the distribution of concept occurrence. Support vector machine classifiers are constructed to detect scene categories.

Oliva and Torralba [2] argue that recognizing a scene not necessarily needs modeling object information but global configurations. They propose the GIST descriptor to model shape of a scene, and assume that images coming from the same scene category have similar configurations. This idea has been proven effective in recognizing outdoor scenes, e.g., mountain and coast. However, the performance decreases significantly for indoor scenes. Based on census transform, Wu and Rehg [1] propose a simple yet effective visual descriptor to model global configurations of scenes. They demonstrate that shape information can be effectively described by comparing the intensity value of a pixel with its eight neighboring pixels. Comprehensive studies were provided in [1] to show the histograms of census transform values, at multiple levels, can provide superior performance over SIFT and GIST in most cases.

Most works in the literature focus on describing scenes in grey-level images, because shape or texture information can be effectively extracted from them. Much fewer studies have been conducted to investigate how color information affects scene categorization. The work in [11] is one of the few studies that investigate color descriptors. From their reported results, color information consistently brings performance increment if it is appropriately incorporated into visual descriptors. Van de Sande et al. [13] evaluate color variants of SIFT descriptors on object and scene recognition. Their results also conform to the trend, but only SIFT-based descriptors were evaluated. In this paper, we design a method to incorporate color information into one of the state-of-the-art visual descriptor, i.e., CENTRIST [1], and demonstrate its effectiveness through comprehensive evaluation.

3. DESCRIPTORS

3.1 CENTRIST

To handle with scene categorization, Wu and Rehg describe desired properties of appropriate visual descriptors [1]:

- Holistic representation: Exactly knowing objects in a scene does not necessarily benefit scene categorization. Oliva and Torralba therefore propose a holistic representation of spatial envelope [2].
- Capturing the structural properties: The desired descriptor is expected to capture general structural properties such as rectangular shapes and flat surfaces, while suppressing detailed texture.
- Rough geometry is useful: Variations of scenes would be higher than that of objects. Rough geometrical constraints are helpful in categorizing scenes.
- Generalizability: A good descriptor would be compact within a category even under large visual variations, and would be distinct for different scene categories.

By considering the properties mentioned above, Wu and Rehg propose a visual descriptor called CENsus TRansform hISTogram (CENTRIST), which is a holistic representation modeling distribution of local structures. Rough geometrical information is captured by describing CENTRIST extracted from spatial pyramids [3]. In the following, we briefly review conventional CENTRIST before we propose its color version.

Census transform [4] compares the intensity value of a pixel with that of its eight spatially neighboring pixels. An example is shown in Figure 1. If the intensity of the center pixel is larger than one of its neighbors, a bit 1 is set in the corresponding position. Otherwise, a bit 0 is set. From left-top to right-bottom, these bits are concatenated to form a binary representation, which can be evaluated to a base-10 number called Census Transform value (CT value) for the center pixel.

$$\begin{array}{ccc} 32 & 64 & 96 \\ 32 & \mathbf{64} & 96 \\ 32 & 32 & 96 \end{array} \Rightarrow \begin{array}{ccc} 1 & 1 & 0 \\ 1 & & 0 \\ 1 & 1 & 0 \end{array} \Rightarrow CT = (11010110)_2 = 214$$

Figure 1. An example of census transform [1].

After evaluating the CT value for each pixel, the histogram of CT values is constructed to form the CENTRIST descriptor. Note that CENTRIST is 256-dimensional because there are 256 different types of CT values. In [1], the authors discuss properties of CT values and CENTRIST descriptors.

CENTRIST can only encode global shape structure in a small image patch. To capture rough global shape structure in an image, the spatial pyramid scheme [3] is used, as illustrated in Figure 2. The image is split into $2^2 \times 2^2 = 16$ blocks at level 2. These blocks are also shifted (dash line blocks) to avoid artifacts caused by nonoverlapping division. Therefore, there are 25 blocks at level 2, 5 blocks at level 1, and 1 block at level 0. From each block, the CENTRIST descriptor is extracted, and descriptors from all blocks are concatenated to describe the image. Different dimensions of the CENTRIST descriptor are not independent, and thus Wu and Rehg [1] use principal component analysis (PCA) to reduce dimensionality of CENTRIST to 40. This compact representation is called spatial Principal component Analysis of Census Transform (spatial PACT) histogram, or sPACT, in [1]. In this case, an image with level 2 pyramids is thus described by a $40 \times (25 + 5 + 1) = 1200$-dimensional descriptor.

