一种支持所有权认证的客户端图像模糊去重方法

李丹平[1)] 杨 超[1)] 姜 奇[1)] 马建峰[1)] 李成洲[2)]

[1)](西安电子科技大学网络与信息安全学院 西安 710071)

[2)](西安电子科技大学电子工程学院 西安 710071)

摘 要 由于图像具有数量大、冗余多的特性,所以需要重复数据删除技术的支持,但已有的基于哈希值的文件重复删除技术不适用于图像去重.针对图像的去重面临以下三个挑战:需要支持图像模糊去重;需要对相似图像进行所有权认证;需要进行图像感知质量评价.针对以上挑战文中提出了一种支持所有权认证的客户端图像模糊去重方法.该方案的核心包括:采用高准确度感知哈希算法,以满足图像相似性的高准确度检查;采用新设计的协议进行相似图像的所有权认证;采用无参考通用图像质量评价方法,以完成图像感知质量的评价.经过安全性分析,结果表明,新方案达到了可证明其安全的安全强度,这是图像去重领域的新突破;同时,经过大量仿真测试,结果表明,新方案可以准确地进行相似检测,还可以对多种失真图像进行感知质量评价,满足了新的技术挑战;另外,性能测试结果表明,新方案的时间开销较小,能快速高效地去重,节省了大量存储资源和网络带宽.

关键词 所有权认证;重复数据删除;感知哈希;离散余弦变换;图像感知质量评价

中图法分类号 TP391 **DOI 号** 10.11897/SP.J.1016.2018.01267

A Client-Based Image Fuzzy Deduplication Method Supporting Proof of Ownership

LI Dan-Ping[1)] YANG Chao[1)] JIANG Qi[1)] MA Jian-Feng[1)] LI Cheng-Zhou[2)]

[1)](*School of Cyber Engineering*, *Xidian University*, *Xi'an* 710071)

[2)](*School of Electronic Engineering*, *Xidian University*, *Xi'an* 710071)

Abstract In the Internet age, a large number of users store their data on the cloud server. A promising technology that keeps storage cost down is deduplication. It stores only a single copy of repeating data on the server. Researchers have proposed a lot of client-based file deduplication methods. Because images have the characteristics of large quantity and redundancy, it is necessary to perform client-based deduplication of images. However, most of the existing client-based file deduplication technologies are hash-based, which cannot apply to the client-based image deduplication. The client-based deduplication of images faces many new challenges. It needs to support image fuzzy deduplication. However, most of the existing deduplication methods use hash values of files to check duplicates precisely. These methods cannot respond to the new challenge of checking duplicates fuzzily. The client-based image deduplication needs to support proof of ownership for similar images. However, none of the existing literatures takes the ownership verification of similar images into consideration. The client-based image deduplication needs to assess image perceptual quality. However, perceptual image quality assessment is not taken into consideration in all of the researches about the client-based image deduplication. Aiming at problems of above

收稿日期:2016-10-11;在线出版日期:2017-07-12.本课题得到国家自然科学基金面上项目(61672415,61671360,61672413)、移动互联网安全 111 创新引智基地基金(B16037)、国家自然科学基金青年科学基金(61303219)资助.**李丹平**,女,1992 年生,博士研究生,主要研究方向为云计算与云存储的安全.E-mail:danpingli@stu.xidian.edu.cn.**杨 超**(通信作者),男,1979 年生,博士,教授,博士生导师,中国计算机学会(CCF)会员,主要研究领域为大数据与云计算的安全、移动智能计算的安全.E-mail:chaoyang@xidian.edu.cn.**姜 奇**,男,1981 年生,博士,副教授,中国计算机学会(CCF)会员,主要研究方向为大数据与云计算的安全、移动智能计算的安全.**马建峰**,男,1963 年生,博士,教授,博士生导师,中国计算机学会(CCF)会员,主要研究领域为信道编码、密码学、无线和移动安全、系统可生存性.**李成洲**,男,1992 年生,硕士研究生,主要研究方向为智能信息处理.

methods and these new challenges, we propose a scheme named a client-based image fuzzy deduplication method supporting proof of ownership. The scheme consists of three parts: duplicate check, proof of ownership and quality comparison. Aiming at the challenge of image fuzzy deduplication, we find that the perceptual image hash can generate similar hash values for similar images and dissimilar hash values for dissimilar images. The discrete cosine transform based perceptual image hash (DCT-phash) has a high discriminative capability to images. It is widely used in image retrieval. However, the accuracy of the DCT-phash is not enough in the deduplication scenario. Therefore, we design the DAN Perceptual Hash Algorithm (DAN-phash) to meet the high accuracy requirement of image similarity checking. Aiming at the challenge of ownership verification for similar images, our idea is that, according to the random requests of the server, the client's image and the image stored on the server are executed the same transformation to generate two images. Then the server judges the similarity between the two transformed images. We firstly propose a new protocol named Verify the Second Phash (VSP). Aiming at the challenge of perceptual image quality assessment, because the deduplication is client-based, the scheme needs a no-reference image quality assessment method. We adopt a distortion-generic no-reference image quality assessment method to complete the evaluation of the image perceptual quality. Security analysis results demonstrate that security strength of the scheme is provable, which is a breakthrough and creation in the field of client-based image deduplication. A large number of simulation results show that the scheme can check image similarity accurately and assess image perceptual quality of a variety of distortions, meeting these new technical challenges; moreover, performance test results show that time cost of the scheme is small, accomplishing image deduplication quickly and efficiently, and saving a lot of bandwidth and space needed to upload and store duplicated data.

Keywords proof of ownership; deduplication; perceptual hash; discrete cosine transform; image perceptual quality assessment

1 引 言

互联网时代里,大量用户将数据存储在云服务器中,为节省这种存储模式下的磁盘空间以及网络带宽,研究人员提出了多种客户端重复数据删除技术[1-3].针对内容相同的文件,无论多少用户想将该文件存储在云服务器上,只有第一位用户需要上传,其他用户直接链接到该文件,无需再上传.据报道,商业应用重复删除的比率可达 1∶10～1∶500,这将节省最多 90%的网络带宽和磁盘空间.

随着图像的复制、传播更加方便快捷,互联网上的图像冗余数据不计其数.图像与普通文件的区别在于其占用存储空间更大,传输需要的网络带宽更多,对图像去重可以更大程度上节省存储空间和网络带宽,提高云存储服务器的性能.因此,如何高效、安全地去除重复图像数据就成为云存储领域中亟待解决的重要问题.

与已有的针对文件的客户端重复数据删除技术不同,图像去重面临着诸多新的挑战.

挑战 1. 图像去重需要支持模糊去重.

对于图像,"重复"不仅指完全相同,还应包括由同一图像经过一组可容忍变换生成的不同副本,这些副本的编码完全不同,但是从人类视觉感知角度而言,它们是相似的[4-5].这类视觉感知相似而编码不同的重复图像在互联网上随处可见,因此,图像重复数据删除技术需要支持针对这类图像的去重,即对于相似图像,是需要进行模糊去重的.

但是,大多现有文件客户端重复数据删除技术[1-3]都是通过哈希值进行精确的重复检测,不能对相似图像进行模糊的重复检测,所以根本不支持模糊去重的新挑战.

挑战 2. 图像去重需要对相似图像进行所有权认证.

图像去重不仅需要支持相似图像的模糊去重,而且还需要对相似图像进行所有权认证,即在对客户端图像进行重复检测时,当发现服务器端确实存在相似图像需要去重,此时,客户端需要向服务器证

明自己确实拥有一个与服务器端图像相似但不完全相同的图像，这是已有去重研究中未曾遇到的问题.因此，图像去重还需要支持相似图像的所有权认证，以保证攻击者不可以只通过重复检测就轻易得到服务器端的整个图像数据.

据我们所知，现有文献还没有涉及到相似图像的所有权认证.例如，首次提出“文件所有权认证”问题的文献[6]可以解决文件客户端去重中攻击者只利用文件哈希值就能得到整个文件的问题，但对于相似图像的客户端去重，服务器端的图像与客户端拥有的图像视觉感知相似但编码不同，所以文献[6]中的协议不再适用.因此，相似图像的所有权认证是所有图像去重研究中都未曾涉及的、至关重要并且难以解决的新挑战.

挑战 3. 图像去重需要进行图像感知质量评价.

图像去重不仅需要判断图像相似性进行模糊去重、对相似图像进行所有权认证，而且还需要进行图像感知质量评价，即为减少用户损失，图像去重时一般选择感知质量较高的图像作为服务器保留图像，需要时可以将高质量替换成低质量[7]，因此，图像去重需要一种方法对图像的感知质量进行评价.

然而，当前的图像去重研究[7-11]中，要么不涉及图像感知质量评价，要么采用的评价方法不能满足高效、准确、符合人眼感知等要求，当前图像去重领域还没有成熟的图像感知质量评价方法.

目前，针对图像去重的研究仍处于起步阶段，主要工作都是服务器端的图像去重方案.文献[7]提出了视频去重框架，利用序列形状相似(Sequence Shape Similar)衡量两个视频的相似性，采用帧率和空间分辨率作为视频质量衡量的标准.文献[8]通过人脸识别技术，用人的面部图像进行身份证信息的检索，利用面部特征点来衡量两张图的差异，可以发现身份证信息数据库里一个人拥有多个身份信息的情况并对此进行去重.文献[9]的方案基于 K-Means 的聚类算法，旨在减小搜索空间提高去重的速度.文献[10]采用 CBIR(Content-Based Image Retrieval)技术进行图像去重，利用直方图细化(Histogram Refinement)检测相似图像，其中质量较高图像的选择完全是靠人工.Chen 等人[11]提出了基于哈尔小波(Haar Wavelet)的图像去重法，利用哈尔小波分解提取图像的特征向量，计算特征向量之间的曼哈顿距离，从而判断图像的相似性，用分辨率作为图像质量高低评判的标准.

基于客户端的图像去重最近才有少数研究人员做了初步的研究与探索.文献[12-13]只能对完全相同的图像去重，这与文本文件去重没有太大差别，且去重率太低，根本不能有效解决图像数据冗余过多的问题.文献[14]提出的 SPSD 方案对相似图像利用基于均值的感知哈希算法进行图像特征提取，用汉明距离衡量图像特征之间的相似度.但文献[14]存在以下几点需要更多考虑：(1)利用基于均值的感知哈希算法[15]进行图像相似检测的准确度较低；(2)没有涉及到相似图像的所有权认证；(3)没有涉及到图像质量评价.

