mCENTRIST : a multi-channel feature generation

Xiao, Yang; Wu, Jianxin; Yuan, Junsong

2014

Xiao, Y., Wu, J., & Yuan, J. (2014). mCENTRIST: A Multi-Channel Feature Generation Mechanism for Scene Categorization. *IEEE Transactions on Image Processing*, 23(2), 823-836.

<https://hdl.handle.net/10356/103923>

<https://doi.org/10.1109/TIP.2013.2295756>

© 2014 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works. The published version is available at: [<http://dx.doi.org/10.1109/TIP.2013.2295756>].

Downloaded on 26 Apr 2023 18:30:02 SGT

mCENTRIST: A Multi-channel Feature Generation Mechanism for Scene Categorization

Yang Xiao, Jianxin Wu *Member, IEEE* and Junsong Yuan *Member, IEEE*

Abstract—mCENTRIST, a new multi-channel feature generation mechanism for recognizing scene categories, is proposed in this paper. mCENTRIST explicitly captures the image properties that are encoded jointly by two image channels, which is different from popular multi-channel descriptors. In order to avoid the curse of dimensionality, tradeoffs at both feature and channel levels have been executed to make mCENTRIST computationally practical. As a result, mCENTRIST is both efficient and easy to implement. In addition, a hyper opponent color space is proposed by embedding Sobel information into the opponent color space for further performance improvements. Experiments show that mCENTRIST outperforms established multi-channel descriptors on four RGB and RGB-NIR datasets, including aerial orthoimagery, indoor and outdoor scene category recognition tasks. Experiments also verify that the hyper opponent color space enhances descriptors' performance effectively.

Index Terms—Scene categorization, multi-channel descriptor, CENTRIST, channel interaction, hyper opponent color space.

I. INTRODUCTION

SCENE categorization has been a long standing research topic in the computer vision and image processing community. It has wide applications [1], and a lot of efforts have been devoted to solve this problem [1]–[7]. In the literature, a major focus of attention in scene categorization is to find visual descriptors that have strong discriminative power, e.g., CENTRIST [1], GIST [5], SIFT [8], BIF [9] and many others.

In most scene categorization research, these descriptors are extracted from grayscale images. However, since currently color images are dominating in consumer electronic imaging systems, it is natural to assume that color images will lead to higher categorization accuracy rates, because they contain more information than their respective grayscale versions. This intuition has been scientifically demonstrated, with the conclusion that multi-channel information (e.g., the RGB channels in a color image, an additional infrared channel, etc.) is a powerful clue for scene and object recognition [10]–[12]. Multi-channel images contain complementary information among channels, which leads to more complete descriptions of the target than any of the single channel. As shown in Fig. 1, in the Sobel gradient image the contour of the tree is more prominent in the R channel than that in the near-infrared (NIR) channel, while the situation is reversed for the building. Thus,

Figure 1. R and NIR channels of an image from the RGB-NIR scene dataset [13]. The Sobel gradient images corresponding to the red and green rectangular regions are enlarged. This figure demonstrates the complementarity between R and NIR channels.

it is important to leverage the multi-channel images in an appropriate way.

Research on descriptors for color or other multi-channel images has been, however, less popular than in the grayscale case. Popular multi-channel descriptors are mainly constructed from SIFT, such as CSIFT [14], HSV-SIFT [2], MSIFT [13], OpponentSIFT [12] and HueSIFT [15]. A common theme of these methods is that: they extract a SIFT vector from each (either raw or processed) channel, and then concatenate these vectors to form a multi-channel descriptor directly or with some additional postprocessings such as PCA. That is, after a possible preprocessing step to *decorrelate* the raw channels, descriptive information is extracted from individual channels separately as if the preprocessed channels are *independent*. As a consequence, such descriptors cannot capture the *image properties that are expressed jointly by two or more channels*. In Fig. 1, knowing the image properties simultaneously from both channels provides more hints than either single channel. For example, curved edge structure in R and suppressed edges

Y. Xiao is with the Institute for Media Innovation, Nanyang Technological University, Singapore. E-mail: hustcowboy@gmail.com.

J. Wu is the corresponding author of this paper, and is with the National Key Laboratory for Novel Software Technology, Nanjing University. E-mail: wujx2001@nju.edu.cn.

J. Yuan is with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. E-mail: jsyuan@ntu.edu.sg.

in NIR jointly describe the tree crown, but any individual channel may not.

An analogy from the probability perspective can better explain the above argument. Given an image with two channels X and Y , a generative method would model the image as a joint probability distribution function (pdf) $p_{X,Y}(x,y)$. Independent channel feature extraction methods investigate the image properties using only the marginals $p_X(x)$ and $p_Y(y)$. However, in general $p_{X,Y}(x,y) \neq p_X(x)p_Y(y)$ because the channels are not independent. Thus, we need to jointly extract image information from both channels—that is, from the joint distribution $p_{X,Y}(x,y)$ —to define multi-channel descriptors. In this way, the complementary information between channels can be better maintained for scene categorization than extracting feature from single channels independently. This point of view will be illustrated in details in Sec. II-A.

In this paper, we propose a visual descriptor, mCENTRIST, to *explicitly capture the interactions (joint information) from multi-channel images* for scene categorization.

mCENTRIST is based on CENTRIST [1], [7], a state-of-the-art descriptor for scene category recognition. CENTRIST is grayscale based and ignores color information. In Sec. II, we first show that naively modeling the interaction of multiple channels is computationally impractical. Then, we propose two techniques (sub-mCT and binary channel interaction) to efficiently capture joint channel information. Although there are works that treat categorization of multi-channel images as a multi-view learning problem [16], [17] to combine channels (heterogeneous features in particular) for higher accuracy rates, we focus on effective and efficient multi-channel visual descriptors in this paper. We leave, however, multi-view learning approaches as future directions when a large number of heterogeneous multi-channel features are available.

In Sec. III, we propose a method that extends normal color images with more channels in order to get more complementary information and consequently higher recognition rate. RGB is the most widely used multi-channel image format. This color space is suggested to transform into opponent color spaces [10], [12], [13] to decorrelate the channels.¹ In [1], Sobel gradient image was shown to improve scene categorization. We also find that its correlation with the opponent channels is very low. Thus, we upgrade the opponent color space to a hyper opponent color space by embedding the Sobel channel to improve categorization performance.

Finally, mCENTRIST's discriminative power is verified on four benchmark datasets in Sec. IV. mCENTRIST achieves the highest accuracy, compared to concatenation of SIFTs from all channels [12], msSIFT [13], mGIST [13], concatenation of CENTRISTs from all channels, and color CENTRIST [18] (a recent extension of CENTRIST to multi-channel images.) Since the primary goal of our experiments is to compare various multi-channel descriptors, we use the bag of visual words (BOV) framework to compare them fairly: fix all other components in the BOV system, and only the visual descriptors vary. We want to emphasize, though, mCENTRIST can be readily used in the extended BOV approaches such as

Figure 2. Local neighborhoods corresponding to the same scene location from two channels.

Fisher vector [19] or VLAD [20].

Overall, the contributions of this paper include:

- mCENTRIST: a novel multi-channel feature generation mechanism that captures the joint channel properties simultaneously. It is easy to implement, very efficient, and has high accuracy in scene categorization;
- Hyper opponent color space: providing more channels for RGB or RGB-NIR images, and thus higher accuracy rates, but without incurring additional imaging costs.

The source code and supporting materials for mCENTRIST will be published online.

II. mCENTRIST: A NOVEL MULTI-CHANNEL FEATURE GENERATION MECHANISM

As aforementioned, the complementary information within multi-channel images leads to more complete descriptions of the scene. Thus, we propose mCENTRIST as a multi-channel feature generation mechanism based on *multi-channel joint information encoding* to explicitly capture the joint properties within multi-channel images. For practical applications, some tradeoffs at both feature and channel levels are made.

A. Multi-channel Census Transform

CENTRIST is the histogram of the Census Transform (CT) [21] values in an image (or image patch) to capture both local and global information [1]. For certain pixel, the Census Transform compares its intensity value with its eight neighboring pixels, as illustrated in Eqn. 1:

$$\begin{array}{ccc|ccc} 87 & 19 & 23 & & 0 & 1 & 1 \\ 25 & \mathbf{25} & 10 & \Rightarrow & 1 & 1 & \Rightarrow (01111011)_2 \Rightarrow \mathbf{CT} = (123)_{10} \cdot (1) \\ 67 & 22 & 24 & & 0 & 1 & 1 \end{array} \quad (1)$$

When the central pixel is bigger than (or equal to) one of the neighbors, a bit 1 is set in the corresponding position. Otherwise, 0 is set. Then, the eight bits are combined from top-left to bottom-right and converted to a base-10 number in [0 255] as the Census Transform Value.

Census Transform is originally single channel based. However, this representation has some drawbacks in multi-channel images. Fig. 2 exhibits two 3×3 neighborhoods corresponding to the same scene location of R and NIR channels from a RGB-NIR image in the the 9 class RGB-NIR scene dataset [13]. The Census Transform values of the central pixel are $\mathbf{CT}_1 = (01100000)_2$ and $\mathbf{CT}_2 = (00010111)_2$, respectively. Obviously, the two channels have very different properties at this same scene location. This phenomenon, in fact, manifests quite frequently in multi-channel images. In other words, complementary information indeed exists in different channels of multi-channel images.

¹The same process is also suggested for the RGB-NIR color space [13].

Table I
AVERAGE $comp_{CT}$ BETWEEN CHANNELS. THE MULTI-CHANNEL IMAGES ARE FROM THE 9 CLASS RGB-NIR SCENE DATASET.

