A High-precision Duplicate Image Deduplication Approach

Ming Chen
1. National Engineering Laboratory for Disaster Backup and Recovery, Beijing University of Posts and Telecommunications, Beijing, China
Email: cm19834@163.com

Shupeng Wang2 and Liang Tian3
2. Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
3. College of Computer and Information Engineering, Xinxiang University, Xinxiang, China
Email: wangshupeng@iie.ac.cn, gaa252@gmail.com

Abstract—Deduplication has been widely used in backup systems and archive systems to improve storage utilization effectively. However, the traditional deduplication technology can only eliminate exactly the same images, but it is unavailable to duplicate images which have the same visual perceptions but different codes. To address the above problem, this paper proposes a high-precision duplicate image deduplication approach. The main idea of the proposed approach is eliminating the duplicate images by five stages including feature extraction, high-dimension indexing, accuracy optimization, centroid selection and deduplication evaluation. Experimental results demonstrate: in a real dataset, the proposed approach not only effectively saves storage space, but also significantly improves the retrieval precision of duplicate images. In addition, the selection of the centroid images can meet the requirements of people’s perception.

Index Terms—image deduplication; B+ tree; accuracy optimization; centroid selection; fuzzy synthetic evaluation

I. INTRODUCTION

Recently, with the development of Internet and the popularity of digital products, the volume of global digital resource is growing at an alarming rate. For examples, in 2007, for the first time ever, the total volume of digital resource exceeded the global storage capacity. It is estimated that by 2011 only half of the digital information will be stored [1]. Hence, it is impossible to solve the data explosion problem by blindly increasing storage devices. In order to solve the requirement of storage space, Kai Li Professor of Princeton University presented a new technology called global compression technology or deduplication. Deduplication can identify redundant data, eliminate all but one copy, and create local pointers to the information that users can access. This technology has been widespread concerned by industry and academia [2, 3, 4, 5].

However, the traditional deduplication technology judges two data items redundant only if their underlying bit-streams are identical. This restriction is too strict for many applications [6]. For example, in an image storage platform, according to encoding rules, any tiny transformation will completely change bit-streams of images. So the traditional deduplication technology can only eliminate exactly the same images. It is unavailable to duplicate images which have the same visual perceptions but different codes.

However, in practical applications, due to the requirement of network transmission or the restriction of storage space, users often uploaded the modified images and the same content images often present different versions which are varied in resolution or quality. From a visual angle, the images which have the same visual perceptions but different codes can be seen a redundancy. Therefore in large scale datacenter storages and data-clouds, effectively eliminating the redundant copies of images can significantly improve storage utilization. Namely, the storage optimization will have an important practical significance.

At present, the research of image deduplication doesn’t have satisfactory results. In 2011, Katiyar proposed an application-aware framework for video deduplication [6]. This framework chose ordinal signature to construct video-signatures, and used Sequence Shape Similar (SSS) to measure the similarity of the compared video sequences. Finally, the proper centroid-video was selected in the duplicate video collection by minimizing the compression-ratio and maximizing the quality of compressed videos. But there were two defects in this framework: Firstly, it did not consider deduplication accuracy. Error deduplication would bring losses to user and affect the quality of the service. Secondly, it did not consider the system scalability. For 1017 videos, this involved \( \binom{1017}{2} \) pairwise video-comparisons which would spend nearly 2 hours [6]. In addition, The Targeted Public Distribution System (TPDS) of India was a mechanism for ensuring access and availability of food grains and other essential commodities at subsidized prices to the households [7]. To bogus ration cards appear...
on the market, Ramaiah presented a photo-based deduplication method. This method used color histogram refinement to detection similar images. Then the top 20 matches which had the maximum similarity scores were returned. Finally, the duplicate images were deleted by manual operation. The disadvantage of this method was that the process of deduplication was entirely dependent on human intervention. In the massive data, this would take a lot of human resources, and prone to generate subjective errors in judgment.

For the above analysis, this paper proposes a high-precision duplicate image deduplication approach. This approach uses the 1-norm of gray block features of images to construct B+ tree index, and then detects the possible similar images by range query. Meanwhile, to ensure deduplication accuracy, more fine-grained comparison is performed to judge two images duplicate by comparing the number of the same elements in different images’ edge information. Finally, the centroid-image is selected from duplicate images by the fuzzy comprehensive evaluation method. Experimental results show that, compared with other classic algorithms, this approach not only can accurately find duplicate images, but also the selection of the centroid-images can meet requirements of people’s perception.

