International Journal of Future Generation Communication and Networking
Vol.6, No.6 (2013), pp.71-80
http://dx.doi.org/10.14257/ijfgcn.2013.6.6.08

A Similarity Search Scheme over Encrypted Cloud Images based on Secure Transformation

Zhihua Xia[1,2], Yi Zhu[1,2], Xingming Sun[1,2], and Jin Wang[1,2]

[1]Jiangsu Engineering Center of Network Monitoring, Nanjing University of Information Science & Technology, Nanjing, 210044, China
[2]School of Computer & Software, Nanjing University of Information Science & Technology, Nanjing, 210044, China

Abstract

With the growing popularity of cloud computing, more and more users outsource their private data to the cloud. To ensure the security of private data, data owners usually encrypt their private data before outsourcing the data to the cloud server, which brings incommodity of data operating. This paper proposes a scheme for similar search on encrypted images. In the setup phase, image owner extracts feature vectors to represent the images as usual image retrieval system does. Then, the feature vectors are transformed by an invertible matrix, which not only protect the information of feature vector but also support similarity evaluation between the vectors. The encrypted vectors and image identifies are used to construct inverted index, which is finally uploaded along with the encrypted image to the cloud. In the search phase, with a query image, the authorized image user extracts and encrypts feature vector to generate the trapdoor. The trapdoor is submitted to the cloud and can be used to calculate the similarity with the transformed feature vectors. The encryption on features does not degrade the result accuracy. Moreover, the image owner could update the encrypted image database as well as the secure index very easily.

Keywords: *Image retrieval, similarity retrieval, secure transformation approach*

1. Introduction

Due to strong data storage and management ability of the cloud server, more and more data owners will outsource data to the cloud server. In order to protect the security of private data, data owners need to encrypt their data before uploading the data. Unfortunately, data encryption, if not done appropriately, may reduce the effectiveness of data utilization. For example, content-based image retrieval (CBIR) technique has been widely used in the real world; however, the technologies are invalid after the feature vectors are encrypted.

Currently, searchable symmetric encryption has been widely researched. Song *et al.*, proposed the first practical searchable encryption method [1]. After that, in order to enhance the search flexibility and usability, some researchers proposed works to support similar keyword search which could tolerate typing errors [2-4]. On the other hand, some of the works focused on multi-keyword searches which could return more accurate results ranked according to some predefined criterions [5-12]. However, these works are mainly designed for the search on encrypted texts, and could not be utilized directly for the encrypted images. Inspired by the searchable encryption on texts, Lu *et al.*, proposed a search scheme over encrypted multimedia databases [13]. They extracted visual words from images, based on which they could achieve similar search on encrypted images with the methods that are usually employed by the encrypted text search schemes. However, this work is not suitable for other image features except the visual words, and their index makes the search result less accurate.

In this paper, we propose a scheme that not only ensure the security of the images and features but also support similar search on encrypted images. In the proposed scheme, the encryption on features does not degrade the result accuracy. Moreover, the image owner could update the encrypted image database as well as the secure index quite easily.

The rest of this paper is organized as follows. Section 2 presents the problem formulation. Section 3 introduces preliminaries. Section 4 describes the detail of the proposed scheme. Section 5 discusses the security and performance of the proposed scheme. Section 6 concludes the proposed scheme.

2. Problem Formulation

2.1. System Model

A similarity search problem involves a collection of objects (documents, images, etc.) that are characterized by a collection of relevant features and represented as points in a high-dimensional attribute space. Given queries in the form of points in this space, we are required to find the nearest (most similar) object to the query. The designed scheme can not only support similarity search, but also prevent the information leakage of the database. The proposed scheme includes three different entities: image owner, cloud server, and image user.