	R	G	B	NIR
R	0.000	0.770	0.808	0.863
G	0.770	0.000	0.720	0.817
B	0.808	0.720	0.000	0.848
NIR	0.863	0.817	0.848	0.000

Table II
AVERAGE $dist_{HI}$ BETWEEN CENTRISTS OF DIFFERENT CHANNELS.

	R	G	B	NIR
R	0.000	0.063	0.061	0.099
G	0.063	0.000	0.051	0.078
B	0.061	0.051	0.000	0.095
NIR	0.099	0.078	0.095	0.000

To further evaluate the CT complementarity between two channels, we propose a measurement given by

$$comp_{CT} = \frac{\#Diff_{CT}}{\#Diff_{CT} + \#Same_{CT}}, \quad (2)$$

where $\#Diff_{CT}$ is the number of pixels that have different CT values in two channels (such as the case in Fig. 2), $\#Same_{CT}$ is the number of pixels that have the same CT values, and $comp_{CT} \in [0, 1]$. Intuitively, the higher $comp_{CT}$ is, the more complementary the two channels are. Table I lists the average $comp_{CT}$ extracted from 477 RGB-NIR images in the 9 class RGB-NIR scene dataset. Indeed, RGB-NIR channels are highly complementary to each other as shown in Table I. That is, between any two channels, at least 72% pixels have different CT values ($comp_{CT} \geq 0.72$).

However, after the CT values are pooled together to form a CENTRIST descriptor, the complementary information between channels almost disappears. We measure the complementarity of two channels' CENTRIST representation using the histogram intersection [22] distance defined as

$$dist_{HI} = 1 - \sum_i \min(X_i, Y_i), \quad (3)$$

where X and Y are two ℓ_1 normalized histograms, $dist_{HI} \in [0, 1]$ and a higher value indicates high complementarity. Given the results in Table I, we expect to see high $dist_{HI}$ values too. However, Table II lists the average histogram intersection distance between CENTRISTS from different channels of the same image, also computed from the 9 class RGB-NIR scene dataset. It can be observed that generally CENTRISTS of RGB-NIR channels are very close to each other. *In other words, the histogram operation in CENTRIST tends to diminish the channel complementarity characterized by CT.* It is not difficult to understand this phenomenon. A histogram is a holistic descriptor only recording the occurrence probability of elements in it. Two channels of an image usually have very different CT values at corresponding locations, but at the same time have similar CT value distributions for the whole image.

Thus, simply concatenating CENTRISTS (or other histogram based visual descriptors) from different channels to construct a multi-channel descriptor may lose useful complementary information. Our solution to overcome this difficulty

Table III
COMPONENT OCCURRENCE INFORMATION OF DIFFERENT CHANNELS.

	U	V	W	X	
A	2	3	3	3	11
B	5	0	4	1	10
C	4	4	4	4	16
D	5	5	2	0	12
	16	12	13	8	

is to evolve Census Transform to capture the multi-channel joint properties before computing the histogram. In this way, the complementary information is embedded in the proposed multi-channel descriptor. An example is provided to illustrate the difference between extracting a multi-channel descriptor by simple channel concatenation and by using multi-channel joint information. Suppose we have 2 channels, in which channel 1 takes possible values A, B, C, and D and channel 2 takes possible values U, V, W, and X. The occurrence of these values for one example image is shown in Table III. Thus, the descriptor for channel 1 is a 4-bin histogram, (11, 10, 16, 12), indicating that there are 11 A, 10 B, etc. Similarly, channel 2 descriptor is a 4-bin histogram (16, 12, 13, 8). Simply concatenating the descriptors will generate a 8-bin multi-channel descriptor (11, 10, 16, 12, 16, 12, 13, 8). On the other hand, we propose to use joint channel properties. It concerns, for example, how many pixels are of value A in channel 1 and simultaneously of value W in channel 2 (in this example there are 3 such pixels). In short, the multi-channel joint descriptor is using all the 16 values in Table III: (2, 3, 3, 3, 5, 0, 4, 1, 4, 4, 4, 4, 5, 5, 2, 0).

To describe the joint properties of the two neighborhoods in Fig. 2, a naive way, which we call mCT (*multi-channel Census Transform*), can put the encoding bits of CT_1 and CT_2 together to form a 16-bit base-2 number, and then convert it to a base-10 number in $[0, 65535]$ as the mCT value:

$$\overrightarrow{mCT} = (\overbrace{01100000}^{CT_1} \overbrace{00010111}^{CT_2})_2 = (24599)_{10}. \quad (4)$$

This naive encoding can be extended to more channels easily. However, it faces a critical problem in practical use. The histogram's dimensionality is 2^{8n} for an n -channel image, which is too high.

To achieve a compact descriptor, one way is to apply PCA. However, for mCT, the time and memory consumption of PCA will be prohibitively huge when n is a relatively big number such as 4 or 5. More importantly, the curse of dimensionality makes learning PCA parameters impractical. If $n = 4$, we expect at least 10×2^{32} (more than 40 billion) examples to learn PCA coefficients, according to the rule of thumb in [23]. Thus, we propose *Census Transform pyramid* to reduce the dimensionality, while still simultaneously capture the properties from multiple channels.

As a specific instance, the Census Transform pyramid of Fig. 2(a) is shown in Fig. 3. It is comprised of three levels. Standard CT at Level 0 has been decomposed into sub-CTs on certain directions (encoding bits are still combined from top-left to bottom-right) at Level 1 and 2. If mCT is defined based on sub-CTs, its dimensionality will be much lower.

Figure 3. Census Transform pyramid of Fig. 2(a).

The Level 1 sub-CT values of Fig. 2(a) and Fig. 2(b) are $\mathbf{CT}_1^1 = (0100)_2$, $\mathbf{CT}_1^2 = (1000)_2$, $\mathbf{CT}_2^1 = (0011)_2$, and $\mathbf{CT}_2^2 = (0101)_2$, respectively. Two sub-mCT values in $[0\ 255]$ are given by

$$\overrightarrow{\mathbf{mCT}}_1 = \overbrace{(0100\ 0011)}^{\mathbf{CT}_1^1\ \mathbf{CT}_2^1} = (67)_{10}, \quad (5)$$

and

$$\overrightarrow{\mathbf{mCT}}_2 = \overbrace{(1000\ 0101)}^{\mathbf{CT}_1^2\ \mathbf{CT}_2^2} = (133)_{10}. \quad (6)$$

In this way, two sub-mCT histograms of 256 dimensions will be extracted for a 2-channel image. Compared with the 65536 dimensions of a mCT histogram, sub-mCT histogram's dimensionality has been significantly reduced. Level 2 sub-CTs could achieve even lower dimensionality, but they will not be employed in this paper, because in practice their discriminative ability is not strong. In Sec. IV-E7, it will be empirically demonstrated that Level 1 sub-CTs consistently outperform Level 2 sub-CTs for mCENTRIST.

Using the Census Transform pyramid, an n -channel image will have two sub-mCT histograms of dimensionality 2^{4n} as the multi-channel joint property descriptors. In Sec. IV-E6, we will empirically show that the sub-mCT descriptors (Eqn. 5 and 6) achieve similar recognition accuracy as the high dimensional mCT descriptor (Eqn. 4). In other words, there is no significant information loss from mCT to sub-mCT.

B. Binary interaction mechanism

For real applications, tradeoffs between theoretical elegance and practicality should be made. The Census Transform pyramid does it at the feature level. However, although sub-mCT histogram pair's dimensionality (2×2^{4n}) is significantly lower than that of mCT histogram (2^{8n}), it is still impractical if n is 5 or higher. We further propose a *binary interaction mechanism* to overcome this difficulty at the channel level.

As aforementioned, the dimensionality of a 2-channel sub-mCT histogram is 256, which is easy to handle. Let $V = \{O_1, O_2, O_3, O_4\}$ be a 4-channel image, 6 channel pairs could be derived from V by binary interactions, as $V'_1 = \{O_1, O_2\}$, $V'_2 = \{O_1, O_3\}$, $V'_3 = \{O_1, O_4\}$, $V'_4 = \{O_2, O_3\}$, $V'_5 =$

Figure 4. Spatial pyramid for CENTRIST.

$\{O_2, O_4\}$ and $V'_6 = \{O_3, O_4\}$. Then, 2-channel sub-mCT histogram pair will be extracted for each V'_i respectively and concatenated together as the mCENTRIST descriptor for V .

By using binary interactions, mCENTRIST possesses the advantage of keeping the feature dimensionality in a reasonable range even when the channel number n is relatively big. The dimensionality of mCENTRIST is only $\binom{n}{2} \times 2 \times 256 = n \times (n-1) \times 256$ for an n -channel image. More importantly, although mCENTRIST tends to avoid prohibitively huge feature dimensionality, it can still project the n -channel image into a higher dimensional feature space than concatenating CENTRISTs from all channels directly. Concatenating CENTRISTs leads to $256n$ dimensions, and the ratio between them is $n-1$. This property could enhance the performance of linear classifiers [24] without using costly kernels or high-dimensional feature mapping [19], [25], and is thus another advantage of mCENTRIST.

Overall, mCENTRIST is a multi-channel feature generation mechanism which achieves a balance between computational efficiency and discriminative power. Similar to CENTRIST, it is easy to implement and has few parameters to tune.

C. Spatial representations

In order to capture the rough global information of an image, we use the spatial pyramid of [1], [7], which is an extension of the SPM scheme in [4]. As shown in Fig. 4, it contains 31 blocks of the same size in 3 levels. In [1], CENTRISTs extracted from all the blocks are then concatenated to form the final feature vector.

The same spatial pyramid structure is also applied to mCENTRIST. An n -channel image is split into 31 blocks. For each block, mCENTRIST is extracted and concatenated as the final feature vector. Each sub-mCT histogram is reduced to 40 dimensions by PCA, which results in a feature vector of dimensionality $\binom{n}{2} \times 2 \times 40 \times (25 + 5 + 1) = n \times (n-1) \times 1240$.

