Feature-Based Sparse Representation for Image Similarity Assessment

Li-Wei Kang, *Member, IEEE*, Chao-Yung Hsu, Hung-Wei Chen, Chun-Shien Lu, *Member, IEEE*, Chih-Yang Lin, *Member, IEEE*, and Soo-Chang Pei, *Fellow, IEEE*

***Abstract*—Assessment of image similarity is fundamentally important to numerous multimedia applications. The goal of similarity assessment is to automatically assess the similarities among images in a perceptually consistent manner. In this paper, we interpret the image similarity assessment problem as an information fidelity problem. More specifically, we propose a feature-based approach to quantify the information that is present in a reference image and how much of this information can be extracted from a test image to assess the similarity between the two images. Here, we extract the feature points and their descriptors from an image, followed by learning the dictionary/basis for the descriptors in order to interpret the information present in this image. Then, we formulate the problem of the image similarity assessment in terms of sparse representation. To evaluate the applicability of the proposed feature-based sparse representation for image similarity assessment (FSRISA) technique, we apply FSRISA to three popular applications, namely, image copy detection, retrieval, and recognition by properly formulating them to sparse representation problems. Promising results have been obtained through simulations conducted on several public datasets, including the Stirmark benchmark, Corel-1000, COIL-20, COIL-100, and Caltech-101 datasets.**

***Index Terms*—Feature detection, image copy detection, image recognition, image retrieval, image similarity assessment, sparse representation.**

I. Introduction

IMAGE similarity assessment is fundamentally important to numerous multimedia information processing systems and applications, such as compression, restoration, enhancement, copy detection, retrieval, and recognition/classification. The major goal of image similarity assessment is to design algorithms for automatic and objective evaluation of similarity in a manner that is consistent with subjective human evaluation. A simple and popularly used metric is the peak signal-to-noise ratio (PSNR) or the corresponding mean-squared error (MSE), whose correlation with human judgment has been shown to be not tight enough for most applications [1], [2]. Some advanced approaches, based on the human visual system (HVS), natural scene statistics (NSS), and/or some image distortion model, also have been proposed to improve the PSNR metric. They demonstrate that visual quality of a test image is strongly related to the relative information present in the image and that the information can be quantified to measure the similarity between the test image and its reference image [1], [2].

There is no doubt that these advanced similarity metrics are efficient to measure the "quality" of an image compared with its original version, especially for some image reconstruction applications. Nevertheless, they mainly focus on assessing the similarities between a reference image and its non-geometrically variational versions, such as decompressed and brightness/contrast-enhanced versions. Different from the above, in this paper, we emphasize the "similarity" between two arbitrary images. In several applications, assessment of the similarities between a reference image and its geometrically variational versions, such as translation, rotation, scaling, flipping, and other deformations, is required. On the other hand, one could encounter appearance variabilities of images, including background clutter, different viewpoints, and different orientations. Even if some advanced approaches, such as the structural similarity (SSIM) index and visual information fidelity (VIF) [1], [2], can tolerate slightly geometric variations, their goals still do not devote to the consideration of more comprehensive image variations.

In this paper, motivated by the concept addressed in Sheikh and Bovik's scheme [2], we interpret the image similarity assessment problem as an information fidelity problem. More specifically, we attempt to quantify the information that is present in a reference image and how much of this information can be extracted from a test image to assess the similarity between the two images. The core of the proposed approach, significantly different from that used in [2], can be addressed as follows. In [2], image information is quantified using HVS, NSS, and an image distortion model, while we propose a feature-based approach to quantify the information present in an image, based on robust image feature extraction. That is, we detect the feature points of an image, followed by describing each feature point using a descriptor. Then, we propose to represent all of the descriptors of an image via sparse representation and assess the similarity between two images via sparse coding technique. The merit is that a feature descriptor is sparsely

Manuscript received October 31, 2010; revised March 03, 2011 and May 14, 2011; accepted May 31, 2011. Date of publication June 09, 2011; date of current version September 16, 2011. This work was supported in part by the National Science Council, Taiwan, under Grants NSC97-2628-E-001-011-MY3, NSC98-2631-H-001-013, NSC98-2811-E-001-008, NSC99-2218-E-001-010, NSC99-2811-E-001-006, and NSC 99-2221-E-468-023. A preliminary version of this manuscript was presented in the 2010 IEEE International Conference on Multimedia and Expo [8]. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Ming-Ting Sun.

L.-W. Kang and C.-S. Lu are with the Institute of Information Science, Academia Sinica, 115 Taipei, Taiwan (e-mail: lcs@iis.sinica.edu.tw).

C.-Y. Hsu and H.-W. Chen are with the Institute of Information Science, Academia Sinica, and Graduate Institute of Communication Engineering, National Taiwan University, Taipei, Taiwan.

C.-Y. Lin is with the Department of Computer Science and Information Engineering, Asia University, Taichung, Taiwan.

S.-C. Pei is with the Graduate Institute of Communication Engineering, National Taiwan University, Taipei, Taiwan.

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

Digital Object Identifier 10.1109/TMM.2011.2159197

1520-9210/$26.00 © 2011 IEEE

represented in terms of a dictionary or transferred as a linear combination of dictionary atoms, so as to achieve efficient feature representation and robust image similarity assessment. In this paper, the term "atom" means a typical pattern or basic unit learned from a set of training data. A dictionary consisting of several atoms can provide sparse representations of the data as a linear combination of a few atoms.

In this paper, we adopt the SIFT feature[1] [3] as the basis of our feature-based image similarity assessment scheme. The reasons, in terms of the robustness and applicabilities of SIFT, are briefly described as follows. In the literature, SIFT is one of the most pervasive and robust image features, and it has been widely used in several multimedia applications, such as image retrieval [4], [5], recognition [6]–[8], copy detection [8], [9], and near-duplicate detection [10], [11]. In addition, in a recent performance evaluation, the SIFT descriptor has been shown to outperform other local descriptors [12]. Current SIFT-based image retrieval approaches are usually based on building indices for SIFT feature descriptors that are extracted from local image regions. Then, the descriptors are quantized into visual words defined in a pre-constructed vocabulary. Finally, image retrieval can be achieved through a text retrieval technique [4]. Moreover, for SIFT-based image recognition, an efficient architecture, called a vocabulary tree, was proposed [5]. Based on quantized SIFT feature descriptors, the support vector machine (SVM) or nearest-neighbor (NN) techniques are usually used for image recognition [7].

In this paper, we study sparse representation and matching techniques of SIFT features for realizing our idea of quantifying image information and similarity assessment between images. We also show that the proposed feature-based sparse representation for image similarity assessment (FSRISA) technique can be broadly applied to numerous multimedia applications through proper problem formulations. In Sections I-A–E, we briefly review the SIFT technique and explore the two aspects of the SIFT feature, namely, representation and matching, followed by a presentation of the overview of the proposed scheme.

A. SIFT

SIFT [3] is a powerful technique extensively used in the community of computer vision and pattern recognition to detect and describe local features in images. Roughly speaking, SIFT transforms an image into a large collection of descriptors (feature vectors), each of which is invariant to image translation, scaling, and rotation, is partially invariant to illumination changes, and is robust to local geometric distortion. The main stages of SIFT include: 1) scale-space extrema detection; 2) keypoint localization; 3) orientation assignment; and 4) keypoint descriptor generation.

B. Representation of SIFT Feature

To extract SIFT features from an image, keypoints are localized first, based on scale-space extrema detection. Then, one or more orientations, based on local image gradient directions, will be assigned to each keypoint. Finally, a local image descriptor is built for each keypoint, based on the image gradients in its local neighborhood. In the standard SIFT descriptor representation, each descriptor is a 128-dimensional feature vector [3]. Usually, hundreds to thousands of keypoints may be extracted from an image.

To make the SIFT feature more compact, the bag-of-words (BoW) representation approach quantizes SIFT descriptors via vector quantization technique into a collection of visual words based on a pre-defined codebook, such as visual vocabulary [4] or vocabulary tree [5]. Also, advanced compression for SIFT features has been investigated recently [13].

C. Matching of SIFT Feature

To evaluate the similarity between two images based on their SIFT features, the most straightforward scheme is to perform keypoint matching by treating each image as a set of keypoints and conducting direct keypoint-set mapping [3]. The similarity between the two images is based on the number of matched keypoints, between the two sets of keypoints.

Moreover, for the BoW representation-based approach [4], the similarity between SIFT features can be measured by matching their corresponding visual words via histogram matching [14]. Typically, the computational complexity of the direct keypoint matching approach is higher than that of the BoW-based approach. Nevertheless, the outcomes of the direct keypoint matching approach are usually more reliable than those of the BoW-based approach suffered from quantization loss [11].

D. Overview of the Proposed Scheme

In this paper, a scheme of feature-based sparse representation for image similarity assessment (FSRISA) is proposed. SIFT is adopted as the representative feature detector in our framework. To compactly represent SIFT feature of an image, we propose construction of the basis (dictionary), consisting of the prototype SIFT atoms via dictionary learning that forms the feature, called "dictionary feature," of the image. To assess the similarity between two images based on their dictionary features, we propose formulating the problem as a sparse representation problem, where we perform sparse coding and calculate the reconstruction error for each SIFT descriptor of a test image. Then, based on a voting strategy, we can define a similarity value (matching score) between the two images. We also apply our FSRISA to three multimedia applications, including image copy detection, retrieval, and recognition, by properly formulating them to their corresponding sparse representation problems.