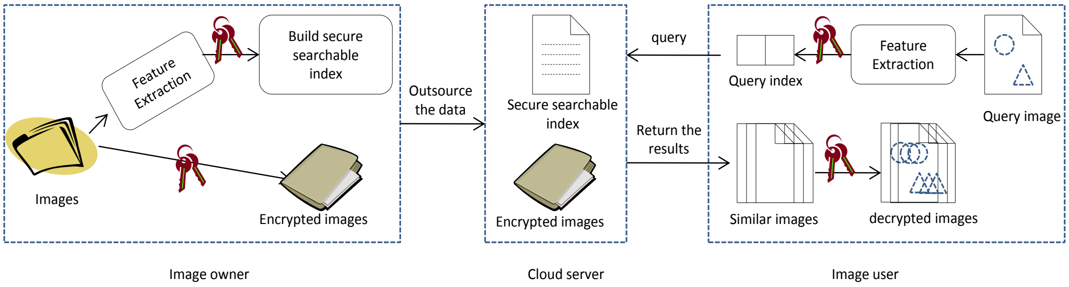


Figure 1. System Model for Secure Image Retrieval

Image owner has a collection of n images $M = \{m_1, m_2, \cdots, m_n\}$ that he wants to outsource to the cloud server in encrypted form. Meanwhile, the image owner wants to keep the capability to search through the images for effective utilization reasons. First, the image owner extracts a feature vector $\mathbf{f} = (f_1, f_2, \cdots, f_l)^T$ from each image as common image retrieval system does. Secondly, the images are encrypted. Thirdly, the image owner builds a secure searchable index I with the set $\{\mathbf{f}_i\}_{i=1}^{n}$. Finally, the encrypted images and the index I are uploaded to the cloud server.

Image user is the authorized ones to use the images. We assume that the authorization between the image owner and image user is appropriately done. In order to query images, the image user extracts the query feature vector $\mathbf{f}_q$ from the query image. Then, the vector $\mathbf{f}_q$ is used to generate a trapdoor $TD(\mathbf{f}_q)$. Finally, the trapdoor $TD(\mathbf{f}_q)$ is submitted to the cloud server for the purpose of searching similar images.

Cloud server stores the encrypted images and the index I for the image owner and processes the query of image users. After receiving a query trapdoor $TD(\mathbf{f}_q)$, cloud server compares the trapdoor $T(\mathbf{f}_q)$ with the items in index I to return *k* most similar images. In the proposed scheme, the cloud server is considered to be 'honest but curious'. The cloud server always tries to learn more information about the data and the user's query. Thus, it is necessary to design a secure similarity search scheme over encrypted images.

2.2. Design Goals

To enable secure and accurate similarity search over encrypted images under the aforementioned model, the proposed scheme tries to achieve the goals as follows.

Security: The scheme must protect the security of sensitive data without leaking information about the image databases M and index I , which is the most important goal in this paper.

Accuracy: The proposed scheme should achieve high retrieval accuracy. The accuracy of the unencrypted image retrieval scheme depends on the feature extraction and similarity evaluation method. Many researchers have done lots of contribution on it. Here, what we particularly concern is that the encryption of the features will not lower the accuracy of the retrieval scheme.

Efficiency: The scheme should reduce the computational complexion and communication spending. In addition, the update of the data should be supported.

3. Preliminaries

3.1. Feature Extraction

Content-based image retrieval (CBIR) has not only received extensive research focus but also been widely adopted by real-world image retrieval system, such as Google image search and Yahoo. CBIR usually involves extraction of features and search on the feature index for similar images. Thus, it is boiled down to two intrinsic challenges. The first challenge is how to mathematically describe an image, which is referred as the feature extraction step. The feature vector can be either globally for the entire image or locally for a small group of pixels, including color [14], texture [15, 16], salient point [17, 18], *etc*. The advantage of global extraction is its high speed for both extracting features and computing similarity. On the other hand, local features based on local invariants such as corner points or interest points, are typically more robust for spatial transformation and usually retrieve more accurate results.