III. HYPER OPPONENT COLOR SPACE

RGB color image is the most widely used multi-channel image format. Recently, NIR image has been demonstrated to have low correlation with RGB channels and employed as an additional valuable clue [13]. In order to improve performance, both RGB and RGB-NIR have been transformed to opponent color space as indicated in [12] and [13]. In this section, opponent color space will be extended to hyper opponent color space by adding Sobel information. We will also refine the multi-channel opponent transformation method in [13].

Figure 5. Example of an RGB kitchen image from the 67 class indoor scene recognition dataset [6] and the corresponding Sobel image of its R channel.

198	197	197	2	2	2
200	198	197	2	2	2
201	199	197	2	2	2

(a) Neighborhood from R.

2	2	2	2	2	2
2	2	2	2	2	2
2	2	2	2	2	2

(b) Neighborhood from Sobel.

Figure 6. Local neighborhoods corresponding to the same scene location from R and Sobel channels.

A. Addition of the Sobel channel

As discussed in [1], Sobel gradient image could suppress fine-scale texture and enhance global spatial structure, which is beneficial for scene categorization. Here, we give a specific example to illustrate this proposition. Fig. 5 exhibits an RGB kitchen image from the 67 class indoor scene recognition dataset [6] and its Sobel image of the R channel. Compared to the RGB image, the main contours of the objects are more salient in the Sobel image, such as the cabinet and the drawer. Meanwhile, Fig. 6 shows two 3×3 neighborhoods corresponding to the same scene location of the R channel and the Sobel channel in Fig. 5(a) and Fig. 5(b). In fact, Fig. 6(a) is almost a homogeneous region. However, because Census Transform is very sensitive to small pixel value variations, Fig. 6(a) will be assigned certain structure property incorrectly. Such a kind of fine-scale texture may confuse the classifier. In contrast, all the Sobel pixels have the same value and the fine-scale texture has been suppressed. Note that although LTP [26] has been used to reduce the effect of such small pixel value fluctuations, it requires a parameter t , which is not easy to tune. An LTP encoding bit is defined as

$$bit_{LTP}(I_N, I_C, t) = \begin{cases} 1, & I_N \geq I_C + t \\ 0, & |I_N - I_C| < t \\ -1, & I_N \leq I_C - t \end{cases}, \quad (7)$$

where I_C is the central pixel; I_N is the neighboring pixel and t is the similarity threshold between I_C and I_N , for which it is not easy to determine a proper value. However, a Sobel image has similar noise reduction effects, but does not require parameter tuning. Thus, we expect that a Sobel channel will provide useful hints for scene recognition.

Another advantage of the Sobel channel is that it is only weakly correlated with the multiple opponent color channels (defined later in this section). That is, it may provide complementary information. For example, Table IV lists the average correlation coefficients between the Sobel image of R channel and the RGB opponent transformed images. 15475 images

Table IV
AVERAGE CORRELATION COEFFICIENTS BETWEEN DIFFERENT CHANNEL IMAGES. O_1 , O_2 AND O_3 ARE THE RGB OPPONENT TRANSFORMED IMAGES AND S IS THE CORRESPONDING SOBEL IMAGE OF R CHANNEL FROM THE 67 CLASS INDOOR SCENE RECOGNITION DATASET [6].

	O_1	O_2	O_3	S
O_1	1.000	0.269	0.083	0.013
O_2	0.269	1.000	-0.062	0.009
O_3	0.083	-0.062	1.000	-0.019
S	0.013	0.009	-0.019	1.000

from the 67 class indoor scene recognition dataset are used to compute these statistics. The correlation coefficient between images X and Y is computed as

$$corr = \frac{\sum_i \sum_j (X_{ij} - \bar{X})(Y_{ij} - \bar{Y})}{\sqrt{\sum_i \sum_j (X_{ij} - \bar{X})^2 \sum_i \sum_j (Y_{ij} - \bar{Y})^2}}, \quad (8)$$

where $\bar{X}$ and $\bar{Y}$ are the mean of all pixel values of X and Y , respectively. Thus, we extend the commonly used opponent color space to hyper opponent color space by adding the Sobel image of the R channel as an additional channel. In this way, more descriptive information of the scene is contained in the new color space without extra imaging cost. As a consequence, both RGB and RGB-NIR images will be described in a higher dimensional space by mCENTRIST when making use of hyper opponent color space than opponent color space. As discussed in Sec. II-B, this phenomenon may enhance the performance of linear classifiers. Actually, hyper opponent color space could also be flexibly constructed by adding Sobel images of other channels for further performance improvement, such as G or NIR channels. However, to maintain the computational efficiency of mCENTRIST, only the R channel is used in this paper. Note that although the choice of the R channel is rather ad hoc, the evaluations in Sec. IV-F2 show that this is a reasonable choice.

B. Refined opponent transformation

As the basis of the hyper opponent color space, the opponent color space is originally suggested for RGB images. Brown *et al.* [13] extended it to the RGB-NIR domain by computing RGB-NIR PCA components. Our multi-channel opponent transformation is defined similar to [13]. However, as proposed in [1], [7], we do not perform mean centering for PCA, in order to achieve better multi-channel information maintenance and higher computational efficiency. For instance, let $\mathbf{X} = [e_1 \ e_2 \ \dots \ e_m]^T$ be an $m \times 4$ RGB-NIR data matrix, where $e_i = (x_{i1} \ x_{i2} \ x_{i3} \ x_{i4}) = (r_i \ g_i \ b_i \ nir_i)$, $i = 1, 2, \dots, m$, is the 4-dimensional RGB-NIR color vector. The covariance matrix of $\mathbf{X}$ is approximately given by

$$\widetilde{cov} = \frac{1}{m-1} \mathbf{X}^T \mathbf{X} = \mathbf{W} \mathbf{\Sigma} \mathbf{W}^T, \quad (9)$$

where $\mathbf{W}$ and $\mathbf{\Sigma}$ contain the eigenvectors and eigenvalues of $\widetilde{cov}$, respectively.

In order to make the transformed values remain within the 4-d unit cube, the linear mapping from [13] is applied to

Figure 7. RGB-NIR image and the corresponding opponent transformed images.

generating the transformed values:

$$x'_{ij} = \frac{1}{\sum_k |w_{kj}|} \sum_k x_{ik} w_{kj} - \frac{\sum_k w_{kj}^{(-)}}{\sum_k |w_{kj}|}, \quad (10)$$

where $(-)$ indicates the corresponding w_{kj} is negative. Then, e_i could be transformed to an opponent color vector as $e'_i = (x'_{i1} \ x'_{i2} \ x'_{i3} \ x'_{i4})$. The column normalized transformation matrix $\widehat{\mathbf{W}}_1$ suggested by [13] and $\widehat{\mathbf{W}}_2$ obtained in this paper are listed below:

$$\widehat{\mathbf{W}}_1 = \begin{bmatrix} \mathbf{0.283} & -0.058 & 0.472 & -0.232 \\ \mathbf{0.283} & -0.072 & 0.025 & 0.482 \\ \mathbf{0.288} & -0.217 & -0.389 & -0.238 \\ \mathbf{0.146} & 0.653 & -0.114 & -0.048 \end{bmatrix}, \quad (11)$$

$$\widehat{\mathbf{W}}_2 = \begin{bmatrix} \mathbf{0.252} & 0.106 & -0.471 & 0.233 \\ \mathbf{0.253} & 0.132 & -0.026 & -0.497 \\ \mathbf{0.248} & 0.259 & 0.402 & 0.231 \\ \mathbf{0.247} & -0.503 & 0.101 & 0.039 \end{bmatrix}. \quad (12)$$

The difference between $\widehat{\mathbf{W}}_1$ and $\widehat{\mathbf{W}}_2$ comes from the fact that in Eqn. 9, we do not subtract the mean of $\mathbf{X}$, which is different from classic PCA. $\widehat{\mathbf{W}}_1$ underplays the NIR channel (only 0.146) in the first principal component. On the contrary, $\widehat{\mathbf{W}}_2$ has almost equal weights for the 4 channels in its first principal component. Because the first principal component contains nearly 90% signal energy, $\widehat{\mathbf{W}}_2$ is better for maintaining multi-channel information. Another intuitive advantage of not using standard PCA is that computational efficiency could be improved by removing mean centering, especially when the sampling size m is huge. Fig. 7 exhibits an RGB-NIR image and its corresponding opponent transformed images according to Eqn. 10 and 12.²

IV. EXPERIMENTS

In experiments, we focus on testing mCENTRIST's discriminative power, which is an important criteria to evaluate multi-channel visual descriptors as addressed in [10]–[12], [14]. Three RGB datasets (21 class land-use classification [27], 67 class indoor scene recognition [6] and 8 class event [28]), and

one RGB-NIR dataset (9 class RGB-NIR scene [13]) are used for testing. If not otherwise specified, the available samples in each dataset are randomly split into a training set and a testing set for 5 times, and the average accuracy is reported. All the images are resized to be no larger than 300×300 pixels.

Following [1], [7], we use 40 eigenvectors in the PCA operation for mCENTRIST without mean centering. The sub-mCT histograms are normalized with zero mean and unit ℓ_2 norm. During mCENTRIST extraction, the pixels corresponding to CT value pair (0, 0) or (255, 255) in each 2-channel image will not be considered. The SPM structure as shown in Fig. 4 is employed for all CENTRIST based multi-channel descriptors.