The major novelties and contributions of this paper include: 1) a feature-based image assessment approach is proposed to quantify how much information present in a reference image can be extracted from a test image by integrating image feature extraction and sparse representation; 2) the inherent discriminative characteristic of sparse representation is exploited to assess the similarity between two images by performing sparse coding with respect to the dictionary integrated from the two dictionary features of the two images, respectively; 3) efficient feature representation can be achieved by representing features in terms of linear combination of dictionary atoms; and 4) the proposed FSRISA provides a bounded similarity score, i.e., $[0, 1]$, for detected features to quantify the similarity between two images.

[1]Nevertheless, any feature descriptors can be used in the proposed framework.

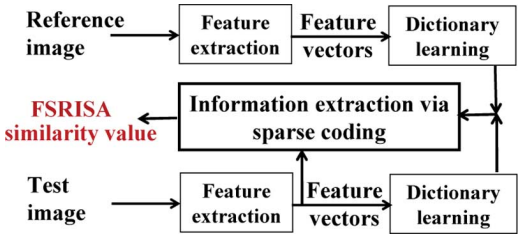

Fig. 1. Concept of the proposed FSRISA framework.

A bounded similarity score should be more suitable for users to conjecture how similar two images are or for a vision system to decide a threshold for image comparison. This metric should be better than just using the number of matched keypoints which may range from zero to thousands of keypoints.

E. Organization of This Paper

The rest of this paper is organized as follows. The proposed FSRISA scheme is addressed in Section II. The applications of FSRISA to image copy detection, retrieval, and recognition are presented in Section III. The simulation results are shown in Section IV, followed by the conclusion presented in Section V.

II. Proposed Feature-Based Sparse Representation for Image Similarity Assessment (FSRISA) Methodology

Sparse representation has resulted in significant impact on computer vision and pattern recognition, usually in unconventional applications where the goal is not just to obtain a compact representation of the observed signal, but also to extract semantic information. The selection of the dictionary plays a key role to achieve this goal. That is, overcomplete dictionaries consisting of (or learned from) the training samples provide the key to attach semantic meaning to sparse signal representations [15].

In this paper, we utilize the sparse representation and dictionary learning techniques to design our framework of feature-based sparse representation for image similarity assessment (FSRISA). As illustrated in Fig. 1, given a reference image, we first apply standard SIFT to detect the keypoints and extract the feature vector for each keypoint in this image. To make the SIFT features more compact, we propose to learn the dictionary consisting of the prototype SIFT atoms to form the "dictionary feature" of the reference image, as described in Section II-A. Similarly, we also extract the dictionary feature for an input test image. Then, we calculate the similarity value between the two images using the proposed FSRISA technique, as described in Section II-B.

A. Dictionary Feature Extraction

Here, we apply the K-SVD dictionary learning algorithm [16] to construct the dictionary for a set of SIFT feature vectors of an image to form its dictionary feature. To learn an overcomplete dictionary for a set of training signals, K-SVD seeks the dictionary leading to the best possible representation of each signal in this set with strict sparsity constraints. The K-SVD algorithm, which generalizes the K-means algorithm, is an iterative scheme alternating between sparse coding of the training signals with respect to the current dictionary and an update process for the dictionary atoms so as to better fit the training signals.

Given a set of K SIFT feature vectors, $y_i \in R^{M\times 1}$, $i = 1, 2, \ldots, K$, we apply K-SVD to find the dictionary D of size $M \times N$, $M < N \ll K$, by formulating the problem as

$$\min_{D,[x_i],\ i=1,2,\ldots,K} \left(\sum_{i=1}^{K} \|y_i - Dx_i\|_2^2 \right) \quad \text{subject to } \forall\, i, \|x_i\|_0 \le L \tag{1}$$

where $x_i \in R^{N\times 1}$ is the sparse representation coefficients of y_i, $\|x_i\|_0$, l_0-norm of x_i, counts the number of nonzero coefficients of x_i, and L is the most desired number of nonzero coefficients of x_i. We apply K-SVD to solve (1) via an iterative manner with two stages: 1) sparse coding stage: apply orthogonal matching pursuit (OMP) [17] to solve x_i for each y_i while fixing D; and 2) dictionary update stage: update D together with the nonzero coefficients of x_i. The two stages are iteratively performed until convergence. It should be noted that the l_0-minimization formulation in (1) can be converted into an l_1-minimization problem [18] and other dictionary learning algorithm, (e.g., the online dictionary learning algorithm [18]) can be also applied in the dictionary feature extraction stage.

The obtained dictionary feature D is an overcomplete dictionary, where $D = \{[d_n]_{M\times 1}\}_{n=1,2,\ldots,N} \in R^{M\times N}$, contains N prototype feature vector atoms as the column vectors in D. Each original feature vector $y_i \in R^{M\times 1}$, $i = 1, 2, \ldots, K$, can be sparsely represented as a linear combination of the atoms defined in D, satisfying $\|y_i - Dx_i\|_2 \le \varepsilon$, where $\varepsilon \ge 0$ is an error tolerance.

B. Sparse Representation-Based Image Similarity Assessment

After obtaining the dictionary feature for each image, we formulate the image similarity assessment based on dictionary feature matching as a sparse representation problem, described as follows.

First, consider the two SIFT feature (column) vectors with length M, y_{1i}, $i = 1, 2, \ldots, K_1$, and y_{2j}, $j = 1, 2, \ldots, K_2$, extracted, respectively, from the two images, I_1 and I_2, where K_1 and K_2 are the numbers of feature vectors of I_1 and I_2, respectively. The dictionary features of I_1 and I_2 are D_1 (of size $M \times N_1$) and D_2 (of size $M \times N_2$), respectively, where $M < N_1$ and $M < N_2$. Hence, $y_{1i} = D_1 x_{1i}$ and $y_{2j} = D_2 x_{2j}$, where x_{1i} and x_{2j} are the two sparse coefficient (column) vectors with length N_1 and N_2, of y_{1i} and y_{2j}, respectively. Obviously, if y_{1i} and y_{2j} can be matched, y_{1i} can be represented sparsely and linearly with respect to D_2. On the other hand, y_{2j} can be represented sparsely and linearly with respect to D_1.

To assess the similarity between a reference image I_1 and a test image I_2, exploiting the discriminative characteristic of sparse representation, we want to quantify how much information present in I_1 can be extracted from I_2. A sparse representation problem for representing each SIFT feature vector y_{2j} of I_2 with respect to the joint dictionary $D_{12} = [D_1|D_2]$ can be defined as

$$\hat{x}_j = \min_{x_j} \|x_j\|_0 \text{ subject to } \|y_{2j} - D_{12}x_j\|_2 \le \varepsilon \tag{2}$$

where x_j with length (N_1+N_2) is the sparse coefficient vector of y_{2j} with length M of I_2. $D_{12}=[D_1|D_2]$ of size $M\times(N_1+N_2)$ is the joint dictionary concatenating D_1 and D_2, and $\varepsilon\geq 0$ is an error tolerance.

To solve the sparsest solution for x_j, (2) can be cast to an l_1-minimization problem as [19]

$$\hat{x}_j=\arg\min_{x_j}\left(\frac{1}{2}\|y_{2j}-D_{12}x_j\|_2^2+\tau\|x_j\|_1\right) \tag{3}$$

where τ is a positive real number parameter. In this paper, we apply an efficient sparse coding algorithm, called the sparse reconstruction by separable approximation (SpaRSA) algorithm [20] to solve (3) in order to find the sparse representation $(\hat{x}_j)$ of y_{2j} with respect to the dictionary D_{12}. SpaRSA is a very efficient iterative algorithm, where each step is obtained by solving an optimization subproblem involving a quadratic term with diagonal Hessian plus the original sparsity-inducing regularizer. Of course, (3) can be directly solved via a greedy algorithm, such as OMP [17] and other l_1-minimization algorithms.

It is expected that the positions of nonzero coefficients in $\hat{x}_j$ (or the selected atoms from D_{12}) should be highly concentrated on only one sub-dictionary (e.g., D_1 or D_2), and the remaining coefficients in $\hat{x}_j$ should be zeros or small enough. Also, it is intuitive to expect that the atoms for sparsely representing y_{2j} should be mostly selected from the sub-dictionary D_2 learned from the feature vectors extracted from the image I_2 itself, instead of D_1. If the parameters for learning the two dictionaries (D_1 and D_2) can be adequately tuned, the manner of atom selection in the sparse coding process may be changed accordingly. That is, we intend to make the sparse coefficients $\hat{x}_j$ (or the used atoms) solved by performing sparse coding for y_{2j} more consistent with our expectation to help for similarity assessment. More specifically, we expect y_{2j} will use more atoms from D_2 to represent it when I_2 and I_1 are visually different. On the other hand, we expect y_{2j} will use more atoms from D_1 to represent it when I_2 and I_1 are visually similar. The details are described in the seventh paragraph of this subsection.

