UDC 62-50
DOI: 10.20535/SRIT.2308-8893.2023.4.09

SURVEY OF IMAGE DEDUPLICATION FOR CLOUD STORAGE

S. CHAUDHARI, R. APARNA

Abstract. Increased growth of real-life communication has motivated the creation, transmission, and digital storage of vast volumes of images and video data on the cloud. The explosive increase in virtual/visual image data on cloud servers requires efficient storage utilization that can be addressed using image deduplication technology. Even though the virtual and visual image properties are different, the existing literature uses a similar approach for deduplication checks, which motivated us to consider both image types for this review. This article aims to provide a detailed survey of state-of-the-art visuals as well as virtual image deduplication techniques in a cloud environment, summarizing and organizing them by developing a five-dimensional taxonomy for analysing the features and performance with several non-overlapping categories in each dimension. These include: 1) location of applying deduplication; 2) image feature extraction; 3) time of application; 4) image data partitioning strategy; 5) involvement of user dataset level. Existing image deduplication techniques are categorized into two main categories based on whether the technique involves security. A comparison of techniques is discussed across a set of functional and performance parameters. The current issues are highlighted with the possible future directions to motivate further research studies on the topic.

Keywords: image deduplication, cloud computing, cloud storage, image copy detection.

INTRODUCTION

With the massive development of electronics and the internet, digital data is increasing at an alarming rate. This includes data in the form of text, images, videos, sketches, etc. All this data comes from different parts of the Internet and hence causes information explosion due to huge velocity of data generation and huge variety of data sources. In 2007, it is said that the total digital resources of the world exceeded the global storage capacity for the very first time. Hence, it was decided that this problem of information explosion cannot be handled by simply increasing the amount of storage. But now it is estimated that by 2025, there will be 163.2 zettabytes of digital data [35].

Primary data generators like social networking platforms, industries and transactional data from various businesses are generating huge volumes of data every day. Due to the sudden increase in volumes of data, it becomes extremely crucial to be able to store this data in a cost-effective manner that optimizes storage. Cloud based infrastructure for on demand service provisioning from anywhere, anytime is the popular solution used. The National Institute of Standards and Technology (NIST) reference architecture for cloud computing has following

five actors: 1) cloud consumer; 2) cloud provider; 3) cloud carrier; 4) cloud auditor; 5) cloud broker. The interaction among these actors is shown in Fig. 1 along with their activities and functions.

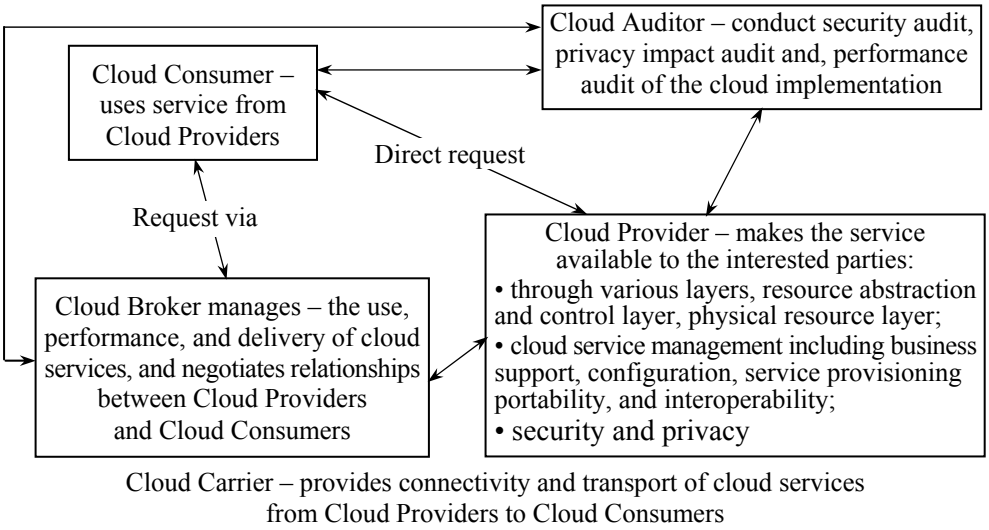

Fig. 1. Cloud Actors Activities and Interaction in Cloud

Storage as a service on Cloud is one of the critical and popular services wherein the cloud storage provider provides cost effective and easy to access storage space on the cloud to the interested customers to host their data instead of maintaining it on their on-premises. The user data stored on the cloud can be in any form like images, audio, video etc. The customers or the data owners cannot rely on public service providers like cloud for data security. So, to provide security to the data, many researchers proposed secure storage techniques for storing the data on cloud.

The cloud services are provided through virtual images whose size is very large requiring large amount of storage space in addition to huge network transmission requirements and reduction in operation time. Virtual images, each with large size starting with 1GB with different configurations may belong to a single cloud user. Almost 80% of the virtual image content is identical among these Virtual Machine (VM) images due to existence of similar data segments [1]. The two primary reasons namely sudden explosion of data and similarity in virtual images apparently induce a need in Cloud Service Providers (CSP) to optimize the storage and network bandwidth used in data transfer of VM images.

Hence, a concept called deduplication was formulated, which could identify duplicates and delete all copies except one (or precisely retain as many copies as specified by deduplication ratio) [36]. It optimally minimizes the storage utilization by deleting redundant data from the cloud storage or data centers and thereby bringing down the unnecessary usage of network bandwidth [40]. It is a lossless data compression technology, which replaces duplicate image copies by using pointers to the unique image object.

The image deduplication [40] process involves four steps to remove duplicate images or parts of images as follows [39]: 1) file chunking wherein the image is divided into fixed or variable size blocks known as chunks; 2) fingerprint generation: the fingerprint will be computed using some transformation algorithm/technique such as hash function; 3) fingerprint lookup: the fingerprint of

already existing images will be compared with this newly created fingerprint in step 2 for identifying the duplicate file/block. If it is found to be same, this new file/block will be discarded otherwise it will be stored in the cloud storage; 4) data storage: store the unique image/block systematically on the cloud storage.

Deduplication can be done at two levels: the single user level and the cross level. In the single user level, the deduplication is done keeping only one user in mind and duplication happens in their own storage. Cross level deduplication is when data is compared by taking from many users, and then redundant data is deleted. It can also be done at either the client side or the server side. At the client side, the data is checked for duplicates by the client itself and is then sent to the server. At the server side, the server collects all the information from the client and then duplicates are found and removed. Deduplication also has two feature-based methods known as global feature-based method and local feature-based method.

There is no detailed investigation done till now to review image data deduplication techniques and its characteristics. Few non-standard articles exist explaining data deduplication survey in unstructured way. The authors of [37] have classified the existing data deduplication techniques into two categories as source deduplication and target deduplication. Source deduplication is further classified into file-based and sub-file-based. Sub file is further considered as fixed or variable length. Target deduplication is further classified as post-process and inline. They have discussed another way for classifying data deduplication namely offline and online deduplication. Even though many research articles exist, the paper discusses about only 10 research articles on data deduplication. The authors of [38] discuss 14 deduplication approaches without any taxonomy or relation among them. They have included 24 research articles under these approaches and compare them in terms of scalability, throughput, efficiency, amount of used bandwidth and cost. Many comparison parameters could have been considered along with some more deduplication techniques. Lack of systematic review/survey motivated us to do this work.

In this survey paper, we survey different types of deduplications or copy detection that have been done for images in a systematic and structured way. As important as it is, traditional deduplication mechanisms can only be used if two images have the same bit stream, that is it can only be used if two images are completely the same. It does not apply for an image that has been cropped, rotated, or edited out. Automatic methods are now getting more attention with the increase in the redundant information. Also, cloud computing has proved to be very flexible and economical service provider that provide to maintain huge amount of data. In this world of immense data, users normally upload similar images in different storages either due to storage restrictions or network restrictions. The aim of this paper is to understand and observe the different techniques used for image deduplication in terms of functional and performance parameters.