In [1], [7], grayscale statistical features such as average value and standard deviation of pixels in a block are appended to CENTRIST to improve performance. To solely evaluate the discriminative power of multi-channel visual descriptors, they are not used in our experiments.

mCENTRIST is compared with four multi-channel descriptors: (1) concatenating CENTRISTs from all channels directly (conCENTRIST), an intuitive way to extend CENTRIST to multiple channels; (2) color CENTRIST (cCENTRIST) [18], a recent extension of CENTRIST to the HSV color space; (3) concatenating GISTs from all channels directly (mGIST [13]);³ (4) concatenating SIFTs from all channels directly (mSIFT) in BOV model as recommended in [12]; and, (5) multi-spectral SIFT (msSIFT) [13] that concatenates SIFTs from all channels with PCA postprocessing.⁴ Different from [13], msSIFT is combined with the densely sampled BOV model in our experiment. We do this for two reasons: first, dense sampling has been demonstrated to be more effective than sparse interest points in scene classification [28] and secondly, in BOV model msSIFT can be evaluated easily using the same classifier (SVM) as the other visual descriptors. More specifically, Gabor filters at 4 scales and 8 orientations per scale are computed for mGIST; mSIFT and msSIFT are densely sampled from patches of size 16×16 in the multi-channel images, and the sampling step size is 6 pixels. SPM in [4] is applied to mSIFT and msSIFT. For each testing dataset, the codebooks of size 4000 are generated respectively by K-means and VQ encoding are applied as in [12]. Following [13], we use 256 eigenvectors in the PCA operation for msSIFT, as well as 128 under the single channel condition. The performance of all descriptors is evaluated corresponding to the same sample split. And the state-of-the-art performances on the four testing datasets are also reported to the best of our knowledge.

The channels in the hyper opponent color space specified in Sec. III for RGB are labeled as O_1 , O_2 , O_3 and S (Sobel image of the R channel) and for RGB-NIR are O_1 , O_2 , O_3 , O_4 and S . The channels of the classic opponent color space

³mGIST is implemented using the GIST Matlab source code released by Torralba [5] at <http://people.csail.mit.edu/torralba/code/spatialenvelope/>.

⁴mSIFT and msSIFT are constructed based on the SIFT Matlab source code released by Yang [29] at <http://www.ifp.illinois.edu/~jyang29/LLC.htm>.

²The comparison between the refined opponent color space and that proposed in [13] will be shown in Sec. IV-F.

Figure 8. 21 different land-use images.

Table V

SCENE CLASSIFICATION RESULTS (%) OF THE 21 CLASS LAND-USE CLASSIFICATION DATASET. THE BEST PERFORMANCE OF EACH DESCRIPTOR IS SHOWN IN BOLDFACE. STANDARD DEVIATIONS ARE IN PARENTHESES.

Channels	Visual Descriptors				
	mCENTRIST	conCENTRIST	mSIFT	msSIFT	mGIST
K	—	73.9(±2.9)	80.7(±1.8)	77.2(±2.5)	50.8(±2.1)
S	—	73.7(±2.3)	71.9(±2.3)	68.9(±3.1)	49.8(±1.7)
RGB	84.5(±0.7)	81.1(±3.0)	80.4(±2.4)	77.3(±1.8)	58.6(±0.9)
RGBS	87.8(±1.3)	84.1(±2.2)	82.1(±2.6)	78.5(±4.0)	68.0(±2.2)
P ₁ P ₂ P ₃	85.3(±1.1)	83.4(±0.4)	80.5(±2.6)	74.4(±3.0)	62.0(±1.9)
P ₁ P ₂ P ₃ S	88.8(±0.5)	85.1(±0.6)	83.7(±3.5)	79.5(±2.6)	71.9(±3.3)
O ₁ O ₂ O ₃	87.3(±1.6)	81.4(±1.8)	83.6(±2.5)	80.0(±2.1)	63.0(±2.1)
O ₁ O ₂ O ₃ S	89.9(±1.2)	84.6(±1.6)	85.0(±2.6)	82.2(±3.6)	72.0(±2.7)
cCENTRIST (HSV)	75.2(±2.4)				
State-of-the-art [31]	77.8				

specified in [12] are labeled as P_1 , P_2 and P_3 given by

$$\begin{pmatrix} P_1 \\ P_2 \\ P_3 \end{pmatrix} = \begin{pmatrix} \frac{R+G+B}{\sqrt{3}} \\ \frac{R+G-2B}{\sqrt{6}} \\ \frac{R-G}{\sqrt{2}} \end{pmatrix}. \quad (13)$$

The grayscale image is marked as K with the transformation: $K = 0.2989 \times R + 0.5870 \times G + 0.1140 \times B$ as in Matlab.

LIBSVM [30] is used as our SVM classifier. Linear kernel is employed in this paper. The penalty factor C is set to 2^{-5} for all cases, following the recommendation in [1].

Experimental results are organized as follows. Sec. IV-A to IV-C report accuracy comparisons on the 4 benchmark datasets, respectively.

Furthermore, in Sec. IV-E, the efficiency (running time) of different descriptors is reported. All these descriptors will also be further compared under the same feature extraction frameworks and color spaces to draw more interpretable and fair conclusions. Moreover, we study the effects of individual components in more details. Sobel channel's effectiveness in the hyper opponent color space, the difference between mCENTRIST and its brute force version (Eqn. 4), and the choice of Census Transform level for mCENTRIST construction will be investigated, respectively.

In the end, more discussions are provided in Sec. IV-F.

A. The 21 class land-use classification dataset

This dataset contains land-use aerial orthoimagery from 21 classes: *agricultural, airplane, baseball diamond, beach, buildings, chaparral, dense residential, forest, freeway, golf*

Table VI
SCENE CLASSIFICATION RESULTS (%) OF THE 67 CLASS INDOOR SCENE RECOGNITION DATASET.

Channels	Visual Descriptors				
	mCENTRIST	conCENTRIST	mSIFT	msSIFT	mGIST
K	—	32.6(±0.4)	36.8(±1.1)	34.8(±1.8)	22.1(±1.1)
S	—	31.2(±0.8)	30.0(±0.8)	26.8(±1.8)	20.5(±1.2)
RGB	36.6(±1.8)	35.9(±0.7)	39.2(±1.9)	36.4(±1.1)	26.2(±1.5)
RGBS	41.5(±1.4)	38.9(±0.9)	40.5(±0.3)	35.0(±2.3)	29.8(±1.5)
P ₁ P ₂ P ₃	38.7(±1.6)	39.2(±1.0)	37.1(±0.9)	33.1(±2.0)	17.2(±6.1)
P ₁ P ₂ P ₃ S	43.5(±1.5)	42.2(±1.1)	39.9(±1.4)	34.2(±2.0)	20.0(±5.9)
O ₁ O ₂ O ₃	40.8(±1.7)	40.6(±1.2)	36.6(±0.4)	33.1(±1.8)	27.2(±1.2)
O ₁ O ₂ O ₃ S	44.6(±1.2)	43.2(±1.4)	39.7(±1.6)	35.5(±2.4)	31.5(±1.4)
cCENTRIST (HSV)	34.1(±0.9)				
State-of-the-art [32]	47.2				

course, harbor, intersection, medium density residential, mobile home park, overpass, parking lot, river, runway, sparse residential, storage tanks and tennis courts (example images in Fig. 8). It is one of the first datasets derived from publicly available high-resolution remote sensing imagery for scene classification.

2100 RGB images are included in this dataset (each class has 100 images). Following [27], five-fold cross validation is executed on this dataset, which randomly splits the samples of each class into five equal sized sets. In each testing round, four of the sets are used for training and the remaining one is for testing. Table V lists the classification results.

According to Table V, it could be observed that:

- On this land-use dataset, hyper opponent mCENTRIST significantly outperforms other multi-channel descriptors. What is more important, mCENTRIST's best performance (89.9%) far exceeds the previous state-of-the-art accuracy (77.8%). And its improvement (16.0%) over grayscale CENTRIST is also prominent. mCENTRIST is suitable for categorizing these remote sensing images;
- conCENTRIST, mSIFT and msSIFT also work well on this dataset with high accuracy. But they are inferior to mCENTRIST in all the cases. mSIFT is stronger than msSIFT. cCENTRIST and mGIST do not perform well;
- The hyper opponent color space is the best choice for most multi-channel descriptors. Sobel channel constantly yields performance improvement for all descriptors, especially for mGIST with nearly 10.0% improvement. This is a strong evidence for the effectiveness of the Sobel channel in scene representation. Another interesting phenomenon is that the Sobel channel is not only effective in hyper opponent color space, but also provides RGB and the classic opponent color space with additional valuable information to obtain further performance improvement.

B. The 67 class indoor scene recognition dataset

Indoor scene categorization is a much more difficult problem than outdoor scene recognition. This challenging dataset covers a large range of indoor scenes from specific categories to generic concepts.

Figure 9. 9 different RGB-NIR scene images.

15620 images are included in this dataset.⁵ In [6], 80 images in each category were randomly chosen for training, and 20 for testing. However, not all the categories have at least 100 available RGB samples. In order to follow [6] as much as possible, we use 80 images per category for training, and at most 20 for testing according to the capacity of the categories. The indoor scene categorization results are shown in Table VI.

It can be summarized from Table VI that:

- Hyper opponent mCENTRIST possesses the strongest discriminative power among the tested descriptors on this challenging dataset with the accuracy of 44.6%. As a comparison, the accuracy of grayscale CENTRIST is only 32.6%. The performance improvement (12.0%) yielded by mCENTRIST is remarkable and beyond that of conCENTRIST. cCENTRIST's performance is 34.1%, which is inferior to mCENTRIST and conCENTRIST significantly;
- mSIFT and msSIFT are generally weaker than mCENTRIST and conCENTRIST for indoor scene recognition task. GIST is not suitable for indoor scene categories as indicated in [1]. Here, mGIST's performance on this dataset is also very limited;
- BOV CENTRIST with large size codebook obtains the highest accuracy rate of 47.2% by using PmSVM as reported in [32]. It is possible that applying mCENTRIST in a similar BOV model may lead to better performance;
- The hyper opponent color space is advantageous over RGB and opponent color space nearly among all descriptors. It is interesting that mSIFT and msSIFT achieve better performance in the RGB color space than opponent color space. Moreover, Sobel channel enhances the descriptors' performance effectively.