Among the existing feature extraction methods, none of them could be regarded as the best. Without loss of generality, the proposed scheme chooses the histogram features which are the most typical and simplest ones for CBIR. We denote $m(x)$ as the gray value at the location *x* in an image *m*. Then, the histogram features can be formulated as

$$f_i = \frac{\sum \mathbf{1}\{m(x)=i\}}{|m|}, \tag{1}$$

where $\mathbf{1}\{m(x)=i\}$ =1, if $m(x)$ equals to *i*, else $\mathbf{1}\{m(x)=i\}$ =0, $|m|$ is the pixel number of the image. The histogram features constitute a feature vector which is used to represent the image. The similarity between two histogram feature vectors can be evaluated by Euclidean distance, defined as

$$D(\mathbf{f}_i, \mathbf{f}_j) = \left\| \mathbf{f}_i - \mathbf{f}_j \right\|_2 = \sqrt{\sum_{k=1}^{l} \left(\mathbf{f}_{i,k} - \mathbf{f}_{j,k} \right)^2}. \tag{2}$$

3.2. Secure Transformation Approach

Image features in plaintext may reveal information about image content. For example, a color histogram with large values for the blue components would indicate the likely presence of sky or ocean. In order to ensure the security, the feature vectors should be encrypted before outsourced to the cloud server. Here, we introduce a secure transformation approach which is widely used in information security field [19]. It can not only prevent the information leaking of the feature vectors but also support the similar search. First, the feature vector $\mathbf{f} = (f_1, f_2, \cdots, f_l)^T$ is extended as

$$\tilde{\mathbf{f}} = (f_1, \cdots, f_l, \|\mathbf{f}\|_2^2)^T, \tag{3}$$

where $\|\mathbf{f}\|_2^2 = \sum_{i=1}^{l} f_i^2$. Then, the modified feature vector is transformed with an $(l+1) \times (l+1)$ invertible matrix $\mathbf{R}$ as

$$\mathbf{f}^{'} = \mathbf{R}^T \cdot \tilde{\mathbf{f}}, \tag{4}$$

where the matrix $\mathbf{R}$ is kept as the secure key by image owner and the authorized image user. In summary, the secure transform algorithm can be written as

$$\begin{aligned} \mathbf{f}^{'} &= SecureTransfrom(\mathbf{R}, \mathbf{f}) \\ &= \mathbf{R}^T \cdot (f_1, ..., f_l, \|\mathbf{f}\|_2^2)^T. \end{aligned} \tag{5}$$

4. The Proposed Scheme

To achieve secure similar search on images outsourced to the cloud, the image owner needs to construct a secure searchable index and outsource it to the cloud server along with the encrypted images. After that, cloud server could perform similar search on the index according to the query requests submitted by image users. The proposed scheme needs to ensure that the cloud server learns nothing about the query, index, and image databases. In this section, we describe our scheme in detail in two phases.

4.1. The Setup Phase

In the setup phase, image owner needs to build a secure index and encrypt the images. Then, the index and the encrypted images are uploaded to the cloud.

Step1: Key Generation.

The image owner generates the private key k_{img} and $\mathbf{R}$ to encrypt the images and the feature vectors respectively.

Step2: Feature Extraction.

The image owner extracts a feature vector $\mathbf{f} = (f_1, f_2, \cdots, f_l)^T$ from each image in the databases M . In the proposed scheme, the features are the histogram features as it is described in subsection 3.1.

Step3: Secure Index Construction.

After the feature vectors are extracted from the image database M , they are utilized to build secure searchable index I . The image owner transforms each $\mathbf{f}$ with private key $\mathbf{R}$ by using the secure transformation method $SecureTransfrom(\mathbf{R}, \mathbf{f})$ so as to generate the corresponding encrypted feature vector $\mathbf{f}^{'}$. Then, the secure index I is constructed as shown in Table 1, where $ID(m_i)$ is the identifier of file m_i that can uniquely locate the actual file.

