SECURE SIFT-BASED SPARSE REPRESENTATION FOR IMAGE COPY DETECTION AND RECOGNITION

Li-Wei Kang, Chao-Yung Hsu, Hung-Wei Chen, and Chun-Shien Lu*

Institute of Information Science, Academia Sinica, Taipei, Taiwan, ROC
Email: {lwkang, cyhsu, hungwei, lcs}@iis.sinica.edu.tw

ABSTRACT

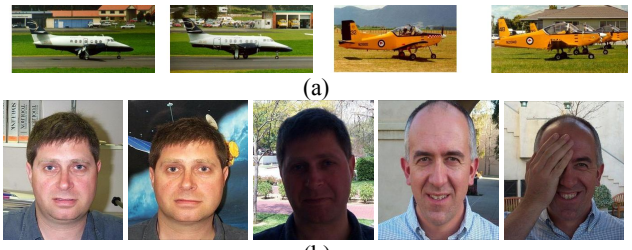
In this paper, we formulate the problems of image copy detection and image recognition in terms of sparse representation. To achieve robustness, security, and efficient storage of image features, we propose to extract compact local feature descriptors via constructing the basis of the SIFT-based feature vectors extracted from the secure SIFT domain of an image. Image copy detection can be efficiently accomplished based on the sparse representations and reconstruction errors of the features extracted from an image possibly manipulated by signal processing or geometric attacks. For image recognition, we show that the features of a query image can be represented as sparse linear combinations of the features extracted from the training images belonging to the same cluster. Hence, image recognition can also be cast as a sparse representation problem. Then, we formulate our sparse representation problem as an l_1-minimization problem. Promising results regarding image copy detection and recognition have been verified, respectively, through the simulations conducted on several content-preserving attacks defined in the Stirmark benchmark and Caltech-101 dataset.

Keywords—Sparse representation, secure SIFT, copy detection, image recognition, compressive sensing.

1. INTRODUCTION

With the increasing availability of digital multimedia data, the integrity verification of image data becomes more and more important [1]-[3]. Digital images distributed through the Internet may suffer from several possible manipulations [4], as illustrative examples shown in Fig. 1. To ensure trustworthiness, image copy detection techniques have emerged to search duplicates and forgeries. Traditionally, image copy detection can be achieved via image hashing [1]-[2] or watermarking [3] techniques. Nevertheless, current hashing techniques may be not very robust to some image manipulations while watermarking techniques will suffer from some distortions induced by data embedding. Recently, SIFT (scale invariant feature transform) [5] has been shown to be invariant to several image variabilities, and efficient to image copy detection [6]. Another conceptually similar problem is image recognition, which is one of the core problems in computer vision and has been extensively investigated [7]-[12]. To correctly recognize images with appearance variabilities induced by background clutter, different viewpoints, orientations, scales, lighting conditions, deformations, or to classify visually different images with the same semantic meaning into the same cluster are challenging, as illustrative examples shown in Fig. 2. Current approaches are usually based on building indices for feature descriptors (usually based on SIFT features), extracted from local image regions. Then, the descriptors are quantized into visual words defined in a pre-constructed vocabulary. Finally, image matching can be achieved via text retrieval technique [7]. Based on the similar idea, a more efficient architecture, called vocabulary tree, was also proposed [8]. Based on quantized feature descriptors, support vector machine (SVM) and/or nearest-neighbor (NN) techniques are also wildly employed for image recognition [9]-[12].

Fig. 1. Some examples of image manipulations defined in Stirmark benchmark [4].



(a)

(b)

Fig. 2. Some examples shown in Caltech-101 [9]: the selected images in the (a) "Airplanes", and (b) "Faces_easy" clusters.

In this paper, we study a secure SIFT sparse representation technology, which can be broadly employed in many applications. Here, we will examine its usefulness in image copy detection and image recognition.

1.1. Representation of SIFT Feature

To extract SIFT features from an image [5], 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 keypoint descriptor representation [5], each descriptor is a 128-D feature vector.

To make SIFT feature more compact, the bag-of-words (BoW) representation approach quantizes SIFT descriptors to a collection of visual words based on a pre-defined visual vocabulary [7] or

*Corresponding author: lcs@iis.sinica.edu.tw

+This work was supported in part by National Science Council, Taiwan, ROC, under Grants NSC 97-2628-E-001-011-MY3 and NSC98-2631-H-001-013.

vocabulary tree [8]. Advanced compression for SIFT features are also investigated recently in [13].