C. The 9 class RGB-NIR scene dataset

This dataset is proposed in [13] to demonstrate the contribution of NIR image for scene classification. It includes 9 scene categories: *country, field, forest, indoor, mountain, old building, street, urban and water* (Fig. 9 shows example images) in RGB-NIR image format.

477 images are contained in this dataset and all of them are used by us for testing. In order to treat all the categories equally in training, 42 images are randomly chosen for training per category, and the rest for testing. This experiment setup is slightly different from [13], which used 11 samples per

⁵15475 available RGB images are picked up for classification. Grayscale images and those RGB images with the same values in all channels (i.e., fake RGB images) are ignored.

Table VII
SCENE CLASSIFICATION RESULTS (%) OF THE 9 CLASS RGB-NIR SCENE DATASET CORRESPONDING TO THE EXPERIMENT SETUP IN THIS PAPER.

Channels	Visual Descriptors				
	mCENTRIST	conCENTRIST	mSIFT	msSIFT	mGIST
K	—	72.2(±3.0)	71.2(±5.3)	67.8(±3.3)	59.1(±5.2)
S	—	73.8(±2.5)	65.2(±4.4)	65.2(±4.0)	60.8(±4.7)
RGB	78.9(±5.1)	76.7(±2.8)	74.1(±3.1)	69.2(±3.3)	60.5(±5.4)
RGBN	79.1(±3.5)	77.8(±3.5)	73.9(±2.8)	71.4(±3.5)	61.9(±4.6)
RGBNS	80.8(±4.0)	79.9(±2.5)	78.2(±2.5)	73.6(±3.6)	65.4(±4.4)
P ₁ P ₂ P ₃	75.4(±3.4)	76.8(±3.9)	74.0(±5.0)	68.6(±2.2)	62.6(±3.9)
P ₁ P ₂ P ₃ N	80.4(±4.9)	78.3(±3.2)	77.6(±4.1)	72.4(±3.2)	64.7(±2.7)
P ₁ P ₂ P ₃ NS	81.2(±4.5)	80.7(±2.5)	80.4(±3.6)	73.5(±6.2)	68.0(±3.0)
O ₁ O ₂ O ₃ O ₄	84.5(±2.1)	81.3(±3.0)	78.6(±2.9)	73.9(±4.6)	67.9(±5.1)
O ₁ O ₂ O ₃ O ₄ S	82.1(±1.9)	81.7(±2.8)	79.3(±1.9)	75.2(±2.6)	69.9(±5.0)
cCENTRIST (HSV)	73.0(±4.1)				

Table VIII
SCENE CLASSIFICATION RESULTS (%) OF THE 9 CLASS RGB-NIR SCENE DATASET CORRESPONDING TO THE EXPERIMENT SETUP IN [13].

Channels	Visual Descriptors				
	mCENTRIST	conCENTRIST	mSIFT	msSIFT	mGIST
O ₁ O ₂ O ₃ O ₄	82.1(±2.5)	80.7(±3.0)	77.1(±3.2)	71.9(±4.2)	67.5(±3.0)
O ₁ O ₂ O ₃ O ₄ S	84.0(±2.3)	80.1(±1.6)	78.9(±3.1)	73.0(±4.7)	68.8(±3.7)
cCENTRIST (HSV)	72.6(±4.4)				
State-of-the-art [33]	87.9(±2.2)				

category for testing and the rest for training. However, the two setups both contain 99 testing images in each train/test random split.

In RGB-NIR images, channels in the hyper opponent color space are labeled as O_1 , O_2 , O_3 , O_4 and S (Sobel image of the R channel). NIR image is marked as N . The classification results are listed in Table VII.

From the table, we can observe that:

- The best classification accuracy of 84.5% on this dataset is achieved by mCENTRIST extracted from the hyper opponent color space but without the Sobel channel, which significantly improves the performance of grayscale CENTRIST (72.2%) by 12.3%. The reason why the Sobel channel does not help in this case may be the lack of enough training samples to match the high-dimensional feature generated by hyper opponent mCENTRIST (24800-dimensional feature but only 378 training samples). However, the addition of Sobel image can improve the descriptors' performance for other cases;
- According to the accuracy rates, mCENTRIST is still a stronger visual descriptor for RGB-NIR scene categorization than conCENTRIST, mSIFT, msSIFT and mGIST. Because cCENTRIST is strictly located in the HSV color space [18], it is not available for RGB-NIR. On this dataset, cCENTRIST does not perform very well in HSV space. Its accuracy improvement over grayscale CENTRIST is trivial.

We also follow the exact experiment setup in [13] to test the descriptors in the proposed opponent, hyper opponent and HSV color space. The corresponding results listed in Table VIII can be summarized as:

- Overall, the hyper opponent mCENTRIST achieves an accuracy of 84.0%. Compared to the best performance of conCENTRIST (80.7%), cCENTRIST (72.6%), mSIFT

Figure 10. 8 different sports event images.

Table IX
SCENE CLASSIFICATION RESULTS (%) OF THE 8 CLASS EVENT DATASET.

Channels	Visual Descriptors				
	mCENTRIST	conCENTRIST	mSIFT	msSIFT	mGIST
K	—	79.5(±0.5)	77.9(±0.8)	75.8(±2.0)	65.7(±1.1)
S	—	76.5(±1.3)	70.0(±2.3)	67.3(±2.7)	62.1(±2.2)
RGB	82.0(±2.0)	82.6(±0.9)	80.0(±1.0)	75.5(±2.0)	71.0(±1.6)
RGBS	83.0(±2.2)	84.5(±1.4)	80.9(±1.0)	77.0(±2.2)	74.1(±0.9)
P ₁ P ₂ P ₃	82.2(±1.2)	81.3(±1.6)	78.7(±1.5)	77.1(±2.0)	70.0(±1.6)
P ₁ P ₂ P ₃ S	84.3(±1.6)	84.0(±1.0)	81.5(±1.5)	78.2(±1.4)	73.8(±1.4)
O ₁ O ₂ O ₃	85.0(±1.5)	84.1(±1.5)	82.4(±1.3)	79.0(±2.0)	73.2(±1.7)
O ₁ O ₂ O ₃ S	86.5(±0.6)	85.4(±1.8)	84.2(±0.7)	80.9(±1.3)	76.2(±1.9)
cCENTRIST (HSV)	82.4(±1.2)				
State-of-the-art [34]	87.1				

(78.9%), msSIFT (73.0%) and mGIST (68.8%), mCENTRIST is significantly better;

- The state-of-the-art performance achieved in [33] was 87.9%. That accuracy was yielded by msSIFT appended with channel intensity statistical features (average value and standard deviation). However, Fisher vector [19] seems to be the key to its performance improvement.

D. The 8 class event dataset

This dataset includes images of 8 sports categories: *badminton*, *bocce*, *polo*, *croquet*, *rock climbing*, *rowing*, *sailing* and *snowboarding* (Fig. 10 shows example images). This dataset is only used for scene classification. And the events are just identified by the scene type.

1579 images are originally contained in this dataset.⁶ Following the setup in [28], 70 images per class are used for training, and 60 for testing. The classification results are listed in Table IX.

Observing the classification results, we find that:

- Among the tested descriptors, the proposed hyper opponent mCENTRIST achieves the highest accuracy of 86.5%. Both mCENTRIST and conCENTRIST extracted from multiple channels can improve the performance over grayscale CENTRIST (79.5%) by a large margin. mCENTRIST outperforms conCENTRIST in most cases. cCENTRIST's accuracy rate is 82.4%. Its improvement over grayscale CENTRIST is relatively limited;
- Although large size 4000 codebooks are applied to mSIFT and msSIFT in BOV model, they are still inferior to mCENTRIST and conCENTRIST, and with much higher time

⁶However, only 1576 RGB images are selected for experiment, because the remaining ones are fake RGB images without difference among the RGB channels.

Table X
TIME CONSUMPTION IN MATLAB (IN SECONDS) OF FEATURE EXTRACTION, PCA OPERATION AND VQ ENCODING OF DIFFERENT DESCRIPTORS IN THE HYPER OPPONENT COLOR SPACE.

Visual Descriptors	Operations		
	Feature extraction	PCA	VQ
mCENTRIST	134.75	0.49	—
conCENTRIST	56.69	0.20	—
cCENTRIST (HSV)	16.55	0.06	—
mSIFT	771.21	—	315.40
msSIFT	771.21	18.45	249.84
mGIST	576.21	—	—

consumption in feature extraction and encoding. Detailed running time comparison will be provided in Sec. IV-E1;

- Among all the tested descriptors, mGIST seems the weakest one on this dataset;
- The best performance on this dataset, as far as we know, is 87.1% [34] achieved by using discriminative compact pyramid, combination of multiple feature types and nonlinear histogram intersection kernel. mCENTRIST (86.5%) is close to this state-of-the-art accuracy rate by only employing single feature type and linear kernel.

E. Comparison of efficiency and various system parts

1) *Time consumption*: As an essential factor, time consumption is always employed to evaluate the visual descriptors. Here, we mainly focus on testing time because it is more important for real applications. The PCA eigenvectors and codebooks could be precomputed offline. The running times are shown in Table X. Time consumption (excluding I/O time) statistics is computed on a computer with Intel (R) Xeon(R) X5670 @ 2.93GHz (only using one core) in Matlab. The hyper opponent color space, HSV color space (for cCENTRIST) and 200 RGB images from the “badminton” category in the 8 class event dataset are used for collecting running time statistics. We can see that mCENTRIST is much faster than mSIFT, msSIFT and mGIST in testing, and only slower than conCENTRIST and cCENTRIST. Since mCENTRIST yields the best results on all testing datasets, it can achieve the balance between computational efficiency and discriminative power.