Table 1. The Secure Searchable Index I

$\mathbf{f}_1^{'}$	$ID(m_1)$
$\mathbf{f}_2^{'}$	$ID(m_2)$
$\mathbf{f}_3^{'}$	$ID(m_3)$
……	……
$\mathbf{f}_n^{'}$	$ID(m_n)$

Step4: Upload.

After constructing the index I , data owner encrypts all of the images in M with the secure key k_{img} . Then, the encrypted images and the secure searchable index I are uploaded to the cloud.

4.2. Search Phase

In search phase, the image user wants to retrieve images that are similar to a query image from the cloud server. In order to avoid the information leakage, the image user generates a secure trapdoor with the query image. Then, the trapdoor is submitted to the cloud server. Utilizing the trapdoor, the cloud server returns k most similar images by searching on the index I .

Step1: Trapdoor Generation.

In order to query images, the image user extracts the query feature vector $\mathbf{f}_q = (f_{q,1}, ..., f_{q,l})$ from the query image with the feature extraction method introduced in the step 2 of setup phase. Then, the query feature vector $\mathbf{f}_q$ is used to generate a trapdoor $TD(\mathbf{f}_q)$ as following.

First, with the $\mathbf{f}_q$, the image user generates

$$\tilde{\mathbf{f}}_q = (-2 f_{q,1}, ..., -2 f_{q,l}, 1)^T . \tag{6}$$

Then, the trapdoor $TD(\mathbf{f}_q)$ is calculated as

$$TD(\mathbf{f}_q) = r\mathbf{R}^{-1} \cdot \tilde{\mathbf{f}}_q , \tag{7}$$

where r is a positive random real number, and $\mathbf{R}$ is the shared secure key. Finally, the trapdoor $TD(\mathbf{f}_q)$ is submitted to cloud server by the image user.

Step2: Search Index.

After receiving a search request $TD(\mathbf{f}_q)$, the cloud server will search on the secure index I , and return k most similar images to the user. The distance between query vector $\mathbf{f}_q$ and the vector $\mathbf{f}_i^{'}, i=1,...,n,$ can be calculated as follows:

$$\begin{aligned}
Dis(TD(\mathbf{f}_q),\mathbf{f}_i^{'}) &= (TD(\mathbf{f}_q))^T \cdot \mathbf{f}_i^{'} \\
&= (r\mathbf{R}^{-1} \cdot (-2f_{q,1},...,-2f_{q,l},1)^T)^T \cdot (\mathbf{R}^T \cdot (f_{i,1},...,f_{i,l},\|\mathbf{f}_i\|_2^2)^T) \\
&= r(-2f_{q,1},...,-2f_{q,l},1) \cdot (\mathbf{R}^{-1})^T \cdot \mathbf{R}^T \cdot (f_{i,1},...,f_{i,l},\|\mathbf{f}_i\|_2^2)^T \\
&= r(-2f_{q,1},...,-2f_{q,l},1) \cdot (f_{i,1},...,f_{i,l},\|\mathbf{f}_i\|_2^2)^T \\
&= r\left(\|\mathbf{f}_i\|_2^2 - 2\sum_{k=1}^{l} f_{q,k}\, f_{i,k}\right) = r\left(\|\mathbf{f}_i\|_2^2 - 2\sum_{k=1}^{l} f_{q,k}\, f_{i,k} + \|\mathbf{f}_q\|_2^2 - \|\mathbf{f}_q\|_2^2\right) \\
&= r\left(\|\mathbf{f}_q - \mathbf{f}_i\|_2^2 - \|\mathbf{f}_q\|_2^2\right).
\end{aligned} \tag{8}$$

For every query, the r and $\|\mathbf{f}_q\|_2^2$ are the same for every $\mathbf{f}_i^{'}$, and the Euclidean distance between $\mathbf{f}_q$ and $\mathbf{f}_i$ is implied in the $Dis(TD(\mathbf{f}_q),\mathbf{f}_i^{'})$. Therefore, with this distance criterion, the cloud server could return the same k most similar results exactly as it does on unencrypted feature vectors.