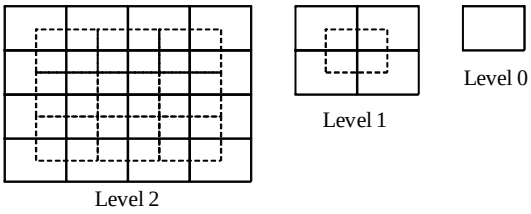

Figure 2. Spatial pyramids of levels 2, 1, and 0 of an image.

Effectiveness of the CENTRIST descriptor on scene categorization was comprehensively studied in [1]. However, the authors also point out some limitations of this descriptor: 1) CENTRIST is not invariant to rotations or scale changes, though requirements of rotation and scale invariance are not critical for scene categorization. 2) CENTRIST is not a precise shape descriptor, which makes it inappropriate for shape retrieval. 3) CENTRIST ignores color information, and thus color information is not fully exploited for scene categorization.

In this paper, we would like to incorporate color information into the CENTRIST descriptor. We will demonstrate that the proposed *color CENTRIST* descriptor effectively enhances performance of scene categorization, through evaluating a wide range of colorful image datasets.

3.2 Color CENTRIST

We represent color by the hue-saturation-value (HSV) color space, which is quantized into M quantized color ranges. To reflect different effects of different color components, we take different quantization granularities for different color components. In the following, we mainly set $M = 256$, for which the hue, saturation, and value components are equally quantized into two, four, and thirty-two ranges, respectively. That is, we index hue, saturation, and value of a pixel by 1, 2, and 5 bits, respectively. To give highest priority of the value component, these color indices are concatenated in the manner (value index, saturation index, hue index). Through this process, each pixel in a color image is represented as an 8-dimensional color index, which value ranges from 0 to 255. Figure 3 shows the flowchart for extracting color CENTRIST.

Although each pixel is represented from 0 to 255 as well, this representation describes color information rather than intensity value used in the conventional CENTRIST. With this representation, we follow the same process illustrated in Figure 1 to transform each pixel into a CT value, and an image's global shape structure, with the consideration of color information, can be represented by a histogram of CT values. We call this histogram *color CENTRIST*. Similarly, we would reduce dimensionality of color CENTRIST by PCA, and model rough global shape structure of an image based on spatial pyramids.

We use Figure 4 and Figure 5 to underline the difference between CENTRIST and color CENTRIST. In Figure 4, one image from the "open country" category and one image from the "coast" category are compared based on CENTRIST. From the second column of Figure 4, we see these two images look similar when they are represented in gray. From the third column, we see census transformed images can effectively represent shape information, which conforms to the description in [1]. We measure similarity between these two images by histogram intersection. Based on CENTRIST, the ratio of histogram intersection is 0.525. On the contrary, from the last column of Figure 5, color CENTRISTs of two images pose higher difference, and the ratio of histogram intersection between these two images is just 0.380. Describing images by more appropriate features provides more clues for scene categorization. This conjecture will be verified in the evaluation section.

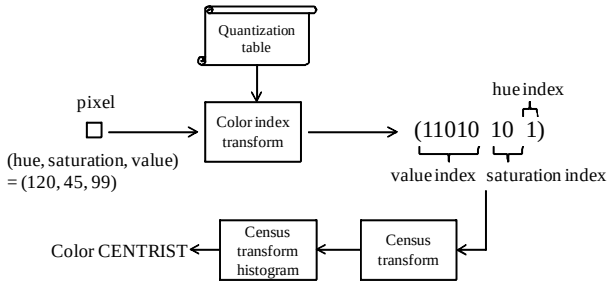

Figure 3. Flowchart for extracting color CENTRIST.

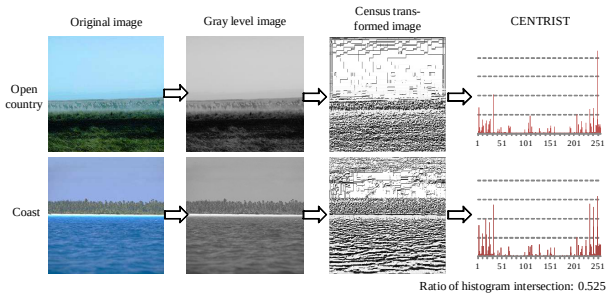

Figure 4. CENTRISTs of two images in different scene categories.