The remainder of this paper is structured as follows: Section II presents the definition of duplicate images. We describe the system architecture of image deduplication in Section III. Section IV gives the experimental results. Section V concludes the paper.

II.DEFINITION

A. Feature Extraction

Image deduplication needs a mechanism to compare image content. And features are the concise representations of image content. So this paper uses the gray block feature to represent images. In the calculation of the gray block feature, each image is uniformly divided into \( n \times n \) blocks. For each block, the average gray value \( I \) is calculated. So an image can be represented as a vector \( F = (I_1, I_2, ..., I_{n\times n}) \). This feature vector not only contains the image’s color information, but also contains the image’s space information. According [12], the gray block feature has strong robustness to scale transform and storage format conversion.

B. High-dimension Indexing

In large-scale image retrieval, because of the influence of “curse of dimensionality”, the traditional indexing technologies fall sharply when facing high-dimensional data. So how to effectively organize index structure to improve the processing capabilities of high-dimensional data has become an urgent problem needed to be solved. This paper introduces a new high-dimensional indexing structure to accelerate the detection of duplicate images.

**Definition 1: Duplicate images.** An image \( O_I \) is an image of an image \( O \), if \( O_I = Tr(O) \), where \( Tr \) is a set of tolerated transformations.

In this paper, the tolerated transformation \( Tr \) mainly include two kinds of variations:

- Scale: Images can be scaled horizontally or vertically. According to Foo’s research [11], the most common duplicate images in web are scaled images.
- Storage format: there are many image storage formats in web, such as JPEG, BMP, PNG, TIFF, etc. When an image is transformed to a different storage format, the corresponding image coding will change though visual perception can be maintained [12].

Other types of image transformation, such as rotation, cropping and so on, are a challenge work out of this paper’s research scope. We will work on that as our future work.

III. SYSTEM ARCHITECTURE

In this section, we describe the detail system architecture of images deduplication. Fig.1 shows the major steps in image deduplication.

![Images Deduplication System Architecture](image)

- **Feature Extraction**
- **High-dimension Indexing**
- **Accuracy Optimization**
- **Centroid Selection**
- **Deduplication Evaluation**

**Figure 1. Images Deduplication System Architecture**

- **Definition 2: Similar measure.** \( D(X,Y) \) is a distance between vectors \( X \) and \( Y \), \( x_i \) represents the value of the \( i^{th} \) component in the vector \( X \).

\[
D_p = \left( \sum_{i=1}^{n} |x_i - y_i|^p \right)^{1/p}
\]

When \( p=1 \), \( D_1(X,Y) \) is known as Manhattan distance.

\[
D_1(X,Y) = \sum_{i=1}^{n} |x_i - y_i| \leq T
\]

**According to the nature of algebra, any two n-dimensional feature vectors \( X, Y \), which can be regarded as points in space, have the (3):**

© 2013 ACADEMY PUBLISHER
From (2) and (3), we can deduce (4).
\[
\sum_{i=1}^{n} |x_i| - T \leq \sum_{i=1}^{n} |y_i| \leq \sum_{i=1}^{n} |y_i| + T
\]  
(4)

Equation (4) guarantees that if two images are similar, the corresponding 1-norms are relatively close. So that sequential scan can transform into range scan [13] in B+ tree.

In this paper, the basic idea of the index and the clustering is that the points in high-dimension space are map to the points in 1-dimension space by 1-norm. 1-norm and its corresponding feature vector are stored in B+ tree as a node. After that, we carry out range query in B+ tree according to equation (4).

Query process is as follows: after computing the 1-norm of query point \( A \), the next step of range query is to compute the lower and the higher bound. For each point in B+ tree which 1-norm is in the interval \([\text{lower limit}, \text{higher limit}]\), we sequentially compute its distance with the query point \( A \). Until we reach \( \text{higher limit} \), we end query.

\[
\begin{align*}
\text{lower limit} & = \|A\| - T \\
\text{higher limit} & = \|A\| + T
\end{align*}
\]  
(5)

C. Accuracy Optimization

Although the gray block feature can effectively identify duplicate images, but it is a coarse comparison based on block units, and the image details can’t be well handled. According to this situation, we adopt a more fine-granular algorithm to improve the retrieval precision.

