International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Image Fingerprinting and File Recovery Using Digital Forensics

[1]S. Vara Raj Kumar, [2]S. Lalitha, [3]S. Siva Sankar, [4]S. Kiran Kumar, [5]Sobia Shabath, [6]T. Jaya Sai

[1,2,3,4,5,6]Department of CSE, GMR Institute of Technology, Rajam. Andhra Pradesh, India

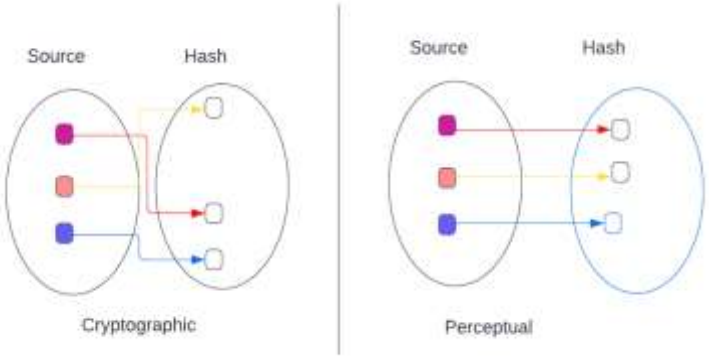
ABSTRACT:

Cyber threat is a process of exploiting sensitive information by losing the golden triad (Confidentiality, Integrity, and Availability). The threats may include Trojans, Phishing, DDOS attacks. Phishing works as a social engineering technique, where in the malicious attacker attempts in such a way that the appearance to the user would seem legitimate thereby causing the user to fall in a trap which may lead to the compromising of the user's sensitive information by means of illegitimate URLs or links etc. According to previous records approximately 18% of businesses lost their financial information due to phishing attack. Phishing can exploit data such as credentials, personal data, medical, etc. In this work we proposed a hybrid stacking model where the URL is identified as a legitimate or illegitimate URL. For Data preprocessing vectorization, tokenization and word embedding techniques used, where words will be converted to vectors or numbers. In stacking, k-fold cross validation with a hybrid model which includes machine learning models for feature selection along with deep learning for the prediction. We compare the metrices such as accuracy, recall, precision, and F1-score with other state of art models.

Keywords: Data Pre-processing, Tokenization, Word Embedding, Stacking, Machine learning, Deep learning.

INTRODUCTION:

Since electronic crime is becoming more prevalent globally nowadays, it is essential to enhance all existing ways for preventing it. One of the key techniques in the investigation of computer crimes is digital forensics. Numerous news stories and academic studies discuss brand-new innovations every hour that are aimed at improving our society. As a result, the issues brought on by this rapid growth call for an equally speedy reaction. The solution is digital forensics. Digital forensics must constantly advance in order to identify data manipulation, safeguard data from attackers, and retrieve destroyed evidence due to the abundance of easily accessible software tools that may change digital media (particularly photos and video).



Political, artistic, and social news are frequently published using images, thus it's critical to be able to confirm if the information is accurate or not, who has permission to post it, and who is responsible for the spread of false information. This topic is handled by the discipline of image forensics. It processes photos using a variety of techniques to find evidence and recover erased evidence. The pictures have individual fingerprints, much like people do. An image fingerprint is a sequence of numbers or characters that an algorithm creates depending on the characteristics of the image. A person's fingerprint may be used to assess them, and we can also use a fingerprint to discover photographs. The widely used and straightforward fingerprint extraction approach known as the image perceptual hash allows for the easy extraction of characteristics. Hamming distance may be used to compare photos that are almost identical but have minor differences. The fingerprints left by photos remain unchanged when they are resized, compressed, or extended. However, hamming distance is necessary if the image has been cropped or was captured from a differing point of view. Recovery and comparison of identical pictures may benefit from the use of the CBIR (content-based image retrieval) approach and the FTK (forensic tool kit).

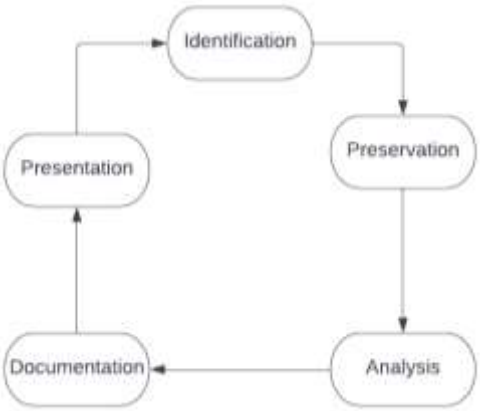


Fig. PHASES OF DIGITAL FORENSICS

The widely used and straightforward fingerprint extraction approach known as the image perceptual hash allows for the easy extraction of characteristics. The identification of identical photos can be aided by various hashing algorithms. In order to acquire reliable results, hashing techniques like Perceptual hash, Average hash, MD5, and SHA256 are highly helpful. Hamming distance may be used to compare photos that are almost identical but have minor differences. The fingerprints left by photos remain unchanged when they are resized, compressed, or extended. However, hamming distance is necessary if the image has been cropped or was captured from a different viewpoint. Recovery and comparison of identical pictures may benefit from the use of the CBIR (content-based image retrieval) approach and the FTK (forensic tool kit).