1.2. Matching of SIFT Features

To evaluate the similarity between two images based on their SIFT features, the most straightforward way is to perform keypoint matching [5]. For each feature vector of an image, the distances between the vector and all feature vectors of another image used for comparison will be calculated. For the BoW-based approach, the similarity between SIFT features can be measured via matching their corresponding visual words via text retrieval [7] or tree search [8]. Typically, the BoW-based approach achieves better search efficiency, but less recognition performance suffered from quantization loss.

1.3. Security of SIFT Feature

To the best of our knowledge, the security of SIFT, usually ignored in the literature, was first addressed in [6], where it has been shown that the SIFT keypoints of an image can be successfully removed or inserted while preserving acceptable visual quality. Nevertheless, all SIFT-based multimedia systems may suffer from such kind of attack. In [6], a secure-SIFT technique is proposed in that a secret key-based transformation is applied to an image before performing SIFT extraction. Then, the SIFT features of an image will be extracted from the secure-transformed domain, instead of the original spatial domain. Even if an image has been successfully attacked via removing or inserting keypoints, the SIFT features extracted from its secure domain are still valid. In addition, the quality of a transformed image is visually meaningless, which is also suitable for secure/privacy-preserving image retrieval [14].

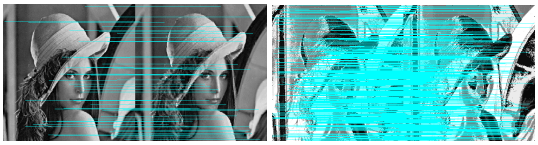
1.4. Overview of our Proposed Scheme

In this paper, a novel secure SIFT-based sparse representation scheme is proposed and used to formulate the image copy detection and image recognition problems. To simultaneously consider the compact representation and security of SIFT, we propose to transform an image to its secure-SIFT domain and extract the SIFT features from the transformed image. Then, we construct the basis consisting of the prototype SIFT atoms to form the final feature (called basis feature in this paper) of the image. To measure the similarity between two images based on basis feature matching, we propose to formulate the problem based on the sparse representations and reconstruction errors of the image basis features. Our major contribution is that robust, secure, and compact features can be simultaneously satisfied.

2. SECURE SIFT-BASED IMAGE FEATURE EXTRACTION AND MATCHING

2.1. Secure SIFT-based Feature Extraction

For an image I, we first apply the secret key-based transformation proposed in [6] on I to get the transformed image I', followed by applying SIFT [5] to I' to get its local descriptors (or feature vectors). The transformation consists of two steps: bit reversing and local encryption. The bit reversing step is to make standard SIFT keypoint detector fail while the local encryption step aims to secure SIFT keypoint detection. In the original image domain, attackers are hard to detect and remove keypoints because the keypoints have been changed, and can be re-generated and detected in the secure domain with the secret key. Please refer to [6] for more details about secure SIFT. An illustrative example for applying the secret key-based transformation to the *Lena* image is shown in Figure 3.



(a) (b)

Fig. 3. An illustrative example for secure SIFT: (a) a failed SIFT keypoint matching (58 matches) between the original *Lena* image (left) and its keypoint-removed version (PSNR = 34.72dB) (right); and (b) a successful SIFT keypoints matching (782 matches) between the secure-SIFT transformed image and its attacked version.

Consider K secure SIFT feature vectors with length M (M = 128 [5]) extracted from an image I, the number K of feature vectors (or the number of keypoints) usually ranges from hundreds to thousands. To make the SIFT feature more compact, we propose to construct the basis consisting of the prototype SIFT atoms to form the (compressed) basis feature of the image. A good basis can provide a sparse representation for each feature vector which can be sparsely represented as a linear combination of the atoms in the basis.

Here, we apply the K-SVD algorithm proposed in [15] to construct the basis for a set of feature vectors of an image to form its basis feature. Given a set of K training feature vectors, $\mathrm{y}_i \in \mathrm{R}^{M\times 1}$, i = 1, 2, …, K, the K-SVD algorithm seeks the dictionary (or basis) leading to the best possible representation for each vector in this set with strict sparsity constraints.