Finally, the cloud server returns k most similar results with minimum distance to the query vector to image user, who could decrypt the images with the shared key k_{img} .

5. Security and Performance

5.1. Security Analysis

(1) *Confidentiality of the data*: In the proposed scheme, the image database, index, and query are encrypted. The cloud server can not access the original images and feature vectors without the secret key k_{img} and $\mathbf{R}$.

(2) *Query unlinkability*: By introducing the random value r in trapdoor generation, the same query requests will generate different trapdoors. Thus, query unlinkability is better protected. However, the proposed scheme does not randomize the query results. Thus, the same query would be found by analyzing the retrieved results from queries. Intuitively, the queries with the same results are likely to be the same ones.

(3) *Privacy of query feature vector*: By using the trapdoor and the transformed feature vectors, the cloud can obtain the exact distances between query image and the images in database. As that in text retrieval schemes, the malevolent cloud server may be able to deduce the information about the query vector by analyzing the distribution of the distances, although such attacks to secure image retrieval scheme is much more difficult than that to text scheme and is rarely discussed by now.

5.2. Performance

An image database consisting of 1000 images is used to test the performance of the proposed scheme [20]. The data base includes 10 categories, and the images in the same category are thought as similar images. The entire secure search system is implemented using C++ language on a Windows Server with Intel Core(TM) Duo Processor 2. 93 GHz.

(1) *Result accuracy*: This criterion is used to evaluate the correction of the returned results, is evaluated by precision defined as follows:

$$precision = \frac{num_{hit}}{num_{return}}, \tag{9}$$

where num_{return} is the number of the returned images, num_{hit} is the number of correct similar images. The accuracy of the scheme is mainly decided by the feature extraction method in common image retrieval systems. The proposed scheme holds the same result accuracy as the common schemes that do not encrypt the feature vectors according to the formula (8). The accuracy of a query in our scheme is shown the Figure 1. The results are averaged from ten queries, each of which aims at an image category.

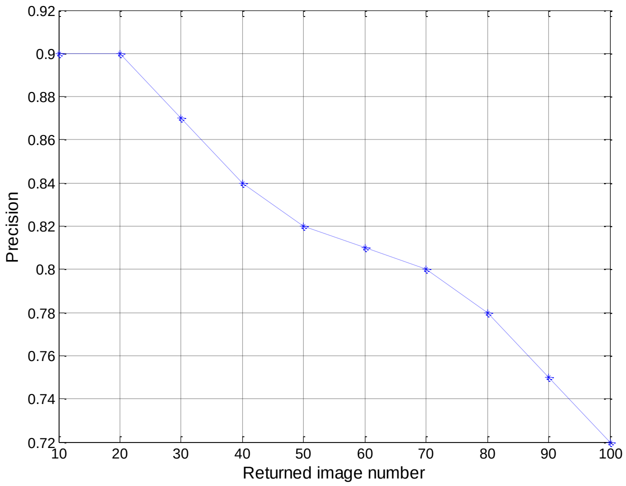


Figure 1. The Precision of the Scheme

(2) *Time complexity*: The process of index construction includes feature extraction and feature vector transformation. The time cost of calculation of histogram is$O(|m|\cdot n)$. Here, $|m|$is pixel number of the image, and n is number of images. The transformation of feature vectors involves a multiplication of a $(l+1)\times(l+1)$ matrix, and thus, the time cost is $O((l+1)^2\cdot n)$. In summary, the time complexity of index construction is$O((|m|+(l+1)^2)\cdot n)$. The search process includes trapdoor generation and search, the time costs of which are$O(l^2)$and$O(l\cdot n)$, respectively. In summary, the time complexities of index construction and query are determined by the size of database *n*, and are shown in Figure 2.