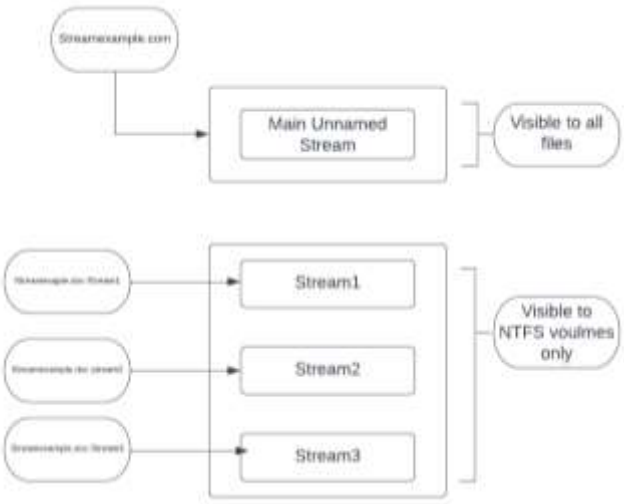


Fig. NTFS file recovery system

2. Literature Survey

[1]. In this paper, the authors, R.S. Khalaf and A.Varol focused on the features of the digital forensic specialty, focusing on images, and also interested in explaining the image of how it is stored on digital devices and how it is deleted, and what the specifications of the operating systems is when it comes to deleting an image. They also mention about the Image file headers, file fragments and also the Photogrammetry forensics as well to identify the images found in the investigation. The authors focused on pixelating the images using the Bitmap images and as well as the vector images, also made use of the image array types of binary images, gray scale images and multispectral images as well. They have also utilized the services of digital forensics in order to gather the digital evidence and also some of the digital forensic tools such as X-ways forensics and Evidor.

[2]. In this paper, the authors, S. Agarwal and K.H Jung deals with the image forensics using the optimal normalization, in this author tried to mention about the digital forensics that are need to be done to assure the authenticity of the image as people may edit and create a fake image that resemble the exact image. Image filtering is one of the novel techniques used in this paper, a novel technique is applied here for detection of median, averaging and Gaussian filtering in the images. The proposed method in this paper deals mainly with the optimal normalization on the images to gather and find the better statistical information. First a image is normalized using the optimal range to obtain its statistical information and then difference of array are calculated by proposing the threshold. This threshold difference is used to extract the features from the thresholding differences and this proposed model has gained significant approval even for the low-quality JPEG compressed as well.

[3]. In this paper, the authors, Y. Wang, A. Sui and W. Fu, introduced a research based on image retrieval technology which is based on image fingerprinting and color features. This paper focus mainly on the image color features and the image fingerprint extracted by the improved perceptual hash algorithm. In this the author proposed a perceptual hash algorithm based on block gray histogram, which generates a long hash sequence and require

a lot of storage space. This paper also tried the locality sensitive hashing algorithm which is very simple and easy to work with but the accuracy is not high and practically poor. Although the paper doesn't concentrate the image fingerprinting based on image scaling and gray processing rather focus only on the image color histogram feature which only consider statistical colors. Main improvement required in this paper is it doesn't take into account regarding the distribution of the characteristics of color in the image. So, the results may vary for rich color images. Image retrieval effect of this algorithm is much better than some other processes on many image categories as the results suggest in this paper, but for a few image categories, whose content color is rich and not obvious, the precision is relatively low.

[4]. In this paper, the authors Y. Li, D. Wang and L. Tang emphasized on using the Neural Networks for the efficient model that can extract feature and percepts the image using the image fingerprinting which doesn't involve the expert knowledge. Here the neural network is trained automatically to discover the optimal mapping from image to fingerprinting through the previously fed data driven image fingerprinting algorithm. The fingerprinting is done through the layer- by-layer manner to progressively improve the robustness against the content-preserving distortions. The proposed training algorithm follows a simple-to-complex principle, which proceeds from pre-training single layer to fine-tuning the whole of the network. This proposed algorithm mainly helpful in reducing the amount of information leakage in output fingerprinting.