Based on the obtained solution $\hat{x}_j$ of (3), we can calculate the reconstruction error as $\|y_{2j}-D_{12}\hat{x}_j\|_2$. By letting the elements in $\hat{x}_j$, corresponding to the atoms from D_2, be zeros, we can get the reconstruction error E_{1j}, using only the atoms from D_1 for reconstructing y_{2j}. On the other hand, by letting the elements in $\hat{x}_j$, corresponding to the atoms from D_1, be zeros, we can get the reconstruction error E_{2j}, using only the atoms from D_2 for reconstructing y_{2j}. If $E_{1j}<E_{2j}$, it is claimed that the atoms from D_1 are more suitable for representing y_{2j} than those from D_2, and D_1 will get a vote. Otherwise, if $E_{2j}<E_{1j}$, y_{2j} is more suitable to be represented by the atoms from D_2 (the dictionary learned from the feature vectors y_{2j} itself) than D_1, and D_2 will get a vote. Considering all the SIFT feature vectors of I_2, y_{2j}, $j=1,2,\ldots,K_2$, the obtained percentage of votes from D_1 and D_2 are denoted by V_1 and V_2, $0\leq V_1, V_2\leq 1$, respectively. Based on the voting strategy, we define the similarity between the two images, I_1 and I_2, as

$$\mathrm{Sim}(I_1,I_2)=\frac{(V_1-V_2+1)}{2} \tag{4}$$

where the range of $(V_1\text{-}V_2)$ is $[-1,1]$, which can be shifted to $[0,1]$, resulting in the $\mathrm{Sim}(I_1,I_2)$ defined in (4). Larger $\mathrm{Sim}(I_1,I_2)$ indicates that more atoms from D_1 learned from I_1 can well represent the feature vectors y_{2j} extracted from I_2. This implies that a considerable amount of information (denoted by D_1) presented in I_1 can be extracted [via spare coding by solving (3)] from I_2. On the other hand, smaller $\mathrm{Sim}(I_1,I_2)$ indicates most suitable atoms for representing y_{2j} extracted from I_2 are from D_2 learned from I_2 itself. This implies that less/no information presented in I_1 can be extracted from I_2. Hence, the larger the $\mathrm{Sim}(I_1,I_2)$ is, the more similar the images I_1 and I_2 are.

Obviously, if I_1 is visually very different from I_2, V_2 is larger than V_1. Nevertheless, if I_1 is visually similar to I_2, V_2 will not be larger than V_1 in all instances. That is, better (or similar) reconstruction performance for y_{2j} may be achieved using D_1 as the dictionary than using D_2 due to some feature vectors extracted from I_2 being able to be matched by the feature vectors extracted from I_1. To achieve this goal, we propose three rules for tuning the parameters used by K-SVD for learning the two dictionaries, D_1 and D_2: 1) the number of the atoms in D_1 should be larger than that in $D_2(N_1>N_2)$; 2) the number (J_1) of iterations K-SVD performs for learning D_1 should be larger than that (J_2) for learning $D_2(J_1>J_2)$; and 3) the number of the target sparsity (L_1), i.e., the number of nonzero coefficients for representing each feature vector for learning D_1, should be larger than that (L_2) for learning $D_2(L_1>L_2)$. According to the rules designed above, when I_1 is visually similar to I_2 and D_1 is finer than D_2, the l_1-minimizer for solving (3) may prefer more promising atoms from D_1 than D_2 to reconstruct y_{2j}, resulting in $V_1>V_2$ and larger $\mathrm{Sim}(I_1,I_2)$. Otherwise, when I_1 is visually different from I_2, most atoms for reconstructing y_{2j} will still be selected from D_2, resulting in $V_1<V_2$ and smaller $\mathrm{Sim}(I_1,I_2)$. The proposed FSRISA technique is summarized in Algorithm I and illustrated in Fig. 2.

The major goal of performing sparse coding with respect to the dictionary consisting of D_1 and D_2, instead of only D_1, can be addressed as follows. When I_1 (reference image) is visually different from I_2 (test image), D_1 and D_2 are significantly different. In this scenario, the idea behind our FSRISA is somewhat related to that of sparse coding-based image classification approach [15], [21] or sparse coding-based image decomposition approach [22]. We use the similar concept to quantify the similarity between I_2 and I_1, which may be interpreted as either 1) classifying I_2 into I_1 or I_2 itself; or 2) decomposing I_2 into the components of I_1 and/or those of I_2 itself. When I_1 is visually similar to I_2, D_1 and D_2 are similar, which is enforced to that D_1 is finer than D_2 in FSRISA. Hence, the above discussions are also valid in this scenario. Moreover, why we do not perform sparse coding with respect to only one dictionary D_1 can be explained as follows. When performing sparse coding for the feature vectors of I_2 with respect to a dictionary D_1 consisting of atoms which may be not suitable for sparsely representing them, the sparse coding procedure still attempts to minimize the reconstruction errors. Based on our experience, it is usually not well distinguishable from reconstruction errors obtained with respect to either related or unrelated dictionaries. On the other hand, it is not easy to de-

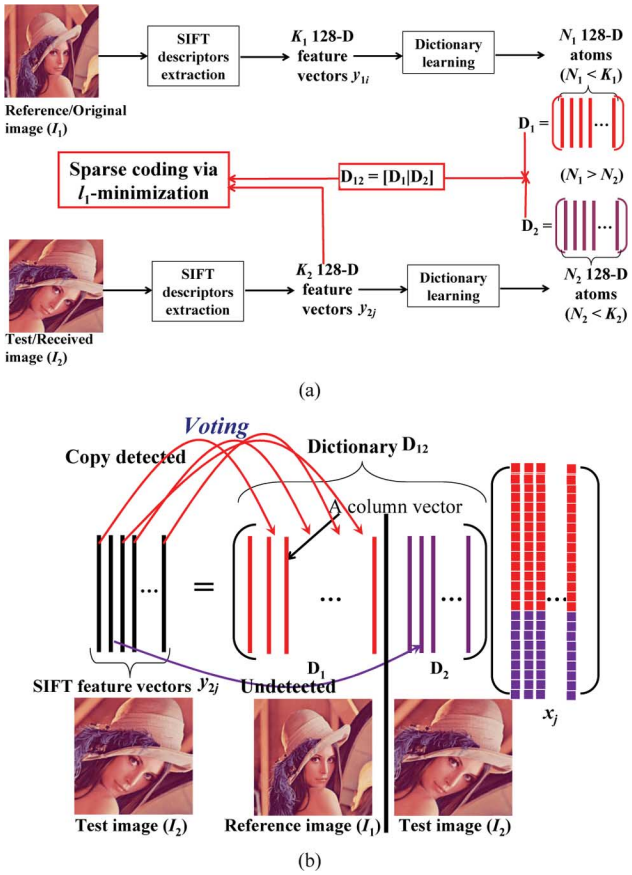

Fig. 2. Illustrated example of the proposed FSRISA framework. (a) Extraction of dictionary features for the reference and test images, respectively. (b) Matching of the two images via sparse coding and voting.

fine a bounded score based on reconstruction error obtained by only one dictionary.

Algorithm I: Proposed FSRISA

Input: A reference image I_1 and a test image I_2.

Output: The similarity value between I_1 and I_2, i.e., $\mathrm{Sim}(I_1, I_2)$.

1. Extract the SIFT feature vectors y_{1i}, $i = 1, 2, \ldots, K_1$, from I_1, followed by learning the dictionary feature D_1 sparsely representing y_{1i}.

2. Extract the SIFT feature vectors y_{2j}, $j = 1, 2, \ldots, K_2$, from I_2, followed by learning the dictionary feature D_2 sparsely representing y_{2j}.

3. Perform l_1-minimization by solving **(3)** for y_{2j}, $j = 1, 2, \ldots, K_2$, with respect to $D_{12} = [D_1|D_2]$.

4. Calculate the reconstruction errors, E_{1j} and E_{2j}, for y_{2j}, $j = 1, 2, \ldots, K_2$, with respect to D_1 and D_2, respectively.

5. Perform voting by comparing E_{1j} and E_{2j}, for y_{2j}, $j = 1, 2, \ldots, K_2$, and get the percentages of votes, V_1 and V_2, with respect to D_1 and D_2, respectively.

6. Calculate $\mathrm{Sim}(I_1, I_2) = (V_1 - V_2 + 1)/2$ (4).

C. Computational Complexity Analysis of FSRISA

The computational complexity of the proposed FSRISA can be analyzed as follows. The computational complexity for extracting the dictionary feature of an image includes the complexities of performing SIFT feature extraction and K-SVD dictionary learning. For an image I with KM-dimensional SIFT feature vectors, the computational complexity for learning a dictionary of size $M \times N$, $M < N \ll K$, using K-SVD [16] can be derived to be around [23]

$$T_{KSVD}(K, L, M, N, J) = [K \times (L^2 \times N + 2 \times M \times N)] \times J \tag{5}$$

where L is the target sparsity and J is the number of training iterations. Hence, the approximate computational complexity of the dictionary feature extraction for an image is obtained as

$$T_{Dict\_Feature}(S, K, L, M, N, J) = T_{SIFT}(S) + T_{KSVD}(K, L, M, N, J) \tag{6}$$

where $T_{SIFT}(S)$ roughly denotes the computational complexity (proportional to S) of SIFT feature extraction, in terms of the size (S) of an image.