上述方法远不能满足图像去重中新的挑战和需求，针对以上图像去重方案存在的不足和所面临的新挑战，本文提出了一种支持所有权认证的客户端图像模糊去重方法(A Client-based Image Fuzzy Deduplication Method supporting Proof of Ownership，CIFD).方案采用改进版高准确度感知哈希算法(DAN Perceptual Hash Algorithm，DAN-phash)，以满足图像相似性的高准确度检查；采用全新的相似图像所有权认证协议(Verify the Second Phash Protocol，VSP)，在不需要过多计算和时间开销的前提下进行协议交互，并且保证只有真正拥有图像的客户端才能通过验证进行去重；采用无参考通用质量评价方法(Image Perceptual Quality Assessment supporting Distortion-Identification，PQA)，以完成图像感知质量的评价.安全性分析结果表明，新方案达到了可证明安全的安全强度，这是图像去重领域的新突破；同时，大量仿真测试结果表明，新方案可以准确地进行相似检测，还可以对多种失真图像进行感知质量评价，满足了新的技术挑战；另外，性能测试结果表明，新方案的时间开销较小，能快速高效地去重，节省了大量存储资源和网络带宽.

2　系统模型与符号缩写

2.1　系统模型

系统针对的“重复”图像指无失真的原始图像在经过某种失真操作后产生的失真图像，这些失真图像应满足三个要求：用户愿意接受用原始图像代替失真图像；失真图像具有某种特定的失真类型，包括快速衰落、高斯模糊、JPEG2000 压缩、JPEG 压缩；失真图像与原始图像内容一致，唯一不同的只有感知质量.

系统包含两个实体，云存储服务客户端(Client)和云存储服务器(Server).

云存储服务器提供一个托管平台来存储客户端的图像数据，支持对相同以及相似图像进行安全高

效去重,并可以保留内容相同的所有相似图像中感知质量最好的图像.

云存储服务客户端将自己的图像数据存储在服务器中,然后删除本地的原始图像.之后客户端可以随时从服务器拿到自己的图像或者与自己图像内容相同但感知质量更好的图像.

2.2 敌手模型

假设服务器是诚实可靠的,不会主动泄露图像信息或者破坏图像信息的完整性,也不会被攻击者攻破导致隐私泄露.系统考虑的安全威胁主要发生在客户端,有以下 3 种:

(1)采用感知哈希算法进行重复检测时,若感知哈希算法准确度不够高,攻击者用与服务器端图像内容不同的图像进行重复检测,通过后得到服务器端图像的所有权;

(2)使用 SHA1 散列函数生成的图像散列值是 160 位的二进制值,若用这个很短的散列值作为拥有相同图像的依据,攻击者容易通过非法手段得到该散列值,与服务器交互进行相同重复检测,通过后得到图像的所有权;

(3)对图像进行特征提取生成的特征标签是一个 64 位二进制值,若用这个很短的特征标签作为拥有相似图像的依据,攻击者容易通过非法手段得到这个特征标签,与服务器交互进行相似重复检测,通过后得到图像所有权.

假设攻击者不会伪造图像感知质量,用感知质量差的图像去替换服务器端感知质量好的图像.

2.3 符号缩写

I 表示客户端待去重的原始图像;

I' 表示服务器端的相似图像;

I_2 表示客户端生成的用于所有权认证的混合图像;

I_2' 表示服务器端生成的用于所有权认证的混合图像;

τ 表示客户端计算的图像 I 的散列值,用于上传进行相同检测,$\tau=SHA1(I)$;

p 表示客户端计算图像 I 的感知哈希值得到的特征标签,用于上传进行相似检测,$p=DAN\text{-}phash(I)$;

p' 表示服务器端存储的图像 I' 的特征标签,用于与客户端的特征标签 p 进行匹配;

p_2 表示客户端计算混合图像 I_2 的感知哈希值得到的证据标签,用于向服务器证明自己的所有权,$p_2=DAN\text{-}phash(I_2)$;

p_2' 表示服务器端计算混合图像 I_2' 的感知哈希值得到的认证标签,用于认证客户端的所有权,$p_2'=DAN\text{-}phash(I_2')$;

$\boldsymbol{F}$ 表示对图像 I 计算离散余弦变换得到的系数矩阵,$\boldsymbol{F}=\boldsymbol{C}I\boldsymbol{C}^{\mathrm{T}}$,其中 $\boldsymbol{C}$ 指变换矩阵;

$\{dt,q\}$ 表示客户端计算的图像失真类型和感知质量,用于决定保留哪张图像,$\{dt,q\}=PQA(I)$;

$d(p,p')$ 表示 p 与 p' 之间的汉明距离,用于相似判断,$d(p,p')=HD(p,p')$,HD 为求汉明距离.

3 CIFD 方案

本节将详细介绍 CIFD 方案,该方案可以对相同以及相似图像进行可靠的重复检测,进行相同及相似图像的所有权认证并保留感知质量最优图像于服务器端.

3.1 CIFD 方案总体设计

CIFD 方案核心包括三部分:(1)对客户端图像进行重复检测;(2)发现重复后对客户端图像进行所有权认证;(3)认证通过后对客户端和服务器的图像进行感知质量比较,最终服务器端保留所有相似图像中感知质量最好的.方案系统框架如图 1 所示.

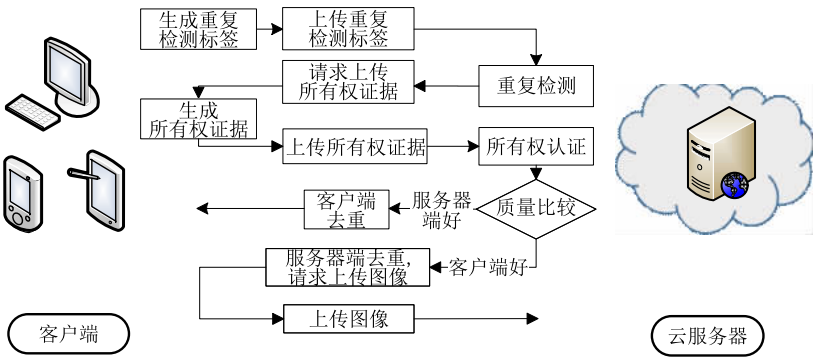

图 1 CIFD 方案系统框架

3.1.1　重复检测子方案设计

重复检测阶段面临挑战 1——图像去重需要支持模糊去重，即对相似图像进行模糊的重复检测. 重复检测的目标有两个：将相似图像都判定为相似；将不相似图像都判定为不相似.

图像感知哈希函数[14-17]根据图像的视觉特征产生哈希值，可用于分辨图像. 文献[14]中采用基于均值的感知哈希算法(AVG-phash)计算图像感知哈希值，利用汉明距离[18]计算两幅图像感知哈希值间的距离. 经测试，AVG-phash 会出现将不相似图像判断为相似、将相似图像判断为不相似的情况，误判太多. 基于离散余弦变换的感知哈希算法(DCT-based Perceptual Hash，DCT-phash)[15,17]对图像的分辨能力高，特别是 JPEG 压缩图像，如果一个图像离散余弦变换低频系数绝对值很小，除非图像发生显著变化，否则低频系数不可能变大，所以低频系数可用于判断图像相似性，该算法目前是图像检索最常用的算法之一，其特点比较适用于 CIFD 方案重复检测的设计与实施.

但是，通过对 DCT-phash 算法进行大量测试，发现虽然 DCT-phash 用于图像相似性检测的准确度比 AVG-phash 高，但是 DCT-phash 依旧会出现不准确的情况，进一步进行原理分析，结果表明 DCT-phash 算法如果用于 CIFD 方案重复检测的设计，则存在以下不足：(1) 对不相似的两幅图像，DCT-phash 算法通过离散余弦变换产生两个差距较大的系数矩阵，但是这两个系数矩阵的系数均值都很大，会出现两个系数矩阵的系数之间差距较大但都小于系数均值的情况，从而使得两个感知哈希值的很多二进制位相同，不相似图像被判断为相似，最终导致方案出现错误的去重结果；(2) 在 32×32 的系数矩阵中取左上角的 8×8 来达到去高频留低频的效果，该做法不能充分过滤表达图像之间差异的高频细节信息，相似图像会因为这些细节信息而被判定为不相似，最终导致方案的去重率降低.

针对上述问题，本文基于 DCT-phash 算法的思想，重新设计了高准确度的感知哈希算法 DAN-phash，其设计思想是：

(1) 将截尾均值[19-20]作为系数均值. 在 DCT-phash 算法中，系数矩阵(0，0)处的直流系数值很大，使得系数均值很大，不能准确代表系数矩阵的大部分系数. DAN-phash 算法采用截尾均值，计算截尾均值时选择将直流系数丢弃掉，不相似图像系数矩阵之间差距较大的系数与系数均值相比会被判定为不同的结果. 同时，无论系数均值如何变化，两个相似图像系数矩阵的系数之间差距很小，其与系数均值相比都会被判定为相同的结果.

(2) 在 64×64 的系数矩阵中取左上角 8×8 的低频系数矩阵. 与 DCT-phash 算法相比，DAN-phash 算法的低频系数矩阵过滤掉更多体现图像之间差异的高频细节信息，相似图像低频系数矩阵之间的差距更小，最终相似图像被准确地判定为相似. 而对于不相似图像，低频信息之间的差异足够用来判断不相似，过滤掉部分更能体现图像差异的高频细节也不会影响对不相似图像的判断.

DAN-phash 算法的特点在于，其采用截尾均值计算感知哈希值，能提高不相似图像的重复检测准确度，但不会影响相似图像的重复检测准确度，同时，该算法选择在 64×64 的系数矩阵中取左上角 8×8 的低频系数矩阵，能提高相似图像的重复检测准确度，但不会影响不相似图像的重复检测准确度.

采用 DAN-phash 算法使得 CIFD 方案的错误率足够低从而使得方案足够可靠，并且使得方案的去重率足够高从而可以节约大量资源. 本文挑选了权威的图像数据库 LIVE[21-22]、TID2013[23]和 Quality Image① 对该算法进行大量测试，结果表明，重复检测阈值为 1 时，DAN-phash 算法的错误率为 0，比 DCT-phash 算法大幅度降低，同时去重率比 DCT-phash 算法大幅度提高(详细说明见第 5 节). 同时，考虑到错误率和去重率是两个互相制约的量，去重之前可以提醒用户选择：如果用户对可靠性的需求更高，则为其提供阈值为 1、错误率为 0、去重率为 69.39%的去重；如果用户对去重率的要求更高，则为其提供阈值为 5、去重率为 95.71%、错误率为 0.006%的去重，与此同时服务器可以对错误率可能造成的风险给予用户存储容量上的补偿. 综上所述，新设计的 DAN-phash 算法可以很好地应对挑战 1，非常适用于 CIFD 方案的重复检测.