2) *Descriptor comparison in the BOV framework*: Previously, the tested descriptors are not compared in an unified feature extraction framework. That is, mSIFT and msSIFT make use of BOV model, while mCENTRIST, conCENTRIST, cCENTRIST and mGIST are directly fed into SVM with or without PCA postprocessing. To evaluate the descriptors more fairly, here we apply BOV to all of them except mGIST. The reason why mGIST is excluded is that it is strictly holistically defined [5], [13], and for an image patch of small size (i.e., 16×16) its discriminative power is limited. The experiments are executed on the 9 class RGB-NIR scene dataset. All descriptors are densely sampled from patches of size 16×16 , and the sampling step size is 8. The codebook size is set to 200. The proposed opponent, hyper opponent and HSV (for cCENTRIST) color space are employed for feature extraction. Experimental results are shown in Table XI. We can see that in the unified BOV framework, mCENTRIST still achieves the best performance. The results demonstrate that although

Table XI

PERFORMANCE COMPARISON IN THE BOV FRAMEWORK ON THE 9 CLASS RGB-NIR SCENE DATASET.

Channels	Visual Descriptors			
	mCENTRIST	conCENTRIST	mSIFT	msSIFT
$O_1O_2O_3O_4$	74.7(± 1.2)	71.6(± 3.1)	74.0(± 4.2)	67.0(± 5.8)
$O_1O_2O_3O_4S$	76.1(± 4.0)	74.8(± 2.0)	74.3(± 3.7)	71.3(± 4.7)
cCENTRIST (HSV)	73.7(± 4.1)			

Table XII

PERFORMANCE COMPARISON IN THE PCA FRAMEWORK ON THE 9 CLASS RGB-NIR SCENE DATASET.

Channels	Visual Descriptors				
	mCENTRIST	conCENTRIST	mSIFT	msSIFT	mGIST
SPM					
O ₁ O ₂ O ₃ O ₄	84.5(±2.1)	81.3(±3.0)	62.0(±5.0)	61.2(±2.0)	67.4(±3.5)
O ₁ O ₂ O ₃ O ₄ S	82.1(±1.9)	81.7(±2.8)	58.5(±3.8)	62.2(±5.6)	69.5(±3.2)
cCENTRIST (HSV)	73.0(±4.1)				
No SPM					
O ₁ O ₂ O ₃ O ₄	76.6(±4.4)	60.0(±4.0)	56.1(±4.5)	55.7(±5.8)	54.4(±2.3)
O ₁ O ₂ O ₃ O ₄ S	77.8(±3.9)	63.9(±3.8)	54.2(±1.8)	56.5(±3.4)	58.4(±3.2)
cCENTRIST (HSV)	57.6(±4.4)				

mCENTRIST is originally proposed as a holistic multi-channel descriptor, its discriminative power is still strong when applied as a local descriptor for scene categorization.

3) *Descriptor comparison in the PCA framework:* Besides comparing all descriptors in the BOV framework, we also test them using the PCA framework. That is, PCA is employed as the feature postprocessing approach, instead of BOV method. Under this framework, all descriptors will be extracted holistically from the whole images or SPM blocks. For fair comparison, the same amount (40) of eigenvectors are used for them during PCA. Following Sec. IV-E2, opponent, hyper opponent and HSV color space are employed for testing. Because SPM is not originally applied to mGIST [5], [13], for better comparison all descriptors will be tested with and without SPM. Table XII lists the categorization results. We can see that in the holistic PCA framework, mCENTRIST is evidently the strongest one. Another interesting phenomenon is that mCENTRIST's performance advantage (at least 13%) over other descriptors is tremendous when SPM is not applied. And, its performance drop from "SPM" to "No SPM" is not very big. Thus, for some applications demanding real-time processing, SPM can be removed from mCENTRIST. mSIFT and msSIFT do not perform well as holistic descriptors. As demonstrated by many previous works [1], [4], SPM can enhance the performance of the descriptors significantly. In addition, although SPM is applied to mGIST, its performance is still inferior to mCENTRIST with or even without SPM.

4) *Descriptor comparison in the HSV and hyper HSV color space:* Because cCENTRIST is strictly located in the HSV color space, it is not compared with other descriptors in the same color space previously. To make a better comparison, here all descriptors will be evaluated in the HSV color space on the 9 class RGB-NIR scene dataset. Additionally, we also extend HSV color space to hyper HSV color space by embedding the Sobel channel for further comparison. The experiment setup and feature extraction mechanism are the same as Table VII. The categorization results are listed in Table XIII. It can be observed that mCENTRIST is still evi-

Table XIII

PERFORMANCE COMPARISON IN THE HSV AND HYPER HSV COLOR SPACE ON THE 9 CLASS RGB-NIR SCENE DATASET.

Channels	Visual Descriptors				
	mCENTRIST	conCENTRIST	mSIFT	msSIFT	mGIST
HSV	78.0(± 4.0)	78.8(± 2.9)	69.8(± 5.1)	67.7(± 4.3)	66.8(± 1.9)
HSVS	79.8(± 5.2)	79.8(± 4.1)	74.8(± 5.1)	71.9(± 5.5)	67.7(± 1.9)
cCENTRIST (HSV)	73.0(± 4.1)				

Table XIV

DEMONSTRATION OF THE EFFECTIVENESS OF SOBEL CHANNEL IN HYPER OPPONENT COLOR SPACE.

Visual Descriptors	Channels		
	O_1	O_2	O_3
CENTRIST	80.5(± 1.1)	69.7(± 0.9)	69.6(± 2.2)
	O_1S	O_2S	O_3S
mCENTRIST	81.5(± 1.4)	76.7(± 1.7)	78.1(± 0.6)

dently stronger than cCENTRIST and other descriptors in the HSV color space, but only slightly inferior to conCENTRIST. And, cCENTRIST performs better than mSIFT, msSIFT and mGIST. Sobel channel is also available for HSV color space to enhance the descriptors' performance, especially for mSIFT and msSIFT. mCENTRIST and conCENTRIST yield the best result in the hyper HSV color space.

5) *Effectiveness of Sobel channel:* For all tested descriptors, the hyper opponent color space enhances their performance more prominently than RGB and opponent color space. To further investigate the effect of the Sobel channel in hyper opponent color space, we make a comparison between CENTRIST extracted from O_1 , O_2 and O_3 and mCENTRIST from O_1S , O_2S and O_3S respectively on the 8 class event dataset. Table XIV lists the result. Evidently, the performance of O_1 , O_2 and O_3 can also be improved by adding the Sobel channel, especially for O_2 and O_3 . These experiments further demonstrate the effectiveness of the Sobel channel for scene categorization.

6) *Comparison between mCENTRIST and bmCENTRIST:* We also concern about the difference between our proposed mCENTRIST and the brute force mCENTRIST (bmCENTRIST) according to Eqn. 4. Because the time and memory consumption is prohibitively huge for making PCA on bmCENTRIST if the channel number is big, they are compared only on binary channel combinations derived from hyper opponent color space, such as O_1O_2 , O_1O_3 , O_1S and so on. Actually, nearly 30G memory is needed by bmCENTRIST for PCA, even when we only use 2 channels. The result on the 8 class event dataset is shown in Table XV. bmCENTRIST is generally better than mCENTRIST. However, the difference is not significant. PCA for high dimensional data requires a huge set of training images (e.g., 10×65536 according to [23]). Since this condition is impractical in most situations, using sub-mCT to replace mCT does not lead to significant information loss (as suggested by the results in Table XV).

7) *Comparison between mCENTRIST-L1 and mCENTRIST-L2:* The proposed mCENTRIST is constructed based on the Level 1 sub-mCTs derived from the Census Transform pyramid. Actually, Level 2 sub-mCTs can also be used for mCENTRIST extraction and support more channels to interact

Table XV
COMPARISON BETWEEN mCENTRIST AND bmCENTRIST ON THE 8 CLASS EVENT DATASET.

Visual Descriptors	Channels		
	O_1O_2	O_1O_3	O_2O_3
mCENTRIST	81.5(± 1.3)	80.9(± 1.3)	73.5(± 2.5)
bmCENTRIST	81.0(± 1.3)	82.2(± 1.0)	74.9(± 1.4)
	O_1S	O_2S	O_3S
mCENTRIST	81.5(± 1.4)	76.7(± 1.7)	78.1(± 0.6)
bmCENTRIST	82.6(± 1.7)	77.0(± 1.6)	77.2(± 1.7)

Table XVI
PERFORMANCE COMPARISON BETWEEN mCENTRIST-L1 AND mCENTRIST-L2.

Channels	Visual Descriptors	
	mCENTRIST-L1	mCENTRIST-L2
21 class land-use classification dataset		
$O_1O_2O_3$	87.3(± 1.6)	79.1(± 1.5)
$O_1O_2O_3S$	89.9(± 1.2)	86.2(± 1.6)
8 class event dataset		
$O_1O_2O_3$	85.0(± 1.5)	76.8(± 1.3)
$O_1O_2O_3S$	86.5(± 0.6)	81.4(± 2.4)
9 class RGB-NIR scene dataset		
$O_1O_2O_3O_4$	84.5(± 2.1)	78.6(± 4.8)
$O_1O_2O_3O_4S$	82.1(± 1.9)	80.3(± 3.7)

simultaneously (e.g., $n = 4, 5$) with a feasible feature dimensionality ($4 \times 4^n \times 31$) for PCA. The reason why they are not chosen is that their discriminative power is weaker than the Level-1 sub-mCTs. We denote the mCENTRISTs based on the Level 1 and Level 2 sub-mCTs as mCENTRIST-L1 and mCENTRIST-L2, respectively. Suppose n channels are used for classification, mCENTRIST-L2 is constructed using n -channel interaction. Table XVI exhibits the performance comparison between mCENTRIST-L1 and mCENTRIST-L2 in the proposed opponent and hyper opponent color space. The experiments are executed on the 21 class land-use classification dataset, 8 class event dataset and 9 class RGB-NIR scene dataset. We can see that mCENTRIST-L1 consistently outperforms mCENTRIST-L2 significantly.