On the other hand, the computational complexity of performing l_1-minimization using SpaRSA can be approximately derived as [20]

$$T_{SpaRSA}(P) = O(P^{\beta}) \tag{7}$$

where P is the number of atoms in a dictionary and β is a constant. It has been shown that the complexity of SpaRSA is approximately linear, i.e., β is very close to 1.

Based on (6) and (7), the overall computational complexity for assessing the similarity between two images, I_1 and I_2, by performing the proposed FSRISA can be derived as

$$\begin{aligned} T_{FSRISA}(S_1, S_2, K_1, K_2, L_1, L_2, N_1, N_2, J_1, J_2, M) &= T_{Dict\_Feature}(S_1, K_1, L_1, M, N_1, J_1) \\ &\quad + T_{Dict\_Feature}(S_2, K_2, L_2, M, N_2, J_2) \\ &\quad + T_{SpaRSA}(N_1 + N_2) \end{aligned} \tag{8}$$

where K_1 and K_2 denote the number of SIFT feature vectors for I_1 and I_2, respectively, L_1 and L_2 denote the target sparsities for learning D_1 and D_2 (the dictionary features of I_1 and I_2, respectively), respectively, N_1 and N_2 denote the number of atoms in D_1 and D_2, respectively, J_1 and J_2 denote the number of iterations for learning D_1 and D_2, respectively, and M denotes the length of a SIFT feature vector $(M = 128)$.

III. FSRISA for Multimedia Applications

In this section, we introduce three multimedia applications, including image copy detection, retrieval, and recognition, of the proposed FSRISA.

A. Image Copy Detection via FSRISA

Digital images distributed through the Internet may suffer from several possible manipulations, such as (re)compression, noising, contrast/brightness adjusting, and geometrical operations. To ensure trustworthiness, image copy detection

techniques have emerged to search for duplicates and forgeries [24], [25]. Image copy detection can be achieved via content-based copy detection approach, which measures the similarity/distance between an original image and its possible copy version through comparing their extracted image features, where SIFT-based features have been recently investigated [8], [9]. In this section, we study content-based image copy detection by applying the proposed FSRISA approach.

A user can perform image copy detection to detect possible copies of her/his original image from the Internet or an image database. To detect whether a test image I_2 is actually a copy of a query image I_1 with the dictionary feature D_1 of size $M \times N_1$, we first extract the SIFT feature vectors y_{2j}, $j = 1, 2, \ldots, K_2$, and learn the dictionary feature D_2 of size $M \times N_2$ of I_2, such that D_1 is finer than D_2. Then, we perform l_1-minimization by solving (3) for each y_{2j} with respect to $D_{12} = [D_1|D_2]$, and voting to get the percentages of votes, V_1 and V_2, with respect to D_1 and D_2, respectively. Finally, based on (4), the similarity between I_1 and I_2 can be calculated as $\mathrm{Sim}(I_1, I_2)$. Given an empirically determined threshold λ, if $\mathrm{Sim}(I_1, I_2) \geq \lambda$, then I_2 can be determined as a copy version of I_1. Otherwise, I_1 and I_2 can be determined to be unrelated. The computational complexity for performing the FSRISA-based image copy detection can be also similarly analyzed via (8).

B. Image Retrieval via FSRISA

The most popular image retrieval approach is content-based image retrieval (CBIR), where the most common technique is to measure the similarity between two images by comparing their extracted image features.

In the proposed scheme, for a query image, we extract its dictionary feature (with N atoms) and transmit it to an image database, where each image is stored together with its dictionary feature and original SIFT feature vectors. For comparing the query image I_Q and each database image I_{Di}, $i = 1, 2, \ldots$, *size_of_database*, where *size_of_database* denotes the total number of database images, we apply the proposed FSRISA scheme to perform the l_1-minimization (3) for each SIFT feature vector of I_{Di} with respect to the dictionary consisting of the dictionary features of the two images. Then, we calculate the reconstruction errors of all the stored feature vectors of I_{Di} and perform voting to get the similarity value between the two images (4) to be the score of I_{Di}. Finally, we retrieve the top Q database images with the largest scores. Similarly, the computational complexity of comparing the two images can be analyzed via (8), except that the complexity for extracting the dictionary feature of each database image can be excluded due to the fact that the process can be performed in advance during database construction.

C. Image Recognition via FSRISA

Consider a well-classified image database, where each class includes several images with the same object, but with different variations. Given a query image, a user may enquire to which class the image belongs. For image recognition, sparse representation techniques have been extensively used [21], [28]. The major idea is to exploit the fact that the sparsest representation is naturally discriminative. Among all of the subsets of atoms in a dictionary, it selects the subset that most compactly expresses the input signal and rejects all of the others with less compact representations. More specifically, image recognition/classification can be achieved by representing the feature of the query image as a linear combination of those training samples from the same class in a dictionary. Moreover, the conclusions in [21] claimed that their sparse representation-based face recognition algorithm should be extended to less constrained conditions (e.g., variations in object pose or misalignment). In order not to incur such a constraint, both variability-invariant features and sparse representation should be properly integrated. In addition, in [21], a dictionary consists of several subsets of image features (down-sampled image pixels were used), where each subset contains the features of several training images belonging to the same class. Nevertheless, if the number of classes, the number of training images in each class, or the feature dimension of a training image is too large, the dictionary size will be very large. It will induce very high computational complexity in performing sparse coding for the feature vector(s) of a query image.

In this paper, we propose an image recognition approach, where we assess the similarity between a query image and each class of training images based on the proposed FSRISA. In the training stage, for the ith image class, $i = 1, 2, \ldots, C$, where C denotes the number of classes in an image database, we extract the SIFT feature vectors for each image as the training samples. Then, we apply K-SVD [16] to learn the dictionary D_{Ci} of size $M \times N_{Ci}$ to be the "dictionary feature" of the ith image class, where $M(= 128)$ denotes the length of a SIFT feature vector and N_{Ci} denotes the number of atoms of D_{Ci}.

In the recognition stage, for a query image I_Q, we extract the SIFT feature vectors y_{Qj}, $j = 1, 2, \ldots, K_Q$, where K_Q denotes the number of SIFT feature vectors, and the dictionary feature D_Q. Then, we apply FSRISA to assess the similarity between I_Q and the ith image class, $i = 1, 2, \ldots, C$, by performing the l_1-minimization [similar to (3)] to obtain the sparse representation coefficients x_{Qj} for y_{Qj} of I_Q, with respect to the dictionary D_{Ci_Q} consisting of D_{Ci} and D_Q. Then, we calculate the reconstruction errors for y_{Qj} with respect to D_{Ci} and D_Q, respectively, and perform voting for each y_{Qj}. Based on (4), we can calculate the similarity between I_Q and the ith image class, denoted by Sim(I_Q, *Class-i*). Finally, the query image I_Q can be determined to belong to the ith class with the largest Sim(I_Q, *Class-i*).

Moreover, the computational complexity for assessing the similarity between the query image I_Q and the ith class of images can be also approximately analyzed based on (8), where the dictionary feature extraction for the ith class of images can be performed in advance during database construction. In our image recognition approach, the sparse coding procedure should be performed for a query image and each image class, which is indeed computationally expensive. The complexity of our approach (slightly cheaper than or similar to that of the two-stage approach proposed in [28]) can be improved by applying more efficient sparse coding techniques, such as multi-core OMP [18].

It is also worth noting that, if the feature size of a query image I_Q is crucial for online applications, each SIFT feature vector y_{Qj} can be further compressed via compressive sensing [19].