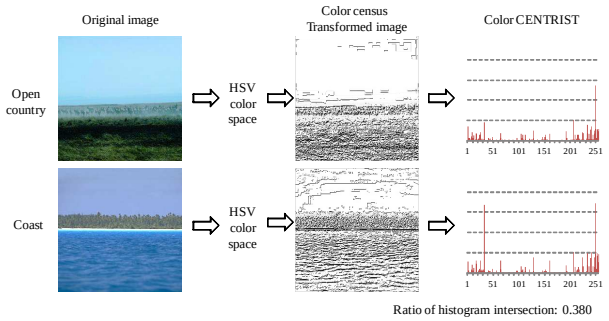

Figure 5. Color CENTRISTs of two images in different scene categories.

4. EXPERIMENTS

In this section, we first show how different color space quantization settings influence the scene categorization accuracy. After finding the best setting, the color CENTRIST visual descriptor is tested 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 [12][15]. In each dataset, data are randomly split into a training set and a testing set, with detailed settings described later. The random splitting is repeated five times, and the average accuracy is reported.

In the following experiments, we remove the two bins with CT values equal to 0 and 255 in both color CENTRIST and conventional CENTRIST, and normalize them such that they have unit norms. Similar to sPACT in [1], to reduce dimensionality of

color CENTRIST, 40 eigenvectors corresponding to 40 largest eigenvalues are found, and 256-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, average and standard deviation of intensity values in a block are added to the sPACT [1]. We analogize this setting and add average and standard deviation of color indices in a block to 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 are applied to conduct scene categorization[1].

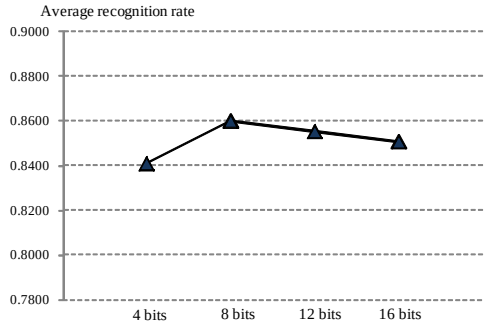

Figure 6. Average recognition rates based on 8-class scene dataset [2], when the quantization levels of the HSV color space are represented by 4 bits, 8 bits, 12 bits, and 16 bits.

4.1 Color Quantization

To represent color information, we quantize the HSV color space into a number of color ranges, and describe each pixel by concatenating quantization indices with respect to hue, saturation, and value. To determine the number of bits to describe quantization levels, we examine the average scene recognition rates for the 8-class scene dataset, by using 4 bits, 8 bits, 12 bits, and 16 bits to describe quantization levels, respectively. For the setting of 8 bits, for example, we test different allocation schemes and calculate the average recognition rate (c.f. Table 1). Figure 6 shows that 8-bits setting, i.e. quantizing the color space into 256 levels, most appropriately describes color information and achieves the best recognition performance. Using more bits to describe color information does not necessarily achive better performance. This may be because quantizing too finely makes the influence of noise apparent. In scene categorization, we are not willing to accurately distinguish light green and dark green, for example.

Table 1 shows detailed results of different allocation schemes of the 8-bits setting. The result in the seventh 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. Therefore, the second row of this table 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 fourth row is similar to CENTRIST (but not exactly the same) because only (quantized) intensity is considered to do census transform.

By comparing the second to the fourth rows, we clearly see that intensity values still play the most important role in scene description. However, by considering hue and saturation, and appropriately quantizing different color channels, better performance can be further achieved. From Table 1, we see that the allocation scheme (H-S-V 1-2-5) gives the best performance. Therefore, as we describe in Sec. 3.2, the hue, saturation, and value components are equally quantized into two, four, and thirty-two ranges, respectively. To give highest priority of the value component, color indices are concatenated in the manner (value index, saturation index, hue index). This setting is used in the following experiments.

Table 1. Recognition rates under different bit allocation strategies.