[5]. In this paper, the author A. Al-Sabawi has focused on the special tools that are required to create and retrieve images from the infected disk or the memory. The author focused mainly on the commands from the command prompt to analyze and retrieve the images from the hard disk rather than writing the program. By using the tools such as forensic tool kit imager which is a standalone disk imaging program is used because it confirms the integrity of the data along with calculating the MD5 hash values before closing the files, and it stores image file in several formats. This paper also uses the image identification commands, image info along with the methods such as process memory which has procdump and memdum commands and networking as well with commands such as connscan. The author of this paper has also focused on the security of the deleted file by adding the malware analysis as well by using the Svcscan command.

[6]. In this paper, the author, R.Zhang and Z.Zhang used probabilistic approach, , where the query image and the images in database are first segmented into homogeneous colour-texture regions. Later on, the properties are extracted from every corner by incorporating multiple semantics-related features like shape, colour, texture properties. For image retrieval performance evaluation, overall 1500 images are randomly chosen from all categories that represents as a query set. The proposed model in this paper assumes that the regions, hidden semantic concepts, and images are random variables and the objective is to discover concept distributions with samples from the region and image distributions. As a future work, the authors want to improve the image segmentation algorithm. With more image segmentation algorithm, the model accuracy and effectiveness of retrieval will be expected more. Content based image retrieval may also be helpful in solving problems like automatic image annotation, classification of image and image database summarization.

[7]. In this study, the authors, G. Carneiro, A. B. Chan, P. J. Moreno and N. Vasconcelos, worked with two goals. , first one is the automatic annotation of previously unseen images. Second one is, the retrieval of images from database that depend on semantic queries. A set of training images was collected and also a binary classifier for classifying all database images. Those classifiers are trained in "one-vs-all" mode. So, they called this semantic labelling framework as supervised OVA. Comparison between supervised multiclass labelling (SML) and Core15k was made. They have introduced a binding together perspective on state-of-the-art methods for semantic-based image annotation and retrieval. The experimental evaluation has shown that performance of SML is robust for variability in parameters like dimension of feature space.

[8]. In this paper, the authors, V. A. Wankhede and P. S. Mohod used two algorithms. They are Content based image retrieval and Annotation based image retrieval. The main goal of this paper is, the user may give input as image query or text query. If the input is image query then the image will be retrieved using content based image retrieval(CBIR). If the input is text query then retrieved image using annotation based image retrieval(ABIR). There are 2 types of retrieval of images. They are query-by-text i.e, the query will be text and the target will be the retrieval of image. Another one is, query-by-image i.e, the query will be image and the target will also be the retrieval of image. Content based image retrieval is mainly used to retrieve the image and Annotation based image retrieval is used to retrieve the text. The main goal is to implement a multiquery image retrieval that display the result in less tie with more efficiency.

[9]. In this paper, the authors, S. K. P. N, D. P, A. G, Saritha, S. K and D. R. U, presented an image independent automated algorithm for pixel-to-pixel matching between digital images which are obtained from different microscope imaging devices. The proposed algorithm was tested on a macro set of 1,69,000 digital images that are obtained from 250 pap smear samples. There are mainly two types of image matching methods. One is related to counting the color value i.e, based on color features. Another one is related to feature points from each image i.e, based on image features. They used the image features algorithm to match the images. They proposed an automated image independent pixel to pixel matching algorithm. They used speeded up robust features (SURF) algorithm for extracting feature points and its descriptors.

[10]. In this paper, the author, Q. Kai studied the importance of visual and semantic features of images synthetically and also realize the classification and retrieval of image database based on semantic classification of image. Before extracting image visual features, this paper firstly analyzed several algorithms for extraction of image color, shape and texture features. In order to express the image content fully, this paper takes the help of semantic and visual features of image as training data, establishes the category model images and achieves the classification and retrieval of image database. The retrieval system is mainly classified into three systems: Database generation subsystem, semantic classification subsystem and query subsystem. The experiments shows that the proposed method improved the retrieval efficiency and achieves high accuracy.

[11]. This paper mainly proposes the Randomized Attacker Simulation Model (RCSM) as checklist for maintain the organization security and privacy.In this paper also proposes the Baseline Data Classification Model (BDCM) as a pre-forensic data categorization model.Prudent application of both models can protect against counterattacks across all sectors. Various hardware and software performance metrics are discussed, how it is better than conventional/traditional cryptography.BDCM model helps organizations in categorizing corporate data more effectively with significance to digital

forensics, while labeling data classification as a mandatory exercise for organizations security.RCSM model is a checklist to assist organizations in performing spontaneous attacker drills as a routine part of their cybersquatting audit culture.These proposed models optimize organizations approaches to the professional application of digital forensics as a cybersquatting component .