另外，考虑到有的用户主观上不愿意服务器对自己的某些图像进行去重，在去重前让用户将不愿去重的图像标记出来，服务器对这些图像只进行精确去重，对其余图像先进行精确去重，不存在相同图像时再进行模糊去重.

3.1.2　所有权认证子方案设计

所有权认证阶段面临挑战 2——图像去重需要

① http://live.ece.utexas.edu/research/quality
http://commons.wikimedia.org/wiki/Commons:Quality_images/

对相似图像进行所有权认证，即重复检测阶段发现存在重复图像后，客户端向服务器证明自己确实拥有一个与服务器端图像相似的待去重原始图像. 所有权认证时客户端需要向服务器提供更多能代表原始图像的信息，且这些信息只有客户端才拥有，攻击者不能轻易得到.

本文设计了一个全新的相似图像所有权认证协议(Verify the Second Phash，VSP)，即服务器和客户端都将自己的图像与同一个任意的辅助图像混合，双方都产生了一个混合图像，然后再次对两个混合图像计算感知哈希值，服务器计算感知哈希值之间的距离来判断是否相似，如果两个混合图像依旧相似，认为通过验证，否则认为没通过. VSP 协议中图像混合过程如图 2 所示.

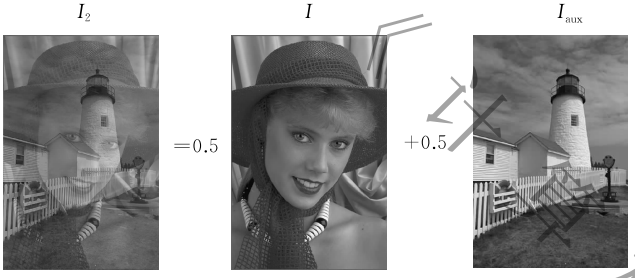

图 2 图像混合过程

VSP 协议的特点在于，客户端证明自己不是只拥有原始图像感知哈希值而是拥有整个原始图像的证据，是混合图像的感知哈希值. 图像拥有者很容易即可生成混合图像，计算出混合图像的感知哈希值通过验证，然而对于攻击者，假设原始图像的感知哈希值已经泄露，攻击者通过了重复检测，因为没有原始图像，攻击者不能生成混合图像，所以不能得到混合图像的感知哈希值，最终不能通过验证.

采用 VSP 协议进行所有权认证的安全性足够高且认证过程简单，如果攻击者要在已经得到原始图像的感知哈希值但没有原始图像的情况下通过验证，他有两个选择，利用原始图像的感知哈希值推导出混合图像的感知哈希值，或者猜测混合图像的感知哈希值，第 4 节安全性分析部分将会证明攻击者通过认证的概率足够小，可以忽略不计.

3.1.3 质量比较子方案设计

质量比较阶段面临挑战 3——图像去重需要进行图像感知质量评价，即所有权认证阶段客户端通过认证后，需要对客户端和服务器端的图像进行感知质量[21-28]评价和比较.

全参考评价方法需要作为参考图像的无失真“原始图像”可以提前获取；部分参考评价方法从参考图像中提取一定的特征作为参考，需要有一定的信息传输量；无参考评价方法中，假设存在一个高品质的“原始图像”，被评估的图像是这个“原始图像”的失真表征，该方法中用于描述高品质图像的假设的统计知识并不局限于单一的“原始图像”，而是用来解释可能的图像空间中所有高品质自然图像的概率分布，通过测量失真图像和“原始图像”的概率分布之间的偏差来完成图像质量评价[24]；比较评价方法[25]建立在评价者拥有两个待比较图像的前提条件之上，而客户端去重场景中，服务器只拥有待比较图像中的一个. 综合考虑，CIFD 方案需要一种无参考质量评价方法.

BRISQUE[26]是一种无参考通用型图像质量评价方法. 该算法不局限于特定的失真类型，只是量化由于失真的存在而导致的“自然”的损失，它比著名的全参考 SSIM[21]算法和 PSNR 算法要优秀，与很多无参考通用质量评价方法相比，在计算效率方面也有很高的竞争力，可以用于 CIFD 方案的图像质量评价和比较.

然而在去重中，同种失真的图像之间进行质量比较才有意义，若直接使用 BRISQUE 评价方法，就需要在假定所有图像都属于同一种失真类型的前提下进行去重，这给去重带来了很大的局限性.

本文设计的 PQA 方法采用 BRISQUE 评价方法的思路，具有失真鉴别和质量评价两种功能，先使用 BRISQUE 的失真鉴别功能对图像产生一个失真鉴别结果，然后使用 BRISQUE 的质量评价功能对同一失真的图像进行质量评价，评价得分越低说明图像质量越高. 与直接使用 BRISQUE 评价方法相比，PQA 方法可以处理多种失真类型，使得 CIFD 方案的去重范围更广.

3.2 CIFD 方案实施过程

CIFD 方案的具体实施包括两大部分：初始上传过程和后续上传过程. 初始上传过程只包含重复检测阶段，后续上传过程是方案研究的重点，其中包含重复检测、所有权认证、质量比较 3 个阶段. 方案协议如图 3 所示.

(1) 初始上传

初始上传执行流程如图 3 中初始上传部分所示，左侧为客户端，右侧为服务器，中间为客户端和服务器的交互过程，具体流程如下：

① 客户端. 客户端计算待去重原始图像 I 的散列值 τ，$\tau = SHA1(I)$，计算 I 的特征标签 p，$p = DAN\text{-}phash(I)$. 将 τ、p 上传后，客户端计算 I 的失

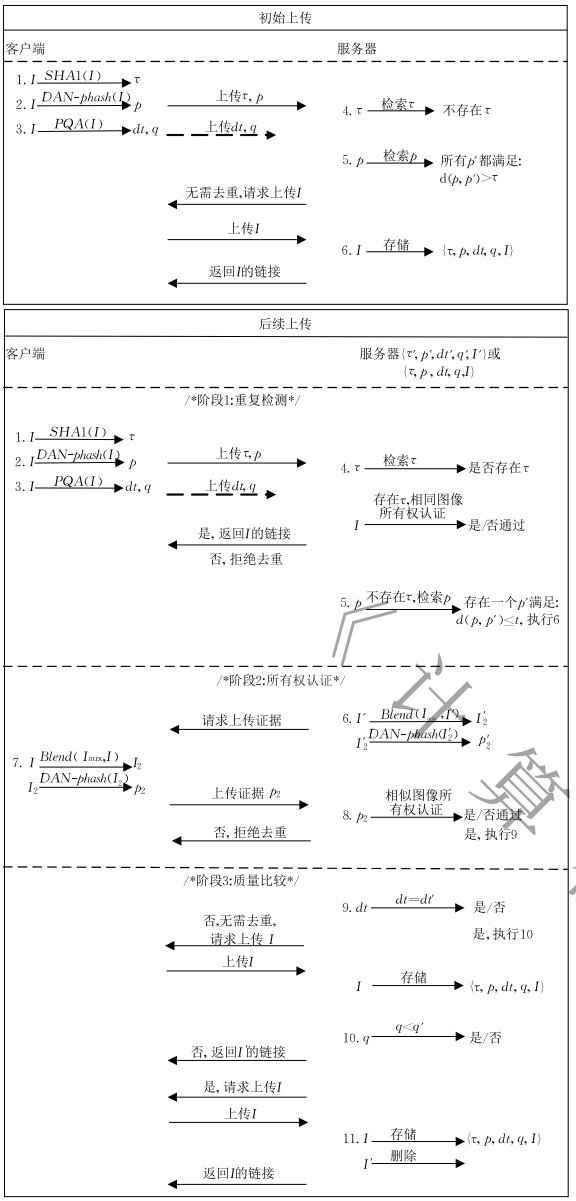

图 3　CIFD 方案协议

真类型 dt 和质量 q，$\{dt,q\}=PQA(I)$，将 $\{dt,q\}$ 上传至服务器. 客户端计算上传 $\{dt,q\}$ 和服务器检索散列值 τ 以及特征标签 p 是同步进行的(图 3 中虚线表示).

② 服务器端. 服务器在数据库中检索散列值 τ，发现不存在与 τ 相同的散列值，说明不存在相同图像. 然后服务器计算存储的所有特征标签 p' 与客户端上传的特征标签 p 之间的汉明距离，发现所有汉明距离都大于设定的阈值 t，说明不存在相似图像. 根据对 DCT-phash 的性能分析，结合 CIFD 方案场景下对 DAN-phash 的大量测试，这里的阈值取 $t=1$. 服务器请求客户端上传原始图像 I，收到 I 之后将原始图像 I 以及散列值 τ、特征标签 p、失真类型 dt 和质量 q 存储在数据库并向客户端返回 I 的链接.

(2) 后续上传

后续上传执行流程如图 3 中后续上传部分所示，具体如下：

阶段 1.　重复检测

① 客户端. 客户端计算待去重原始图像 I 的散列值 τ，$\tau=SHA1(I)$，计算 I 的特征标签 p，$p=DAN\text{-}phash(I)$. 将 τ、p 上传后，客户端计算 I 的失真类型 dt 和质量 q，$\{dt,q\}=PQA(I)$，将 $\{dt,q\}$ 上传至服务器. DAN-phash 算法是重复检测阶段的核心，其主要步骤如算法 1 所示.

算法 1.　高准确度感知哈希算法(DAN-phash).

输入：原始图像 I

$img\_bgr=imread(I)$；

$img\_gray=cvColor(img\_bgr)$；

$img\_dst=resize(img\_gray,Size(64,64))$；

$\boldsymbol{F}=dct(img\_dst)$；

$\boldsymbol{F}^{lf}=\boldsymbol{F}(0,\cdots,7;0,\cdots,7)$；

$AVG=\left(\sum\boldsymbol{F}^{lf}(i,j)-\boldsymbol{F}^{lf}(0,0)\right)/63, 0\leq i\leq 7, 0\leq j\leq 7$；

IF $\boldsymbol{F}^{lf}(i,j)\geq AVG$, $p(i,j)=1$；

ELSE $p(i,j)=0$；

$p=p(i,j)$.