Setting	Recognition rates
H-S-V (8-0-0)	74.45±1.16
H-S-V (0-8-0)	83.11±0.42
H-S-V (0-0-8)	85.32±0.80
H-S-V (0-1-7)	85.76±0.42
H-S-V (0-2-6)	86.69±0.49
H-S-V (1-1-6)	86.74±0.44
H-S-V (0-3-5)	86.85±0.76
H-S-V (1-2-5)	**86.92±0.58**
H-S-V (0-4-4)	85.39±1.05
H-S-V (1-3-4)	85.74±1.19
H-S-V (2-2-4)	85.87±1.10
H-S-V (2-3-3)	84.71±0.98

4.2 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 context) based on this smaller dataset. This data set contains a wide range of scene categories in outdoor environments, such as coast, forest, mountain, and etc. Figure 7 shows some sample images. All these color images are normalized to 256×256 pixels, and there are 260 to 410 images in each category.

The five-fold cross validation scheme is used to evaluate performance. In each fold, 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. We compare CENTRIST with cCENTRIST, in the representation of level 0, the representation of level 1 with PCA, and the representation of level 2 with PCA. Table 2 shows the experimental results. We see that the proposed cCENTRIST stably has superior performance over CENTRIST at all levels. These results verify that color information provide extra benefit over shape for scene categorization. Another observation is that level 2 representation with PCA provides better performance over levels 0 or 1 for both CENTRIST and cCENTRIST. This conforms to the trend reported in [1] and [3].

[1] The software for extracting color CENTRIST is available at: http://www.cs.ccu.edu.tw/~wtchu/projects/cCENTRIST/index.html

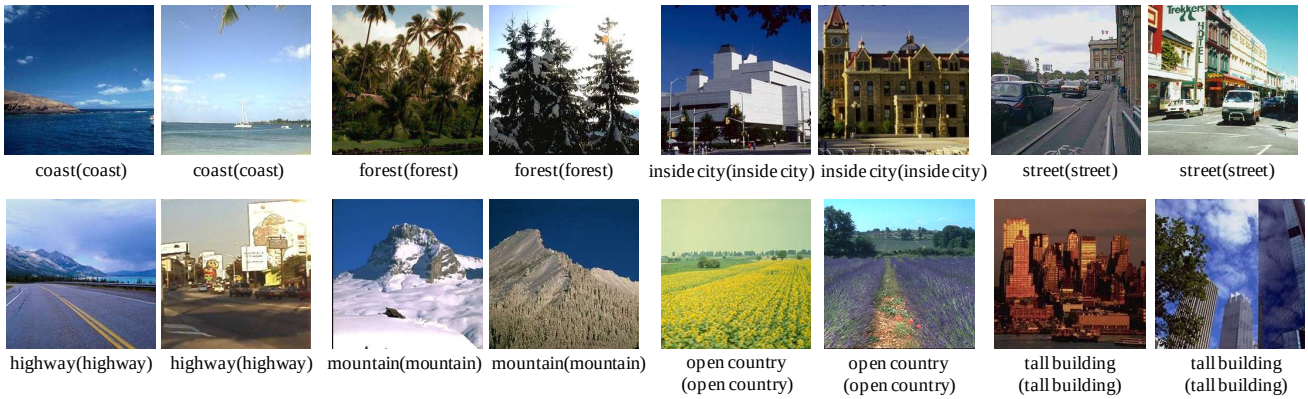

Figure 9. Examples of correctly recognized images.

Table 2. Recognition rates on the 8-class scene dataset.

L	Method	Feature type	Rates
0	CENTRIST	CENTRIST, not using PCA	77.70±1.04
0	cCENTRIST	cCENTRIST, not using PCA	**79.19±1.12**
1	sPACT	CENTRIST, 40 eigenvectors	83.75±0.66
1	sPAcCT	cCENTRIST, 40 eigenvectors	**85.53±0.77**
2	sPACT	CENTRIST, 40 eigenvectors	84.63±1.08
2	sPAcCT	cCENTRIST, 40 eigenvectors	**86.92±0.58**

Figure 7. A sample image from each of the 8 scene categories. These categories are coast, forest, highway, inside city, mountain, open country, street, and tall building, respectively (from top to bottom, and from left to right).

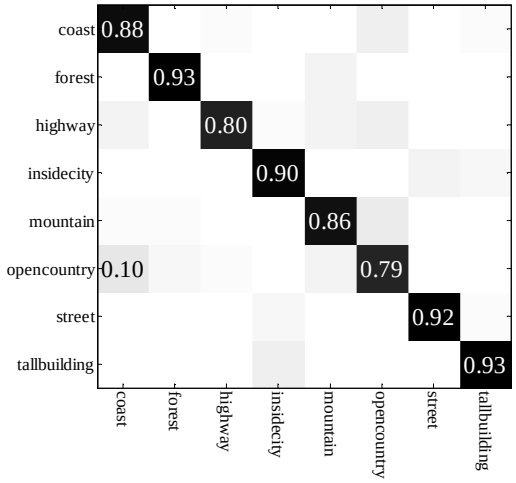

Figure 8. Confusion matrix of the 8-class scene data set. Only rates higher than 0.1 are shown in the figure.