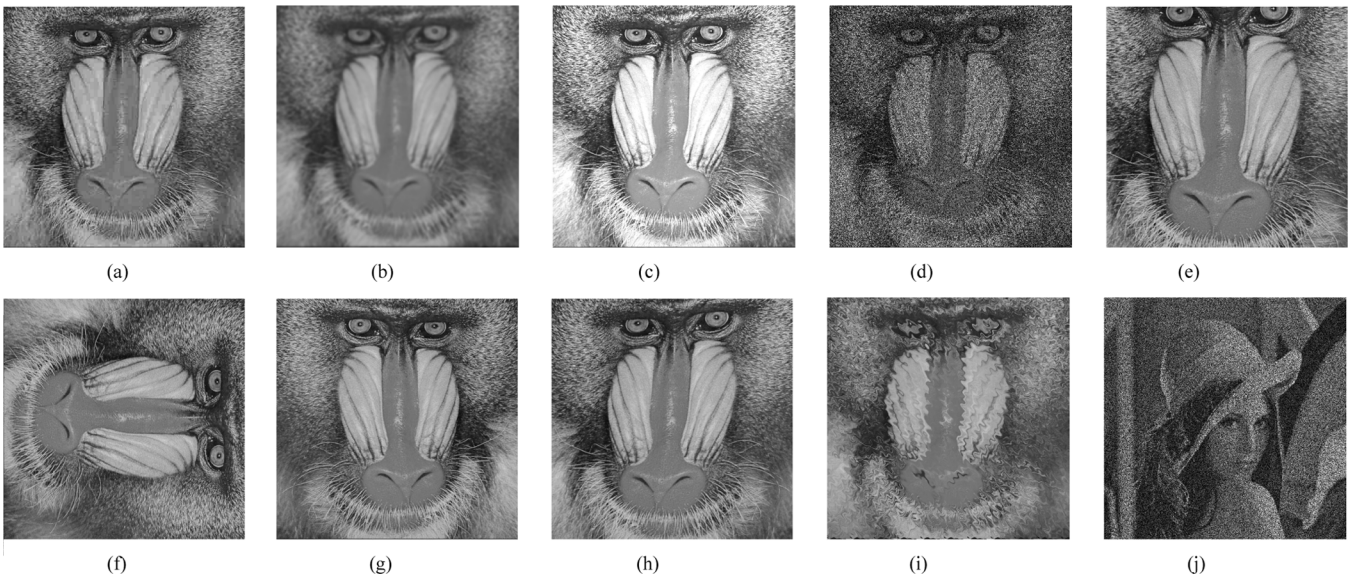

Fig. 3. Similarity values obtained from the PSNR, VIF, and our FSRISA between the *Baboon* image and its manipulated images. (a) JPEG compression ($\mathrm{PSNR} = 23.71\ \mathrm{dB}, \mathrm{VIF} = 0.28, \mathrm{FSRISA} = 0.70$). (b) Blurring ($\mathrm{PSNR} = 21.29\ \mathrm{dB}, \mathrm{VIF} = 0.14, \mathrm{FSRISA} = 0.66$). (c) Brightness and contrast adjusting ($\mathrm{PSNR} = 15.23\ \mathrm{dB}, \mathrm{VIF} = 1.07, \mathrm{FSRISA} = 0.74$). (d) Noising ($\mathrm{PSNR} = 11.20\ \mathrm{dB}, \mathrm{VIF} = 0.07, \mathrm{FSRISA} = 0.54$). (e) Scaling and cropping ($\mathrm{PSNR} = 14.16\ \mathrm{dB}, \mathrm{VIF} = 0.01, \mathrm{FSRISA} = 0.55$). (f) Rotation ($\mathrm{PSNR} = 13.26\ \mathrm{dB}, \mathrm{VIF} = 0.01, \mathrm{FSRISA} = 0.68$). (g) Flipping ($\mathrm{PSNR} = 14.89\ \mathrm{dB}, \mathrm{VIF} = 0.01, \mathrm{FSRISA} = 0.56$). (h) Shearing ($\mathrm{PSNR} = 16.32\ \mathrm{dB}, \mathrm{VIF} = 0.03, \mathrm{FSRISA} = 0.61$). (i) Rippling ($\mathrm{PSNR} = 18.27\ \mathrm{dB}, \mathrm{VIF} = 0.03, \mathrm{FSRISA} = 0.54$). (j) Irrelevant image ($\mathrm{PSNR} = 9.72\ \mathrm{dB}, \mathrm{VIF} = 0.01, \mathrm{FSRISA} = 0.03$).

Then, x_{Qj} can be also solved by performing the l_1-minimization based on the received measurements for y_{Qj} (compressed y_{Qj}).

IV. Simulation Results

In this section, we present simulations conducted on publicly available benchmarks or datasets for evaluation of the proposed FSRISA scheme in the fundamental issue of image assessment and three multimedia applications. Then, we address some experimental comparisons between sparse coding-based and traditional approaches to demonstrate the advantage of the proposed scheme.

A. Evaluation of FSRISA-Based Image Similarity Assessment

To evaluate the efficiency of FSRISA for assessing the similarity between two images, we use several examples of image manipulations (including signal processing and geometric distortions) defined in the Stirmark benchmarks [26] and compare FSRISA of similarity range $[0, 1]$ with well-known metrics, PSNR of range $[0, \infty]$, and VIF of range $[0, 1]$ (for image contrast enhancement, VIF value may be larger than 1) [2]. In the three evaluated metrics, the larger the value is, the more similar the two evaluated images are.

In our simulation, the size of each image is 280×280. Nevertheless, it should be noted that our scheme can work for images of different sizes. Based on the principle for tuning the KSVD parameters described in Section II-B, we set the following parameters to ensure that D_1 is finer than D_2. For a reference image I_1 of size $S_1 = 280 \times 280$, the parameters are shown as follows. We set the number of atoms in the dictionary feature D_1 to $N_1 = 0.5 \times K_1$, where K_1 denotes the number of SIFT feature vectors for I_1, the number of iterations K-SVD performs for learning D_1 to $J_1 = 30$, and the target sparsity to $L_1 = 0.3 \times N_1$. For a test image I_2 of size $S_2 = 280 \times 280$, the parameters are $N_2 = 0.3 \times K_2$, $J_2 = 10$, and $L_2 = 0.1 \times N_2$. If $N_1 < N_2$, we force $N_1 > N_2$ by properly adjusting the factors to be multiplied by K_1 and K_2, respectively.

The similarity values obtained from the PSNR, VIF, and FSRISA for some examples of image manipulations for the *Baboon* and *Lena* images, respectively, are shown in Figs. 3 and 4. It can be observed from Figs. 3 and 4 that for image similarity assessment, FSRISA is more robust to several image manipulations, especially for geometrical manipulations. The similarity values between an image and its related manipulated versions for FSRISA are usually higher than 0.5, while the ones between an image and unrelated versions are usually far lower than 0.5. Nevertheless, the VIF value between an image and its manipulated versions are usually lower than 0.5, except for the brightness and contrast adjusting manipulation, which enhances the image quality. For some manipulations (e.g., scaling, cropping, and flipping), the VIF value is almost indistinguishable from that for an unrelated image. Moreover, PSNR is obviously not good for similarity assessment. Hence, the discriminability of FSRISA is usually better than the other two metrics used for comparisons. In this paper, we focus on "similarity" assessment between images and hence, the similarity scores for different kinds of manipulations are somewhat similar. The major goal is also the "discriminability" between different images. Nevertheless, for the "quality" issue, the differences between a test image and its reference image (ground truth) becomes more critical, which is not the focus of this paper.

B. Evaluation of FSRISA-Based Image Copy Detection

To evaluate the proposed FSRISA-based image copy detection scheme, ten 280×280 images, *Baboon*, *Boat*, *Clock*, *Girl*, *House*, *Lena*, *Monarch*, *Pepper*, *Splash*, and *Tiffany*, were used. Each image was manipulated by 204 manipulations defined in the Stirmark benchmarks [26]. These image manipulations are also very similar to the ones used to evaluate image copy or

Fig. 4. Similarity values obtained from the PSNR, VIF, and our FSRISA between the *Lena* image and its manipulated images. (a) JPEG compression ($\mathrm{PSNR} = 30.87\ \mathrm{dB}, \mathrm{VIF} = 0.31, \mathrm{FSRISA} = 0.73$). (b) Blurring ($\mathrm{PSNR} = 25.17\ \mathrm{dB}, \mathrm{VIF} = 0.13, \mathrm{FSRISA} = 0.60$). (c) Brightness and contrast adjusting ($\mathrm{PSNR} = 15.41\ \mathrm{dB}, \mathrm{VIF} = 1.06, \mathrm{FSRISA} = 0.67$). (d) Noising ($\mathrm{PSNR} = 12.05\ \mathrm{dB}, \mathrm{VIF} = 0.08, \mathrm{FSRISA} = 0.50$). (e) Scaling and cropping ($\mathrm{PSNR} = 13.71\ \mathrm{dB}, \mathrm{VIF} = 0.02, \mathrm{FSRISA} = 0.67$). (f) Rotation ($\mathrm{PSNR} = 11.66\ \mathrm{dB}, \mathrm{VIF} = 0.01, \mathrm{FSRISA} = 0.77$). (g) Flipping ($\mathrm{PSNR} = 11.83\ \mathrm{dB}, \mathrm{VIF} = 0.02, \mathrm{FSRISA} = 0.53$). (h) Shearing ($\mathrm{PSNR} = 18.00\ \mathrm{dB}, \mathrm{VIF} = 0.07, \mathrm{FSRISA} = 0.57$). (i) Rippling ($\mathrm{PSNR} = 22.18\ \mathrm{dB}, \mathrm{VIF} = 0.07, \mathrm{FSRISA} = 0.49$). (j) Irrelevant image ($\mathrm{PSNR} = 9.69\ \mathrm{dB}, \mathrm{VIF} = 0.02, \mathrm{FSRISA} = 0.04$).

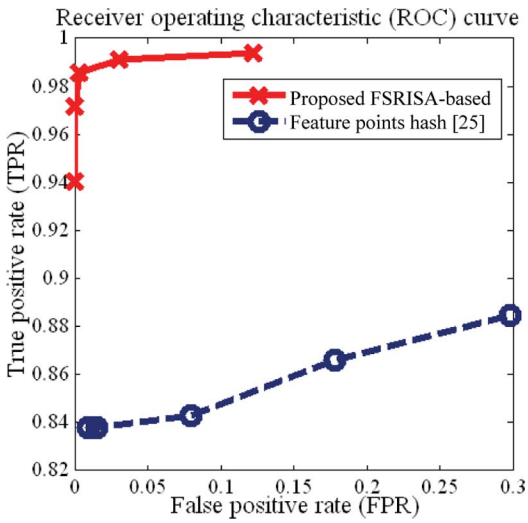

Fig. 5. Comparison of ROC curves obtained using the proposed FSRISA-based image copy detection scheme and the "feature points hash" scheme [25].