Given a set of K secure SIFT feature vectors extracted from an image I, we apply the K-SVD algorithm to find the basis D of size $M\times N$, $N << K$, to be the basis feature for the image I. The basis D is an over-complete dictionary matrix, where $\mathrm{D} = \{[\mathrm{d}_n]_{M\times 1}\}_{n=1,2,\ldots,N} \in \mathrm{R}^{M\times N}$, $M < N << K$, contains N prototype feature vector atoms as the column vectors in D. Each original feature vector $\mathrm{y}_i \in \mathrm{R}^{M\times 1}$, i = 1, 2, …, K, can be sparsely represented as a linear combination of the atoms defined in D, satisfying $\|\mathrm{y}_i - \mathrm{D}\mathrm{x}_i\|_2 \le \varepsilon$, where $\mathrm{x}_i \in \mathrm{R}^N$ is the sparse representation coefficients of y_i and $\varepsilon \ge 0$ is an error tolerance. After obtaining the basis feature for each image, we will formulate the image similarity measurement based on basis feature matching as a sparse representation problem, described in Sec. 2.2.

2.2. Sparse Representation-based Basis Feature Matching

Sparse signal representation techniques have been extensively studied in the computer vision and pattern recognition community [12], [15]-[18]. In [17], it is argued that in many problems of interest, the sparsest linear representation of a signal in terms of a dictionary actually exists, and the sparse representation can be efficiently solved via l_1-minimization.

In this subsection, we formulate the similarity measurement between two images based on their basis feature matching as a sparse representation problem. Assume the two secure SIFT column feature vectors of length M, y_{1i}, i = 1, 2, .., K_1, and y_{2j}, j = 1, 2, .., K_2, are extracted, respectively, from the two images, I_1 and I_2,

whose basis features are D_1 and D_2, respectively. Hence, $y_{1i} = D_1 x_{1i}$ and $y_{2j} = D_2 x_{2j}$, where D_1 and D_2 are of size $M \times N_1$ and $M \times N_2$, respectively, and x_{1i} and x_{2j} are two sparse coefficient column vectors with length N_1 and N_2, respectively. The idea behind our approach is that if y_{1i} and y_{2j} can be matched, y_{1i} can be also represented as the linear combination of the atoms in D_2. It is also valid for y_{2j}.

To measure the similarity between I_1 and I_2 exploiting the discriminative characteristic of sparse representation, we want to quantify how much information presented in I_1 can be extracted from I_2. A sparse representation problem for representing each feature vector y_{2j} of I_2 with respect to the dictionary $D = [D_1|D_2]$ can be defined as:

$$\hat{x}_j = \min_{x_j} \left\| x_j \right\|_0 \text{ subject to } \| y_{2j} - Dx_j \|_2 \le \varepsilon, \quad (1)$$

where $\|x_j\|_0$ denotes the l_0-norm of x_j, counting the number of nonzero entries in x_j with length (N_1+N_2), which is the sparse coefficient vector corresponding to y_{2j} with length M of I_2. $D = [D_1|D_2]$ of size $M \times (N_1+N_2)$, *i.e.*, concatenation of D_1 and D_2, and $\varepsilon \ge 0$ is an error tolerance.

To solve the sparsest solution for x_j in Eq. (1) based on recent development in the emerging theory of sparse representation and compressive sensing, the l_0-minimization problem can be equivalent to the l_1-minimization problem as:

$$\hat{x}_j = \min_{x_j} \left\| x_j \right\|_1 \text{ subject to } \|y_{2j} - Dx_j\|_2 \le \varepsilon. \quad (2)$$

To solve Eq. (2), the problem can be formulated as a convex unconstrained optimization problem which can be solved via the "sparse reconstruction by separable approximation (SpaRSA)" algorithm [19], which has been shown to be very efficient. It is expected that the positions of nonzero coefficients in $\hat{x}_j$ (or the selected atoms from D) should be highly concentrated in only one sub-dictionary (*e.g.*, D_1 or D_2), and the rest coefficients in x_i should be zeros or small enough.

Based on the obtained solution $\hat{x}_j$, we can calculate the reconstruction error as $\|y_{2j} - Dx_j\|_2$. By letting the elements in $\hat{x}_j$, corresponding to D_2 to zeros, we can get the reconstruction error E_{1j} using only the elements in D_1 for reconstructing y_{2j}. On the other hand, by letting the elements in $\hat{x}_j$, corresponding to D_1 to zeros, we can get the reconstruction error E_{2j} using only the elements in 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, y_{2j} is more suitable to be represented by D_2 (the original basis itself) than D_1, and D_2 will get a vote. Considering all SIFT feature vectors of I_2, y_{2j}, $j = 1, 2, \ldots, K_2$, the obtained votes of D_1 and D_2 are denoted by V_1 and V_2, respectively. Based on voting strategy, we define the similarity between I_1 and I_2 as

$$\text{Sim}(I_1, I_2) = V_1/V_2. \quad (3)$$

Larger $\text{Sim}(I_1, I_2)$ indicates that more feature atoms from I_1 can well represent the feature vectors extracted from I_2. This implies that much information presented in I_1 can be extracted from I_2. Hence, the larger the $\text{Sim}(I_1, I_2)$ 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 be not always larger than V_1. That is, better (or similar) reconstruction performance for y_{2j} may be achieved by using D_1 as the basis than D_2 due to some of the feature vectors extracted from I_2 can be matched by the feature vectors extracted from I_1.