Our contributions. Consequently, this significant amount of published research on Image deduplication requires some categorization to provide convenient overview of the current state of the art. To this end, we have developed multi-dimensional taxonomy to classify the Image deduplication research based on the properties supported in the research work as described in Section 2. Even though multi-dimensional taxonomy is used popularly for defining image deduplication techniques, we categorize them into two main categories based on whether secu-

rity is incorporated or not. Non-secure techniques are further categorized based on image type as virtual image or pixel image. Techniques in each category is discussed across a set of functional and performance parameters. The presented taxonomy allows us to analyze the Image deduplication research trends over time and various features supported in the work. To illustrate the usefulness of the provided classification, we discuss a detailed survey of the collected research articles from extensive databases available online where Image deduplication-based references can be explored according to the designed dimensions and categories of the presented taxonomy.

Our specific contributions are as follows: 1) design and discuss multi-dimensional taxonomies for comparison of the various parameters used in image deduplication; 2) explain the image deduplication research trends across two main categories based on whether security is considered or not. Non-secure techniques are further categorized based on image type as virtual image or pixel image; 3) compare the discussed image deduplication schemes in each category in terms of functional and performance parameters.

The remaining part of this article is structured in various sections as follows. Section 2 explains the methodology for creating the taxonomy of Image deduplication research work with its dimensions and categories. It also explains the Image deduplication-related research articles to analyze and provide trends on the distribution across the proposed dimensions. Section 3 presents a detailed survey of the key research findings and related comparison with respect to a set of functional and performance parameters related to Image deduplication. Section 4 addresses the scope of the research on Image deduplication. Finally, conclusions are drawn in Section 5.

DESIGN OF IMAGE DEDUPLICATION TAXONOMY AND CLASSIFICATION

The taxonomy is aimed at classifying the work carried out in Image deduplication to have an in-depth understanding of the topic. Taxonomy construction varies from topic to topic, but all works in one class given in the taxonomy should be similar in the features or properties. The classification categories should be non-overlapping with well-defined limits between them. The taxonomy designed for Image deduplication related research for analyzing the features and performance includes five dimensions with several non-overlapping categories in each dimension. These include: 1) location of applying deduplication; 2) image feature extraction; 3) time of application; 4) image data partitioning strategy; 5) involvement of user dataset level.

Each dimension consists of a set of categories used to classify the existing image deduplication related articles. The presented taxonomy allows us to analyze the image deduplication research trends over time and various features supported in the work. A given article may not be mutually exclusive to the category as it may belong to one or more categories per dimension. The illustration of image deduplication taxonomy in graphical form is shown in Fig. 2. We have tried to minimize the possible overlap between the existing image deduplication techniques as per the proposed dimensions in this early stage of defining the classification categories.

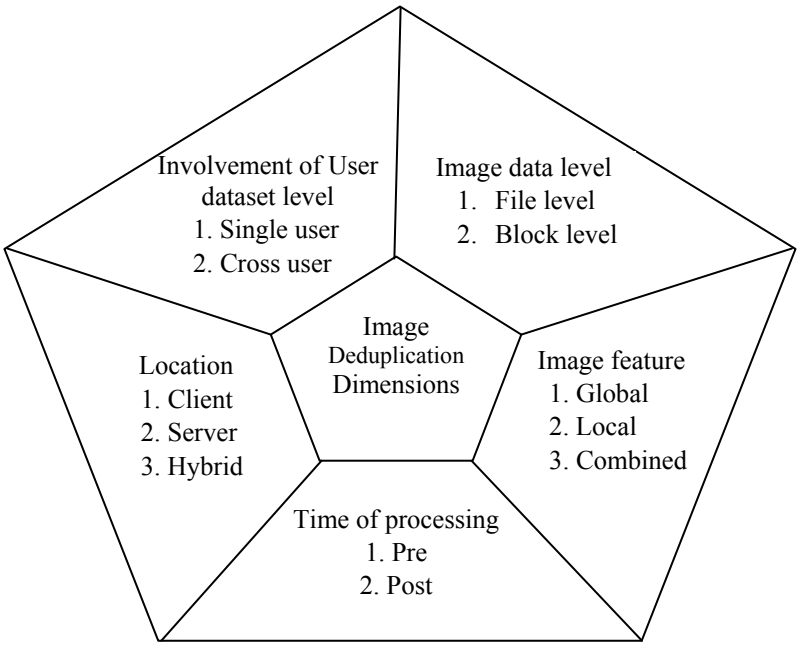

Fig. 2. Taxonomy of Image Deduplication Techniques

The first dimension namely location of deduplication in proposed image deduplication taxonomy further classifies the existing research works into three categories depending upon the place where deduplication is carried out. Since all the papers are based on client server architecture, cloud being the serving platform, the process of deduplication can be executed at the client side, server side or partly in both the places. We categorize the image deduplication techniques with respect to location of deduplication dimension into three classes as explained next: 1) server-side image deduplication: users upload the images to the cloud server in server-side image deduplication category and then the cloud service provider will perform the image deduplication check on its cloud storage to identify whether newly arrived image already exists on the server or not; 2) client-side image deduplication: client will identify the existence of similar data on the cloud server before sending it entirely to the cloud in client-side image deduplication category. Server-side image deduplication reduces computational cost at the client side but with high bandwidth requirements; 3) hybrid location-based image deduplication: Image deduplication check may be partially done at client side whereas remaining check will be done at the server side in hybrid image deduplication category of this dimension.

Second dimension named as image feature based, as proposed in image deduplication taxonomy further classifies the existing research works into three categories depending upon the usage of local and global features of images for identification of deduplication. Image features are numerical values extracted from images that are used as discriminating information to distinguish various images or parts of images. Features are extracted for reducing the processing overhead as they are small when compared to image data. Global features of image describe the whole image to generalize the image data while local image features describe small group of pixels in image. Combination of both improves the accuracy of image recognition with the side effect of computational overhead. We categorize the image deduplication techniques in image feature-based dimensions

into three classes namely: 1) global feature-based image deduplication: global features of an image are used to identify the image deduplication; 2) local feature-based image deduplication: local features of an image are used to identify the image deduplication; 3) combined feature-based image deduplication: local and global features of an image are used to identify the image deduplication.

As per the proposed image deduplication taxonomy, third dimension identified as Time of duplicate removal processing, further classifies the existing research works into two categories depending upon the time at which the duplicate data is removed for identified deduplicated images, as explained next: 1) inline image deduplication processing: the identification of deduplication is immediately started when cloud server receives the image without storing it. The deduplicated image/block of image is deleted before storing for achieving unique image data copy; 2) post-image deduplication processing: the received image will be stored in buffer on the cloud server first, then the deduplication check will be performed to identify the duplicate image/block of image. Only the unique images/blocks will be stored on the cloud server database/storage.

This dimension namely Time of duplicate removal with respect to virtual image deduplication can be categorized into three categories namely deduplication before backup, deduplication during backup and deduplication after backup. In the first case namely deduplication before backup, duplicate check is done before performing the backup operation so that the size of the data transmitted would be that of the compressed image size. Here, both the fingerprint calculation and index lookup operation must be performed by the host node. In the second case namely deduplication after backup, deduplication check is performed after backing up the image. Since whole image is transmitted, the data transmission size would be large. In this case, storage node is the location for the fingerprint calculation and index lookup operation. The third case namely deduplication operation during backup aims at balancing the resource overhead at both the host side and storage side.