Figure 10. Examples of incorrectly recognized images.

The confusion matrix of scene recognition based on level 2 sPAcCT is shown in Figure 8, where rows are true labels and columns are predicted labels. We obtain the best performance for forest and tall building categories. CENTRIST also works best for forest but doesn't work that well for tall building [1]. There is clear shape 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/coast, which also conforms to the trend reported in [1] and [3].

Figure 9 and Figure 10 show images that are correctly and incorrectly recognized, respectively. The caption coast(coast), for example, means the corresponding image is detected as coast, while the true label is coast. From Figure 9 we see cCENTRIST achieves reliable performance even there is significant intra-class variation. On the other hand, in Figure 10, some cases that may also confuse humans still annoy the proposed descriptor.

4.3 The 8-Class Event Dataset

The 8-class event dataset includes images of eight sports: badminton, bocce, croquet, polo, rowing, rock climbing, sailing, and snowboarding (see Figure 11 for example images from each category). Although this dataset was designed for event recognition, in this experiment we classify events by classifying the scenes, and do not attempt to recognize objects or persons.

In contrast to the 8-scene dataset, 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-fold cross validation scheme, 70 images per class are randomly selected for training, and the remaining images are for testing. Similarly, we respectively construct multiclass SVM classifiers with the RBF kernel, based on CENTRIST or cCENTRIST in the representation of level 0, the representation of level 1 with PCA, and the representation of level 2 with PCA.

Table 3 shows experimental results. Similar to the results for the 8-class scene dataset, cCENTRIST achieves better performance over CENTRIST with all levels of representations. But interestingly, the performance superiority of cCENTRIST decreases as level increases, which is opposite to results in Table 2. Comparing sample images in Figure 7 with Figure 11, the reason for such trend may be less regular-texture regions in images of the 8-class event dataset. Moreover, even in the same sports game, color of different players' uniforms may be significantly different. This also diminishes usefulness of color information.

Figure 12 shows the confusion matrices of scene recognition based on level 2 sPAcCT (top) and level 2 sPACT (bottom), respectively. In both matrices, the most confused case is croquet/bocce, which is reasonable because the pair of events shares very similar scenes or backgrounds. Comparing these two matrices, sPAcCT works better for discriminating sailing/rowing. Both rowing and sailing have a flat background such as water or sky. Color information in sPAcCT helps in distinguishing water and sky.

Table 3. Recognition rates on the 8-class event dataset.

L	Method	Feature type	Rates
0	CENTRIST	CENTRIST, not using PCA	65.24±1.78
0	cCENTRIST	cCENTRIST, not using PCA	**67.12±1.06**
1	sPACT	CENTRIST, 40 eigenvectors	77.37±1.37
1	sPAcCT	cCENTRIST, 40 eigenvectors	**78.16±0.53**
2	sPACT	CENTRIST, 40 eigenvectors	79.82±0.75
2	sPAcCT	cCENTRIST, 40 eigenvectors	**79.88±0.59**

Figure 11. Sample images from 8-class event dataset. The categories are badminton, bocce, croquet, polo, rowing, rock climbing, sailing, and snowboarding, respectively (from top to bottom, and from left to right).

4.4 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 totally 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 achieved about 21 percent average recognition accuracy on this challenging data set. By jointly considering local information, the accuracy was improved to 25 percent.

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-fold cross validation 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 4 shows the experimental results. The average recognition accuracy for level 2 sPAcCT is 36.09±0.70%, while the average recognition accuracy for level 2 sPACT is 34.48±0.98%. In this challenging indoor scene recognition problem, the performance of sPAcCT derived from cCENTRIST is better than GIST and sPACT.