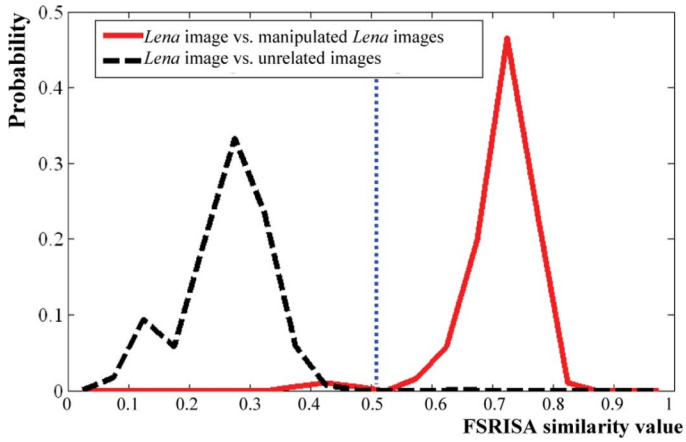

Fig. 6. Probability distributions of the FSRISA similarity values between the *Lena* image and its 204 manipulated versions and those between the *Lena* image and the 204 manipulated versions of each of the other nine test images.

near-duplicate detection in [9], [10], [24], and [25]. We treat each of the ten images as a query image and its 204 manipulated versions as the test images. The parameter settings for applying FSRISA to each query image and each test image are the same as those settings for the reference image and test image, respectively, used in Section IV-A.

To evaluate the true positive rate (TPR), the proposed scheme was conducted between each original image and its 204 manipulated versions. To evaluate the false positive rate (FPR), the proposed scheme was conducted for each original image and the 204 manipulated versions of each of the other nine images. The receiver operating characteristic (ROC) curves (TPR-FPR curve) obtained from the proposed scheme by adjusting the threshold λ, and the "feature points hash" scheme [25] with public source code available for the ten images are shown in Fig. 5. It can be observed from Fig. 5 that the performance of the proposed FSRISA-based scheme can significantly outperform that of the "feature points hash" scheme.

On the other hand, we also give an example for demonstrating the discrimination of our FSRISA. In Fig. 6, the probability distributions of the FSRISA similarity values between the *Lena* image and its 204 manipulated versions, and those between the *Lena* image and the 204 manipulated versions of each of the other nine images, are displayed. It can be observed from Fig. 6 that our FSRISA can indeed discriminate between related and unrelated images.

It is particularly worth considering the flipping manipulation. Standard SIFT keypoint matching [3] can only match a few keypoints due to the orientations of the corresponding feature vectors being switched to the opposite (e.g., only six pairs of keypoint can be correctly matched for the *Lena* image). Nevertheless, even if the proposed FSRISA is based on SIFT, FSRISA can usually provide a reasonable similarity score between an image and its flipped version, as examples shown in Figs. 3 and 4. That is, each feature vector of a flipped image and the corresponding vector of its original (unflipped) version usually only have different signs (with the same magnitudes). Let us consider a dictionary integrated from the two dictionary features of the two images, respectively. When performing sparse coding for each feature vector of the flipped image with respect to the dictionary, most of the feature vectors of the image can be still well sparsely and linearly represented by the dictionary feature of its original version.

C. Evaluation of FSRISA-Based Image Retrieval

To evaluate the proposed FSRISA-based image retrieval scheme, we construct an image database consisting of the ten test images together with their respective 204 manipulations used in Section IV-B (total 2050 images), and the Corel-1000 image dataset [27] (total 1000 images from ten classes), resulting in a total of 3050 images. The parameter settings for applying our FSRISA to each query image and each database image are the same as those settings for reference image and test image, respectively, used in Section IV-A. In this subsection, we conduct two kinds of experiments, including "copy image retrieval" and "general image retrieval."

In this paper, we just consider the simplest scenario for image retrieval, where the score between a query image and each database image is individually calculated using the proposed FSRISA. Then, we retrieve the top Q images with the largest scores, where Q is the desired number of retrieved images. Here, we neither consider performing any indexing or clustering techniques to re-organize an image database nor integrating multiple features for efficient image retrieval.

The main reasons for constructing this database and conducting such two kinds of experiments to evaluate our scheme can be described as follows. We focus on investigating sparse representation of SIFT features and finding its usefulness in image retrieval application. Hence, the database includes several images and their variations for "copy image retrieval" evaluation (similar to the test dataset collected in [9] and [10]). On the other hand, without integrating multiple features, we just want to test some "pure" query images (without overly complex scenes or with a clear background) to retrieve the images with the same semantic meaning, with different appearance for "general image retrieval" evaluation.

For "copy image retrieval," we use the ten original images as the query images and evaluate the precision rates for retrieving the top 205 database images with the largest scores for each query image, as shown in Table I, where the average precision is 99.01%. Such simulation settings are similar to the ones used for near-duplicate image retrieval performed in [10]. It can be observed from Table I that FSRISA is efficient for retrieving the copy versions of a query image.

TABLE I
PRECISION RATES FOR RETRIEVING THE TOP 205 DATABASE IMAGES WITH THE LARGEST SCORES FOR EACH QUERY IMAGE

Query image	Precision	Query image	Precision
Baboon	98.95%	*Lena*	99.48%
Boat	97.91%	*Monarch*	98.43%
Clock	98.95%	*Pepper*	98.95%
Girl	98.95%	*Splash*	100%
House	99.48%	*Tiffany*	98.95%

TABLE II
AVERAGE PRECISION RATES OF THE "DINOSAURS" QUERY IMAGES FOR RETRIEVING THE TOP 10, 20, AND 40 IMAGES FROM THE COREL-1000 DATASET

Retrieved number of images	10	20	40
Proposed FSRISA-based	**100%**	**100%**	**95%**
Visually significant point features [29]	100%	95%	92%

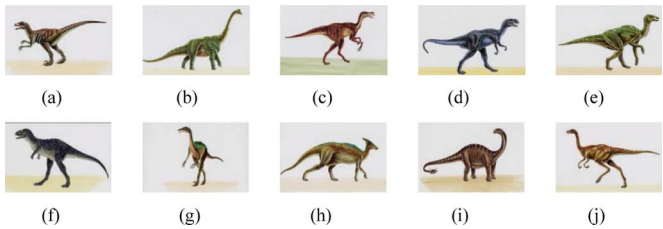

Fig. 7. Retrieved top ten images. (a) Query image and the retrieved 1st image and (b)-(j) the retrieved 2nd through 10th images.

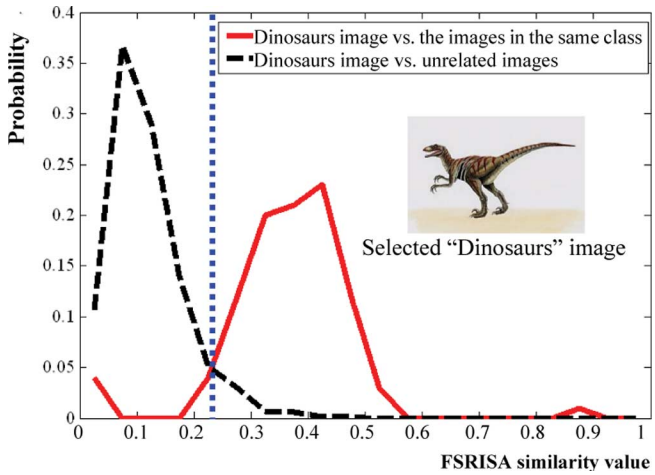

Fig. 8. Probability distributions of the FSRISA similarity values between a selected "Dinosaurs" image and the 100 images in the same class, and those between the image and the other 900 images in Corel-1000 dataset.

For "general image retrieval," we randomly select ten images from the "Dinosaurs" class of the Corel-1000 image dataset and evaluate the average precision for retrieving the top 10, 20, and 40 images, respectively, with the largest scores from the dataset. We compare the results with the best ones reported in [29] (denoted by "visually significant point feature"), as shown in Table II. Some retrieved images for a query image are illustrated in Fig. 7. Moreover, the probability distributions of the FSRISA similarity values between a selected "Dinosaurs" image and the 100 images in the same class, and those between the image and the other 900 images in Corel-1000 dataset, are displayed in Fig. 8. It can be observed from Table II and Figs. 7 and 8, for images without overly complex scenes, FSRISA can be still efficient for retrieving images with similar se-

mantic meanings. It should be noted that the FSRISA values between an image and its related images, shown in Fig. 8, are smaller than those shown in Fig. 6. This is because, in Fig. 8, FSRISA is used to assess the similarities between a query image and the images with the same semantic meaning, but different appearances in the same class, instead of the manipulated versions of the query image.