Through a large number of experiments to duplicate images, we found the edge information of duplicate images contains a large number of the same elements (the same elements may be in different locations). The main reason is that duplicate images contain a large number of the same key points. Therefore, for improving deduplication accuracy, when \( D_1 \leq T \), we need to further judge whether there are a number of the same elements in different edge information.

This paper exploits Haar wavelet decomposition to extract the edge information of images [14]. The specific algorithm is as follows:

Step1: We adjust the image resolution to 128*128 and convert the image into the grayscale.

Step2: We exploit Haar wavelet decomposition to the target image.

Step3: After Haar wavelet decomposition, we extract 60 elements which have the largest absolute values from the image matrix.

Step4: We replace the 60 elements with their one-dimensional subscripts \( (V[i]/=i*64+j) \), and if an element is less 0, the corresponding one-dimension subscript multiplies by -1.

Figure 2. 1-NB+ tree

Step5: The subscripts are sorted and a partial sequence of sorted subscripts is selected to generate a feature vector. The selection position of the partial sequence will be described in section IV.

When the number \( v \) of the same elements of feature vectors is more than threshold \( t \), we think the two images are duplicate images.

D. Centroid Selection

In the course of images deduplication, the collection of duplicate images can be regarded as a cluster. We only store the highest perceptual-quality representative image that we call the centroid-image. A centroid-image is regarded as the center point in a cluster. Other images in the cluster, which are created logical points to the centroid-image, can be derived from the centroid-image by using the standard image transformations like down-scaling, down-quality, storage format conversion, etc.

The selection of the centroid-image is determined by the attributes of images. Each image has a lot of attributes. Different attributes have different values, scopes, and meanings. How to use these attributes to select the optimal image as the centroid-image is a problem that we need to solve. This paper exploits the fuzzy comprehensive evaluation method [15] to select the centroid-image.

The fuzzy comprehensive evaluation method is a comprehensive evaluation method based on fuzzy mathematics. Namely, using fuzzy mathematics make an
overall evaluation to an object which is restricted by many factors. The specific method is as follows:

Step1: We determine the factor set \( U \) which affects the image quality. In this study, taking into account the purpose of image deduplication and the human visual effect, we select \( \text{resolution}, \text{clarity}, \text{size} \) as the main factors to measure the perceived quality of an image. Here \( U = \{ \text{resolution}, \text{clarity}, \text{size} \} \).

Step2: We determine the evaluation set \( V = \{ \text{low}, \text{middle}, \text{high} \} \), and establish the fuzzy relationship matrix \( R \) between the factor set \( U \) and the evaluation set \( V \).

\[
R = \begin{bmatrix}
    r_{11} & r_{12} & r_{13} \\
    r_{21} & r_{22} & r_{23} \\
    r_{31} & r_{32} & r_{33}
\end{bmatrix}
\]  

(6)

Among them, \( r_{ij} \) represents the degree of factor \( i \) belonging to evaluation \( j \). The calculation of membership exploits a trigonometric function, as shown in fig.3.

![Figure 3. The trigonometric membership function](image)

Step3: The weight \( W \) of various factors in the factor set \( U \) is calculated by Analytic Hierarchy Process (AHP) [16-17]. Here \( W = (w_1, w_2, w_3) \) with \( \sum_{i=1}^{3} w_i = 1.0 \leq w_i \leq 1 \), and the weight is assigned depending on the importance of each factor.

Step 4: The fuzzy comprehensive evaluation result \( Z \) is calculated by selecting the proper fuzzy operator, such as \( M(\land, \lor), M(\bullet, \vee), M(\land, \oplus) \) and so on. Here \( Z = W \circ R = (z_1, z_2, z_3) \), and “ \( \circ \) ” represents the fuzzy operator.

\[
Z = W \circ R = (z_1, z_2, z_3) = (w_1, w_2, w_3) \circ \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{bmatrix}
\]  

(7)

\[
M(\bullet, +): \ z_j = (w_1 \cdot r_{1j}) + (w_2 \cdot r_{2j}) + (w_3 \cdot r_{3j})
\]  

(8)

5. The exact comprehensive evaluation result \( Z \) is calculated by defuzzification. Here \( \text{value} = Z \bullet S \), “ \( \bullet \) ” represents defuzzification, and \( S \) represents a fuzzy subset.