[12]. This paper mainly discusses about the research done on gathering the information in forensics and its different domains,anti-forensic techniques, and also analysis of current status of forensics.The main advantage of digital forensics is it's direct relation to data recovery and data carving. this field struggles with the rapid increase in volume of data.The main struggles faced this field is with the rapid increase in volume of data.Data storage and data recovery techniques contain data mining, machine intelligence, deep learning,algorithms and mobile forensic which includes android programming knowledge.The major challenge faced by forensic examiners and researchers is to implement cloud forensics and mobile forensics as a service.The further work includes the designing and implementing integrated forensic tool that supports all the platforms.

[13]. This paper gives a complete view on various approaches such as, state-of-the-art methods that are implemented by different organizations in different domains to satisfy needs in IOT devices. Furthermore, this paper also discusses several machine learning and deep learning algorithms utilized for IoT device. This paper us a thorough view of various profiling methods, such as device type, device instance, anomaly detection and multiple security issues in the field.limitations mentioned in this paper are availability of limited resources in IoT devices , they have little to no security , that makes them vulnerable on the network that are hard to identify.We can achieve better performance by choosing a set of features with the highest practicality and highest efficiency. This could increase the performance of both of accuracy and scalability of the approach.The machine learning classifier that are used by the state-of-the-art approaches for profiling the IoT devices has shown the best performance .

[14]. This paper mainly discusses about the research done on various techniques and proprietary forensics applications used by investigators to examine the copy of digital devices, searching hidden, deleted, encrypted, or damaged files or folders.The aim of this paper is to discover and analyse patterns of fraudulent activities and evidence found are carefully analyzed and documented in a "finding report" in preparation for legal proceedings to be peresented in court of law .NLP techniques are used to extract and represent unstructured information into a standardized format and it requires to perform Information Extraction and Retrieval tasks to correlate the information coming from different sources.This approach brings an advantage to easily make relationships between the entities belonging to a specific domain .Future works involves the improvement of analysis , performances and the support of a wider range of devices and sources for the collection phase.

[15]. This paper mainly , focuses on role of machine learning in digital forensics and how machine learning has been used in Digital Forensics to collect digital evidence .This paper proposes , a technique called Network forensics for tracking down cyber criminals by analysing and tracing back network data.Network forensics examining network traffic for the purpose of detecting intrusion and determining how the crime occurred, i.e., setting up a crime scene for analysis . Network traffic classification is also important for network surveillance, security analysis, and digital forensics.In this paper , network traffic collection tools like Iris, Net Intercept, Net Witness, SoleraDS5150, and Xplico are be deployed to examine the network traffic .The random forests classifier produces the most promising results, while the proposed improved statistical protocol recognition approach offers an interesting trade-off between higher efficiency and slightly lower accuracy, according to the experiments.

[16]. In this study, hash functions are used to detect duplicate images, and the hash pool structures are constructed using a feature vector derived from fingerprints. The dataset consists of 11,080 photos, where 80 of the original images were edited to serve as comparison queries. An established process or mathematical function known as a hash transforms a huge amount of data into a small amount of data. In a huge dataset, related or duplicate entries can be found using hash methods. The pre-processing phase, the index generation phase, and the fingerprint query phase are the three phases of the system. A feature vector that can represent a feature of the input picture is produced during the pre-processing stage. Utilizing hash functions, the index generation process creates the fingerprint image feature vector database. When authentication verification is required, the fingerprint query phase checks the suspect photos with the database images. In the first experiment, the accuracy rate and recall rate are both greater than 85%, while in the second experiment, the average recall rate and precision rate are both higher than 85%.

[17]. In this study, for indexing local image descriptors, a novel hash function is suggested in this research. The range neighbor method was employed by the author. The dataset is divided into two sets, the first of which includes 30k photographs and the second of which has 510k images. The author used 50 unique photos in this article as searches. An approximative closest technique is K-means clustering. Only points in the class of the query point will be evaluated as neighbors at query (or directly declared as neighbors). As was already said, the fundamental disadvantage of the k-means-based method is that it is not designed for datasets that are updated often. Our approach is based on unique hash functions that were motivated by several closest neighbor methods. Since there is no learning step, the outcomes are independent of any off-line processes used on a specific dataset. Recall and time(ms) on the bag-of-words method, range neighbor's algorithm, and K-means based vocabulary are the performance parameters that are calculated here; range neighbor's algorithm has a higher recall rate of 0.974 in comparison to the other techniques.