Fourth dimension named as Image data level in the proposed image deduplication taxonomy further classifies the existing research works into three categories depending upon the whole image or part of image being used for identification of duplicates. The categories in this dimension are given as follows: 1) file-level image deduplication: the same image existing on the cloud server will be checked using the hash value created for each file based on the specific hash function. If the received image hash value and one of the existing image hash values is same, then the received image will not be stored otherwise it will be stored on cloud server database; 2) block level image deduplication: the received image will be divided into blocks. Hash value is calculated for each block using specific hash function. The hash value for the block is called as block fingerprint. Only one block will be stored on cloud server for two or more blocks with same fingerprint. Otherwise, all blocks are stored on the cloud server; 3) hybrid-level image deduplication: both file – level and block level hash are checked for image deduplication check.

Fifth dimension named as involvement of user dataset level in proposed image deduplication taxonomy further classifies the existing research works into two categories depending upon the usage of user databases being scanned for checking identical images. Dataset used for checking image deduplication may belong to specific user or may have permission to store data of multiple users. We cate-

gorize the image deduplication techniques based on involvement of user dataset level dimensions into two classes namely: 1) single user level image deduplication: image dataset belonging to a user is scanned to check the duplicate images for that user alone; 2) cross-user level image deduplication: Image databases of multiple users are scanned to check the duplicate image. Even though cross user level image deduplication generates higher deduplication ratio and is attractive in terms of storage cost in comparison with single user level image deduplication, it affects the privacy and security concern for users.

Even though multi-dimensional taxonomy is used popularly for defining image deduplication techniques in the literature in a scattered way, we categorize them into two main categories based on whether security is incorporated or not. Non-secure techniques are further categorized based on image type as virtual image or pixel image as shown in Fig. 3.

LITERATURE SURVEY ON IMAGE DEDUPLICATION TECHNIQUES

This section discusses the two main categories designed in our taxonomy as shown in Fig.3 based on whether the techniques incorporate security or not. Non-secure approaches are further categorized for virtual image or pixel image types. Non-secure image deduplication techniques are described in Section 3.1 and Section 3.2 for virtual image types and pixel types respectively while Section 3.3 discusses all secure techniques.

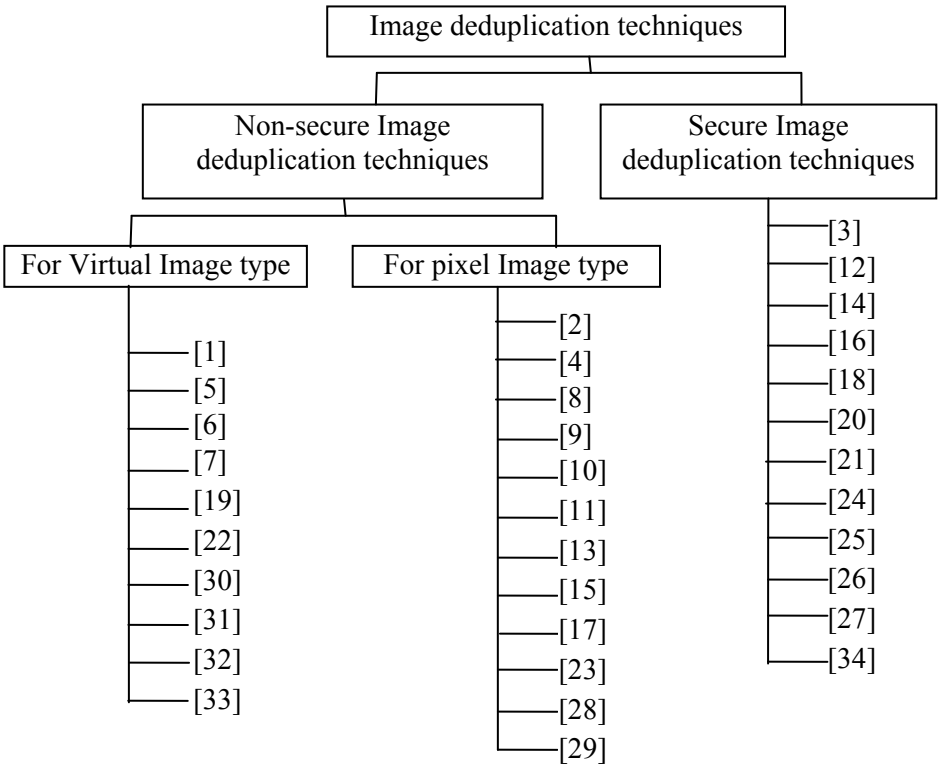

Fig. 3. Proposed Taxonomy for Image Deduplication Techniques

As the proposed image deduplication techniques are not mutually exclusive to any specific dimension, one technique may belong to one or more dimensions.

They are compared in terms of functionality based on the dimensions and performance parameters in the respective Section. Table 1 provides the functionality based on the dimensions and performance parameters used for comparison of the image deduplication techniques.

NON-SECURE IMAGE DEDUPLICATION FOR VIRTUAL IMAGE TYPE

Virtualization is a very important part of cloud computing. It allows multiple servers running on a single host and all disk contents are encapsulated in a single Virtual Machine (VM) Image. But this mechanism can store redundant data and cause storage issues. A lightweight virtual machine image deduplication backup approach in cloud environment is a technique used to eliminate this problem [1]. The process includes dividing the VM image into chunks and checking if the chunk has a fingerprint. The fingerprint is compared with the existing fingerprints in the fingerprint index table and if it exists then it is not entered otherwise the chunk is added in the storage system. The two problems faced are as follows: 1) if the fingerprint index table is long then it will take longer to compare the fingerprints; 2) the process can interrupt the other processes as different virtual machines take the same runtime. This paper gives a classification method to reduce the fingerprint search by converting global duplication to local duplication and to improve index lookup. Two sampling methods are used to find the proper group to perform the deduplication operation in the virtual machine image. A numerical method is also used to calculate the ample space size. Deduplication rate is 10.2%.

Table 1. Comparison parameters for image deduplication techniques

SN	Functional parameter	Remark
1	Matching Algorithm Used	Scale invariant feature transform (SIFT), principal component analysis (PCA) for SIFT, min-hash algorithm, feature extraction, high dimension indexing, accuracy optimization, centroid selection, deduplication evaluator, mean median, standard deviation, hash, map reduce, CRC, pixel based, special layout, visual similarity
2	Location	Client side, server side, hybrid
3	Processing time	Post-process, inline
4	Feature-based	Local, global, structural features, visual model features, and feature points
5	Image data level	File, Block,Hybrid
6	Block size	Fixed length, variable length
7	Image content type	Pixel image (PI) or Virtual Machine Image (VMI)
8	Metadata overhead	The extra information required to be stored along with actual image data
9	Optimization objective	The goal of the proposed image deduplication technique
Cloud environment parameter		
10	Cloud/OS - Cloud software	If any specific cloud software used
11	Cloud type	Public, private and hybrid cloud
12	Number of cloud images	Finite number of images existing in the cloud database
13	User level	Single user or cross user
14	Security provisioning Protocol	Usage of cryptographic protocol, adaptation of cryptographic protocol

An improved k-means clustering method is implemented in Clustering-based acceleration for virtual machine image deduplication in the cloud environment [5]. This is the first paper to have image layout taken into consideration and propose the method of small group merging and periodical triggering to store the virtual machine deduplication. Experimental results show the robustness and efficiency of this method. Deduplication rate is 89.74%.

The number of virtual machine and images grows very rapidly and takes a lot of storage space out of which 90% of the data is redundant. The storage problems caused is studied an Improved Image File Storage Method Using Data Deduplication [6] and discussions about employing deduplication and evaluation. A reference count for image is added to show reliability of image libraries.

IM-dedup proposed in [7] transmits the unique blocks of image to the cloud server for reducing transmission time. The kernel file system with deduplication functionality in the image storage helps to manage the duplicated blocks through indexing. Client and server communicate within each other during the process of image deduplication.