To achieve this goal, we propose to apply K-SVD algorithm [15] to train D_1 with different parameters from those used for training D_2 to make D_1 be finer than D_2. We set that the number of the atoms in D_1 should be larger than that in D_2, *i.e.*, $N_1 > N_2$. We also set that the number of iterations K-SVD performs for training D_1 should be larger than that for training D_2. When I_1 is visually similar to I_2 and D_1 is finer than D_2, the l_1-minimization for solving Eq. (2) may seek more promising atoms from D_1 than D_2 to reconstruct y_{2j}, resulting in $V_1 > V_2$ and larger $\text{Sim}(I_1, I_2)$. Otherwise, when I_1 is visually different from I_2, most atoms for reconstructing y_{2j} will be selected from D_2, resulting in $V_1 < V_2$ and smaller $\text{Sim}(I_1, I_2)$. This similarity measure metric can be directly applicable to image copy detection described in Sec. 3.

3. SECURE SIFT-BASED SPARSE REPRESENTATION FOR IMAGE COPY DETECTION

A valid image user can perform image copy/duplicate detection to verify whether a received image is a duplicate or not. For the sake of security and privacy, the user can only receive some compact feature/hash extracted from the original image together with a secret key for copy detection purpose. In the proposed scheme, the "secret key" contains the parameters for basis feature extraction, including the key for transforming the image to its secure-SIFT domain, the dictionary size, the number of iterations for K-SVD dictionary training.

Consider an original image I_1 and a possibly manipulated version of I_1, denoted by I_2. After receiving the basis feature D_1 of size $M \times N_1$ of I_1 and the secret key from the image owner, the user/server can extract the secure SIFT feature vectors y_{2j} with length M, $j = 1, 2, .., K_2$, from I_2, and the basis feature D_2 of size $M \times N_2$ of I_2. Based on the received "secret key" from the owner, the user can extract D_2 of size $M \times N_2$ of I_2 to maintain that D_2 is coarser than D_1 by letting $N_2 < N_1$ and the number of training iterations for D_2 is smaller than that for D_1. Then, the dictionary D can be formed as $D = [D_1|D_2]$ of size $M \times (N_1+N_2)$. The user/server can solve Eq. (2) via the l_1-minimization algorithm proposed in [19], followed by calculating the respective reconstruction errors, E_{1j} and E_{2j}, for reconstructing y_{2j} with respect to the solved coefficients $\hat{x}_j$ corresponding to the atoms in D_1 and D_2, respectively. Finally, based on Eq. (3), the similarity between I_1 and I_2 can be calculated as $\text{Sim}(I_1, I_2)$. Given an empirically determined threshold τ, if $\text{Sim}(I_1, I_2) \ge \tau$, I_1 and I_2 can be determined to be relevant. Otherwise, I_1 and I_2 can be determined to be irrelevant.

4. SECURE SIFT-BASED SPARSE REPRESENTATION FOR IMAGE RECOGNITION

In this section, we consider the image recognition problem as follows. Consider a well-classified image dataset, where each cluster includes several images with the objects of the same semantic meaning, but with appearance variabilities. For example, in the Caltech 101 dataset [9], there are 101 image categories, where most categories have about 50 images, as illustrative examples shown in Fig. 2. Given a query image, a user may enquire whether the image or related images with the same semantic meaning have been in the database. If the answer is "yes," the user may want to know which cluster the query image belongs to, retrieve the relevant images from the same cluster, or insert the query image to the same cluster. The related applications include object recognition, image classification, and data

update/insertion in image database.

On the other hand, the conclusions in [17] pose a problem that whether their sparse representation-based face recognition approach can be useful for object detection and recognition. It is claimed that their algorithm should be extended to less constrained conditions (*e.g.*, variations in object pose or misalignment). We believe that our approach exploiting both variability-invariant features and sparse representation has addressed this problem and solved it to a certain extent, as elaborated as follows.

