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

Zhihua Xia, Yi Zhu, Xingming Sun, and Jin Wang

Jiangsu Engineering Center of Network Monitoring, Nanjing University of Information Science & Technology, Nanjing, 210044, China
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 based on a secure transformation method. The transformation on features protects the information about features, and does not degrade the result accuracy. Moreover, the image owner could update the encrypted image database as well as the secure index very easily.

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 ensure 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 [5-8]. 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 [9]. 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

NGCIT 2013, ASTL Vol. 27, pp. 103 - 109, 2013

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.

2 Problem Formulations

The proposed scheme includes three different entities: image owner, cloud server, and image user.

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.

3 Preliminaries

3.1 Feature Extraction

Content-based image retrieval usually involves extraction of features and search on the feature index for similar images. 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

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

where $^{x}\{m(x) = i\}$ =1, if $m(x)$ equals to i, else $^{x}\{m(x) = i\}$ =0, $|m|$ is the pixel number of 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. 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:

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

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 secure 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.

5.2 Performance

(1) *Result accuracy*: This criterion is used to evaluate the correction of the returned results. 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).

(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.

6 Conclusion

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.

References

1. D. X. Song, et al., "Practical techniques for searches on encrypted data," in Security and Privacy, 2000. S&P 2000. Proceedings. 2000 IEEE Symposium on, ed: IEEE, 2000, pp. 44-55.
2. C. Wang, et al., "Achieving usable and privacy-assured similarity search over outsourced cloud data," in INFOCOM, 2012 Proceedings IEEE, pp. 451-459, 2012.
3. J. Li, et al., "Fuzzy keyword search over encrypted data in cloud computing," in INFOCOM, 2010 Proceedings IEEE, pp. 1-5, 2010.
4. M. Chuah and W. Hu, "Privacy-aware bedtree based solution for fuzzy multi-keyword search over encrypted data," in Distributed Computing Systems Workshops (ICDCSW), 2011 31st International Conference on, pp. 273-281, 2011.
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. C. Ning, et al., "Privacy-preserving multi-keyword ranked search over encrypted cloud data," in INFOCOM, 2011 Proceedings IEEE, pp. 829-837, 2011.
7. W. Sun, et al., "Privacy-preserving multi-keyword text search in the cloud supporting similarity-based ranking," in Proceedings of the 8th ACM SIGSAC symposium on Information, computer and communications security, pp. 71-82, 2013.
8. X. Jun, et al., "Two-Step-Ranking Secure Multi-Keyword Search over Encrypted Cloud Data," in Cloud and Service Computing (CSC), 2012 International Conference on, pp. 124-130, 2012.
9. W. Lu, et al., "Enabling search over encrypted multimedia databases," in IS&T/SPIE Electronic Imaging, pp. 725418-725418-11, 2009.