输出：特征标签 p

具体步骤如下：

(a) 读入待去重的原始图像 I，得到 RGB 图像 img_bgr；

(b) 将 img_bgr 转化为灰度图像 img_gray，这样做是为了简化计算，使得后续的离散余弦变换只对图像灰度通道进行，对三个通道都进行变换的运算量太大；

(c) 将灰度图像 img_gray 尺寸转化为 64×64，得到缩小图像 img_dst，对整个图像做离散余弦变换运算量太大且没有必要，缩小的图像包含足够的内容信息可用于提取有鉴别能力的标识信息，DCT-phash 一般将图像缩小到 32×32，为了提高算法对相似图像的判断能力，DAN-phash 选择的缩小图像尺寸为 64×64；

(d) 对缩小图像 img_dst 做离散余弦变换，得到 64×64 的系数矩阵 $\boldsymbol{F}$，离散余弦变换具有很强的“能量集中”特性，图像的能量几乎都集中在左上角的低频系数上，所以系数矩阵 $\boldsymbol{F}$ 从左上角到右下角，频率越来越高；

(e) 选取系数矩阵 $\boldsymbol{F}$ 左上角的前 8 行与前 8 列，得到 8×8 的低频系数矩阵 $\boldsymbol{F}^{lf}$，尽管低频部分

的数据量比高频部分的数据量要小的多,但低频部分的信息量大于高频部分的信息量,因此低频系数矩阵用少量数据代表了图像的大部分信息;

(f)将低频系数矩阵 $\boldsymbol{F}^{lf}$ 在位置(0,0)处的直流系数去掉,对其余的系数计算系数均值 AVG,因为直流系数会影响算法对不相似图像的判断能力;

(g)将低频系数矩阵 $\boldsymbol{F}^{lf}$ 的 64 个系数中大于等于系数均值 AVG 的设为 1,小于 AVG 的设为 0,生成 64 位的二进制感知哈希值 $p(i,j)$,这是最主要的一步,$p(i,j)$不能体现实际的低频系数是什么,只能粗略地显示系数跟系数均值的相对大小,只要图像的整体结构保持不变,$p(i,j)$就不会改变;

(h)将二进制感知哈希值 $p(i,j)$从左到右、从上到下转换为一维数组,并将转换后的一维数组设定为原始图像 I 的特征标签 p.

②服务器端. 服务器在数据库中检索散列值 τ,如果发现存在与 τ 相同的散列值,说明存在相同图像,此时服务器对客户端进行相同图像的所有权认证. 这里的所有权认证与对普通的文件进行所有权认证一样,采用的方法是基于 MerkleTree 的 PoW[46]. 如果验证通过,服务器通知客户端进行去重,并将原始图像 I 的链接返回给客户端;否则不进行去重. 如果服务器检索发现不存在与 τ 相同的散列值,说明不存在相同图像,然后服务器计算存储的所有图像的特征标签 p' 与客户端上传的图像特征标签 p 之间的汉明距离 $d(p,p')$,计算公式如下:

$$d(p,p')=HD(p,p')=\sum_{p_i\neq p'_i}1,$$

其中,$1\leq i\leq 64$.

若服务器发现存在一个图像 I' 的特征标签 p' 与客户端上传的特征标签 p 之间的汉明距离小于阈值 t,说明 I' 与 I 相似,客户端和服务器进入所有权认证阶段.

阶段 2. 所有权认证

通过 VSP 协议进行相似图像的所有权认证,VSP 协议核心步骤如过程 1 所示.

过程 1. 相似图像所有权认证协议(VSP).

服务器:

选择辅助图像 I_{aux},请求客户端上传证据;

读入 I' 和 I_{aux};

调整 I_{aux} 尺寸,$Size(I_{aux})=Size(I')$;

设 $\alpha=0.5$,计算 $I'_2=Blend(I',I_{aux},\alpha)=\alpha I'+(1-\alpha)I_{aux}$;

计算 $p'_2=DAN\text{-}phash(I'_2)$;

客户端:

接收到请求后,读入 I 和 I_{aux};

调整 I_{aux} 尺寸,$Size(I_{aux})=Size(I)$;

设 $\alpha=0.5$,计算 $I_2=Blend(I,I_{aux},\alpha)=\alpha I+(1-\alpha)I_{aux}$;

计算 $p_2=DAN\text{-}phash(I_2)$,上传 p_2;

服务器:

接收到 p_2 后,计算 $d(p_2,p'_2)=HD(p_2,p'_2)\sum_{p_{2_i}\neq p'_{2_i}}1$;

IF $d<t$,认证通过;

ELSE,认证不通过.

具体如下:

(a)服务器从辅助数据库中随机选定一个用于帮助服务器完成认证过程的辅助图像 I_{aux},将 I_{aux} 的编号发送给客户端,请求客户端上传所有权认证的证据,辅助图像可以为任何尺寸大于 64×64 的图像;

(b)服务器分别读入辅助图像 I_{aux} 和相似图像 I',调整辅助图像 I_{aux} 的尺寸,使其与相似图像 I' 大小一致,每一次认证过程 I_{aux} 尺寸都根据 I' 的尺寸而改变,这是后续的图像混合 $Blend$ 运算需要的条件;

(c)服务器进行 $Blend$ 运算:设定图像混合参数 $\alpha=0.5$,由相似图像 I' 和辅助图像 I_{aux} 产生服务器端的混合图像 I'_2,α 取 0.5 时的混合图像最符合方案需要的安全强度,这一点在第 4 节安全性分析部分将给出证明;

(d)服务器计算混合图像 I'_2 的感知哈希值,并将该感知哈希值设定为认证标签 p'_2,p'_2 用于对客户端的证据标签进行认证;

(e)客户端收到辅助图像的编号后,从辅助数据库中选定该编号对应的辅助图像 I_{aux},分别读入原始图像 I 与辅助图像 I_{aux},客户端和服务器在一次认证过程中采用的辅助图像必须相同,以保证辅助图像不会对所有权认证过程产生副作用,不同认证过程采用的辅助图像尽量保持不同;

(f)与服务器一样,客户端调整辅助图像 I_{aux} 的尺寸,使其与原始图像 I 大小一致;进行 $Blend$ 运算,设定图像混合参数 $\alpha=0.5$,由原始图像 I 和辅助图像 I_{aux} 产生客户端的混合图像 I_2;计算混合图像 I_2 的感知哈希值,并将该感知哈希值设定为证据标签 p_2,上传 p_2 至服务器端;

(g)服务器收到证据标签 p_2 后,计算证据标签 p_2 与认证标签 p'_2 之间的汉明距离 $d(p_2,p'_2)$;判断汉明距离 $d(p_2,p'_2)$ 是否小于给定的阈值 t,若是,则认证通过,进入质量比较阶段;若否,则认证不通过,拒绝客户端进行去重.

阶段 3. 质量比较

收到客户端上传的失真类型 dt 和质量 q 后，服务器从数据库中存储的 n 个图像中找到相似图像 I' 对应的失真类型 dt' 和质量 q'. 服务器判断 dt' 是否等于 dt，若否，则无需去重，服务器请求客户端上传原始图像 I，收到 I 之后将原始图像 I 以及散列值 τ、特征标签 p、失真类型 dt 和质量 q 存储在数据库并向客户端返回 I 的链接；若是，则进行质量比较，如果 $q\geq q'$，说明服务器端图像质量更好，服务器通知客户端删除原始图像 I，并将相似图像 I' 的链接返回给客户端，如果 $q<q'$，说明客户端的图像 I 质量更好，服务器请求客户端上传原始图像 I，接收到 I 后，服务器删除之前存储的相似图像 I'，将原始图像 I 以及散列值 τ、特征标签 p、失真类型 dt 和质量 q 存储在数据库并向客户端返回 I 的链接.

PQA 方法利用对局部归一化亮度系数的场景统计来量化由于失真存在而导致的“自然”的损失. 具体如下：

(a) 计算原始图像 I 的多尺度的去均值对比度归一化系数；

(b) 对这些系数及其沿不同方向的相关系数进行非对称广义高斯拟合，得到参数作为特征；

(c) 用支持向量机进行失真类型识别，产生失真鉴别结果 dt；

(d) 对特定失真类型建立支持向量回归分析模型，将图像特征映射到质量分数 q.

以上是整个 CIFD 方案的流程，其中主要存在四种情况，首位上传会遇到服务器端不存在相同图像也不存在相似图像这一种情况，此时需要上传图像. 后续上传者会遇到服务器端存在相同图像、服务器端存在相似图像且服务器端图像质量更好、服务器端存在相似图像且客户端图像质量更好这三种情况，前两种情况都不需要上传图像，最后一种情况才需要上传图像. 由此可见，CIFD 方案既能节省网络带宽，又能节省服务器的存储空间，可以防止用户图像数据的非授权访问，还可以将质量差的图像去重而保留质量好的图像.

4　CIFD 方案安全性分析

本节将对 CIFD 方案的安全性进行理论分析证明，主要针对相似图像的所有权认证.

定义 1. 如果图像 I 的低频系数矩阵 $\boldsymbol{F}^{lf}$ 在位置 (i,j) 处的系数 $\boldsymbol{F}^{lf}(i,j)$ 和图像 I_2 的低频系数矩阵 $\boldsymbol{F}_2^{lf}$ 在位置 (i,j) 处的系数 $\boldsymbol{F}_2^{lf}(i,j)$ 满足 $\boldsymbol{F}_2^{lf}(i,j)>\boldsymbol{F}^{lf}(i,j)$，或 $\boldsymbol{F}_2^{lf}(i,j)\leq\boldsymbol{F}^{lf}(i,j)$，称 $\boldsymbol{F}_2^{lf}(i,j)$ 与 $\boldsymbol{F}^{lf}(i,j)$ 大小关系确定，记作 $C(\boldsymbol{F}_2^{lf}(i,j),\boldsymbol{F}^{lf}(i,j))$；否则，称 $\boldsymbol{F}_2^{lf}(i,j)$ 与 $\boldsymbol{F}^{lf}(i,j)$ 大小关系不确定，记作 $UC(\boldsymbol{F}_2^{lf}(i,j),\boldsymbol{F}^{lf}(i,j))$.