F. Further discussions

1) *Student's t-test*: For the 4 testing datasets, both mCENTRIST and conCENTRIST can improve the performance of grayscale CENTRIST significantly. cCENTRIST and mGIST are significantly inferior to them. mCENTRIST outperforms conCENTRIST in almost all cases and is the best choice for all 4 datasets. Compared with mSIFT and msSIFT, both mCENTRIST and conCENTRIST keep the advantage over classification performance and computational efficiency.

It can be observed that conCENTRIST is the strongest competitor of mCENTRIST. Here, paired Student's t -test is executed for further comparison between them on their best performing setups. The resulting p -values of all datasets are listed in Table XVII. Using the significance level 0.05, mCENTRIST is statistically significantly better than conCENTRIST on the 21 class land-use classification dataset, 67 class indoor scene recognition dataset and the 9 class RGB-NIR scene dataset.

2) *The choice of Sobel channel*: By embedding the Sobel channel, the hyper opponent color space works better than

Table XVII
 P -VALUES OF STUDENT'S t -TEST ON THE 4 TESTING DATASETS.

	Testing datasets	
	21 class land-use	67 class indoor
p -value	0.0003	0.0294
	9 class RGB-NIR	8 class event
p -value	0.0119	0.1196

Table XVIII
SCENE CLASSIFICATION RESULTS (%) OF THE 9 CLASS RGB-NIR SCENE DATASET CORRESPONDING TO DIFFERENT SOBEL CHANNELS.

Channels	Visual Descriptors				
	mCENTRIST	conCENTRIST	mSIFT	msSIFT	mGIST
S_R	—	73.8(± 2.5)	65.2(± 4.4)	65.2(± 4.1)	60.8(± 4.7)
S_N	—	73.9(± 4.3)	63.5(± 3.5)	63.4(± 4.9)	60.9(± 4.9)
S_K	—	72.1(± 2.2)	66.1(± 4.6)	62.9(± 4.1)	60.6(± 4.9)
S_{O_1}	—	75.3(± 3.4)	63.5(± 5.2)	63.7(± 5.1)	59.3(± 4.5)
$O_1O_2O_3O_4S_R$	82.1(± 1.9)	81.7(± 2.8)	79.3(± 1.9)	75.2(± 2.6)	69.9(± 5.0)
$O_1O_2O_3O_4S_N$	82.2(± 2.8)	80.3(± 3.4)	78.6(± 2.2)	72.6(± 3.3)	70.2(± 4.1)
$O_1O_2O_3O_4S_K$	83.4(± 2.2)	81.3(± 3.6)	79.4(± 0.7)	73.5(± 3.6)	70.7(± 4.2)
$O_1O_2O_3O_4S_{O_1}$	83.0(± 1.9)	80.2(± 3.2)	79.0(± 2.5)	74.0(± 4.2)	70.1(± 4.4)

RGB, RGB-NIR and opponent color space. Moreover, Sobel channel is always readily available for all color spaces. As aforementioned, besides R channel the Sobel images of other channels (e.g., K , NIR and O_1) can also be applied to constructing the hyper opponent color space. However, which one is optimal is still difficult to judge. So far, we find that the performance difference between these Sobel channels is not big. And, no single one is the best choice for all descriptors. Sobel image of R channel is chosen in previous experiments. Table XVIII⁷ shows the classification results of the 9 class RGB-NIR scene dataset corresponding to different Sobel channels. The Sobel images from R, K, N and O_1 channels are labelled as S_R , S_K , S_N and S_{O_1} . It can be observed that for all descriptors, different Sobel channels generally achieve similar performance both under single and multiple channel conditions. According to the listed results, which Sobel channel is optimal for the hyper opponent color space cannot be well inferred. This phenomenon, in fact, also happens to other datasets.

3) *The optimal channel interaction mechanism*: The proposed mCENTRIST is extracted based on binary channel interaction, while ternary and quaternary channel interaction mechanisms are also applicable to mCENTRIST. However, binary channel interaction is optimal from both the computational efficiency and classification accuracy perspective. Here, we denote the mCENTRISTs based on binary, ternary and quaternary channel interaction mechanisms as mCENTRIST-B, mCENTRIST-T and mCENTRIST-Q. Suppose the channel number is n ($n \geq 4$), without PCA operation the feature dimensionality of the three descriptors is $\binom{n}{2} \times 2 \times 16^2 \times 31$ (mCENTRIST-B), $\binom{n}{3} \times 2 \times 16^3 \times 31$ (mCENTRIST-T) and $\binom{n}{4} \times 2 \times 16^4 \times 31$ (mCENTRIST-Q), respectively. In the hyper opponent color space, n is usually 4 or 5. Under this condition, the feature dimensionality of mCENTRIST-T and mCENTRIST-Q is much higher than mCENTRIST-B. Thus, their time and space computational complexity are also much

⁷The experiment setup is the same as Table VII.

Table XIX
PERFORMANCE COMPARISON ON mCENTRISTs BASED ON DIFFERENT
CHANNEL INTERACTION MECHANISMS.

Channels	Visual Descriptors		
	mCENTRIST-B	mCENTRIST-T	mCENTRIST-Q
8 class event dataset			
O ₁ O ₂ O ₃	85.0(±1.5)	81.3(±1.7)	—
O ₁ O ₂ O ₃ S	86.5(±0.6)	85.0(±2.3)	82.2(±2.9)
9 class RGB-NIR dataset			
O ₁ O ₂ O ₃ O ₄	84.5(±2.1)	81.4(±3.3)	77.1(±4.9)
O ₁ O ₂ O ₃ O ₄ S	82.1(±1.9)	83.1(±3.1)	78.7(±3.8)

Table XX
PERFORMANCE COMPARISON ON mCENTRIST BETWEEN THE OPPONENT
COLOR SPACE PROPOSED IN THIS PAPER AND THAT PROPOSED IN [13].

Channels	Testing datasets			
	21 land-use	67 indoor	9 RGB-NIR	8 event
T ₁ T ₂ T ₃	86.6(±2.4)	40.0(±1.5)	—	83.9(±1.2)
O ₁ O ₂ O ₃	87.3(±1.6)	40.8(±1.7)	—	85.0(±1.5)
T ₁ T ₂ T ₃	89.4(±1.7)	44.2(±1.2)	—	85.7(±1.1)
O ₁ O ₂ O ₃	89.9(±1.2)	44.6(±1.2)	—	86.5(±0.6)
T ₁ T ₂ T ₃ T ₄	—	—	84.2(±2.6)	—
O ₁ O ₂ O ₃ O ₄	—	—	84.5(±2.1)	—
T ₁ T ₂ T ₃ T ₄ S	—	—	82.9(±3.6)	—
O ₁ O ₂ O ₃ O ₄ S	—	—	82.1(±1.9)	—

greater than those of mCENTRIST-B during the phase of PCA (especially for the eigenvector computation). Additionally, to achieve satisfactory PCA performance, more training samples are required by them. Table XIX lists the classification results of the three kinds of mCENTRIST descriptors on the 9 class RGB-NIR scene dataset and 8 class event dataset.⁸ The proposed opponent and hyper opponent color space are used for categorization. As a matter of fact, mCENTRIST-B outperforms mCENTRIST-T and mCENTRIST-Q nearly in all cases, especially in the opponent color space.

4) *Opponent color space comparison:* We also make a further comparison on the opponent color space proposed in this paper and that proposed in [13] (the channels are labeled as T₁, T₂ and T₃ for RGB and T₁, T₂, T₃ and T₄ for RGB-NIR) with and without the Sobel channel in mCENTRIST. The results on all the 4 testing datasets are listed in Table XX. It can be observed that our opponent color space outperforms the color space in [13] (i.e., Eqn. 11 vs. Eqn. 12) almost consistently: although the difference is not large, Eqn. 12 wins 15 times in the 16 tested experiments.

5) *Some other issues:* Finally, we conclude our discussions with two closely related methods: cCENTRIST and LBP. In Tables V, VI, VII, IX, it is obvious that cCENTRIST consistently outperforms the grayscale version of CENTRIST, which corroborates the conclusion in [18]. However, when compared with mCENTRIST and conCENTRIST, we find that cCENTRIST is not as effective in utilizing the color information. Like mCENTRIST, cCENTRIST also employs multi-channel joint information for scene categorization, but in a different way. It uniformly quantizes the hue, saturation and value components into 2, 4, and 32 levels respectively, using 1, 2 and 5 encoding bits. The 8 bits are then put together to encode all HSV color channels in one single byte. It is

⁸Because of the memory limitation, the same experiment cannot run on the other two larger datasets.

worth noting that color space quantization leads to serious information loss. For example, the hue components in [0 127] are all quantized as 0. Consequently, a lot of descriptive scene properties vanish. However, since cCENTRIST only uses 8 bits to encode a pixel's color value, it might have advantages in resource constrained applications, such as image retrieval in devices with limited resources (e.g., mobile phone).