4.1. Sparse Representation-based Image Recognition

Based on the discussions in Sec. 2.2, if two secure SIFT feature vectors, y_{1i} and y_{2j}, respectively, extracted from the two images, I_1 and I_2, can be matched, then y_{1i} can be linearly represented with respect to the atoms in the basis D_2 of I_2. Consider a dataset consisting of several images categorized into S clusters, where each cluster contains T selected training images. Based on the respective basis features of the T training images in a cluster, the basis feature of the cluster can be extracted via training the basis whose atoms can linearly represent any feature vectors of each training image belonging to this cluster. Here, we apply the K-SVD algorithm [15] to generate the basis feature of each cluster in the dataset. The basis features of all the clusters can form the dictionary for the dataset as $D_{dataset} = [d_{1,1}, d_{1,2}, \ldots, d_{1,N}, d_{2,1}, d_{2,2}, \ldots, d_{2,N}, \ldots, d_{s,1}, d_{s,2}, \ldots, d_{s,N}, \ldots, d_{S,1}, d_{S,2}, \ldots, d_{S,N}]$ of size $M\times(S\times N)$, where $d_{i,j}$ denotes the j-th atom in the basis feature (with N atoms) of the i-th cluster, $i = 1, 2, \ldots, S$, $j = 1, 2, \ldots, N$.

Consider a query image I which belongs to, but may be not exactly the same as any training images in the s-th cluster, each secure SIFT feature vector y_i with length M, $i = 1, 2, .., K$, of I can be represented with respect to the atoms in the basis feature of the s-th cluster as:

$$y_i = x_{i,s,1}d_{s,1} + x_{i,s,2}d_{s,2} + \ldots + x_{i,s,N}d_{s,N}, \quad (4)$$

where $x_{i,s,j}$ denotes the sparse coefficient corresponding to the atom $d_{s,j}$, $j = 1, 2, \ldots, N$. Then, Eq. (4) can be further sparsely and linearly represented as:

$$y_i = D_{dataset}x_i, \quad (5)$$

where $x_i = [0, 0, \ldots, 0, x_{i,s,1}, x_{i,s,2}, \ldots, x_{i,s,N}, 0, 0, \ldots, 0]^T$ with length $S\times N$ is the sparse coefficient vector corresponding to y_i. Here x_i can be solved using l_1-minimization formulated as:

$$\hat{x}_i = \min_{x_i} \|x_i\|_1 \text{ subject to } \|y_i - D_{dataset}x_i\|_2 \le \varepsilon, \quad (6)$$

where x_i is the sparse coefficient vector corresponding to y_i, and $\varepsilon \ge 0$ is an error tolerance. Ideally, the entries of $\hat{x}_i$ should be zeros except that those with respect to the cluster that the image I belongs to may have some nonzero coefficients. We solve Eq. (6) via SpaRSA algorithm [19] and calculate the respective reconstruction errors, $E_{1i}, E_{2i}, \ldots$, and E_{Si}, for reconstructing y_i with respect to the solved coefficients in $\hat{x}_i$ corresponding to the atoms in the basis features of the 1-st, 2-nd, …, and S-th clusters, respectively. The one corresponding to the minimum reconstruction error is the cluster that y_i belongs to. Similar to the voting strategy employed in Sec. 2.2, by considering the voting statuses for all y_i, $i = 1, 2, .., K$, it can be recognized that the query image I belongs to the cluster with the most votes, as an illustrative example shown is Fig. 4.

The proposed sparse representation-based image recognition scheme is in spirit similar to the sparse representation-based face recognition scheme presented in [17], where the key is that the distribution of multiple clusters can be modeled as a mixture subspace model with one subspace for each cluster. Nevertheless, the major difference distinguishing between [17] and our scheme is that the former employs the image feature that is not variability-invariant (*e.g.*, down-sampled image pixel data), and hence, image data are assumed to be well-aligned, while our scheme presents secure SIFT image feature that is essentially variability-invariant and privacy preserved, which can be applicable for object recognition in privacy preserving video surveillance applications [20].

It is worth noting that if the feature size of a query image is critically crucial for applications in wireless sensor networks [18], each feature vector y_i of length M can be further compressed via the compressive sensing technique [21] via $z_i = \Phi y_i$, where z_i of length m is the measurement vector of y_i, $m < M$, and Φ is a sampling matrix (whose entries are drawn randomly from a distribution), which should be incoherent with $D_{dataset}$. Then, Eq. (5) can be re-expressed by

$$z_i = \Phi y_i = \Phi D_{dataset}x_i = Ax_i, \quad (7)$$

where $A = \Phi D_{dataset}$, and x_i can be similarly solved via solving l_1-minimization problem.