定义 2. 在位置 (i,j) 处，如果图像 I 的感知哈希值的二进制位 $p(i,j)=1$，且图像 I 和图像 I_2 的低频系数矩阵中的系数 $\boldsymbol{F}^{lf}(i,j)$ 和 $\boldsymbol{F}_2^{lf}(i,j)$ 满足，$\boldsymbol{F}_2^{lf}(i,j)>\boldsymbol{F}^{lf}(i,j)$ 即 $C(\boldsymbol{F}_2^{lf}(i,j),\boldsymbol{F}^{lf}(i,j))$，或者 $p(i,j)=0$，且 $\boldsymbol{F}^{lf}(i,j)$ 与 $\boldsymbol{F}_2^{lf}(i,j)$ 满足 $\boldsymbol{F}_2^{lf}(i,j)\leq\boldsymbol{F}^{lf}(i,j)$ 即 $C(\boldsymbol{F}_2^{lf}(i,j),\boldsymbol{F}^{lf}(i,j))$，这两种情况，称在位置 (i,j) 处 $p_2(i,j)$ 由 $p(i,j)$ 可推导，记作 $CBD(p_2(i,j),p(i,j))$.

定义 3. 如果图像 I 的感知哈希值 p 与图像 I_2 的感知哈希值 p_2 中至少有一个位置（第一位除外）满足 $CBD(p_2(i,j),p(i,j))$，称得到 p_2 的概率为推导概率 P_D；否则，称得到 p_2 的概率为猜测概率 P_G.

定理 1. 如果得到图像 I_2 的感知哈希值 p_2 的概率为推导概率 P_D，则 $P_D=1/2^{(N-K)}$，其中 N 为 p_2 中需要得到的位数，K 为满足 $CBD(p_2(i,j),p(i,j))$ 的位数；如果得到 p_2 的概率为猜测概率 P_G，则 $P_G=1/2^N$.

证明. 如果得到 p_2 的概率为推导概率 P_D，K 为 p_2 中满足 $CBD(p_2(i,j),p(i,j))$ 的位数，$K\geq 1$，则这 K 位已知，此时要得到 p_2 只需猜测其余位的值，即 $P_D=1/2^{(N-K)}$；如果得到 p_2 的概率为猜测概率 P_G，则 p_2 中所有位置都不满足 $CBD(p_2(i,j),p(i,j))$，此时要得到 p_2 需猜测 p_2 每一位的值，即 $P_G=1/2^N$. 证毕.

定理 2. 在 CIFD 方案中，客户端待去重原始图像 I 的感知哈希值为 p，客户端混合图像 I_2 的感知哈希值为 p_2，p 和 p_2 长度皆为 64，假设 p 已经泄露给恶意攻击者. p_2 第一位确定为 1，服务器进行所有权认证时通过的阈值为 1，即 p_2 有两位攻击者不需要得到，攻击者需要得到的位数 $N=62$，则攻击者通过验证的概率 $P\leq P_G$，$P_G=1/2^{62}$.

证明. 假设攻击者可以在拥有 p 的前提下，以不可忽略的概率 P ($P>P_G$) 得到 p_2，通过所有权认证，则攻击者要在没有原始图像 I 的情况下得到 p_2，使 p_2 与服务器端混合图像的感知哈希值的汉明距离小于阈值. 考虑攻击者可能有以下方法：

(1) p_2 中至少有一个位置满足 $CBD(p_2(i,j),p(i,j))$，攻击者通过 p 推导 p_2，有以下可能：

(1a) 攻击者没有原始图像 I 的任何信息的前提

下,通过 p 推导 p_2;

(1b)攻击者已知原始图像 I 的部分信息的前提下,通过 p 推导 p_2;

(2) p_2 中所有位都不满足 $CBD(p_2(i,j),p(i,j))$, p 对攻击者没有任何意义,攻击者通过暴力攻击或者碰撞攻击得到 p_2,这里的碰撞攻击指攻击者得到混合图像 I_2 低频系数矩阵 $\boldsymbol{F}_2^{lf}$ 或者混合图像 I_2,使混合图像 I_2 的感知哈希值 p_2 和服务器端混合图像感知哈希值满足阈值,有以下可能:

(2a)通过暴力攻击猜测 p_2;

(2b)通过猜测混合图像 I_2 低频系数矩阵 $\boldsymbol{F}_2^{lf}$ 中每一个系数对系数均值的分布得到 p_2;

(2c)通过猜测混合图像 I_2 低频系数矩阵 $\boldsymbol{F}_2^{lf}$ 的每一个系数得到 p_2;

(2d)在已知原始图像 I 的部分信息的前提下,通过猜测混合图像 I_2 得到低频系数矩阵 $\boldsymbol{F}_2^{lf}$ 的每一个系数,从而得到 p_2;

(2e)通过猜测混合图像 I_2 得到 p_2;

攻击者采用方法(1a)时,K 为 p_2 中满足 $CBD(p_2(i,j),p(i,j))$ 的位数,则 p 和 p_2 对应的低频系数矩阵在这 K 个位置上的系数 $\boldsymbol{F}^{lf}(i,j)$ 和 $\boldsymbol{F}_2^{lf}(i,j)$ 满足 $C(\boldsymbol{F}_2^{lf}(i,j),\boldsymbol{F}^{lf}(i,j))$,即混合图像 I_2 低频系数矩阵中有 K 位与原始图像 I 低频系数矩阵中对应位的值的大小关系是确定的.

由于辅助图像 I_{aux} 与原始图像 I 是任意的两幅图像,且攻击者没有 I,任意位置 (i,j) 处 I 和 I_{aux} 的像素值 $I(i,j)$ 和 $I_{aux}(i,j)$ 满足 $UC(I_{aux}(i,j),I(i,j))$. 图像混合时 $I_2(i,j)=0.5\times I_{aux}(i,j)+0.5\times I(i,j))$,如果 $I(i,j)\leq I_{aux}(i,j)$,则 $I_2(i,j)\geq I(i,j)$,否则,$I_2(i,j)<I(i,j)$,即 $I(i,j)$ 和 $I_2(i,j)$ 的大小关系取决于 $I(i,j)$ 和 $I_{aux}(i,j)$ 的大小关系,则 $I(i,j)$ 与 $I_2(i,j)$ 满足 $UC(I_2(i,j),I(i,j))$,即原始图像 I 任意位置处的像素值与混合图像 I_2 任意位置处的像素值大小关系不确定. 由离散余弦变换公式可得

$$\boldsymbol{F}=\boldsymbol{C}\boldsymbol{I}\boldsymbol{C}^{\mathrm{T}},\ \boldsymbol{F}_2=\boldsymbol{C}\boldsymbol{I}_2\boldsymbol{C}^{\mathrm{T}},$$

$$I=\begin{bmatrix}x_{11}&x_{12}&\cdots&x_{1n}\\x_{21}&x_{22}&\cdots&x_{2n}\\\vdots&&&\\x_{n1}&x_{n2}&\cdots&x_{nn}\end{bmatrix},\ I_2=\begin{bmatrix}y_{11}&y_{12}&\cdots&y_{1n}\\y_{21}&y_{22}&\cdots&y_{2n}\\\vdots&&&\\y_{n1}&y_{n2}&\cdots&y_{nn}\end{bmatrix},$$

$$C=\begin{bmatrix}c_{11}&c_{12}&\cdots&c_{1n}\\c_{21}&c_{22}&\cdots&c_{2n}\\\vdots&&&\\c_{n1}&c_{n2}&\cdots&c_{nn}\end{bmatrix},$$

$$CI(i,j)=c_{i1}x_{1j}+c_{i2}x_{2j}+\cdots+c_{in}x_{nj},$$

$$CI_2(i,j)=c_{i1}y_{1j}+c_{i2}y_{2j}+\cdots+c_{in}y_{nj},$$

因为任意位置 (i,j) 处 $I(i,j)$ 与 $I_2(i,j)$ 满足 $UC(I_2(i,j),I(i,j))$,则 $CI(i,j)$ 与 $CI_2(i,j)$ 满足 $UC(CI_2(i,j),CI(i,j))$ 则 $CI(i,j)\boldsymbol{C}^{\mathrm{T}}$ 和 $CI_2(i,j)\boldsymbol{C}^{\mathrm{T}}$ 满足 $UC(CI_2(i,j)\boldsymbol{C}^{\mathrm{T}},CI(i,j)\boldsymbol{C}^{\mathrm{T}})$,即 $UC(F_2(i,j),F(i,j))$,原始图像 I 系数矩阵任意位置处的值与混合图像 I_2 系数矩阵任意位置处的值大小关系不确定. 这与"混合图像 I_2 低频系数矩阵中有 K 位与原始图像 I 低频系数矩阵中对应位的大小关系是确定的"相矛盾,即攻击者不能通过方法(1a)得到 p_2.

攻击者采用方法(1b)时,攻击者拥有原始图像 I 的部分信息,假设攻击者由原始图像 I 的部分信息得到混合图像 I_2 的部分信息,由于 I 和 I_2 系数矩阵任意位置处系数值的大小关系与图像每个位置上的灰度值有关,拥有 I 的部分信息不能帮助攻击者确定 I 和 I_2 系数矩阵的大小关系,因此,攻击者不能通过 p 推导 p_2.

攻击者采用方法(2a)和(2b)时,$P=1/2^{62}$,这与假设的 $P>P_G$ 矛盾.

攻击者采用方法(2c)时,一般情况下,低频系数矩阵内的系数空间大概在 2 位到 4 位双精度浮点型数值之间,为了便于量化,充分简化模型,假设系数空间在 10～1000 之间,则通过猜测 $\boldsymbol{F}_2^{lf}$ 的每一个系数得到 p_2 的概率是 P 满足,$1/1000^{62}<P<1/10^{62}$,这与假设的 $P>>P_G$ 相矛盾.

攻击者采用方法(2d)时,假设原始图像 I 的尺寸为 $m\times n$,攻击者拥有 I 中尺寸为 $a\times b$ 的灰度值,并假设攻击者由已知的 I 的信息得到混合图像 I_2 的中尺寸为 $a\times b$ 的灰度值(实际上攻击者得到辅助图像才可以做到这一点). 由于 DCT 变换得到的系数矩阵中任何一个系数都与图像每个位置上的灰度值有关,因此,要确定 $\boldsymbol{F}_2^{lf}$ 就需要确定混合图像 I_2 每个位置上的灰度值,则 $P=1/256^{m\times n-a\times b}$,又因为 $P>P_G$,则

$$1/256^{m\times n-a\times b}>1/2^{62}>1/2^{64},$$

$$8\times(m\times n-a\times b)<64,$$

$$a\times b/m\times n>(m\times n-8)/m\times n.$$

即攻击者拥有的灰度值占所有灰度值的百分比大于 $(m\times n-8)/m\times n$ 时才可以以 $P>P_G$ 的概率得到 p_2 通过所有权认证,假设图像尺寸为 640×480,则该百分比的值为 99.9974%,此时认为攻击者已经拥有图像.