E. Deduplication Evaluation

**Definition 3: Deduplication Evaluation.** There are two main parameters about deduplication evaluation:

Deduplication rate = \( \frac{C_{DS}}{C_{TS}} \)

Deduplication accuracy = \( \frac{C_{DDN}}{C_{DN}} \)

\( C_{DS} \) represents the size of the deduplicated images

\( C_{TS} \) represents the size of the total images

\( C_{DDN} \) represents the number of the deduplicated duplicate images

\( C_{DN} \) represents the number of the deduplicated images.

The selection of feature extraction algorithms, similar metric algorithms and accuracy optimization algorithms has a very large impact on the final deduplication rate and deduplication accuracy. So our aim is to select proper algorithms to make deduplication rate and deduplication accuracy meet actual requirements.

In this paper, due to the particularity of image deduplication, we hope to get higher deduplication accuracy as much as possible to reduce the burden of human resources and avoid subjective judgment errors.

**F. Deduplication Comparison**

The course of image deduplication is similar with the course of text deduplication, as shown in table 1. When a new image comes, client sides firstly extract the image fingerprinter and sent the image fingerprinter to storage sides. If the storage side can’t find the matching fingerprinter in index tables, it will store the new coming image from client sides and add the image fingerprinter to index tables. Otherwise, the store side will retain the higher perception-quality image as the centroid-image and create a point pointing to the centroid-image. So that users can access the image as need.

The difference is that image deduplication is based on the similar matching of image content, and text deduplication is based on exact hash matching of binary bit streams. Moreover in the course of text deduplication, due to the duplicate data having exactly the same content, there is no such problem of centroid selection.

**Definition 4: Image fingerprinter.** Image fingerprinter is comprised of a triple \( < F, N, P > \):

\( F \) represents a feature vector.

\( N = \| F \| \), which is the 1-norm of the vector \( F \).

\( P \) represents the properties of an image. This paper refers to the image’s storage format, \( \text{resolution} \), \( \text{clarity} \), \( \text{size} \).

**Definition 5: Deduplication rules (approximate matching rules).** Images, which satisfy the following rules, shall be treated as duplicate images.

\[
\begin{cases}
\text{rule1: } D_i \leq T \\
\text{rule2: } v \geq t
\end{cases}
\]  

(9)

Rule1 means the Manhattan distance \( D_i \) of gray block features is less than the threshold \( T \).

Rule2 means the number \( v \) of the same elements in edge information is more than the threshold \( t \).
TABLE I.
COMPARISON BETWEEN TEXT DEDUPLICATION AND IMAGE DEDUPLICATION

<table>
<thead>
<tr>
<th></th>
<th>text deduplication</th>
<th>image deduplication</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>read text file</td>
<td>read image file</td>
</tr>
<tr>
<td>2</td>
<td>data partition</td>
<td>Image preprocess</td>
</tr>
<tr>
<td>3</td>
<td>hash computation</td>
<td>feature extraction</td>
</tr>
<tr>
<td>4</td>
<td>index lookup: an exact matching</td>
<td>index lookup: an approximate matching, $D_i &lt; T$</td>
</tr>
<tr>
<td>5</td>
<td>accuracy optimization: comparison byte by byte</td>
<td>accuracy optimization: compare the number of the same elements</td>
</tr>
<tr>
<td>6</td>
<td>storage one copy</td>
<td>Centroid selection and storage the centroid-image</td>
</tr>
</tbody>
</table>

IV. EXPERIMENTAL RESULTS AND ANALYSIS

In order to verify the effect of the proposed approach, this paper conducts a large number of related experiments, including: 1) Parameter estimation. We analyze experimental results to select the proper parameters of image deduplication. 2) Algorithms comparison. We compare the anti jamming capability of the proposed approach with other classic algorithms. 3) experiment verification. We evaluate the actual deduplication effect on a real dataset.