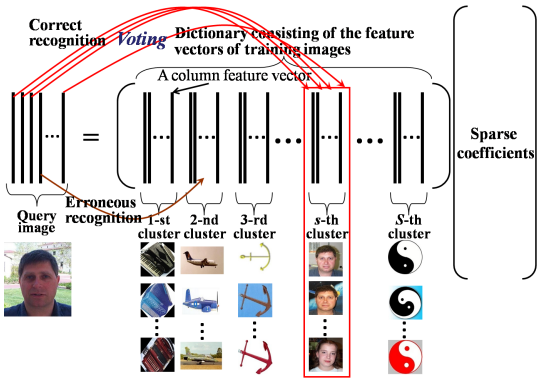


Fig. 4. Illustration of the proposed image recognition scheme.

4.2. Recognition Strategy based on Hierarchical Parallel Structure

In practice, if the size of the dictionary $D_{dataset}$ is too large, the complexity for solving Eq. (6) will be very expensive and the performance may be degraded. Based on the compressive sensing theory [21], it is stated that a f-sparse signal x with length r, *i.e.*, $\|x\|_0 = f$, can be accurately reconstructed by taking only $m = O(f\times\log(r/f))$, where $f < m << r$, linear and non-adaptive measurements from the random projection $z = \Phi x$, where Φ is an $m\times r$ sampling matrix with entries drawn from a standard Gaussian distribution. This theory is for the projection matrix with random Gaussian distribution, different from the employed redundant/over-complete dictionary in our case, which cannot be directly applied. Nevertheless, it can still be observed that the number of measurements (in our case, it is the length M of a secure SIFT feature vector, $M = 128$) required for sparse signal recovery is highly related to the length of the original signal (the number of atoms in $D_{dataset}$, *i.e.*, the length of x_i in our case). If the number of atoms in $D_{dataset}$ is too large, it will induce that the length of the sparse coefficient is too long. To recover such lengthy signal, it is required that each secure SIFT feature vector should be longer. Nevertheless, the feature vector length is fixed to 128 based on [5]. On the other hand, if we reduce the basis feature size for each cluster to keep the size of $D_{dataset}$ to be smaller, we will lose a certain amount of information and sacrifice recognition performance. In the sparse representation-based face recognition

scheme presented in [17], each training image has only one simple feature vector with length 120 due to the considered images are all well-aligned face images without needing complex variability-invariant features. In one of the simulations conducted in [17], in the total 38 clusters, each one consists of 32 vectors (with length 120 for each), resulting in total 1,216 training feature vectors. Hence, the size of training dataset is not an important issue in [17]. Nevertheless, in our simulations (presented in Sec. 5), *e.g.*, in the total 101 clusters, we extract 200 basis feature vectors (with length 128 for each) for each cluster, resulting in total 20,200 training feature vectors. Hence, memory requirement and computation complexity will be important for our scheme to fit practical use.

In view of this, we propose a hierarchical parallel structure for performing our sparse representation-based image recognition. In the first hierarchy, we divide the dictionary $D_{dataset}$ into several sub-dictionaries with equal size, *i.e.*, equal number of clusters. Then, given a query image, we perform image recognition via solving Eq. (6) individually for each sub-dictionary. This process can be performed in a concurrent manner if parallel processors are supported. For each recognition process, we can get a cluster candidate the query image belongs to. For each sub-dictionary, if the votes from the query image almost equally distribute over all clusters in it, this sub-dictionary will be directly rejected due to its indistinguishableness. Otherwise, the candidate cluster obtained from this sub-dictionary will be further considered in the next hierarchy. In the second hierarchy, all the candidate clusters selected from the previous hierarchy will be divided into several sub-dictionaries with equal size. Then, the recognition process will be individually and parallelly performed again for each sub-dictionary. Similarly, all the selected candidate clusters will be kept for the next hierarchy. The recognition process can be hierarchically and parallelly performed until only one candidate cluster is obtained, which can be decided to be the final recognition result.

5. SIMULATION RESULTS

Simulations conducted on publicly available benchmarks for evaluation of image copy detection and recognition are presented.

5.1. Evaluation of Image Copy Detection

To evaluate the proposed image copy detection scheme, eight 512×512 standard test images, *Barbara*, *Baboon*, *Boat*, *F16*, *Goldhill*, *Lena*, *Pepper*, and *Sailboat* images were used. Each image was manipulated by 204 attacks defined in the Stirmark 3.1 and 4.0 benchmarks (*e.g.*, signal processing attacks and geometric attacks) [4].