As explained in [1], the Census Transform is similar to LBP_{8,1} [35], [36]. They differ by a few factors: whether pixel values are interpolated or not, direction of pixel value comparison, order of collecting the 8 bits, etc. However, our multi-channel feature generation mechanism is independent of these differences. That is, when we extract multi-channel features for LBP, similar accuracy boost is expected.

V. CONCLUSIONS

In this paper, mCENTRIST (multi-channel CENTRIST) is proposed as a multi-channel feature generation mechanism for scene categorization. Different from popular color descriptors, mCENTRIST captures the intrinsic joint properties within multi-channel images explicitly. Both feature and channel level tradeoffs have been made to avoid the curse of dimensionality in practical applications. Meanwhile, mCENTRIST is also very easy to implement and has nearly no parameter to tune.

The Sobel gradient image is demonstrated to have low correlation with channels in the opponent color space and be beneficial for scene classification. Thus, the opponent color space is upgraded to hyper opponent color space by embedding Sobel channel to integrate more valuable scene information.

On four RGB and RGB-NIR datasets, including aerial orthoimagery, indoor and outdoor scene categorization tasks, mCENTRIST improves grayscale CENTRIST's performance remarkably. And it achieves higher accuracy than other multi-channel descriptors. Meanwhile, mCENTRIST's computational efficiency is much higher than mSIFT, msSIFT and mGIST. The effectiveness of the hyper opponent color space has also been demonstrated.

In the future, we plan to further enhance mCENTRIST's discriminative power by overcoming the negative effects yielded by feature and channel level tradeoffs, while maintaining its high computational efficiency. Furthermore, we plan to research on how to jointly encode other features (e.g., SIFT or ORB [37]) from multiple channels.

ACKNOWLEDGMENT

We thank the anonymous reviewers for their useful comments and suggestions, which have helped significantly improve this paper. J. Wu is the corresponding author of this paper, and is partially supported by the National Science Foundation of China under Grant No. 61321491.

Y. Xiao and J. Yuan are supported in part by IMI Seed Grant M4081078.B40, and Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office.

REFERENCES

- [1] J. Wu and J. M. Rehg, "CENTRIST: A visual descriptor for scene categorization," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 33, no. 8, pp. 1489–1501, 2011.
- [2] A. Bosch, A. Zisserman, and X. Muñoz, "Scene classification using a hybrid generative/discriminative approach," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 30, no. 4, pp. 712–727, 2008.
- [3] L. Fei-Fei and P. Perona, "A Bayesian hierarchical model for learning natural scene categories," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, vol. 2, 2005, pp. 524–531.
- [4] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, vol. 2, 2006, pp. 2169–2178.
- [5] A. Oliva and A. Torralba, "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, 2001.
- [6] A. Quattoni and A. Torralba, "Recognizing indoor scenes," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2009, pp. 413–420.
- [7] J. Wu and J. M. Rehg, "Where am I: Place instance and category recognition using spatial PACT," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2008, pp. 1–8.
- [8] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.
- [9] D. Song and D. Tao, "Biologically inspired feature manifold for scene classification," *IEEE Transactions on Image Processing*, vol. 19, no. 1, pp. 174–184, 2010.
- [10] G. J. Burghouts and J. M. Geusebroek, "Performance evaluation of local colour invariants," *Computer Vision and Image Understanding*, vol. 113, no. 1, pp. 48–62, 2009.
- [11] J.-M. Geusebroek, R. van den Boomgaard, A. W. M. Smeulders, and H. Geerts, "Color invariance," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 23, no. 12, pp. 1338–1350, 2001.
- [12] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek, "Evaluating color descriptors for object and scene recognition," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1582–1596, 2010.
- [13] M. Brown and S. Sussstrunk, "Multi-spectral SIFT for scene category recognition," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2011, pp. 177–184.
- [14] A. E. Abdel-Hakim and A. A. Farag, "CSIFT: A SIFT descriptor with color invariant characteristics," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, vol. 2, 2006, pp. 1978–1983.
- [15] J. van de Weijer, T. Gevers, and A. D. Bagdanov, "Boosting color saliency in image feature detection," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 28, no. 1, pp. 150–156, 2006.
- [16] W. Liu and D. Tao, "Multiview hessian regularization for image annotation," *IEEE Trans. on Image Processing*, vol. 22, no. 7, pp. 2676–2687, 2013.
- [17] X. Wang, W. Bian, and D. Tao, "Grassmannian regularized structured multi-view embedding for image classification," *IEEE Trans. on Image Processing*, vol. 22, no. 7, pp. 2646–2660, 2013.
- [18] W.-T. Chu and C.-H. Chen, "Color CENTRIST: a color descriptor for scene categorization," in *Proc. ACM Conf. on Multimedia Retrieval*, 2012.
- [19] F. Perronnin and C. Dance, "Fisher kernels on visual vocabularies for image categorization," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2007, pp. 1–8.
- [20] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2010, pp. 3304–3311.
- [21] R. Zabih and J. Woodfill, "Non-parametric local transforms for computing visual correspondence," in *Proc. European Conf. on Computer Vision*, vol. 801, 1994, pp. 151–158.
- [22] J. Wu, "Efficient HIK SVM learning for image classification," *IEEE Trans. on Image Processing*, vol. 21, no. 10, pp. 4442–4453, 2012.
- [23] J. W. Osborne and A. B. Costello, "Sample size and subject to item ratio in principal components analysis," *Practical Assessment, Research & Evaluation*, vol. 9, no. 11, 2004.
- [24] O. Vinyals, Y. Jia, L. Deng, and T. Darrell, "Learning with recursive perceptual representations," in *Proc. Advances in Neural Information Processing Systems*, 2012, pp. 2834–2842.
- [25] B. Schölkopf and A. J. Smola, *Learning with kernels: Support vector machines, regularization, optimization and beyond*. Cambridge, Massachusetts: The MIT Press, 2002.
- [26] X. Tan and B. Triggs, "Enhanced local texture feature sets for face recognition under difficult lighting conditions," *IEEE Trans. on Image Processing*, vol. 19, no. 6, pp. 1635–1650, 2010.
- [27] Y. Yang and S. Newsam, "Spatial pyramid co-occurrence for image classification," in *Proc. IEEE Conf. on Computer Vision*, 2011, pp. 1465–1472.
- [28] L.-J. Li and L. Fei-Fei, "What, where and who? Classifying events by scene and object recognition," in *Proc. IEEE Conf. on Computer Vision*, 2007, pp. 1–8.
- [29] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2010, pp. 3360–3367.
- [30] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, pp. 27:1–27:27, 2011, software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
- [31] Y. Jiang, J. Yuan, and G. Yu, "Randomized spatial partition for scene recognition," in *Proc. European Conf. on Computer Vision*, 2012, pp. 730–743.
- [32] J. Wu, "Power mean SVM for large scale visual classification," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2012, pp. 2344–2351.
- [33] N. Salamat, D. Larlus, and G. Csurka, "Combining visible and near-infrared cues for image categorisation," in *Proc. British Machine Vision Conf.*, 2011, pp. 49.1–49.11.
- [34] N. M. Elfiky, F. S. Khan, J. van de Weijer, and J. González, "Discriminative compact pyramids for object and scene recognition," *Pattern Recognition*, vol. 45, no. 4, pp. 1627–1636, 2012.
- [35] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, 2002.
- [36] J. Ren, X. Jiang, and J. Yuan, "Noise-resistant local binary pattern with an embedded error-correction mechanism," *IEEE Trans. on Image Processing*, 2013 (in press).
- [37] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," in *Proc. IEEE Conf. on Computer Vision*, 2011, pp. 2564–2571.

Yang Xiao received his BS, MS and PhD degrees from Huazhong University of Science and Technology, China. He is currently a research fellow in the Institute for Media Innovation at Nanyang Technological University, Singapore. Previously, he was a research fellow in the School of Computer Engineering at Nanyang Technological University, Singapore. His research interests involve computer vision, image processing and machine learning.

Jianxin Wu received his BS and MS degrees in computer science from Nanjing University, and his PhD degree in computer science from the Georgia Institute of Technology. He is currently a professor in the Department of Computer Science and Technology at Nanjing University, China, and is associated with the National Key Laboratory for Novel Software Technology, China. He was an assistant professor in the Nanyang Technological University, Singapore. His research interests are computer vision and machine learning. He is a member of the IEEE.

Junsong Yuan received Ph.D. from Northwestern University, Illinois, USA and M.Eng. from National University of Singapore. Before that, he graduated from the Special Class for the Gifted Young of Huazhong University of Science and Technology and received his B.Eng. in 2002.

He joined Nanyang Technological University (NTU) as a Nanyang Assistant Professor in Aug. 2009. Dr. Yuans research interests include computer vision, video analytics, large-scale visual search and mining, human computer interaction, biomedical image analysis, etc. He received Nanyang Assistant Professorship and Tan Chin Tuan Exchange Fellowship from Nanyang Technological University, Outstanding EECS Ph.D. Thesis award from Northwestern University, Best Doctoral Spotlight Award from IEEE Conf. Computer Vision and Pattern Recognition Conference (CVPR'09), and National Outstanding Student from Ministry of Education, P.R.China. He co-chairs workshops at IEEE Conf. Computer Vision and Pattern Recognition (CVPR'12'13) and IEEE Conf. on Computer Vision (ICCV'13), and serves as Area Chair for IEEE Winter Conf. on Computer Vision (WACV'14) and IEEE Conf. on Multimedia Expo (ICME'14). He is Organizing Chair and Area Chair of Asian Conf. on Computer Vision (ACCV'14). He currently also serves as associate editor for The Visual Computer journal (TVC) and Journal of Multimedia. He recently gives tutorials at IEEE ICIP'13, FG'13, ICME'12, SIGGRAPH VRCAI'12, and PCM'12, has filed three US patents and two provisional US patents.