[18]. In this This study suggests a DNA replacement and hash table structure scrambling-based picture encryption technique. The technique employs the traditional "scrambling-diffusion" method, with a pseudo-random sequence utilized in each step. Hyperchaotic systems feature more control parameters, more complex dynamic properties, and more complicated beginning circumstances than one-dimensional chaotic systems. By combining them with picture encryption methods, the key space may be efficiently expanded. The process of scrambling starts by creating two sequences with no repeated values using the closed hash technique in the hash table structure for chaotic sequences, and then scrambling the plain picture twice in accordance with the two sequences. Then, during the diffusion phase, DNA replacement rules, coding rules, and decoding rules are created in accordance with the chaotic sequences. DNA is used to dynamically encode, replace, and decode the picture. The stronger the encryption effect, the closer the Number of Pixels Change Rate (NPCR) and Unified Average Changing Intensity (UACI) are to 99.6049% and 33.4635%, respectively, two effective measures of the algorithm's resistance to attack.

[19]. In this paper, this work uses two-dimensional chaotic mao and six-dimensional hyperchaotic map for image hashing and colour picture encryption. First, the colour image is pre-processed, and the image hashing technique extracts the hash sequence, which is utilized as the chaotic system's starting value and control parameter. Second, the enhanced two-dimensional chaotic map is used to replace the pixels in the colour picture RGB after synthesizing the three colour channels into a two-dimensional matrix. The encrypted picture is then produced using a 6D hyperchaotic system that generates random sequences for DNA dynamic coding and arithmetic operations on colour images. The Unified Average Changing Intensity (UACI) and Number of Pixels Change Rate (NPCR) are used to quantify the sensitivity of an encryption system, respectively. The NPCR and UACI values are computed and repeated 200 times to arrive at an average value.

[20] In this paper, FAT32 and NTFS, which run on Windows computers, are the suggested file systems for file recovery. Because of this, this paper focuses on the fundamental ideas of data recovery on Windows FAT32 and NTFS and provides a solution. The DOS boot record (DBR) sector, FAT, and data area are the three components that make up a FAT32 partition, which can accommodate partitions bigger than 2G. There is no distinct directory area in FAT32 since it organizes directories like files, and all file directory entries are kept in the data area. The cluster number of a file inside a chain is recorded in a file allocation table, or file allocation table. There are two FAT tables, one of which is the backup table and the other of which is the basic table, to provide security. There must be at least one file directory entry for each file or directory, which contains information such as the file name, directory name, extension name, file attributes, file size, first cluster number of the file or of file directory entries of the next level directory, creation and modification times, and other data. The usefulness and reliability of the suggested methodologies have been confirmed by extensive experimental findings, and the models developed within the suggested framework have shown to be more generalizable to new styles and topics.

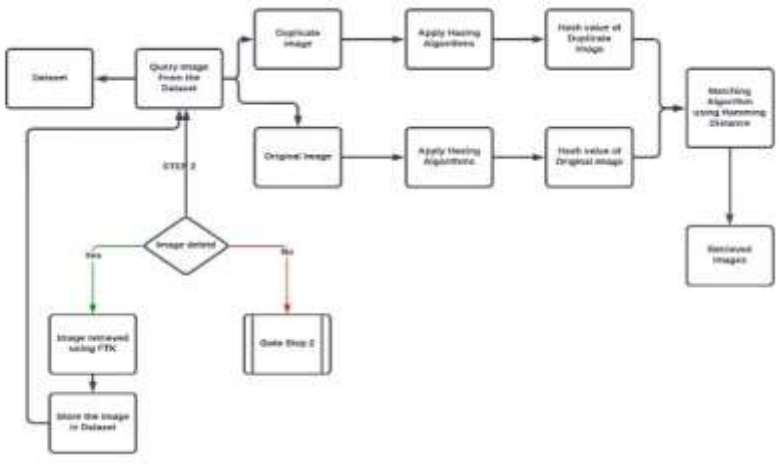
3. Methodology

Image fingerprinting is used for detecting duplicate images ,the main idea for proposing this model is to identify the copy or duplicate images in a large data sets.This method is mainly depend on hashing which Is very efficient in time because using hash is better than any algorithm .In this paper we included image recovery technique ,because when any image is deleted then its useful for recovering that image .we took 101_objectcategories dataset it containd 9144 real time images with 300*200 pixels size.

Duplicate image detection use usually we seen in copy rights strike in social media ex:in youtube some channels getting copyrights strike for copying original content .In this paper we proposed mainly four techniques for detecting copy images .

1) A hashing
2) Perseptua hashing
3) D hashing
4) Haar wavlet hashing

3.1. Flow chart:



Dataset:

The dataset used in this paper is a very large dataset which can be used for image fingerprinting. The dataset contains the data items for various objects which can be used for matching the similar images. The dataset used is 101_ObjectCategories which contains 101 different types of objects and have the data items of 9,144 real world objects in it. Each object has about 40 to 750 images. When compared with the pixels of each image in the dataset is roughly around 300X200 pixels. This dataset is mostly regarded and widely considered for the finding the nearly duplicate images through fingerprinting.