A. Parameter Estimation

In order to verify the precision of the proposed approach, this paper selects 10000 images from Corel datasets as a testing dataset which contains a large number of similar image groups. Then we randomly select 100 images from the testing dataset to be as the query images. To generate copies, each query image is processed by scaling transformation or storage format conversion using Photoshop.

¬ Threshold $t$

The setting of the threshold $t$ directly affect the anti jamming capability of image features. When the threshold $t$ is too small, it is difficult to play the role of distinction which will lead to a higher false detection. But when $t$ is too large, it will treat some duplicate images as non-duplicate images to filter out, which will reduce deduplication rate. So we need to choose an appropriate threshold $t$ according to experimental results. This paper obtains the optimal threshold $t$ through a receiver operating characteristic or ROC curve. ROC represents the relation between false rejection (fr) and correct rejection (cr) under different threshold $t$. When the curve is closest to the upper-left, it will get the optimal detection effect. As shown in fig.4, we select $t=8$.

\[
fr = \frac{\text{the number of undetected copies}}{\text{the number of total copies}} \tag{10}
\]

\[
cr = \frac{\text{the number of detected non-copies}}{\text{the number of total non-copies}} \tag{11}
\]

- Threshold $T$

To obtain the optimal threshold $T$, we introduce Precision-Recall (PR) curve to measure the performance of the duplicate image detection. Intuitively, the ideal duplicate image detection algorithm should be able to distinguish various similar images and resist various distortions. Namely, it means precision and recall should meanwhile tend to be 100%. But in fact, due to the influence of image transform noise and the imprecision of similar matching, precision and recall are a check and balance. As shown in fig.5, when the threshold $T$ varies between 8 and 10, precision and recall have a better balance. Here we select $T=9$.

\[
\text{recall} = \frac{\text{the number of detected copies}}{\text{the number of total copies}} \tag{12}
\]

\[
\text{precision} = \frac{\text{the number of detected copies}}{\text{the number of detected images}} \tag{13}
\]

- The selection of partial sequence
In section III, we need to select a partial sequence from 60 elements to generate a feature vector. As shown in table 2, the selection position of partial sequence has an important impact on deduplication results. If we select 10 elements in the middle, the probability of collision, which means the probability that the non-duplicate images are wrong treated as the duplicate images, will be higher than both ends. This is consistent with our knowledge. According section III, the elements in both ends are bigger than the elements in the middle. This means the elements in both ends mainly locate in HL or HH quadrant which presents the edge information of an image. And there is the same edge information between duplicate images. So the elements in both ends have more discriminative than the elements in the middle. Here we select the left 10 elements to generate a feature vector.

TABLE II.
THE SELECTION OF THE PARTIAL SEQUENCE

<table>
<thead>
<tr>
<th>selection position</th>
<th>the number of error deduplicated images</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>t=8</td>
</tr>
<tr>
<td>0-9</td>
<td>142</td>
</tr>
<tr>
<td>10-19</td>
<td>318</td>
</tr>
<tr>
<td>20-29</td>
<td>2318</td>
</tr>
<tr>
<td>30-39</td>
<td>2088</td>
</tr>
<tr>
<td>40-49</td>
<td>287</td>
</tr>
<tr>
<td>50-59</td>
<td>134</td>
</tr>
</tbody>
</table>

B. Algorithms Comparison

In this section, we compare the proposed approach to other three classic algorithms based on the testing dataset. Algorithm 1: Wang et al. proposed a large-scale duplicate detection for web image search [12]. Among them, the projection matrix is trained by the testing dataset. And hamming distance is set to 0. Because there is no threshold selection, so the result in fig.6 is a bar graph rather than a curve graph; Algorithm 2: Kim proposed to used ordinal measure of 35 DCT coefficients of an 8×8 sub-image to detect unauthorized copies of images [8].

As shown in fig.6, the proposed approach has obvious advantages than other algorithms: 1) Under the condition of the same recall, the precision of the proposed approach is closest to the ideal value (100%). 2) When the recall raise, the precision of Kim’s algorithm and Bhat’s algorithm decline sharply. Although the precision of the proposed approach also decline, the magnitude is small and stable. So the proposed approach has more excellent performance.

It should be pointed out that, due to the testing dataset containing a large number of similar images, the retrieval results of three classic algorithms are not ideal. But it also from the side reflects that the proposed approach is more in line with the precision requirement. It is a high-precision duplicate image deduplication approach.