Deduplication-Enabled P2P based VM Image distribution protocol is introduced to speed up the provisioning in the VM [19]. Peer 1 contacts a tracker which sends it the list of its peers which also has an image of file A, it also shows the similarity Matrix between two peers.

Scalable read/write throughput in RAID with deduplication capability to Ext4 file system is proposed in [22] as deduplication file system named as ScaleDFS. Parallel processing for fingerprints computation on multiple CPU core improves the write throughput. Deduplication cache improves read throughput and retrieve identical blocks easily. Reduced memory usage cache more fingerprint information in memory. Deduplication is focused for single storage partition file system.

Authors of [30] have proposed an adaptive deduplication mechanism, which performs fixed length and variable length block-level deduplication for reducing VM disk image file size is used in variable length block-level deduplication is implemented using Rabin–Karp rolling hash algorithm. Multithreading in AKKA framework is used to perform the deduplication and streaming since live migration of VM disk image files is a bulky operation.

QuickDedup [31] algorithm is proposed by authors to perform optimal deduplication of VM disk images by reducing the number of hash computations and comparisons and by storing minimal metadata thereby reducing the overall deduplication time. In this approach, a novel byte comparison scheme to create various categories of blocks so that further the QuickDedup algorithm performs the calculation of hashes and their comparisons within the respective categories only. Hence, hash storage space is minimized and comparing within categories speeds up the deduplication of VM disk images much.

Authors of [32] propose a highly parallel deduplication cluster (HPDV) which optimizes VM images by considering the foreground quality of VM services and the background performance of deduplication for VM images. Generally, chunk-based deduplication process involves four sub processes namely chunking, fingerprinting, fingerprint indexing and data storing. Authors of HPDV have parallelized the chunking, fingerprinting tasks which are compute-intensive and fingerprint indexing, which is I/O-intensive task, using the servers in the clus-

ter. Quality of the foreground VM services is ensured while parallelization is in progress by proposing a resource-aware scheduler (RAS) in this work.

Authors of [33] have done an extensive review of several deduplication strategies and come up with a deduplication algorithm for VM images called DedupCloud. The VM images are divided into blocks and stored sequentially in the preprocessing stage. Then the blocks are categorized based on hashes of blocks derived using SHA-3 hash function.

We compare the above discussed non-secure image deduplication techniques for virtual image type in the form of three table. Table 2 gives the comparison in terms of parameters related to algorithms used and perspectives of various dimensions in the image deduplication technique. The comparison parameters include algorithm used for deduplication check, location where the deduplication takes place, application of deduplication, usage of image features, granularity of image data level, block size for block level granularity, content of the image, metadata overhead and objective function of the proposed technique.

Table 2. Comparison of non-secure Image Deduplication techniques for virtual image type

Paper	Matching Algorithm	Place	Processing time	Feature-based	Data level	Block size	Metadata overhead	Optimization objective
1	Improved *k*-means clustering algorithm with fingerprint	Hybrid	Post	Hybrid	Block	Fixed	Fingerprints of fixed size chunks	In memory deduplication by using clustering and sampling of fingerprints – fingerprint search space optimization
5	Improved *k*-means clustering with statistical indices	Server	Post	Local	Block	4KB	Fingerprints of fixed size chunks	Reduce the fingerprint search space and improve the index lookup performance
6	MD5 index	Server	Post	Local	Hybrid	Fixed 4/8KB	MD5 code of entire image file, size of image file, the number of image blocks, file name, storage address of each image block and storage address of the final block	Storage space reduction of VM images
7	MD5 and SHA-1	Client	Inline	Local	Block	Fixed 4KB	Array of fingerprints and the reference counter	Reduction in VMI storage and transmission time
19	Bloom filter's hash function, Rabin fingerprinting scheme	Hybrid	Post	Global	Block	Variable	File block id, hash, and control messages	Minimize data access or transfer during VMI distribution in data centers
22	a POSIX-compliant, kernel-space driver module	Server	42 VM images of different Linux distribution	Cross	Block	Both fixed and variable	Cryptographic fingerprints of blocks, locality (hash) table that holds the full fingerprints and block numbers of the data blocks that correspond to the most recently accessed fingerprint block	Scalable read/write throughput in RAID to provide increased capacity, reliability, and performance. for storage

Table 2. Comparison of non-secure Image Deduplication techniques for virtual image type

Paper	Matching Algorithm	Place	Processing time	Feature-based	Data level	Block size	Metadata overhead	Optimization objective
30	Rabin–Karp rolling hash algorithm	Server	Inline	Cross	Block	Both fixed and variable	Fingerprints of fixed size chunks and thread related metadata	Reduction in image storage space and total migration time and improvement in deduplication rate
31	SHA-1 Hash based on byte comparisons, categorizes the blocks	Server	Post	Local block metadata	Block	Fixed	block numbers, hashes	Least number of hashes and comparisons, minimum metadata, and fast retrieval of VMs for deployment
32	Parallel fingerprint sub-indexes	Server	Post	Global	Block	Fixed	Fingerprints of fixed size chunks and thread related metadata	Parallelizing chunking and fingerprinting tasks with multiple threads to speed up the tasks and Superior throughput with minimum interference on the foreground VM services
33	SHA-1 Hash based on byte comparisons, categorizes the blocks	Server	Post	Local block metadata	Block	Variable	Fingerprints of file chunks	Time required for the deduplication of VMI and storage of VMI and metadata

The performance analysis environment is discussed in Table 3 in terms of cloud software used, type of cloud, number of images in the cloud dataset, and user level involvement for accessing this database. Table 4 gives the advantages and disadvantages of the corresponding method.

Table 3. Cloud type/ environment Parameters for non-secure Image Deduplication techniques

Paper	Cloud/OS – Cloud software	Cloud type	Number of cloud images	User level
1	VM - Amazon EC2	Private	584 VMI	Single
5	Aliyun - largest cloud of China and ISCAS – own cloud	Aliyun-Public ISCAS - private	Aliyun-Variable ISCAS- 584	Cross
6	Own cloud on a PC	Private	11 VMI	Cross
7	Openstack	Private	35 VMI	Cross
19	PeerSim P2P simulator	Private	Variable- max 30	Cross
22	Openstack	Private	102 VMI	Cross
30	OpenStack image registry with a standard configuration of 2GB memory and 10GB hard disk in CloudSim simulator	Private	4 types of virtual images - VDI, VMDK, VHD, Raw, qcow2 images in total 2,426,552,114	Cross
31	Own configuration on Ubuntu 14.04 (64-bit)	Private	10 VMI for Operating system	Single
32	Own setup with 9-servers and 16 desktops as clients running Ubuntu 12.10 64 bit with Linux kernel version 3.5.0-17	Private	276 VMI	Cross
33	Own configuration on Ubuntu 14.04 (64-bit)	Private	10 VMI for Operating	Single

Table 4. Advantages and disadvantages of non-secure Image Deduplication techniques

Paper	Advantages	Disadvantages or Limitation
1	Reduce the virtual machine image deduplication backup time	Slight storage space waste. 2 groups can have a high sample hit rate so when done in local deduplication, it can lead to duplicated blocks again
5	Work in both VHD and raw formats; accelerate the backup process	Focuses on preprocessing phase than deduplication phase; little increment of disk space usage
6	Deletion rate for image groups which have the same version of operating systems, but different versions of software applications is up about 58%	No backup system or Rapid indexing method
7	Uses the memory filter to reduce the overhead of disk index; improves the locality of data by centralizing fingerprints in disk to achieve a higher IO throughput rate with the limited memory occupancy rate	Optimization of image download process not clearly given
19	30% performance gain. Image blocks are trades in two swarms. It also deals with hash collisions	Lacks real-world environment
22	Parallel deduplication, deduplication cache and reduced memory	Cloud platform is distributed environment, but ScaleDFS is single storage partition-based deduplication
30	Very good overall reduction in image storage space and total migration time are achieved when compared with the existing image management systems	The reduction in size is dependent on the dataset and the applications running on the VM
31	Reduction in the metadata storage overhead and the number of hash computations thereby a smaller number of comparisons will be made so that overall deduplication time is reduced for the VM disk images	Dataset used in not standard
32	Parallelization of compute-intensive chunking and fingerprinting, and the I/O-intensive fingerprint indexing will speed up the deduplication process	Setting up of a cluster of deduplication servers is a costly investment
33	DedupCloud minimizes the number of hash value computations and comparisons within similar categories by using byte comparison technique	Dataset used in not standard

NON-SECURE IMAGE DEDUPLICATION FOR PIXEL IMAGE TYPE

A High-precision duplicate image deduplication 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 [2]. It compares the number of same elements in two images edges information. The fuzzy comprehensive evaluation method is used to select duplicate images by finding the centroid image. The size ratio of deduplicated images and total images is 9.7%.