For each original image I_1, the owner extracts the basis feature D_1 of size $M \times N_1$ of I_1, where $M = 128$ [5] and $N_1 = 200$. In applying K-SVD for training D_1, the number of iterations was set to 20. For each possibly manipulated version I_2 of I_1, the user/server extracts the basis feature D_2 of size $M \times N_2$, where $N_2 = 100$ and the K-SVD iteration number was set to 10. Hence, D_1 is kept to be finer than D_2.

To evaluate the true positive rate (TPR), the proposed scheme was conducted between each image and its 204 manipulated versions. To evaluate the false positive rate (FPR), the proposed scheme was conducted for each image and the 204 manipulated versions of each of the other seven images. The receiver operating characteristic (ROC) curves (TPR-FPR curve) obtained from the proposed scheme by adjusting the threshold τ, and the "feature points hash" scheme [2] for image copy detection for the eight test images are shown in Fig. 5. The "feature points hash" scheme uses an iterative feature detector to extract visually significant feature points which are invariant under perceptually insignificant distortions. It can be observed that the performance of the proposed scheme significantly outperforms the "feature points hash" scheme [2].

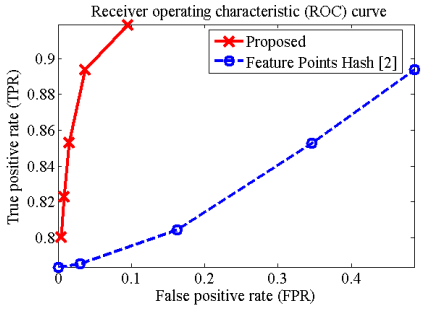


Fig. 5. Comparisons of ROC curves obtained using the proposed scheme and the "feature points hash" scheme [2] for image copy detection.

To consider the flipping manipulation defined in the Stirmark benchmark, standard SIFT keypoint matching can only match few keypoints due to the directions of the original corresponding feature vectors become to be opposite. Nevertheless, by applying our scheme, the feature vectors of one image can be sparsely and linearly represented by the basis feature of its flipped version.

On the other hand, a SIFT removal attack [6] can remove most keypoints from an image while keeping acceptable visual quality, resulting in mismatch between the original and attacked images (as shown in Fig. 3). Nevertheless, by applying our scheme, for an image whose keypoints have been removed, the feature vectors can still be extracted from its secure-SIFT transformed domain.

5.2. Evaluation of Image Recognition

To evaluate the proposed image recognition scheme, we used Caltech-101 dataset [9] consisting of 101 image categories (including animals, vehicles, flowers, and etc.) with high shape variability, where there are 31 to 800 images per category. Most categories have about 50 images. We followed the common setup for testing Caltech-101 dataset. We randomly selected 5, 15 and 30 training images per category, respectively, and tested on the rest images. We repeated the simulations 10 times with different randomly selected training images and averaged the recognition rate obtained from each run.

In the proposed scheme, we extract the basis feature (with length 128 for each atom/vector) with 100 atoms from each training image. For each image cluster, we extract the basis feature with 200, 250, and 300 atoms for the number of training images set to 5, 15, and 30, respectively, via applying KSVD [15] to all the basis feature atoms of the training images belonging to this cluster. That is, the parameters used in Sec. 4.1 were $M = 128$, $N = 200$, 250, or 300, $S = 101$, and $T = 5$, 15, or 30. The number of KSVD training iterations for extracting basis feature from each cluster for 5, 15, and 30 training images were set to 10, 30, and 50, respectively. For simplicity, we divided the dictionary $D_{dataset}$ into 20 sub-dictionaries with the basis feature atoms from 5 clusters for each, and an additional 1 cluster. For each query image, we apply the proposed hierarchical parallel structure to perform the recognition process for each sub-dictionary. Then, we can get 20

candidate clusters which can be further divided into 4 sub-dictionaries of 5 clusters for each. Finally, we perform the recognition process again for the dictionary formed by the 4 candidate clusters and the additional 1 cluster to find the final cluster result.

The recognition rates obtained by selecting 5, 15, and 30 training images per cluster using the proposed recognition scheme, the SVM-KNN (support vector machine-K nearest-neighbor) [10], NBNN (naive-Bayes nearest-neighbor) [11], and ScSPM (linear spatial pyramid matching that uses linear kernel on spatial-pyramid pooling of SIFT sparse codes) [12] schemes are shown in Table 1. It can be observed from Table 1 that the performances obtained from the proposed scheme can outperform or be comparable to those of the schemes used for comparisons.