攻击者采用方法(2e)时，假设攻击者需要猜到图像的所有灰度值，则 $P=1/256^{m\times n}$，设图像像素为 640×480，$P=1/256^{640\times480}$，远小于 P_G，与假设的 $P>P_G$ 矛盾. 证毕.

另外，方案的重复检测阈值为 $t=1$，这是权衡了方案可靠性和去重率的结果，此阈值与碰撞成功的可能性有关，阈值 t 越大，碰撞成功的可能性越高.

因此，攻击者通过所有权认证的概率 $P\leq1/2^{62}$，我们认为此概率值足够小可以忽略，攻击者不能通过所有权认证.

5　CIFD 方案性能测试与分析

本节从以下两个角度对方案进行性能对比分析：对 CIFD 和图像去重方案 SPSD[14] 进行对比；对 CIFD 和传统的文件客户端去重技术进行对比.

5.1　测试环境与测试方案

本地搭建客户端，配置参数如表 1 所示.

表 1　客户端参数配置

项目	配置
硬盘	40GB
CPU	单核
内存	1GB
操作系统	UBUNTU15.10 64 BIT

租用云服务器，配置参数如表 2 所示.

表 2　服务器端参数配置

项目	配置	项目	配置
服务器	阿里云	CPU	单核
位置	青岛	内存	4GB
服务器 ID	i-281kfyxga	操作系统	Ubuntu 14.04 64BIT
硬盘	40GB	带宽	5MBPS

使用的阿里云服务器部署于青岛，由百度地图可以测得客户端所在地西安与云服务器所在地青岛的距离为 1058.6 km，如图 4 所示.

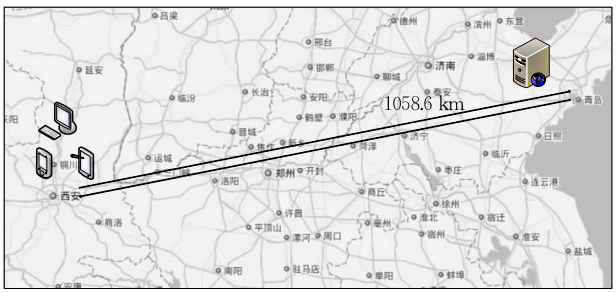

图 4　测试方案总体示意图

服务器端和客户端测试程序均由 C++语言编写完成，图像处理使用 OpenCV2.4.12 库.

Quality Image 数据库中的图像是符合特定质量标准的图像，其中囊括了各类图像且每类图像包含了许多内容有少量差异的图像，这些图像在检索领域里被认为是相似的，但不属于 CIFD 方案的范围，因此在去重场景下应该被认为是不重复图像.

LIVE[21-22] 数据库是由 LIVE 实验室建立的，TID2013[23] 数据库是 Ponomarenko 等人建立的，这两个数据库常用于图像质量评价，因其包含的参考图像多，每类失真中包含着大量的失真图像，所以成为图像研究领域的权威主流数据库，得到了广泛使用，这两个数据库中的失真图像符合 CIFD 方案对重复图像的范围设定.

5.2　CIFD 方案与 SPSD 方案对比分析

CIFD 和 SPSD 方案特点总体对比如表 3 所示.

表 3　CIFD 和 SPSD 方案特点总体对比

方案	重复检测	所有权认证	质量比较
SPSD	有	无	无
CIFD	有	有	有

主要对 CIFD 方案和 SPSD 方案从重复检测阶段、所有权认证阶段进行测试和对比分析.

5.2.1　重复检测阶段

CIFD 和 SPSD 两种方案重复检测阶段的核心都是判断图像相似性的感知哈希算法，CIFD 采用 DAN-phash 算法，SPSD 采用均值哈希算法(AVG-phash). 用两种感知哈希算法进行重复检测的特点总体对比如表 4 所示(其中时间是 20 组去重中 DAN-phash 和 AVG-phash 时间开销的平均值，详细对比在后文中一一介绍).

表 4　CIFD 和 SPSD 重复检测特点总体对比

方案	时间均值/s	错误率	去重率
SPSD	0.0005	高	低
CIFD	0.0011	低	高

(1) 重复检测所用感知哈希算法时间对比

对 LIVE 数据库中的快速衰落，高斯模糊，JPEG2000 压缩，JPEG 压缩四种失真类型进行时间测试，这几种失真类型的测试是类似的，由于数据太多，这里只展示 JPEG 失真的测试结果. 选取 20 组内容不同的 JPEG 失真图像进行 20 组去重，每组去重分别测试 DAN-phash 和 AVG-phash 的时间开销，结果如表 5 所示，根据表 5，比较两种方案中感知哈希算法的时间开销，结果如图 5 所示.

表 5 CIFD 和 SPSD 重复检测阶段感知哈希算法时间开销

去重测试编号	时间/ms		去重测试编号	时间/ms	
	AVG-phash	DAN-phash		AVG-phash	DAN-phash
1	0.888	1.669	11	0.448	0.846
2	0.644	1.722	12	0.476	1.090
3	0.558	1.283	13	0.478	0.792
4	0.544	1.148	14	0.429	1.272
5	0.537	1.193	15	0.426	1.416
6	0.536	1.605	16	0.450	1.460
7	0.502	1.341	17	0.413	0.805
8	0.512	1.249	18	0.402	0.848
9	0.512	1.058	19	0.377	0.722
10	0.530	1.138	20	0.377	0.738

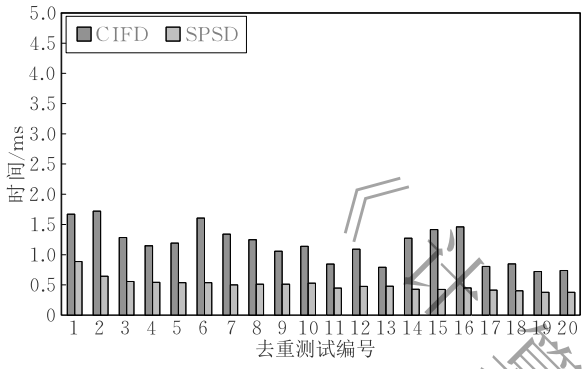

图 5 CIFD 和 SPSD 重复检测感知哈希算法时间对比

根据图 5 可得，DAN-phash 比 AVG-phash 时间开销稍大，但是最多大不到 1ms，这个时间对于整个去重过程而言非常小.

(2) 重复检测错误率和去重率对比

重复检测的错误率主要针对不相似图像，测试其是否会被判断为重复导致方案出现严重错误，测试对象包括两种：完全不相似的图像，选择 LIVE 和 TID2013 数据库的参考图像，这些图像应该被判断为不重复；有针对性的不相似图像，选择 Quality Image 中的内容有少量不同的图像，这些图像应该被判断为不重复，其具有的特点非常适用于测试算法的错误率. 去重率主要针对相似图像，测试其能否准确地被判断为重复，测试对象选择 LIVE 和 TID2013 数据库中的失真图像，这些图像应该被判断为重复.

错误率测试一方面选择 LIVE 和 TID2013 数据库的参考图像，测试结果显示所有图像都被判断为不重复，没有出错. 另一方面，从 Quality Images 数据库选取 100 张同类图像作为一个数据集，分别采用 AVG-phash、DCT-phash 和 DAN-phash 算法对这 100 张图像各计算 100 个感知哈希值，用汉明距离衡量这 100 个感知哈希值之间的相似度，按照上述步骤一共重复选取 10 个不同的数据集，所有参与该测试的图像共有 1000 张，3 个算法各产生 49500 个汉明距离. 分别计算 AVG-phash、DCT-phash 和 DAN-phash 算法产生的 49500 个汉明距离中小于阈值 t 的汉明距离数量，计算其占总共的汉明距离的比例，对阈值 $t=1$ 和 $t=5$ 分别进行上述测试，得到 3 个算法的错误率，结果如表 6、表 7 所示.

表 6 阈值 $t=5$ 的重复检测结果

算法	错误率/%	去重率/%
DCT-phash	0.810	93.39
SPSD (AVG-phash)	0.970	81.31
CIFD (DAN-phash)	0.006	95.71

表 7 阈值 $t=1$ 的重复检测结果

算法	错误率/%	去重率/%
DCT-phash	0.036	57.66
SPSD (AVG-phash)	0.040	29.98
CIFD (DAN-phash)	0	69.39

去重率测试选择 LIVE 数据库中的 29 张参考图像在失真作用下产生的 982 张失真图像，TID2013 数据库中的 25 张参考图像在失真作用下产生的 3000 张失真图像，分别采用 AVG-phash、DCT-phash 和 DAN-phash 算法计算所有图像的感知哈希值，对每种失真类型分开处理，用汉明距离衡量内容相同但质量不同的图像的感知哈希值之间的相似度，所有参与该测试的图像共有 3982 张，3 个算法各产生 8907 个汉明距离. 分别计算 AVG-phash、DCT-phash 和 DAN-phash 算法产生的 8907 个汉明距离中小于阈值 t 的汉明距离数量，计算其占所有的汉明距离的比例，对阈值 $t=1$ 和 $t=5$ 分别进行上述测试，得到 3 个算法的去重率，如表 6、表 7 所示.

由表 6、表 7 可得，采用 AVG-phash、DCT-phash 和 DAN-phash 算法进行重复检测的去重率依次提高，错误率依次降低，阈值为 1 时 DAN-phash 算法的错误率已经降低到 0，其较于其他两个算法的优势显而易见. 同时，阈值减小会使得重复检测更加严格，虽然这会使得方案去重率有所降低，但是随之而减小的错误率可以使得方案更加可靠.

另外，错误率从客户端的角度而言代表了方案的可靠性，从攻击者的角度而言代表了方案的抗攻击能力，SPSD 方案重复检测时会将不相似图像判断为重复，即攻击者若用大量 Quality Image 中的图像进行重复检测，在阈值 $t=1$ 时有 0.04% 的可能会通过，在阈值 $t=5$ 时有 0.97% 的可能会通过，最终

攻击者可能会得到服务器端图像的所有权，而 CIFD 方案在阈值 $t=1$ 完全不会受到攻击者的威胁.