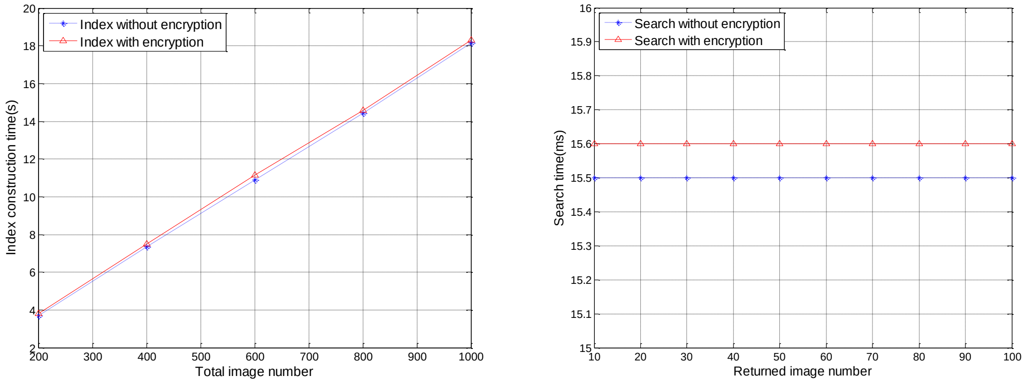


Figure 2. Time Complexity: (a) Time of Index Construction, (b) the Averaged Time of Ten Queries

(3) *Update*: The index used in the proposed scheme is the typical invert index, on which the deletion and insertion of the data is quite easy.

6. Conclusions

A basic similarity search scheme over encrypted images is proposed based on a secure transformation approach. The proposed scheme protects the confidentiality of image database, feature vectors, and user's query. Meanwhile, the proposed scheme possesses the same accuracy as the schemes which use the same feature extraction method but do not encrypt the features. However, the proposed scheme is by no means the optimal one. It does not bedim the search pattern and access pattern, and thus may suffer from statistic attacks. In addition, the time complexity of query on invert index is $O(n)$, which can be further improved by using better index. In future, we will improve our scheme in these two aspects.

Acknowledgements

This work is supported by the NSFC (61232016, 61103141, 61070195, 61070196, 61173141, 61173142, 61173136, 61103215, 61373132, 61373133, and 61073191), National Basic Research Program 973 (2011CB311808), 2011GK2009, GYHY201206033, 201301030, 2013DFG12860, SBC201310569, Research Start-Up fund of NUIST (20110428), and PAPD fund. It was also supported by the Industrial Strategic Technology Development Program (10041740) funded by the Ministry of Trade, Industry and Energy (MOTIE Korea), and by the Natural Science Foundation of Jiangsu Province (No. BK2012461).

References

[1] D. X. Song, "Practical techniques for searches on encrypted data", Security and Privacy, S&P 2000. Proceedings, 2000 IEEE Symposium on, ed: IEEE, **(2000)**, pp. 44-55.

[2] C. Wang, "Achieving usable and privacy-assured similarity search over outsourced cloud data", INFOCOM, 2012 Proceedings IEEE, **(2012)**, pp. 451-459.

[3] J. Li, "Fuzzy keyword search over encrypted data in cloud computing", INFOCOM, 2010 Proceedings IEEE, **(2010)**, pp. 1-5.

[4] M. Chuah and W. Hu, "Privacy-aware bedtree based solution for fuzzy multi-keyword search over encrypted data", Distributed Computing Systems Workshops (ICDCSW), 2011 31st International Conference, **(2011)**, pp. 273-281.

[5] D. Boneh and B. Waters, "Conjunctive, subset, and range queries on encrypted data," in Theory of cryptography, ed: Springer, 2007, pp. 535-554.

[6] X. Zhiyong, Efficient Multi-Keyword Ranked Query on Encrypted Data in the Cloud", Parallel and Distributed Systems (ICPADS), 2012 IEEE 18th International Conference, **(2012)**, pp. 244-251.

[7] J. Katz, "Predicate encryption supporting disjunctions, polynomial equations, and inner products", Advances in Cryptology–EUROCRYPT 2008, ed: Springer, **(2008)**, pp. 146-162.