Image Pre-processing:

Images need to pre-processed in order to find the nearly duplicate images through the technique of fingerprinting. Similar images are need to be identified through the process of fingerprinting or popularly known as the hashing technique. We pre-process the image before generating the image hash of all the images, pre-processing is required for the images as it may be difficult to find the duplicate images if the brightness, contrast of the images may be different so we have to normalize these values, for that we need to perform the image pre-processing steps as Decolorizing the images, normalizing the pixel values and down-scaling the image. We decolorize the image in order to recognize the image present in the grayscale format. We also reduce the pixel size of the image through normalization process from 24 bits per pixel to 8 bits per pixel which in-turn is more computationally faster and efficient usage of memory and we finally perform the down scaling the images by choosing the 64-bit hash so we down scale the image to 8X8 pixels.

Analysis and working of A-hash on the dataset:

Each pixel block is compared to the average (as the name implies) of all the pixel values in the image after the decolorizing and scaling steps. The image is scaled to 88 pixels since we will create a 64-bit hash. The pixel block receives a value of 1 (white) if the value is greater than the average, and a 0 otherwise (black). The binary array is flattened into a vector to produce the picture hash. When we apply the average hash function to the 10 changed photos, we can observe that there are only slight variations in the image hash. The picture hash only begins to differ when the brightness is increased by 50%. In each combination, the image hash varies by an average of 2.1.

For any hashing technique we have to follow the steps to generate hash values .

1. 1.Convert to gray scale
2. 2.Resize
3. 3.Colmpute the difference
4. 4.Build the hash

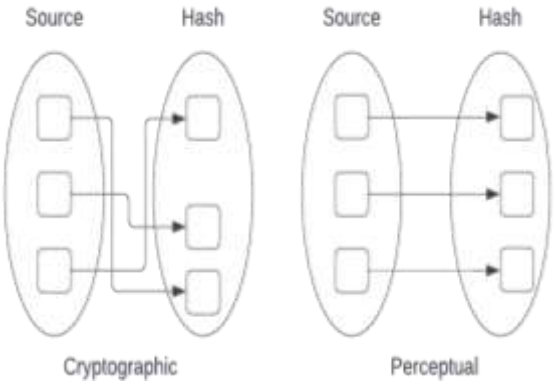
Results shown by A-hash (Average Hashing):

On the basis of the 8(64-bit) input hash size, the average hash function identified 135 groups that might be connected to 335 photos. Even though identical photographs were found, the majority of the groupings had crashes, like the top and bottom left, or almost identical images, like the motorcycles. When the hash size was increased to 16 (256 bits), 28 groups for 64 photos could be found. Although there were no crashes, similar pictures, like the motorcycles, were found.

Analysis and working of P-hash on the dataset:

Simply decolorizing the image is all that is done in the perceptual hash function before applying a Discrete Cosine Transform (DCT); initially per row and then later per column. The high frequency pixels have now been reduced to an 8 × 8-pixel size. The median of all the grey values in the image is then used to compare each pixel block against. The value in the pixel block receives a value of 1 if it is greater than the median and a value of 0 otherwise. The binary array is flattened into a vector after the last picture hash.

Different minor variations in picture hashes are found when the perceptual hash function is applied to the 10 photos with certain modifications. The picture hash begins to diverge significantly more in the event of a 50% brightness increase. In each combination, the picture hash changes four times on average.

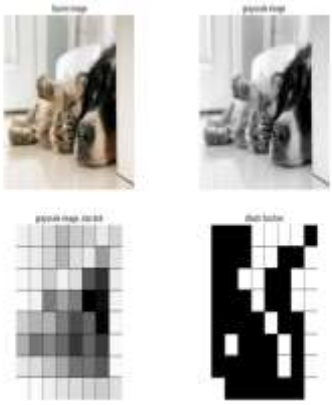


Results shown by P-hash (Perceptual Hashing):

38 groups were found using the perceptual hash function that could be connected to 41 photographs that had the same hash value (threshold=0). Visual investigation revealed no crashes, yet similar pictures, like the motorcycle, were found despite this. With a hash size increase to 16 (256 bits), 10 groups for 20 photos were found. All of the photos are aesthetically similar, and no collisions or nearly identical images are found.

Analysis and working of D-hash on the dataset:

Decolorization and scaling are the first steps. Later the pixels will be serially compared to their neighbour to the right. The image will be scaled down to 9*8 pixels in order to create 64-bit hash. If the byte at position x is less than the byte at position x+1, it gives a value of 1 otherwise 0. It is applied on 10 images with some alterations, the resultant image hash results in some small variations. Increase of bright 3.8 changes in hash image overall the combinations.