5.2.2　所有权认证阶段

对 LIVE 数据库中的快速衰落、高斯模糊、JPEG2000 压缩、JPEG 压缩四种失真类型各选取 20 张内容不同的图像，用 CIFD 方案进行 80 组去重. 测试 80 组去重中所有权认证阶段的时间开销，结果如表 8 所示. 然而，SPSD 方案没有涉及相似图像的所有权认证，客户端仅需要通过相似性检测就可以进行去重. 根据表 8，分析 CIFD 方案所有权认证阶段的时间开销范围，结果如图 6 所示.

表 8　CIFD 所有权认证阶段时间开销

去重测试编号	时间/ms	去重测试编号	时间/ms	去重测试编号	时间/ms	去重测试编号	时间/ms
1	54	21	50	41	54	61	46
2	47	22	56	42	49	62	48
3	49	23	59	43	53	63	54
4	54	24	52	44	48	64	47
5	52	25	50	45	50	65	48
6	55	26	56	46	51	66	54
7	57	27	50	47	53	67	51
8	54	28	45	48	55	68	51
9	50	29	50	49	50	69	49
10	48	30	49	50	50	70	51
11	50	31	48	51	50	71	49
12	50	32	53	52	48	72	47
13	51	33	54	53	46	73	47
14	49	34	47	54	46	74	47
15	43	35	45	55	41	75	45
16	90	36	48	56	50	76	44
17	67	37	45	57	43	77	49
18	49	38	44	58	53	78	48
19	48	39	47	59	51	79	47
20	52	40	45	60	46	80	48

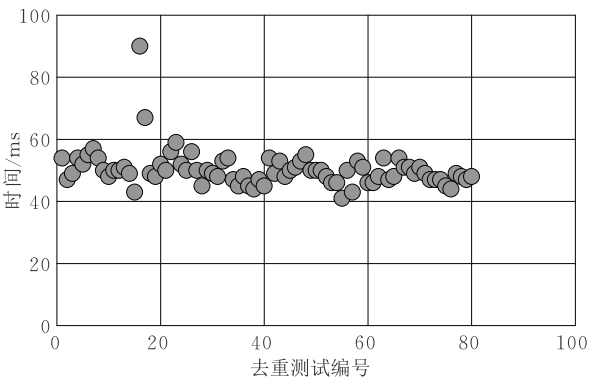

图 6　CIFD 所有权认证时间开销范围

由图 6 可得，所有权认证阶段时间开销基本在 40～60 ms 之间，在整个方案中只占一小部分时间，但是却起了很重要的作用. 采用 VSP 认证协议可以准确验证客户端对图像的所有权，防止恶意攻击者的非授权访问，而这部分工作在之前的相似图像客户端去重研究中都没有做到.

5.2.3　质量比较阶段

SPSD 方案不包括质量评价和比较，因此本文主要测试 CIFD 方案质量评价和比较的准确性. 由于 CIFD 方案主要针对 JPEG2000 压缩、JPEG 压缩、高斯模糊和快速衰落四种类型的失真，而 LIVE 和 TID2013 数据库都拥有 JPEG2000 压缩、JPEG 压缩和高斯模糊这三种失真，为了方便将 LIVE 数据库与 TID2013 数据库的测试结果进行对比，本文只针对两个数据库共同拥有的三种失真类型进行测试分析. 使用斯皮尔曼等级次序相关系数(SROCC)来对 PQA 算法评价结果与数据库中图像主观评分的一致性进行衡量. PQA 算法的特点在于其可以同时进行失真鉴别和质量评价，其与 BRISQUE 算法的评价准确度是相同的. PQA 算法产生的质量分数越小，说明图像质量越高. TID2013 数据库用平均主观得分(Mean Opinion Score，MOS)表示主观图像质量，MOS 值越大，说明图像质量越高. LIVE 数据库用平均主观得分差异(Differential Mean Opinion Score，DMOS) 表示主观图像质量，DMOS 值越小，说明图像质量越高.

TID2013 数据库中有 24 张自然参考图像，对 JPEG2000 压缩、JPEG 压缩和高斯模糊三种失真，分别选取 24 张自然参考图像产生的 120 张失真图像，共 360 张图像. PQA 算法评价结果与 MOS 值的相关性测试结果如表 9 所示.

表 9　LIVE 和 TID2013 数据库上的 SROCC

数据库	JPEG 压缩	JPEG2000 压缩	高斯模糊
LIVE	0.9572	0.9441	0.9753
TID2013	−0.9065	−0.9130	−0.8945

LIVE 数据库有 29 张参考图像，选取 JPEG2000 压缩图像 227 张，JPEG 压缩图像 233 张，高斯模糊图像 174 张，共 604 张图像. PQA 算法评价结果与 DMOS 值的相关性测试结果如表 9 所示.

从表 9 中可得，PQA 算法具有较小的数据库依赖性，其在 LIVE 和 TID2013 数据库中的 SROCC 皆能保持在 0.9 左右，这说明 PQA 算法评价结果与两个数据库中图像主观评分的一致性较高.

5.3　CIFD 方案与传统文件去重方案对比分析

由于 CIFD 方案与传统文件去重方案不存在从每个阶段进行对比的必要性，这里主要从整体方案的节约资源角度进行对比. CIFD 方案去重时主要会出现以下四种情况：服务器端存在相同图像(same)；服务器端不存在相同图像也不存在相似图

像(no similar);服务器端存在一个相似图像且服务器端图像质量更好(server better);服务器端存在一个相似图像且客户端图像质量更好(client better).

从客户端和服务器角度,分别测试以上四种情况 CIFD 方案对 LIVE 数据库中的快速衰落、高斯模糊、JPEG2000 压缩、JPEG 压缩四种失真类型进行去重的总时间开销,这里只展示 JPEG 失真的测试结果,如表 10、表 11 所示,根据表 10、表 11,比较 same、server better、no similar、client better 四种情况的总时间开销,结果如图 7、图 8 所示.

表 10 CIFD 客户端四种情况总时间开销

去重测试编号	图像大小/MB	时间/s			
		same	server better	no similar	client better
1	1.12	0.246	0.281	2.145	2.028
2	1.12	0.271	0.301	2.356	2.180
3	1.12	0.272	0.311	2.139	2.014
4	1.12	0.255	0.289	1.929	2.173
5	1.12	0.235	0.289	2.536	2.166
6	1.12	0.251	0.294	2.537	2.403
7	1.12	0.247	0.300	2.336	2.401
8	1.12	0.245	0.294	2.114	2.417
9	1.12	0.265	0.297	2.545	2.184
10	1.12	0.223	0.288	2.169	2.395
11	0.98	0.233	0.286	1.883	1.550
12	0.98	0.251	0.290	1.907	2.172
13	0.94	0.245	0.281	1.505	1.739
14	0.94	0.250	0.282	1.921	1.544
15	0.92	0.239	0.271	1.892	1.757
16	0.91	0.237	0.273	1.495	1.742
17	0.87	0.224	0.277	1.481	1.531
18	0.85	0.217	0.267	1.699	1.139
19	0.80	0.237	0.271	1.280	1.520
20	0.79	0.228	0.252	0.856	1.541

表 11 CIFD 服务器端四种情况总时间开销

去重测试编号	图像大小/MB	时间/s			
		same	server better	no similar	client better
1	1.12	0.144	0.191	3.373	3.457
2	1.12	0.160	0.193	3.416	3.417
3	1.12	0.158	0.200	3.324	3.380
4	1.12	0.161	0.187	3.322	3.397
5	1.12	0.138	0.193	3.351	3.394
6	1.12	0.151	0.189	3.361	3.434
7	1.12	0.141	0.199	3.354	3.423
8	1.12	0.142	0.190	3.669	3.438
9	1.12	0.136	0.195	3.573	3.416
10	1.12	0.135	0.194	3.396	3.580
11	0.98	0.142	0.190	3.071	2.996
12	0.98	0.158	0.187	3.095	3.154
13	0.94	0.140	0.191	2.928	2.973
14	0.94	0.155	0.178	2.936	2.979
15	0.92	0.123	0.184	2.872	2.773
16	0.91	0.129	0.178	2.734	2.765
17	0.87	0.131	0.183	2.716	2.752
18	0.85	0.131	0.172	2.723	2.742
19	0.80	0.137	0.175	2.510	2.538
20	0.79	0.127	0.157	2.492	2.553

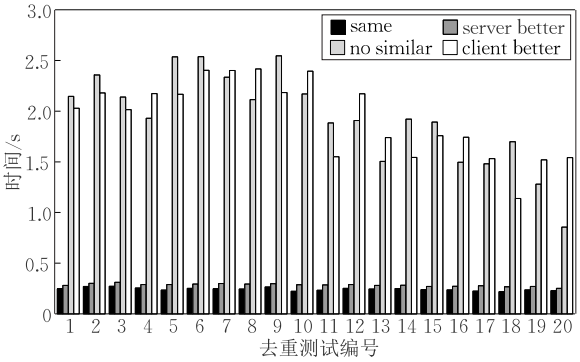

图 7 CIFD 客户端四种情况总时间对比

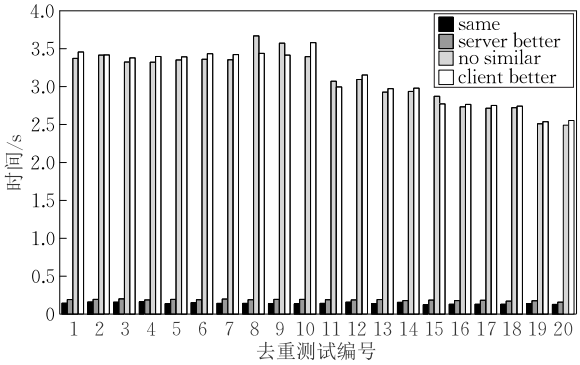

图 8 CIFD 服务器端四种情况总时间对比

根据图 7 和图 8 可得,用 CIFD 方案进行去重,“服务器端存在相同图像”和“服务器端存在相似图像且服务器端图像质量更好”这两种情况下客户端不需要上传图像,总时间开销(协议交互)在 0.2s 左右,“服务器端不存在相似图像”和“服务器端存在相似图像且客户端图像质量更好”这两种情况下客户端需要上传图像,总时间开销(协议交互和上传图像)在 2~3s 左右.由此可见,CIFD 方案中协议交互的时间很小,图像去重(两种情况不用上传)比不去重(所有情况都上传)节约时间的多少主要取决上传图像耗时的多少,即取决于图像本身的大小.

对于传统的文件去重,一般方案的协议交互耗时也远小于上传文件耗时,文件去重(后续上传者不用上传)比不去重(所有情况都上传)节约时间的多少主要取决上传文件耗时的多少,即取决于文件本身的大小.