D. Evaluation of FSRISA-Based Image Recognition

To evaluate the proposed FSRISA-based image recognition scheme, we used the COIL-20 [30] and COIL-100 [31] datasets. We followed the setup for simulations provided in [32], where randomly selected 36 images from each class were used for training samples and the remaining 36 images were used for testing. We repeated the simulations for ten times with different randomly selected training images and averaged the recognition rate obtained from each run. We also evaluated our scheme by using the Caltech-101 dataset consisting of 101 image categories with high shape variability [33]. We followed the common setup to randomly select 5, 15, and 30 training images per category, respectively, and test on the rest images. We repeated the simulations for ten times with different randomly selected training images and averaged the recognition rate obtained from each run. It should be noted that similar to our image retrieval application, using only single feature (SIFT-based feature) may not work very well for recognizing objects with overly complex background. Hence, we only employed the three above-mentioned datasets to investigate sparse representation of SIFT features and find its usefulness in image recognition application.

The parameter settings for applying FSRISA to each query image and each image class are the same as those settings for the reference image and test image, respectively, used in Section IV-A. That is, we set the parameters N_{Ci}, J_{Ci}, L_{Ci}, N_Q, J_Q, and L_Q based on the settings for N_1, J_1, L_1, N_2, J_2, and L_2, respectively, used in Section IV-A, where K_1 and K_2 are replaced by K_{Ci} and K_Q, respectively (K_{Ci} denotes the number of SIFT feature vectors for the class i).

The number of SIFT feature vectors extracted from an image in the COIL-20/100 dataset is usually small, and hence, K_{Ci} is not large. When applying our scheme to the Caltech-101 dataset, we adopted the online dictionary learning algorithm to learn a dictionary [18] and the implementation of OMP provided in [18] to perform sparse coding, which are both highly efficient multicore implementations.

For testing the COIL-20 dataset, the recognition rates obtained using FSRISA, the best ones reported in [34] (denoted by "neighborhood-preserving projections") and [35] (denoted by "invariant moment"), respectively, are shown in Table III for comparison. For testing the COIL-100 dataset, the recognition rates obtained using FSRISA, the best ones reported in [36] (denoted by "distributed sparse representation"), [32] (denoted by "SVM-based"), and [37] (denoted by "bipartite graph matching"), respectively, are shown in Table IV for comparison. For testing the Caltech-101 dataset, the recognition rates obtained using FSRISA, the results reported in [38] (denoted by "SVM-KNN"), [39] (denoted by NBNN), and [7] (denoted by ScSPM), are shown in Table V for comparison. It can be observed from Tables III–V that the performances obtained using the proposed FSRISA can outperform or be comparable to those of the schemes used for comparisons.

TABLE III
RECOGNITION RATES FOR EVALUATING COIL-20 DATASET

Evaluated schemes	Recognition Rate
Proposed FSRISA-based	**98.50%**
Neighborhood-preserving projections [34]	98.13%
Invariant moment [35]	50.20%

TABLE IV
RECOGNITION RATES FOR EVALUATING COIL-100 DATASET

Evaluated schemes	Recognition Rate
Proposed FSRISA-based	**95.38%**
Distributed sparse representation [36]	95.00%
SVM-based [32]	96.45%
Bipartite graph matching [37]	93.30%

TABLE V
RECOGNITION RATES FOR EVALUATING CALTECH-101 DATASET

Schemes	5 training images	15 training images	30 training images
Proposed FSRISA-based	**58.09%**	**69.06%**	**74.51%**
ScSPM [7]	-	67.00%	73.20%
NBNN [39]	50.00%	65.00%	70.40%
SVM-KNN [38]	45.10%	59.10%	66.20%

E. Experimental Comparisons Between Sparse Coding-Based and Traditional Approaches

In this subsection, we experimentally demonstrate the advantage from sparse coding techniques by evaluating the following two kinds of comparisons. First, to evaluate the impact of sparse coding-based feature representation strategy, we compare two kinds of approaches denoted by: 1) "BoW-Traditional:" BoW-based feature $\mathrm{representation} + \mathrm{traditional\ matching}$; and 2) "Sparse-Traditional:" sparse coding-based feature $\mathrm{representation} + \mathrm{traditional\ matching}$, as follows. A classical example of "BoW-Traditional" approach realized by using BoW and SVM can be found in [6], where the best recognition rate (53.90%) for the Caltech-101 dataset was reported. A good example of "Sparse-Traditional" approach realized by using sparse coding and SVM can be found in [7], where the reported best recognition rate (73.20%) for the same dataset can significantly outperform the one reported in [6].

Second, to evaluate the impact of sparse coding-based matching strategy, we compare two kinds of approaches denoted by: 1) "Sparse-Traditional;" and 2) "Proposed:" sparse coding-based feature $\mathrm{representation} + \mathrm{sparse}$ coding-based matching. Based on Table V, "Proposed" approach can slightly outperform "Sparse-Traditional" approach (denoted by "ScSPM" [7]). Another example for "Sparse-Traditional" also realized by using sparse coding and SVM for testing COIL-100 dataset can be found in [36]. Based on Table IV, "Proposed" approach can slightly outperform "Sparse-Traditional" approach (denoted by "Distributed sparse representation" [36]). Even if the improvement of the proposed scheme seems to be limited, a unique characteristic of our scheme is to define a similarity assessment metric, which can be widely applicable in several multimedia applications.

V. Conclusions

In this paper, we have proposed a scheme of FSRISA. The core is to propose a feature-based image similarity assessment technique by exploring the two aspects of a feature detector in terms of representation and matching in our FSRISA framework. Then, we properly formulate the image copy detection, retrieval, and recognition problems as sparse representation problems and solve them based on our FSRISA. The future works may focus on reducing the computational complexities for dictionary feature extraction and image matching by performing sparse coding, which can be further reduced via novel techniques, such as the online dictionary learning algorithm [18], efficient greedy algorithm [18], or fast l_1-minimization algorithm [40]. On the other hand, for FSRISA-based image retrieval applications, further indexing techniques should be also studied. The proposed FSRISA-based image copy detection scheme may be extended to video copy detection by learning the "dictionary feature" for each video sequence/clip. Moreover, incorporated with our secure SIFT techniques [41], [42], the three FSRISA-based applications can be performed in the encrypted domain and are suitable for privacy-preserving applications.

References

[1] Z. Wang and A. C. Bovik, "Mean squared error: Love it or leave it?—A new look at signal fidelity measures," *IEEE Signal Process. Mag.*, vol. 26, no. 1, pp. 98–117, Jan. 2009.

[2] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 430–444, Feb. 2006.

[3] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[4] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *Proc. IEEE Int. Conf. Computer Vision*, Nice, France, Oct. 2003, vol. 2, pp. 1470–1477.

[5] D. Nistér and H. Stewénius, "Scalable recognition with a vocabulary tree," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2006, pp. 2161–2168.

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

[7] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Jun. 2009.

[8] L. W. Kang, C. Y. Hsu, H. W. Chen, and C. S. Lu, "Secure SIFT-based sparse representation for image copy detection and recognition," in *Proc. IEEE Int. Conf. Multimedia and Expo*, Singapore, Jul. 2010.

[9] Z. Xu, H. Ling, F. Zou, Z. Lu, and P. Li, "A novel image copy detection scheme based on the local multi-resolution histogram descriptor," *Multimedia Tools Appl.*, Jan. 2010.

[10] Y. Ke, R. Sukthankar, and L. Huston, "Efficient near-duplicate detection and sub-image retrieval," in *Proc. ACM Multimedia*, 2004.

[11] W. L. Zhao and C. W. Ngo, "Scale-rotation invariant pattern entropy for keypoint-based near-duplicate detection," *IEEE Trans. Image Process.*, vol. 18, no. 2, pp. 412–423, Feb. 2009.

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

[13] V. Chandrasekhar, M. Makar, G. Takacs, D. M. Chen, S. S. Tsai, R. Grzeszczuk, and B. Girod, "Survey of SIFT compression schemes," in *Proc. Int. Workshop Mobile Multimedia Processing*, Turkey, 2010.

[14] O. Pele and M. Werman, "A linear time histogram metric for improved SIFT matching," in *Proc. Eur. Conf. Computer Vision*, 2008, vol. 5304, pp. 495–508.

[15] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T. Huang, and S. Yan, "Sparse representation for computer vision and pattern recognition," *Proc. IEEE*, vol. 98, no. 6, pp. 1031–1044, Jun. 2010.

[16] M. Aharon, M. Elad, and A. M. Bruckstein, "The K-SVD: An algorithm for designing of overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, Nov. 2006.

[17] S. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. Signal Process.*, vol. 41, no. 12, pp. 3397–3415, Dec. 1993.

[18] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding," *J. Mach. Learn. Res.*, vol. 11, pp. 19–60, 2010.

[19] E. Candes and M. Wakin, "An introduction to compressive sampling," *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 21–30, Mar. 2008.

[20] S. J. Wright, R. D. Nowak, and M. A. T. Figueiredo, "Sparse reconstruction by separable approximation," *IEEE Trans. Signal Process.*, vol. 57, no. 7, pp. 2479–2493, Jul. 2009.

[21] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, Feb. 2009.

[22] J. M. Fadili, J. L. Starck, J. Bobin, and Y. Moudden, "Image decomposition and separation using sparse representations: An overview," *Proc. IEEE*, vol. 98, no. 6, pp. 983–994, Jun. 2010.

[23] R. Rubinstein, M. Zibulevsky, and M. Elad, Efficient Implementation of the K-SVD Algorithm Using Batch Orthogonal Matching Pursuit, Technion—Israel Institute of Technology, 2008, CS Tech. Rep.