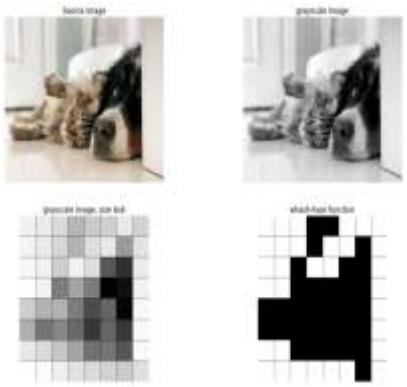


Results shown by D-hash (Difference Hashing):

The differential hash function verified 28 images that might be connected to 31 images with an identical hash(threshold=0). There is no collisions but near-similar images were detected. If the size of hash is increased to 16(256-bit), 16 images for 8 groups were detected. There are no collisions but some similar images, such as motorbikes. If hash size is increased to 16, 8 groups for 16 images were detected. Then no collisions and no similar images are detected, all images are visually identical. The final hash image will be followed by flattening the binary array to vector.

Analysis and working of W-hash (Wavelet Hashing):

After the step of decolorizing and scaling, a -dimensional wavelet transform is carried out to the image. Each pixel block is then compared to the median of all grey values of the photo. If the value inside the pixel block is bigger than the median, it gets cost 1 and in any other case a 0. The final picture hash is followed by using pulling down the array right into a vector. If we evaluate the image hash of the ten altered pictures, we can see that the haar-wavelet hash is quite strong regardless of a few minor changes throughout the altered pictures. On average there are 2.5 adjustments in picture hash across all combos.



Results shown by W-hash (Wavelet Hashing):

The wavelet hash function detected 141 agencies that might be linked to 513 images with an equal hash (threshold=0) based totally on the enter hash length of 8 (64-bit). A visual inspection confirmed that almost all companies contained both collisions or near-equal photographs (Figure 12). Who had recognised that a strawberry ought to have a similar image hash as the motorcycle? By increasing the hash size to sixteen (256-bit), 25 organizations for fifty-one photographs have been detected. No collisions have been present but close to-same photographs are detected, inclusive of the motorbikes.

In model stacking, we create predictions using number of distinct models rather than just one, and then use those predictions as features in a higher-level meta model. Since each type of lower-level model brings a unique set of strengths to the meta model, it can significantly benefit a variety of lower-level learner types. There are numerous ways to construct model stacks, and there is no one right approach to use stacking. With more layers, weights, averaging. A stacked model basically functions by passing the output of several models via a "meta-learner" (often a linear regressor/classifier, but it can also be another model like a decision tree). The meta-learner makes an effort to increase the strengths and decrease the weaknesses of each particular model. Usually, the outcome is a fairly solid model that generalises well to new data.

Analysis and working of File Recovery mechanism:

The major development in this paper is adding of the file recovery mechanism to the existing work. Furthermore, developments took in the recent past suggest to have a recovery mechanism for the deleted images which can be (accidental/intentional) and these images can be retrieved using simple python code which is run through the command prompt and gives result for the significant image type extension and that can be changes according to the use convenience using the file signature. We have also developed a user-friendly GUI (Graphical User Interface) which decreases the complexity of the project and enables the usage for the common people in real time application to find the nearly duplicate images and recovering the deleted images as well.

RESULTS

In the case of image experiments, the results for each hash function were fairly stable. However, using a real data set with 10,000 data items (similar, unsimilar images) and 9,000 images, it becomes clear that both the average hash and the wavelet hash have many collisions when the hash size is 8 (64 bits). Additionally, group many nearly identical images. The delta hash results were accurate, but also the most conservative. Perceptual hashing, on the other hand, also gave accurate results, but is less conservative. When we increased the hash size to 16 (256 bits), none of the hash functions collided, but nearly identical images were found with the average hash and the wavelet hash. Both perceptual and differential hashes showed no collisions and nearly identical images. Again, perceptual hashes are less conservative and the threshold value being zero it showed very good results wth less number of false positives.

IV. CONCLUSION

In the case of the cat and canine image experiment, the outcomes have been quite solid for every of the hash functions. However, when we use the real-world dataset, it becomes clear that a hash length of eight (sixty-four-bit) outcomes in many collisions for both common hash and wavelet hash. In addition, it corporations many close to-same pictures. The results of differential hash had been accurate but also the maximum conservative. On the other hand, the perceptual hash also confirmed correct outcomes but is much less conservative. When the hash size is improved to sixteen (256-bits), none of the hash capabilities ended in collisions, but close to-identical photographs were detected for average hash and wavelet hash. Both perceptual hash and differential hash showed no collisions nor close to-same photographs. Here again, the perceptual hash is much less conservative.