Cloud-scale image compression through content deduplication deals with combating the issues faced with storing storage costs with exponential increase in data [4]. It presents an image compression technique, which takes advantage by compressing each individual image with GIST nearest neighbor to overcome the scalability state-of-art issues.

Image deduplications check on massive image file storage that includes distributed database and file system is discussed in [8]. It uses MD5 based signature

on features of binary image stream instead of file –level or block level fingerprint check.

To reduce storage space, the authors of [9] focuses on the Haar wavelet decomposition and Manhattan distance to select image duplication. When the number of same elements between two collections is greater than or equal to the preset threshold t, they considered the two images are duplicate images.

Deduplication of electricity bills is done using the content-based image retrieval with block truncation coding [10]. This is used to categorize pictures of the electricity bills and blocks of images with the same sizes are clustered together. Each cluster is checked for duplicates, and they are a part of a big block.

Deduplication image middleware detection comparison in standalone cloud database given in [11; 15] talks about techniques used in image deduplication in a standalone database. Most of the time people pay for more memory due to duplication of images. This paper shows a new framework for the early stages of image deduplication in a cloud service. 11 software taken, which are either use standalone or cloud databases. A plugin is used to detect the duplication, which is still a new topic, but mobile Cloud detection has been around from 2008. In all the software used two out of 10 is that you have high detection of duplication and those are hash and Visual similarity. The focus of the paper is to allow users to select Software and Hardware to give them a better use of the cloud services.

Large Scale Image Deduplication given in [13] deals with the problem of near Duplicate Image detection. Each duplicate in the database is linked with a Feature representation of it, what is called as a bundle. Two bundles join to form a feature of SPIHT, which is a robust technique, but become slower and gradually less accurate when the data in the database becomes larger. Maximally stable extreme regions algorithm is used for clustering as it is told to be better than the KNN means as it can also detect duplication when an image is cropped or rotated.

Authors of [17] propose a similar file extraction method where a file with high similarity is extracted. To extract similar files, average hash method is used for determining file similarity. The execution time of deduplication process can be reduced by using only similar files for comparison. Variable length blocks of files are used in this method. The average hash method is used to find the duplication of images. Morphological analysis and cosine similarity is used for the text Duplication. Results show that as the similarity percentage is increased, exact image duplicates can be determined. But the time taken for deduplication increases with increase in similarity percentage. Experimental results say that this method is very efficient to shorten the execution time.

Recognition built on vocabulary tree with indexing scheme that quantized descriptions from image key points hierarchically, which is used for image similarity indication is described in [23]. Indexing descriptor is computed for local regions. The proposed recognition method handles large number of objects for selecting one of them within the acceptable time. Local image descriptors are based on video frames extracted.

DBTP [28] i.e., Double Bytes Transport Protocol is used where double chunks are sent by the client to request for deduplication checks simultaneously, and the server responds to the deduplication requests. This scheme helps in mitigating the side channel's risk.

DriveHQ [29] is a website developed to perform efficient image deduplication for optimized photo and video viewing. In this concept, images are divided into rectangular blocks and hash value is generated for each block using MD5.When similar images are uploaded, the images are divided into blocks and each block is checked with the stored hash values. If hash value is similar, then the images are considered as near identical images and not stored in the cloud to conserve space.

We compare the above discussed non-secure image deduplication techniques for pixel image type in the form of three table. Table 5 gives the comparison in terms of parameters related to algorithms used and perspectives of various dimensions in the image deduplication technique like Table 2. The performance analysis environment is discussed in Table 6in terms like Table 3. Table 7 gives the advantages and disadvantages of the corresponding method.

Table 5. Comparison of non-secure Image Deduplication techniques for pixel image type

Paper	Matching Algorithm	Place	Processing time	Feature-based	Data level	Block size	Optimization objective
2	1-norm of gray block features to construct B trees	Server	Inline	Hybrid	Block	nxn Image blocks	Duplicate images retrieval precision and deduplication accuracy
4	GIST nearest neighbor for compression	Server	Post	Local	File	NA	Image compression rates and reducing computational effort
8	MD5	Server	Post	Local	Binary stream features	Fixed	Optimization of massive image files storage
9	Manhattan distance and Haar wavelet decomposition	Server	Inline	Local	File	Fixed size file	Higher deduplication ratio, deduplication accuracy
10	Block truncation code	Client	Post	Local	Block	Variable	De-duplication process speed
11	Depend on the existing study technique	Server	Post	Local	File	NA	Storage space reduction
13	SIFT And Maximally Stable Extremal Regions (MSER),	Server	Post	Local	File	NA	Increased deduplication accuracy and performance
15	Deduplication image detector software of existing study or plugin for cloud storage	Server	Post	Local	File	NA	High-precision image deduplication
17	Average hash method, File similarity determination	Server	Post	Local	Hybrid	Fixed/ variable	Minimization of execution time for deduplication
23	Local regions indexing descriptors based on visual vocabulary tree	Server	Inline	Local	Block	Fixed	Improvement in retrieval quality
28	Double Bytes Transport Protocol with double chunks simultaneously	Client	Inline	Local	Block	Fixed	Mitigate the side channel's risk and achieve high bandwidth efficiency of deduplication
29	MD5	Server	Post	Global	Block	Fixed	Storage optimization

NA-Not Applicable

Table 6. Cloud type/ environment Parameters for non-secure Image Deduplication techniques for pixel image type

Paper	Cloud/OS – Cloud software	Cloud type	Number of cloud images	User level
2	Corel image database	Private Corel based	1000 PI	Single/cross
4	Canonical set	Private	Dynamic, millions	Cross
8	Own dataset	Private	Variable	Single
9	Corel image database and selected images from www.picsearch.com	Private	1000 PI	Single/cross
10	Own dataset	Private	Variable	Single
11	11 datasets – standard/own	Private/ public	Variable	Single/cross
13	Two dataset – Dataset of [23] for Accuracy and ILSVRC2010 for Performance	Private	Accuracy-10200; Performance- 1.2M	Cross
15	11 datasets – standard/own	Private/ public	Variable	Single/cross
17	Own cloud on a PC	Private	90 bmp images	Cross
23	Own setup	Public	40000 images of popular music CD's	Cross
28	Python 3.7.6 platform and MySQL database	Private	Variable	Cross
29	Own setup	Private	optimized photo and video viewing	Single

Table 7. Advantages and disadvantages of non-secure Image Deduplication techniques for pixel image type