[8] C. Ning, "Privacy-preserving multi-keyword ranked search over encrypted cloud data", INFOCOM, 2011 Proceedings IEEE, **(2011)**, pp. 829-837.

[9] W. Sun, "Privacy-preserving multi-keyword text search in the cloud supporting similarity-based ranking", Proceedings of the 8th ACM SIGSAC symposium on Information, computer and communications security, **(2013)**, pp. 71-82.

[10] P. Golle, et al., "Secure conjunctive keyword search over encrypted data", Applied Cryptography and Network Security, **(2004)**,pp. 31-45.

[11] X. Jun, "Two-Step-Ranking Secure Multi-Keyword Search over Encrypted Cloud Data", Cloud and Service Computing (CSC), 2012 International Conference, **(2012)**, pp. 124-130.

[12] C. Wang, "Enabling secure and efficient ranked keyword search over outsourced cloud data", Parallel and Distributed Systems, IEEE Transactions, vol. 23, no. 8, pp. 1467-1479, **(2012)**.

[13] W. Lu, "Enabling search over encrypted multimedia databases", IS&T/SPIE Electronic Imaging, **(2009)**, pp. 725418-725418-11.

[14] J. R. Smith and S.-F. Chang, "Tools and techniques for color image retrieval", Electronic Imaging: Science & Technology, **(1996)**, pp. 426-437.

[15] D. Dunn, "Texture segmentation using 2-D Gabor elementary functions", Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 16, no. 2, **(1994)**, pp. 130-149.

[16] B. S. Manjunath and W.-Y. Ma, "Texture features for browsing and retrieval of image data", Pattern Analysis and Machine Intelligence, IEEE Transactions, vol. 18, no. 8, **(1996)**, pp. 837-842.

[17] D. G. Lowe, "Distinctive image features from scale-invariant keypoints", International journal of computer vision, vol. 60, no. 2, **(2004)**, pp. 91-110.

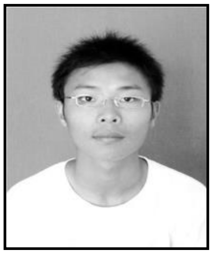
[18] T. Deselaers, "Discriminative training for object recognition using image patches", Computer Vision and Pattern Recognition, 2005. CVPR 2005, IEEE Computer Society Conference, **(2005)**.

[19] N. Cao, "Privacy-preserving multi-keyword ranked search over encrypted cloud data", **(2011)**.

[20] Corel test set. Available: http://wang.ist.psu.edu/~jwang/test1.tar.

Authors

Zhihua Xia Dr. Zhihua Xia received his BE in Hunan City University, China, in 2006, PhD in computer science and technology from Hunan University, China, in 2011. He works as a lecturer in School of Computer & Software, Nanjing University of Information Science & Technology. His research interests include cloud security, and digital forensic.



Yi Zhu Mr. Yi Zhu is currently pursuing his MS in computer science and technology at the College of Computer and Software, in Nanjing University of Information Science and Technology, China. His research interests include cloud security.

Xingming Sun Prof. Xingming Sun received his BS in mathematics from Hunan Normal University, China, in 1984, MS in computing science from Dalian University of Science and Technology, China, in 1988, and PhD in computing science from Fudan University, China, in 2001. He is currently a professor in School of Computer & Software, Nanjing University of Information Science & Technology, China. His research interests include network and information security, digital watermarking.

Jin Wang Dr. Jin Wang received the B.S. and M.S. degree from Nanjing University of Posts and Telecommunications, China in 2002 and 2005, respectively. He received Ph.D. degree from Kyung Hee University Korea in 2010. Now, he is a professor in the Computer and Software Institute, Nanjing University of Information Science and technology. He has published more than 120 journal and conference papers. His research interests mainly include routing protocol and algorithm design, performance evaluation and optimization for wireless ad hoc and sensor networks. He is a member of the IEEE and ACM.