Table 1. Recognition rates (%) tested on Caltech 101 dataset.

Schemes	5 training images	15 training images	30 training images
Ours	**57.08**	**68.04**	**73.91**
ScSPM [12]	-	67.00	73.20
NBNN [11]	50.00	65.00	70.40
SVM-KNN [10]	45.10	59.10	66.20

The performances of the proposed scheme are mainly restricted by the computational complexity. That is, the basis feature size for each cluster cannot be too large in order to keep the overall dataset dictionary size and the computational complexity for solving l_1-minimization to be reasonable. However, the novelties of our scheme includes: (i) applying sparse representation technique to general object recognition without relying on SVM or NN-based classification techniques; and (ii) for object recognition applications in unreliable environment (*e.g.*, SIFT keypoints may be removed [6]), our scheme can still survive.

6. CONCLUSIONS

In this paper, we have proposed a secure SIFT-based image feature extraction technology and incorporated it with sparse representation techniques for image copy detection and image recognition. In our scheme, the feature size for training image can be substantially reduced, compared to the standard SIFT feature size. The feature size for query image can also be reduced via the compressive sensing technique. By exploiting the discriminative property of sparse representation, image copy detection and image recognition can be efficiently achieved. Moreover, the above applications can be readily conducted in a privacy-preserving manner using secure SIFT [6].

7. REFERENCES

[1] C. S. Lu and C. Y. Hsu, "Geometric distortion-resilient image hashing scheme and its applications on copy detection and authentication," *Multimedia Systems*, vol. 11, no. 2, 2005.

[2] V. Monga and B. L. Evans, "Perceptual image hashing via feature points: performance evaluation and tradeoffs," *IEEE Trans. on Image Processing*, vol. 15, no. 11, pp. 3453–3466, 2006 (source codes available from http://users.ece.utexas.edu/~bevans/projects/hashing/software.html).

[3] V. Licks and R. Jordan, "Geometric attacks on image watermarking systems," *IEEE Multimedia*, vol. 12, 2005.

[4] F. A. P. Petitcolas, "Watermarking schemes evaluation," *IEEE Signal Processing Magazine*, vol. 17, no. 5, 2000.

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

[6] C. Y. Hsu, C. S. Lu, and S. C. Pei, "Secure and robust SIFT," in *Proc. of ACM Int. Conf. on Multimedia*, China, 2009.

[7] J. Sivic and A. Zisserman, "Video Google: a text retrieval approach to object matching in videos," *Proc. of ICCV*, 2003.

[8] D. Nistér and H. Stewénius, "Scalable recognition with a vocabulary tree," in *Proc. of CVPR*, 2006.

[9] 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. of CVPR Workshop on Generative-Model Based Vision*, 2004.

[10] H. Zhang, A. Berg, M. Maire, and J. Malik, "SVM-KNN: discriminative nearest neighbor classification for visual category recognition," in *Proc. of CVPR*, 2006.

[11] O. Boiman, E. Shechtman, and M. Irani, "In defense of nearest-neighbor based image classification," in *Proc. of CVPR,* 2008.

[12] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *Proc. of CVPR*, 2009.

[13] M. Makar, C. L. Chang, D. Chen, S. S. Tsai, and B. Girod, "Compression of image patches for local feature extraction," in *Proc. of ICASSP*, 2009.

[14] W. Lu, A. L. Varna, A. Swaminathan, and M. Wu, "Secure image retrieval through feature protection," in *Proc. of ICASSP*, 2009.

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

[16] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T. Huang, and S. Yan, "Sparse representation for computer vision and pattern recognition," to appear in *Proceedings of the IEEE*.

[17] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, 2009.

[18] A. Y. Yang, M. Gastpar, R. Bajcsy, S. S. Sastry, "Distributed sensor perception via sparse representation," to appear in *Proceedings of the IEEE*.

[19] S. J. Wright, R. D. Nowak, and M. A. T. Figueiredo, "Sparse reconstruction by separable approximation," *IEEE Trans. on Signal Processing*, vol. 57, no. 7, pp. 2479-2493, July 2009.

[20] M. Cossalter, M. Tagliasacchi, and G. Valenzise, "Privacy-enabled object tracking in video sequences using compressive sensing," in *Proc. of IEEE Int. Conf. on Advanced Video and Signal Based Surveillance*, Genova, Italy, Sept. 2009.

[21] E. Candes and M. Wakin, "An introduction to compressive sampling," *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 21-30, March 2008.