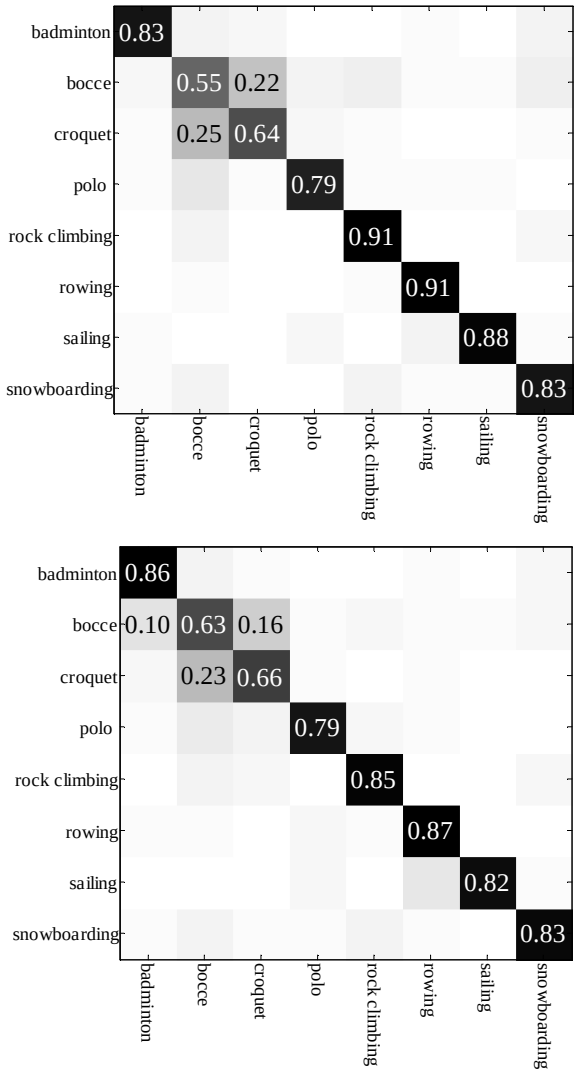

Figure 12. Confusion matrix of the 8-class event dataset. Only rates higher than 0.1 are shown in the figures. Top: sPAcCT; bottom: sPACT [1].

Table 4. Recognition rates on the 67-class indoor scene dataset.

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

4.5 The KTH-IDOL and The KTH-INDECS Dataset

The KTH Image Database for rObot Localization (IDOL) dataset [14] was captured by two robots, Minnie and Dumbo, that took pictures in a five-room office environment, including a one-person office, a two-person office, a kitchen, a corridor, and a printer area. This dataset was designed to recognize which room the robot is in based on a single image.

A robot captured a complete image sequence when it drove through all five rooms. Images were taken under three weather conditions: cloudy, night, and sunny. For each robot and each weather condition, four runs were captured on different days, and thus there are $2 \times 3 \times 4 = 24$ sequences. Resolution of these images is 320×240. There may be walking persons and objects may be moved/added/removed in different image sequences. The first two rows of Figure 13 show sample images captured by Minnie and Dumbo in a one-person office, under different weather conditions.

The KTH-INDECS dataset [15] was captured in the same environment as the IDOL dataset, but images were captured by cameras mounted in several fixed locations inside each room. The third row of Figure 13 shows three sample images in this dataset.

We use 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 same robot under the same weather condition. Run 1 is used for training and run 2 is used for testing, and vice versa.
- Setting 2: Train and test using the same robot but under different weather conditions. This experiment tests generality over variations of object locations and illumination.
- Setting 3: Training set and testing set are under the same weather conditions, but are 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 was used for training, and images from INDECS under different weather conditions were used for testing.

Table 5 shows the average recognition accuracies based on Setting 1. In this experiment, sPACT and sPAcCT have similar performance for cloudy and sunny conditions. However, sPAcCT achieves nearly 1% accuracy behind that of sPACT for the night condition. In the images captured at night, light from fluorescent lamps may cause color shift and influence the robustness of color CENTRIST.

Table 6 shows average recognition accuracies when training and testing data are in different weather conditions (Setting 2). The training and test conditions in the second 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 still has worse 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 7 shows the average recognition accuracies when images taken by different robots were used for training and testing, respectively. From this table we see that sPAcCT has slightly weak robustness for this experimental setting. Table 8 shows the average recognition accuracies for the KTH-INDECS dataset. sPACT and sPAcCT generally have similar performance.

Overall, sPAcCT has similar performance to that of sPACT in the KTH-IDOL and KTH-INDECS datasets. The reason may be that fewer color variations in these datasets. Comparing the results reported in Sections 4.1 to 4.5, we conclude that color CENTRIST benefits scene recognition more when images captured in difference scenes convey more color variations.