Paper	Advantages	Disadvantages or Limitation
2	Reduces workload of users. The fuzzy comprehensive evaluation allows the procession of selection of centroid images by visual reference	The algorithm is unable to work on images which have been rotated, edited, blurred, or have a watermark excreta
4	Image processing rates reduces the effort used for computation by at least one order of magnitude	Ideal Canonical set is not constructed
8	Signature generation and uploading speed is improved and offers an optimization to massive image files storage	Massive image file storage distributed database without considering its deficiency
9	The proposed approach can achieve higher deduplication ratio and deduplication accuracy by setting suitable thresholds	Methods can't be used for images with similar structures
10	A single instance of the image in the database avoids confusion	Error due to entire data compression at one time
11	Evaluation of existing software is given in detail	Pilot test using standalone dataset is performed based on existing image deduplication detector
13	Method can be used even when images are cropped, rotated, or edited	Size of visual word affects performance – too small may give false results and too large will be impossible to match in one SIFT mapping
15	Deduplication image detector such as plugin, middleware or software used for deduplication	Compares standalone image deduplication detector
17	Both duplication and execution time was reduced	Most of discussion is related to text files not images. The time taken for deduplication increases with increase in similarity percentage
23	entropy weighting of the vocabulary tree is defined with video independent of the database.	Describes only retrieval process
28	DBTP implements two-side privacy to avoid side channel attack	The deduplication ratio is a little reduced compared to existing methods
29	Images are divided into blocks and hashes are generated so that duplication check can use these stored hashes and detect near identical images	Does not work for exact duplicates

SECURE IMAGE DEDUPLICATION

Secure image deduplication through image compression in the cloud storages embeds partial encryption to ensure security against a semi honest CSP and unique hashing to identify identical images into SPIHT compression algorithm [3]. Image compression followed by encryption and hashing in sequence reduces the computational overhead, resources, and metadata to be stored.

Authors of [12] discuss how cloud services have had an immense improvement in this year in terms of Secure Image Deduplication. Due to this great development in the services, many people have started storing data, which may also be redundant. Image duplication is necessary to save cost and space. This research has also used encryption called convergent encryption, which is got, by using the Hash Function on the image data the data is encrypted and decrypted with the same keys and hence the duplicates of the image will produce the same cipher text, by recognizing the duplicate of the cipher text, the image duplicates also found.

Secure Image Data Deduplication through Compressive Sensing given in [14] presents a scheme by comparing the Compression Sensing (CS) and the SPIHT technique for image deduplication according to their experimental results it is shown that the CS Technique is more efficient and has more security than the other methods. They have also further studied that this technique can be used in video duplication as well since videos also take up a lot of data space in the cloud.

The authors of [16] discuss an approach where client side takes an image, compresses it using the SPIHT compression, and partially encrypts it. It also takes the hash value of the image, and the user then uploads only the hash on to the server and the server side checks if the hash value is the same as the previous values. If it is not, then it stores the hash value or it removes the hash to eliminate redundancy.

An efficient approach towards image deduplication using Watson proposes a cost-efficient method of image duplication which has proven to reduce the storage of cloud services by one third its uses of WATSON and a MATLAB SSIM algorithm to do so [18]. In this technique when the user uploads an image it is sent to the WATSON visual where it image is given a tag and the image with the highest tag is sent to the database and is checked if other tags with the similar name is told. If it is so it is, then sent to the MATLAB SSIM to check if the images are similar or not. If the images already stored in the database, then it will not be stored again in the clients 'profile in the Cloud Service or image is uploaded on the cloud servers and the user details are updated.

Client-Side Secure Image Deduplication of [20] uses a dice protocol, which finds image deduplication in its block level. This research concludes that with all their experimental results images, which are more like each other, have smaller number of blocks in their storage. This however does not show what happens to images, which have been cropped, are in different lighting or have scaling and any other kind of Editing done on them. It also does not deal with file, which has been compressed into different formats.

Data outsourcing model of [21] uses file level as well as block level deduplication using Dekey convergent key management scheme. A user computes and sends the block tags to the cloud server, which stores only unique tags. The stored tags are informed to the client so that it can secure it and resend back to server.

The indexed information of this secured block is maintained at client also for future access.

Secure data deduplicationusing radix trie and bloom filter discussed in [24] used existing hash-based deduplication technique. It starts with convergent encryption to avoid leakage of data followed by three stages – authorization deduplication using role re-encryption process, proof of ownership and role key update. Roles and keys are mapped with radix trie. Data updation and retrieval of ownership verification is done using bloom filter.

BDKM [25] is a blockchain based approach to ensure confidentiality of outsourced data and reliability of Convergent Key (CK) management is enhanced by adopting an oblivious pseudorandom function to generate the randomized CK. Data reliability in BDKM is achieved by dividing the CK into segments and distributed to blockchain. This work can be extended by employing blockchain to implement a secure and efficient integrity verification on the data deduplication, where a user can verify the integrity of other users' data without knowing any information about the data.

Multistage for coarse to fine deduplication is proposed in [26]. The global features are comparing to find the duplicate initially followed by local features if no match found. Fine deduplication is applied using SHA 256 based Merkle hash tree. Local and global features work at file level while hash tree works at block level. The database is maintained for each file on dataset consisting of global features, local features, and hash tree details.

Authors of [27] propose an in-line block matching-based data deduplication scheme with dynamic user management. Users encrypt their data using convergent encryption. Server uses in-line block matching protocol to generate unique proof by calculating the group key and re-encrypts the file using the group key. Another user uploading the same file will verify the proof against the server and re-encrypts using a new group key. Contents of the file remains confidential and even the server will not be aware of the file contents. Ownership list is maintained, and access control techniques are employed to prevent the access of cipher text from unauthorized users, cloud servers and adversaries. The analysis of the proposed scheme shows that the computational time, communication, and storage overhead is reduced when compared with the existing deduplication schemes.

SEDS [34] scheme is proposed to provide a secure server sided data deduplication scheme for storing data in the cloud. This scheme generates constant size ciphertext, which is independent of the number of key servers, and cloud server performs proxy re-encryption to prevent semi-honest proxy server to transform the ciphertext. This scheme supports both intra-Key Server and cross-Key Server duplication check. Experimental analysis proves that the scheme is efficient compared to previous schemes with respect to computation and communication overheads and security.

We compare the above discussed secure image deduplication techniques in the form of three tables. Like Table 2 and Table 5, Table 8 gives the comparison in terms of parameters related to algorithms used and perspectives of various dimensions in the image deduplication technique. Similarly, the performance analysis environment is discussed in Table 9 wherein additional parameters are added named as security provisioning protocol used in addition to deduplication techniques. Table 10 gives the advantages and disadvantages of the corresponding method.

DISCUSSION AND CONCLUSION

Image deduplication for duplicate check helps to reduce the communication and network transmission cost in the cloud environment. Most of the techniques work towards reduction of algorithm complexity through smaller hash size generation functions. We observed that image deduplication is either possible on pixel-based images or virtual machine images. There is no universal technique, which can be applied on both types of images. We presented taxonomy and classification of existing image deduplication related articles, which clearly shows that image.