因此,图像去重和文件去重节约时间的多少都取决于图像或文件本身的大小,而常用的图像一般都比文件大,图像去重可以更大程度上节约时间,并且在多媒体数据成为主流的互联网中,图像重复数量比文件多,所以,图像去重比传统的文件去重能节约更多资源.

综上所述,可得:(1)与 SPSD 方案相比,CIFD

方案的重复检测错误率更低、抗攻击能力更强，去重率更高；CIFD 方案提出的 VSP 协议可以进行相似图像所有权认证，并且不增加过多计算和时间开销，这是 SPSD 方案没有涉及到的部分；CIFD 方案增加了质量比较，可以保证质量好的图像保留在服务器端.(2)与传统文件去重相比，CIFD 方案可以更大程度上节约网络带宽和存储空间.

6　结　论

由于图像具有数量大、冗余多等特性，因此针对图像的客户端重复数据删除技术是去重领域很具研究价值的新问题.然而图像的客户端重复数据删除存在诸多难题：如何对相似图像进行模糊的重复检测；如何解决相似图像的所有权认证；如何对图像感知质量进行评价.针对上述问题，在研究已有方案的基础上，本文提出了一种支持所有权认证的客户端图像模糊去重方法 CIFD.方案的核心思想是，通过 DAN-phash 算法对客户端图像进行高准确度的模糊重复检测，发现重复后采用 VSP 协议进行相似图像的所有权认证，认证通过后通过 PQA 算法对图像进行感知质量评价，最终将感知质量最好的图像保留在服务器端，删除其余图像.本文对 VSP 协议进行了安全性证明，证明其达到了可证明的安全强度.并且，本文进行了大量测试和性能分析，结果表明，CIFD 方案可以对多种失真类型的相似图像安全高效地去重，节约了大量存储空间和网络带宽，并能在服务器端保留感知质量最好的图像，具有很高的应用价值.另外，在未来的工作中，我们将着力于研究可以保护数据机密性的客户端图像去重方法.

致　谢　感谢国家自然科学基金和移动互联网安全 111 创新引智基地基金的支持.感谢计算机学报编辑和审稿专家的宝贵意见！

参　考　文　献

[1] Pietro R D, Sorniotti A. Boosting efficiency and security in proof of ownership for deduplication//Proceedings of the 7th ACM Symposium on Information, Computer and Communications Security. Seoul, Korea, 2012: 81-82

[2] Zheng Q, Xu S. Secure and efficient proof of storage with deduplication//Proceedings of the 2nd ACM Conference on Data and Application Security and Privacy. New York, USA, 2012: 1-12

[3] Yang C, Ren J, Ma J. Provable ownership of files in deduplication cloud storage//Proceedings of the IEEE Global Communications Conference. Atlanta, USA, 2013: 2457-2468

[4] Joly A, Buisson O, Frelicot C. Content-based copy retrieval using distortion-based probabilistic similarity search. IEEE Transactions on Multimedia, 2007, 9(2): 293-306

[5] Ming C, Wang S, Yun X, et al. FAIDA: A fast and accurate image deduplication approach. Journal of Computer Research and Development, 2013, 50(1): 101-110

[6] Halevi S, Harnik D, Pinkas B, et al. Proofs of ownership in remote storage systems//Proceedings of the 18th ACM Conference on Computer and Communications Security. Chicago, USA, 2011: 491-500

[7] Katiyar A, Weissman J. ViDeDup: An application-aware framework for video de-duplication//Proceedings of the 3rd USENIX Conference on Hot Topics in Storage and File Systems. Portland, USA, 2011: 31-35

[8] Yang X, Su G, Chen J, et al. Large scale identity deduplication using face recognition based on facial feature points//Proceedings of the 6th Chinese Conference on Biometric Recognition. Beijing, China, 2011: 25-32

[9] Xu J, Zhang W, Ye S, et al. A lightweight virtual machine image deduplication backup approach in cloud environment//Proceedings of the 38th IEEE Signature Conference on Computers, Software and Applications. Vasteras, Sweden, 2014: 503-508

[10] Ramaiah N P, Mohan C K. De-duplication of photograph images using histogram refinement//Proceedings of the IEEE International Conference on Recent Advances in Intelligent Computational Systems Recent Advances in Intelligent Computational Systems. Trivandrum, India, 2011: 391-395

[11] Chen M, Wang Y, Zou X, et al. A duplicate image deduplication approach via Haar wavelet technology//Proceedings of the 2nd IEEE International Conference on Cloud Computing and Intelligent Systems. Hangzhou, China, 2012: 624-628

[12] Gang H, Yan H, Xu L. Secure image deduplication in cloud storage. Information and Communication Technology. Cham, Switzerland: Springer, 2015, 9357: 243-251

[13] Rashid F, Miri A, Woungang I. Secure image deduplication through image compression. Journal of Information Security and Applications, 2016, 27: 54-64

[14] Li X, Li J, Huang F. A secure cloud storage system supporting privacy-preserving fuzzy deduplication. Soft Computing, 2015, 20(4): 1437-1448

[15] Zauner C. Implementation and benchmarking of perceptual image hash functions. Revista Musical Chilena, 2011, 65(215): 71-72

[16] Kalker T, Haitsma J, Oostveen J C. Issues with digital watermarking and perceptual hashing//Proceedings of the SPIE 4518 Multimedia Systems and Applications IV. Denver, USA, 2001: 189-197

[17] Fridrich J, Goljan M. Robust hash functions for digital watermarking//Proceedings of the IEEE International Conference on Information Technology: Coding and Computing. Las Vegas, USA, 2000: 178-183

[18] Hamming R W. Error detecting and error correcting codes. Bell Labs Technical Journal, 1950, 29(2): 147-160

[19] Bloch D. A note on the estimation of the location parameters of the Cauchy distribution. Journal of the American Statistical Association, 1966, 61(316): 852-855

[20] Ferguson T S. Maximum likelihood estimates of the parameters of the Cauchy distribution for samples of size 3 and 4. Journal of the American Statistical Association, 1978, 73(361): 211-213

[21] Wang Z, Bovik A C, Sheikh H R, et al. Image quality assessment: From error visibility to structural similarity. IEEE Transactions on Image Processing, 2004, 13(4): 600-612

[22] Sheikh H R, Sabir M F, Bovik A C. A statistical evaluation of recent full reference image quality assessment algorithms. IEEE Transactions on Image Processing, 2006, 15(11): 3440-3451

[23] Ponomarenko N, Jin L, Ieremeiev O, et al. Image database TID2013: Peculiarities, results and perspectives. Signal Processing Image Communication, 2015, 30: 57-77

[24] Wang Z, Bovik A. Modern Image Quality Assessment. New York, USA: Morgan and Claypool, 2006

[25] Zhai G, Kaup A. Comparative image quality assessment using free energy minimization//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Vancouver, Canada, 2013: 1884-1888

[26] Mittal A, Moorthy A K, Bovik A C. No-reference image quality assessment in the spatial domain. IEEE Transactions on Image Processing: A Publication of the IEEE Signal Processing Society, 2012, 21(12): 4695-4708

[27] Zheng Yun-Ping, Chen Chuan-Bo. Study on a new algorithm for gray image representation. Chinese Journal of Computers, 2010, 33(12): 2397-2406(in Chinese)
(郑运平，陈传波. 一种新的灰度图像表示算法研究. 计算机学报，2010，33(12)：2397-2406)

[28] Gao Fei, Gao Xin-Bo. Active feature learning and its application in blind image quality assessment. Chinese Journal of Computers, 2014, 37(10): 2227-2234(in Chinese)
(高飞，高新波. 主动特征学习及其在盲图像质量评价中的应用. 计算机学报，2014，37(10)：2227-2234)

LI Dan-Ping, born in 1992, Ph. D. candidate. Her research interests include cloud computing and storage security.

YANG Chao, born in 1979, Ph. D., professor, Ph. D. supervisor. His research interests include big data and cloud computing security, mobile and smart computing security.

JIANG Qi, born in 1981, Ph. D., associate professor. His research interests include big data and cloud computing security, mobile and smart computing security.

MA Jian-Feng, born in 1963, Ph. D., professor, Ph. D. supervisor. His research interests include channel coding, cryptography, wireless and mobile security, system survival-ability.

LI Cheng-Zhou, born in 1992, M. S. candidate. His research interest is intelligent information processing.

Background

Images have the characteristics of large quantity and redundancy, large bandwidth and space needed to upload and store data. Therefore, how to deduplicate images efficiently and safely has become an important issue for urgent solution. However, there are many new challenges in the client-based image deduplication, which are different from these existing client-based file deduplication technologies. The client-based image deduplication needs to support the image fuzzy deduplication, but most of these existing client-based file deduplication technologies don't support the new challenge of fuzzy deduplication; the client-based file deduplication needs to support proof of ownership for similar images, but these existing literatures are not yet related to proof of ownership for similar images; the client-based image deduplication needs to assess image perceptual quality, but there are no mature image quality assessment method in the field of image deduplication.

Aiming at these above challenges, we proposed the CIFD (A Client-based Image Fuzzy Deduplication Method supporting Proof of Ownership). Security analysis results demonstrate that security strength of the scheme is provable, which is a breakthrough and creation in the field of the image deduplication. A large number of simulation results show that the CIFD can check image similarity accurately and assess

image perceptual quality of a variety of distortions, meeting these new technical challenges; moreover, performance test results show that time cost of the CIFD is small, saving a lot of bandwidth and space needed to upload and store duplicated data, and accomplishing image deduplication quickly and efficiently.

Our research belongs to the project of the National Natural Science Foundation, whose name is "Research on Protection of Security Attributes and Verification of Operation Fidelity of Cloud Big Data". New cloud computing model and the change of big data computing have brought an unprecedented impact to the field of information security: "data" and "computing" are exported to the "cloud", which has no fixed infrastructures and security boundary; thereby causing users lost direct control to "data" and "computing". So, there are a lot of new security issues, which need new methods and techniques to protect big data and computing in cloud. Among them, the most important security issues including the protection of security attributes and verification of operation fidelity of cloud big data.

The results of our research group in this project include: we received a grant from the National Natural Science Foundation, 630 thousand yuan; we published a number of papers, for example, a new verification method of cloud storage disaster recovery in different locations written by Zhou H C et al. and novel cloud data assured deletion approach based on ciphertext sample slice written by Zhang K et al., and so on. This paper aims at a part of our project, which is proof of ownership of encrypted data in scenario of cloud big data duplication.