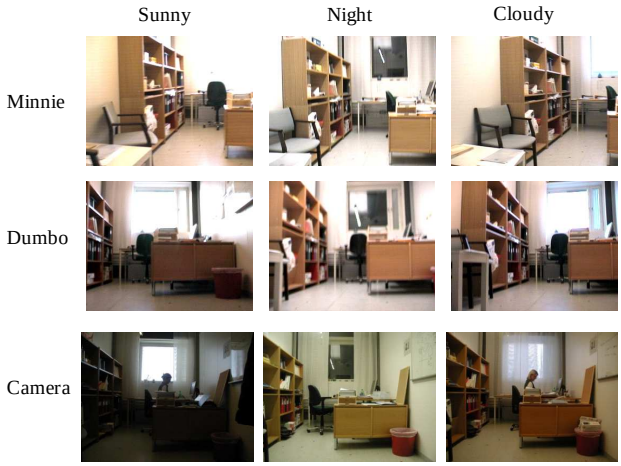

Figure 13. Sample images from the KTH-IDOL dataset (the first and the second rows) and the KTH-INDECS dataset, in different weather conditions. These examples show nearly the same angle of a one-person office.

Table 5. Average recognition accuracies on the KTH-IDOL dataset (Setting 1).

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

Table 6. Average recognition accuracies on the KTH-IDOL dataset (Setting 2).

Train	Test	Train condition	Test Condition	sPACT	sPAcCT
Minnie	Minnie	Sunny1	Night2	**80.69%**	79.76%
Minnie	Minnie	Night1	Sunny2	**86.10%**	83.04%
Minnie	Minnie	Cloudy1	Sunny2	92.93%	**93.40%**
Minnie	Minnie	Sunny1	Cloudy2	91.01%	**91.12%**
Minnie	Minnie	Night1	Cloudy2	**90.39%**	87.81%
Minnie	Minnie	Cloudy1	Night2	**92.72%**	90.09%

Table 7. Average recognition accuracies on the KTH-IDOL dataset (Setting 3).

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

Table 8. Average recognition accuracies on the KTH-INDECS dataset (Setting 4).

Train	Test	Train condition	Test condition	sPACT	sPAcCT
Camera	Camera	Sunny	Night	84.52%	**86.54%**
Camera	Camera	Night	Sunny	87.04%	**89.26%**
Camera	Camera	Cloudy	Sunny	**95.28%**	92.96%
Camera	Camera	Sunny	Cloudy	**93.70%**	92.78%
Camera	Camera	Night	Cloudy	**92.31%**	91.39%
Camera	Camera	Cloudy	Night	89.10%	**91.30%**
Average				90.33%	**90.70%**

5. CONCLUSION

We have presented a new color descriptor, i.e. color CENTRIST, that consistently provides better performance on scene categorization over conventional intensity-based descriptors. By appropriately quantizing the HSV color space, color information is elaborately represented and incorporated into the framework of CENTRIST. After evaluating various datasets, we conclude that this descriptor is especially suitable for images with higher color variations, though it reliably provides performance increment for almost all datasets.

In the future, we would conduct comprehensive studies on comparing color CENTRIST with other color descriptors. The usage of color CENTRIST in other applications will also be investigated. Moreover, because color CENTRIST share the same limitation, i.e. not invariant to rotation and scale, as CENTRIST, we would further enhance descriptor design in the future.

Acknowledgement: The work was partially supported by the National Science Council of Taiwan, Republic of China under research contract NSC 100-2221-E-194-061.

6. 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 on 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, A., 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] Yeung, M., Yeo, B.-L., and Liu, B. 1998. Segmentation of video by clustering and graph analysis. Computer Vision and Image Understanding, vol. 71, no. 1, pp. 94-109.

[17] Rasheed, Z. and Shah, M. 2003. Scene detection in Hollywood movies and tv shows. Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 343-348.

[18] Szummer, M. and Picard, R.W. 1998. Indoor-outdoor image classification. Proceedings of IEEE International Workshop on Content-Based Access of Image and Video Database, pp. 42-51.

[19] Vailaya, A., Jain, A., and Zhang, H.-J. 1998. On image classification: city vs. landscape. Proceedings of IEEE International Workshop on Content-Based Access of Image and Video Libraries, pp. 3-8.

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