Table 8. Comparison of Secure Image Deduplication techniques

Paper	Matching Algorithm	Place	Processing time	Feature-based	Data level	Block size	Content type	Optimization objective
3	SPIHT compression, partial encryption, and hashing	Hybrid	Inline	Global	File	NA	PI	Ensure security against a semi honest CSP and compressed image deduplication
12	attribute-based encryption	Hybrid	Post	Global	File	NA	PI	Confidentiality, Privacy protection and completeness
14	CSP - SHA basedduplicate images removal	Hybrid	Post	Global	File	NA	VMI	Ensure security against a semi honest CSP and optimize storage space
16	Robust image hashing based on SPIHT	Server	Inline	Local	File	NA	PI	Ensures data security against a curious and semi truthful CSP or any malicious user
18	WATSON and MATLAB SSIM algorithm	Server	Post	Local	File	NA	PI	Reduction in time required to perform deduplication
20	Dual Integrity Convergent Encryption protocol	Client	Inline	Local	Block	Fixed-	PI	Optimal block size determination for hashing and optimize storage space
21	DCT compression and convergent encryption	Hybrid	Inline	Local	Hybrid	Fixed - 8x8	PI	Ensure data security and optimize storage space
24	Radix Trie with Bloom Filter (SDD-RT-BF), hash function	Client	Inline	Local	Block	Fixed	PI/ Audio	Maximizing deduplication rate and ensuring security
25	Distributed Blockchain, SHA256, RSA	Client	Inline	Global	File /Block	Fixed	PI / text files	Achieve secure and reliable Convergent Key management and resistance to the brute-force attack and collusion attack launched by the external adversaries
26	Merkle-Hash and Image Features	Client	Inline	Local and global	File/Block	Fixed	PI	Ensure data security and optimize storage space
27	Guillou-Quisquater identification protocol with dynamic ownership management, Convergent encryption	Client, group Key server	Inline	—	Block	Fixed	PI / text files	Reduce network traffic and storage. Better ownership management
34	Convergent encryption and server re-encryption	Server	Inline	—	File	NA	PI / text files	Better performance of deduplication algorithm

NA- Not Applicable

Table 9. Cloud type/ environment Parameters for Secure Image Deduplication techniques

Paper	Cloud/OS - Cloud software	Cloud type	Number of cloud images	User level	Security provisioning Protocol
3	MATLAB environment dataset	Private	252 own set	Single	Partial encryption
12	No specific cloud	Any	Variable	Cross	Convergent encryption
14	MATLAB environment dataset	Private	6	Cross	semi honest CPs
16	MATLAB environment dataset	Private	10 PI	Single	Partial encryption
18	WATSON and MATLAB	Private	Variable	Cross	Password Protection
20	Own JAVA based environment	Private	30 PI	Cross	Dual Integrity Convergent Encryption
21	Not given	Private	Not given	Cross	Convergent encryption
24	Java on Amazon EC2 serve	Public	Variable	Single	SHA-256 with radix trie and bloom filter
25	No specific cloud – own index server	Private	Variable	Cross	SHA 256, RSA for encryption
26	No specific cloud – own index server	Private	Variable	Cross	SHA 256 for Merkle hash tree
27	Own server	Private	Not given	Cross	Convergent encryption
34	Own set up with multiple servers	Private	Variable	Cross	Convergent encryption and server re-encryption

Table 10. Advantages and disadvantages of secure Image Deduplication techniques

Paper	Advantages	Disadvantages or Limitation
3	Save monumental amounts of computational time and resources; Can find duplicate images even when images are extremely similar and compressed	Experimentation results are not derived in a real Cloud Service setting
12	In the paper ensure privacy protection and confidentiality	The steps done by the client are too many
14	Efficient compression scheme makes the CSP store less data	Small set of testing data with small image size
16	Great combination of analysis and security for storage	Only same images can be deduplicated. No proof of Storage protocols
18	Cloud storage space usage after deduplication has been reduced up to one third. Cost-effective	Image is only removed if user has last access to it or only the image details are hidden from the user
20	Reduce communication and bandwidth cost	Lacks real-world environment and diverse image dataset
21	DCT compression reduce storage space	small encoding/decoding overhead
24	Client-side deduplication and Tag consistency preservation with Fault tolerance	other queuing techniques and lightweight cryptographic algorithms could be used to improve performance
25	Blockchain ensures data reliability and secure key management	Secure against the collusion attack with a limited overhead and blockchain can be extended to verify integrity of other users' data without knowing the details
26	CNN is used to compare global and local features of stored images in the database with that of the incoming image and then additionally Merkle hash is used to check for duplicates	Database storage is increased for multiple level comparisons
27	File integrity is achieved by using convergent encryption, in-line block matching protocol and group key management that hides file contents from unauthorized users, cloud servers and adversaries	Generation of group keys and re encryption for subsequent uploads of the same file by other users is the overhead
34	Ensures data confidentiality, possession proof, resistant against tag inconsistency attack, cross-key server duplication check and scalability	The scheme involves multistage key generation and encryption, the process is slower. Cloud server performs proxy re-encryption which is an overhead

Deduplication has considerable potential towards efficient cloud storage usage. The proposed taxonomy has proved a convenient means of grouping the available image deduplication research and giving insight on its contribution in terms of standard features supported in the image deduplication algorithms, cloud environment, advantages, and limitations. This survey explores published research works in greater depth related to the exploitation of features of the technique used for the deduplication check. The existing image deduplication techniques neither use standard dataset as benchmark image dataset for performance evaluation nor have standard metric for similarity computation. Authors have their own way to consider performance environment. The optimization objective of the different algorithms is also listed here for researchers to get an overview of the goal of deduplication. The identified drawbacks can be scope for future research to work further for strengthen this area.

REFERENCES

1. J. Xu, W. Zhang, S. Ye, J. Wei, and T. Huang, "A lightweight virtualmachine image deduplication backup approach in cloud environment," in *2014 IEEE 38th Annual Computer Software and Applications Conference*, pp. 503–508.
2. M. Chen, S. Wang, and L. Tian, "A High-precision Duplicate Image Deduplication Approach," *JCP*, 8(11), pp.2768–2775, 2013.
3. F. Rashid, A. Miri, and I. Woungang, "Secure image deduplication through image compression," *Journal of Information Security and Applications*, 27, pp. 54–64, 2016.
4. D. Perra and J.M. Frahm, "Cloud-scale Image Compression Through Content Deduplication," in *BMVC*, 2014.
5. J. Xu, W. Zhang, Z. Zhang, T. Wang, and T. Huang, "Clustering-based acceleration for virtual machine image deduplication in the cloud environment," *Journal of Systems and Software*, 121, pp.144–156, 2016.
6. Z. Lei, Z. Li, Y. Lei, Y. Bi, L. Hu, and W. Shen, "An Improved Image File Storage Method Using Data Deduplication," in *2014 IEEE 13th International Conference on Trust, Security and Privacy in Computing and Communications*, pp. 638–643.
7. J. Zhang et al., "IM-Dedup: An image management system based on deduplication applied in DWSNs," *International Journal of Distributed Sensor Networks*, 9(7), p.625070, 2013.
8. S. Youjun and Z. Daxing, "Research on deduplication technology for massive image file storage," *Computer Applications and Software*, 4, p. 15, 2014.
9. M. Chen, Y. Wang, X. Zou, S. Wang, and G. Wu, "A duplicate image deduplication approach via Haar wavelet technology," in *2012 IEEE 2nd International Conference on Cloud Computing and Intelligence Systems*, vol. 2, pp. 624–628).
10. A.J. Zargar, N. Singh, G. Rathee, and A.K. Singh, "Image data-deduplication using the block truncation coding technique," in *2015 International Conference on Futuristic Trends on Computational Analysis and Knowledge Management (ABLAZE), IEEE*, pp. 154–158.
11. N. Yusof, A. Ismail, and N.A.A. Majid, *Deduplication image middleware detection comparison in standalone cloud database*.
12. H. Gang, H. Yan, and L. Xu, "Secure image deduplication in cloud storage," in *Information and Communication Technology-EurAsia Conference*, pp. 243–251. Springer, Cham, 2015.
13. T.Y. Wen, *Large Scale Image Deduplication*. Available: http://vision.stanford.edu/teaching/cs231a_autumn1213_internal/project/final/writeup/nondistributable/Wen_Paper.pdf
14. F. Rashid and A. Miri, "Secure image data deduplication through compressive sensing," in *2016 14th Annual Conference on Privacy, Security and Trust (PST), IEEE*, pp. 569–572.