[24] C. Kim, "Content-based image copy detection," *Signal Process.: Image Commun.*, vol. 18, pp. 169–184, 2003.

[25] V. Monga and B. L. Evans, "Perceptual image hashing via feature points: Performance evaluation and tradeoffs," *IEEE Trans. Image Process.*, vol. 15, no. 11, pp. 3453–3466, Nov. 2006.

[26] F. A. P. Petitcolas, "Watermarking schemes evaluation," *IEEE Signal Process. Mag.*, vol. 17, no. 5, pp. 58–64, Sep. 2000.

[27] J. Z. Wang, J. Li, and G. Wiederhold, "SIMPLIcity: Semantics-sensitive integrated matching for picture libraries," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 9, pp. 947–963, Sep. 2001.

[28] Y. Liu, F. Wu, Z. Zhang, Y. Zhuang, and S. Yan, "Sparse representation using nonnegative curds and whey," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, San Francisco, CA, Jun. 2010, pp. 3578–3585.

[29] M. Banerjee, M. K. Kundu, and P. Maji, "Content-based image retrieval using visually significant point features," *Fuzzy Sets Syst.*, vol. 160, no. 23, pp. 3323–3341, Dec. 2009.

[30] S. A. Nene, S. K. Nayar, and H. Murase, Columbia Object Image Library (COIL-20), 1996, Tech. Rep. CUCS-005-96.

[31] S. A. Nene, S. K. Nayar, and H. Murase, Columbia Object Image Library (COIL-100), 1996, Tech. Rep. CUCS-006-96.

[32] H. Cevikalp, "New clustering algorithms for the support vector machine based hierarchical classification," *Pattern Recognit. Lett.*, vol. 31, pp. 1285–1291, 2010.

[33] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition Workshop on Generative-Model Based Vision*, 2004.

[34] T. Zhang, K. Huang, X. Li, J. Yang, and D. Tao, "Discriminative orthogonal neighborhood-preserving projections for classification," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 40, no. 1, pp. 253–263, Feb. 2010.

[35] G. A. Papakostasa, E. G. Karakasisb, and D. E. Koulouriotis, "Novel moment invariants for improved classification performance in computer vision applications," *Pattern Recognit.*, vol. 43, pp. 58–68, 2010.

[36] A. Y. Yang, M. Gastpar, R. Bajcsy, and S. S. Sastry, "Distributed sensor perception via sparse representation," *Proc. IEEE*, vol. 98, no. 6, pp. 1077–1088, Jun. 2010.

[37] K. Riesen and H. Bunke, "Approximate graph edit distance computation by means of bipartite graph matching," *Image Vision Comput.*, vol. 27, pp. 950–959, 2009.

[38] H. Zhang, A. Berg, M. Maire, and J. Malik, "SVM-KNN: Discriminative nearest neighbor classification for visual category recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2006.

[39] O. Boiman, E. Shechtman, and M. Irani, "In defense of nearest-neighbor based image classification," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2008.

[40] A. Yang, A. Ganesh, Y. Ma, and S. Sastry, "Fast l_1-minimization algorithms and an application in robust face recognition: A review," in *Proc. IEEE Int. Conf. Image Process.*, Hong Kong, Sep. 2010.

[41] C. Y. Hsu, C. S. Lu, and S. C. Pei, "Secure and robust SIFT," in *Proc. ACM Int. Conf. Multimedia*, Beijing, China, 2009, pp. 637–640.

[42] C. Y. Hsu, C. S. Lu, and S. C. Pei, "Homomorphic encryption-based secure SIFT for privacy-preserving feature extraction," in *Proc. IS&T/SPIE Media Watermarking, Forensics, and Security*, Jan. 2011.

Li-Wei Kang (S'05–M'06) has been with the Institute of Information Science, Academia Sinica, Taipei, Taiwan, as an Assistant Research Scholar since August 2010. His research interests include multimedia content analysis and multimedia communications.

Dr. Kang served as an Editorial Advisory Board Member for the book *Visual Information Processing in Wireless Sensor Networks: Technology, Trends and Applications* (Hershey, PA: IGI Global, 2011), a Guest Editor of the Special Issue on Advance in Multimedia, *Journal of Computers*, Taiwan, 2010, a Co-organizer, Special Session on Advanced Techniques for Content-Based Image/Video Resizing, 2011 *Visual Communication and Image Processing* (VCIP2011), Special Session on Image/Video Processing and Analysis, 2011 *APSIPA Annual Summit and Conference* (APSIPA2011), and a reviewer/TPC member for several international conferences and journals. He won the two paper awards presented by 2006 and 2007 Computer Vision, Graphics, and Image Processing Conferences, Taiwan, respectively.

Chao-Yung Hsu is currently pursuing the Ph.D. degree in the Graduate Institute of Communication Engineering of National Taiwan University, Taipei, Taiwan.

He has been a research assistant with the Institute of Information Science, Academia Sinica, Taipei, Taiwan, since 2003. His research interests include multimedia signal processing, data hiding, and digital halftoning.

Hung-Wei Chen received the B.S. degree from National Taipei University of Technology, Taipei, Taiwan, in 2006 and the M.S. degree from National Dong-Hwa University, Hualien, Taiwan, in 2008. He is currently pursuing the Ph.D. degree in the Graduate Institute of Communication Engineering of National Taiwan University, Taipei, Taiwan.

He has been a research assistant with the Institute of Information Science, Academia Sinica, Taipei, Taiwan, since 2008. His research interests include image/video compression, multimedia signal processing, and compressive sensing.

Chun-Shien Lu (M'99) received the Ph.D. degree in electrical engineering from National Cheng-Kung University, Tainan, Taiwan, in 1998.

From October 1998 to July 2002, he joined the Institute of Information Science, Academia Sinica, Taipei, Taiwan, as a postdoctoral fellow for his military service. From August 2002 to July 2006, he was an assistant research fellow at the same institute. Since July 2006, he has been an associate research fellow. His current research interests mainly focus on various topics (including signal processing and security) of multimedia, sensor network security, and compressive sensing.

Dr. Lu organized a special session on Multimedia Security in the 2nd and 3rd IEEE Pacific-Rim Conference on Multimedia, respectively (2001–2002). He co-organized two special sessions (in the area of media identification and DRM) in the 5th IEEE International Conference on Multimedia and Expo (ICME), 2004. He was a guest co-editor of *EURASIP Journal on Applied Signal Processing*, special issue on Visual Sensor Network in 2005. He has owned two U.S. patents, three ROC patents, and one Canadian patent in digital watermarking. He won the Ta-You Wu Memorial Award, National Science Council in 2007 and was a co-recipient of a National Invention and Creation Award in 2004. Since July 2007, he has served as a member of the Multimedia Systems and Applications Technical Committee of the IEEE Circuits and Systems Society. He is currently an associate editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING. He is a member of ACM.

Chih-Yang Lin (M'10) received the Ph.D. degree in computer science and information engineering from National Chung-Cheng University, Chiayi, Taiwan, 2006.

After graduating, he served in the Advanced Technology Center of the Industrial Technology Research Institute of Taiwan (ITRI) from 2007 to 2009. Then, he joined the Institute of Information Science (IIS), Academia Sinica, Taipei, Taiwan, as a postdoctoral fellow. Currently, he is an Assistant Professor in the Department of Computer Science and Information Engineering, Asia University, Taichung, Taiwan. His research interests include computer vision, digital rights management, image processing, and data mining.

Soo-Chang Pei (SM'89–F'00) was born in Soo-Auo, Taiwan, in 1949. He received the B.S.E.E. degree from the National Taiwan University, Taipei, Taiwan, in 1970 and the M.S.E.E. and Ph.D. degrees from the University of California, Santa Barbara, in 1972 and 1975, respectively.

Since 1984, he has been a Professor with the Department of Electrical Engineering, National Taiwan University, Taipei, Taiwan. He was an Engineering Officer with the Chinese Navy Shipyard from 1970 to 1971. From 1971 to 1975, he was a Research Assistant with the University of California, Santa Barbara. He was a Professor and the Chairman of the Electrical Engineering Department, Tatung Institute of Technology, Taipei, Taiwan, from 1981 to 1983, and was with the National Taiwan University as the Chairman of the Electrical Engineering Department from 1995 to 1998, and the Dean of the College of Electrical Engineering and Computer Science from 2003 to 2009, respectively. His research interests include digital signal processing, image processing, optical information processing, and laser holography.

Dr. Pei was the President of the Chinese Image Processing and Pattern Recognition Society in Taiwan from 1996 to 1998 and is a member of Eta Kappa Nu and the Optical Society of America. He became an IEEE Fellow in 2000 for his contributions to the development of digital eigenfilter design, color image coding, and signal compression and to the electrical engineering education in Taiwan. He was a recipient of a National Sun Yet-Sen Academic Achievement Award in Engineering in 1984, the Distinguished Research Award from the National Science Council from 1990 to 1998, an Outstanding Electrical Engineering Professor Award from the Chinese Institute of Electrical Engineering in 1998, the Academic Achievement Award in Engineering from the Ministry of Education in 1998, the Pan Wen-Yuan Distinguished Research Award in 2002, and the National Chair Professor Awards from the Ministry of Education in 2002 and 2008, respectively.