C. Experimental Evaluation

◆ Centroid selection

Figure 7. The selection of the centroid-image
In this section, we use the fuzzy comprehensive evaluation method to select the centroid-images, and analyze the select results. As shown in fig.7, fig.7(a) is the original image called Lenna.bmp. Fig.7(b) is transformed from fig.7(a) by color depth changes. Fig.7(h) is transformed from fig.7(a) by storage format conversion. The three images have the same appearance. But due to fig.7(h) occupying the minimum storage space which meet the deduplication purpose, so fig.7(h) obtains a highest evaluation score. Fig.7(d), Fig.7(e), Fig.7(f), Fig.7(g) are transformed from fig.7(a) by scale transformation. Their evaluation scores are proportional to their resolutions. This is consistent with our knowledge. Because the small-scale image transforming to the big-scale image will produce blurring. This will affect visual effect, such as fig.7(e), which obtains a lowest evaluation score. In the course of centroid selection, the clarity of the image is calculated by the energy gradient algorithm. The calculation result of the weight is \( W = \{0.3, 0.3, 0.4\} \), the fuzzy operator \( M(\cdot, +) \) is selected, and the fuzzy subset is \( S = \{0.6, 0.8, 1\} \). Concrete results are shown in table 3.

### TABLE III.

<table>
<thead>
<tr>
<th>image</th>
<th>value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fig.9(a)</td>
<td>0.834</td>
</tr>
<tr>
<td>Fig.9(b)</td>
<td>0.8928</td>
</tr>
<tr>
<td>Fig.9(c)</td>
<td>0.7512</td>
</tr>
<tr>
<td>Fig.9(d)</td>
<td>0.8228</td>
</tr>
<tr>
<td>Fig.9(e)</td>
<td>0.8156</td>
</tr>
<tr>
<td>Fig.9(f)</td>
<td>0.8028</td>
</tr>
<tr>
<td>Fig.9(g)</td>
<td>0.8228</td>
</tr>
<tr>
<td>Fig.9(h)</td>
<td>0.9608</td>
</tr>
</tbody>
</table>

In summary, the fuzzy comprehensive evaluation results are in line with our visual perception and the purpose of image deduplication.

- **Experimental verification**

In order to evaluate the effect of the proposed approach on real datasets, we download 16032 images from HD landscape image set (http://99118.com).

Due to this dataset having no complicate transformation, to simplify the operation we directly select the highest resolution image as the centroid image. Deduplication results are shown in tables 4. Deduplication rate is 9.7%, and deduplication accuracy is 98.8%. The analysis of error deduplicated images reveal these images have the same content compared with the original image, but in the same position the inserted text information have changed. At present it is difficult to distinguish them. So at this stage they must be manually assisted selection. However, compared with other algorithms, our approach not only meets the requirement of deduplication rate, but also has higher deduplication accuracy. This will greatly reduce the workload of participants. Moreover, the fuzzy comprehensive evaluation method can provide a visual reference to the selection of the centroid-images, which can avoid the influence of subjective factors. So the proposed approach has a good practical value.

### V. CONCLUSIONS AND FUTURE WORK

In this study, we propose a high-precision duplicate image deduplication approach. The approach uses the 1-norm of gray block features of images to construct B+-tree index and takes full advantage of the characteristics of duplicate images to improve the deduplication accuracy. The experimental results show that the proposed approach can achieve higher deduplication rate and deduplication accuracy by setting suitable threshold \( T \) and \( t \).

In the future, we will further expand the definition of duplicate images, and use a new algorithm to identify a variety of image transformations such as rotation, shift, blur, noise, watermark and so on. Moreover, the future work also focuses on a new index structure to further improve deduplication speed and scalability.

### ACKNOWLEDGMENT

This research is supported by National Natural Science Foundation of China (No. 61003260, 61202067, 61271275).

### REFERENCES


Ming Chen, born in 1983, Ph.D. candidate. His research interests include big data, duplicate image detection.

Shupeng Wang, born in 1980, Ph.D., Associate professor. His research interests include system survivability and data mining.

Liang Tian, born in 1982, Master. His research interests include computer network and cloud computing