15. N. Yusof, N.A.A. Majid, and A. Ismail, “Framework deduplication image detection assisted multimedia system using multi technique,” in *2016 6th International Workshop on Computer Science and Engineering, WCSE 2016*, pp. 402–406.
16. S.P. Bini and S. Abirami, “Secure image deduplication using SPIHT compression,” in *2017 International Conference on Communication and Signal Processing (ICCSP), IEEE*, pp. 0276–0280.
17. T. Koike, M.Z. Nurshafiqah, and T. Kinoshita, “Data Deduplication for Similar Image Files,” in *Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA)*, pp. 296–301, 2018.
18. R. Aathira and V.P. Poonthottam, “An efficient approach towards image deduplication using WATSON,” in *2017 International Conference on Inventive Computing and Informatics (ICICI), IEEE*, pp. 180–183.
19. C. Lee, S. Kim, and E. Kim, “A Deduplication-Enabled P2P Protocol for VM ImageDistribution,” *IEICE TRANSACTIONS on Information and Systems*, 98(5), pp. 1108–1111, 2015.
20. A. Agarwala, P. Singh, and P.K. Atrey, “Client Side Secure Image Deduplication Using DICE Protocol,” in *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), IEEE*, pp. 412–417.
21. M.S. Soofiya and S.V. Kumar, *DCT Image Compression and Secure Deduplication with Efficient Convergent Key Management*.
22. M. Ma, *Kernel-space Inline Deduplication File Systems for Virtual Machine Image Storage*; Doctoral dissertation, Chinese University of Hong Kong, 2013.
23. D. Nistr and H. Stewnius, “Scalable recognition with a vocabulary tree,” in *IN CVPR*, pp. 2161–2168, 2006.
24. S.E. Ebinazer and N. Savarimuthu, “An efficient secure data deduplication method using radix trie with bloom filter (SDD-RT-BF) in cloud environment,” *Peer-to-Peer Networking and Applications,* 14(4), pp. 2443–2451, 2021.
25. G. Zhang, H. Xie, Z. Yang, X. Tao, and W. Liu, “BDKM: A blockchain-based secure deduplication scheme with reliable key management,” *Neural Processing Letters*, pp. 1–18, 2021.
26. D.P. Akarsha, S. Chaudhari, and R. Apama, “Coarse-to-Fine Secure Image Deduplication with Merkle-Hash and Image Features for Cloud Storage,” in *2021 Asian Conference on Innovation in Technology (ASIANCON), IEEE*, pp. 1–6.
27. V. Kanagamani and M. Karuppiah, “Zero knowledge-based data deduplication using in-line Block Matching protocol for secure cloud storage,” *Turkish Journal of Electrical Engineering & Computer Sciences*, 29(4), pp. 2067–2083, 2021.
28. J. Ouyang, H. Zhang, H. Hu, X. Wei, and D. Dai, “Enhanced Deduplication Protocol for Side Channel in Cloud Storages,” *International Journal of Network Security*, 23(2), pp. 270–277, 2021.
29. S. Vinoth Kumar, L. Kruthika, K. Pooja, H.J. Priyanka, and N.R. Rachana, “Image Deduplication in DriveHQ Cloud,” *Journal of Computational and Theoretical Nanoscience*, 17(9-10), pp. 3895–3898, 2020.
30. N.M. Tyj and G. Vadivu, “Adaptive deduplication of virtual machine images using AKKA stream to accelerate live migration process in cloud environment,” *Journal of Cloud Computing*, 8(1), pp. 1–12, 2019.
31. S. Saharan, G. Somani, G. Gupta, R. Verma, M.S. Gaur, and R. Buyya, “QuickDedup: Efficient VM deduplication in cloud computing environments,” *Journal of Parallel and Distributed Computing*, 139, pp. 18–31, 2020.
32. C. Lin, Q. Cao, J. Huang, J. Yao, X. Li, and C. Xie, “HPDV: A highly parallel deduplication cluster for virtual machine images,” in *2018 18th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID), IEEE*, pp. 472–481.
33. S.S. Patra, S. Jena, J.R. Mohanty, and M.K. Gourisaria, “DedupCloud: an optimized efficient virtual machine deduplication algorithm in cloud computing environment,” *Data Deduplication Approaches: Concepts, Strategies, and Challenges*, 281, 2020.
34. S.K. Nayak and S. Tripathy, “SEDS: secure and efficient server-aided data deduplication scheme for cloud storage,” *International Journal of Information Security*, 19(2), pp. 229–240, 2020.

35. D. Reinsel, J. Gantz, and J. Rydning, “Data Age 2025: The Evolution of Data to Life-Critical,” *Seagate*, an IDC White Paper 2017.
36. Q. He, Z. Li, and X. Zhang, “Data deduplication techniques,” in *2010 International Conference on Future Information Technology and Management Engineering (FITME)*, pp. 430–433.
37. Kirti Ashok Tayade and G.S. Malande, “Survey paper on a secure and authorized de-duplication scheme using hybrid cloud approach for multimedia data,” in *2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)*, IEEE, pp. 2966–2969.
38. Shieh Fatemeh, Mostafa Ghobaei Arani, and Mahboubeh Shamsi, “De-duplication approaches in cloud computing environment: a survey,” *International Journal of Computer Applications*, 120, no. 13, 2015.
39. W. Xia et al., “A comprehensive study of the past, present, and future of data deduplication,” *Proceedings of the IEEE*, vol. 104, pp. 1681–1710, 2016.
40. “Data deduplication in the cloud explained, part one,” *ComputerWorld*. Accessed on: Dec 1, 2021. [Online]. Available: https://www.computerworld.com/article/2474479/data-deduplication-in-the-cloud-explained--part-one.html
41. “Data deduplication in the cloud explained, part two: the deep dive,” *ComputerWorld*. Accessed on: Dec 1, 2021. [Online]. Available: https://www.computerworld.com/article/2475106/data-deduplication-in-the-cloud-explained--part-two--the-deep-dive.html

Received 21.02.2023

INFORMATION ON THE ARTICLE

Shilpa Chaudhari, ORCID: 0000-0001-8659-4214, Ramaiah Institute of Technology, Bangalore, India, e-mail: shilpasc29@msrit.edu

Ramalingappa Aparna, ORCID: 0000-0002-8093-916X, Ramaiah Institute of Technology, Bangalore, India, e-mail: aparna@msrit.edu

ОГЛЯД ДЕДУПЛІКАЦІЇ ЗОБРАЖЕНЬ ДЛЯ ХМАРНОГО ЗБЕРІГАННЯ / Шілпа Чаудхарі, Рамалінгаппа Апарна

Анотація. Посилення комунікацій у реальному житті спонукало до створення, передавання та цифрового зберігання великих обсягів зображень і відеоданих у хмарі. Вибухове збільшення даних віртуальних/візуальних зображень на хмарному сервері потребує ефективного використання сховища, цьому сприяє технологія дедуплікації зображень. Незважаючи на те, що властивості віртуального зображення та візуального зображення розрізняються, наявна література використовує подібний підхід для перевірки дедуплікації, що спонукало розглянути обидва типи зображень для цього огляду. Дослідження має на меті надати детальний огляд найсучасніших візуальних засобів, а також методів дедуплікації віртуальних зображень у хмарному середовищі, узагальнюючи та організовуючи їх шляхом розроблення п’ятивимірної таксономії для аналізу функцій і продуктивності з кількома категоріями, що не перетинаються, у кожен вимір. До них належать: 1) місце застосування дедуплікації; 2) виділення ознак зображення; 3) час звернення; 4) стратегія розподілу даних зображення; 5) залучення рівня набору даних користувача. Наявні методи дедуплікації зображень класифікуються на дві основні категорії залежно від того, чи передбачає цей метод захист чи ні. Порівняння методів виконано за набором функціональних і продуктивних параметрів. Поточні проблеми висвітлюються з можливими майбутніми напрямами для подальших досліджень цієї теми.

Ключові слова: дедуплікація зображень, хмарні обчислення, хмарне сховище, виявлення копій зображень.