Furthermore, the consequences can also rely on the form of enter photographs. When pix have a solid background, which includes for the motorbikes and traffic signs, collisions can happen extra regularly. One of the reasons is that the binary pixel information characterizes the photo and paperwork the photo hash will become less particular. As an instance, images with a solid historical past that contains any darker object inside the middle can without difficulty get an identical photo hash using common hash and additionally wavelet hash.

V. FUTURESCOPE

The future works to this project can be adding the feature of clustering the images which are of similar hash value and can also use the much larger dataset in order to work with higher memory storage and check the efficiency of the hashing algorithms, and also the same approach of finding the nearly duplicate images can be extended to the study of audio and video fingerprinting as well. The other proposed work can be usage of novel approach of fingerprinting which can reduce the number of false positives. Further scope of the project is to work with all kind of image array types such as the (JPEG, PNG, GIF, TIFF) with including the file signatures of those image types for the image recovering techniques. The present work only deals with the JPG formatted images retrieval so it can be extended to other format images as well.

IV. References

[1] R. S. Khalaf and A. Varol, "Digital Forensics: Focusing on Image Forensics," 2019 7th International Symposium on Digital Forensics and Security (ISDFS), 2019, pp. 1-5, doi: 10.1109/ISDFS.2019.8757557.

[2] S. Agarwal and K. -H. Jung, "Image Forensics using Optimal Normalization in Challenging Environment," 2021 International Conference on Electronics, Information, and Communication (ICEIC), 2021.

[3] Y. Wang, A. Sui and W. Fu, "Research on image retrieval technology based on image fingerprint and color features," 2019 International Conference on Audio, Language and Image Processing (ICALIP), 2019, pp. 447-451, doi: 10.1109/ICALIP.2019.7846636.

[4] Y. Li, D. Wang and L. Tang, "Robust and Secure Image Fingerprinting Learned by Neural Network," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 30, no. 2, pp. 362-375, Feb. 2020

[5] A. Al-Sabaawi, "Digital Forensics for Infected Computer Disk and Memory: Acquire, Analyse, and Report," 2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), 2020, pp. 1-7, doi: 10.1109/CSDE50874.2020.9411614.

[11] R. Zhang and Z. Zhang, "Effective Image Retrieval Based on Hidden Concept Discovery in Image Database," in IEEE Transactions on Image Processing, vol. 16, no. 2, pp. 562-572, Feb. 2007, doi: 10.1109/TIP.2006.888350.

[12] G. Carneiro, A. B. Chan, P. J. Moreno and N. Vasconcelos, "Supervised Learning of Semantic Classes for Image Annotation and Retrieval," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 29, no. 3, pp. 394-410, March 2007, doi: 10.1109/TPAMI.2007.61.

[13] V. A. Wankhede and P. S. Mohod, "Content-based image retrieval from videos using CBIR and ABIR algorithm," 2015 Global Conference on Communication Technologies (GCCT), 2015, pp. 767-771, doi: 10.1109/GCCT.2015.7342767.

[14] S. K. P. N, D. P, A. G, Saritha, S. K and D. R. U, "Automated matching of pixel of interest between two digital images from two different microscope imaging devices," 2020 IEEE Recent Advances in Intelligent Computational Systems (RAICS), 2020, pp. 96-100, doi: 10.1109/RAICS51191.2020.9332522.

[15] Q. Kai, "A Painting Image Retrieval Approach Based on Visual Features and Semantic Classification," 2019 International Conference on Smart Grid and Electrical Automation (ICSGEA), 2019, pp. 195-198, doi: 10.1109/ICSGEA.2019.00052.

[21] S. -L. Hsieh, Chuan-Ren Chen and Chun-Che Chen, "A duplicate image detection scheme using hash functions for database retrieval," 2019 IEEE International Conference on Systems, Man and Cybernetics, 2019, pp. 3487-3492, doi: 10.1109/ICSMC.2019.5642432.

[22] A. Auclair, N. Vincent and L. D. Cohen, "Hash functions for near duplicate image retrieval," 2018 Workshop on Applications of Computer Vision (WACV), 2018, pp. 1-6, doi: 10.1109/WACV.2009.5403104.

[23] X. Wang and L. Liu, "Image Encryption Based on Hash Table Scrambling and DNA Substitution," in IEEE Access, vol. 8, pp. 68533-68547, 2020, doi: 10.1109/ACCESS.2020.2986831.

[24] Zhang, Q., Han, J. A novel color image encryption algorithm based on image hashing, 6D hyperchaotic and DNA coding. *Multimed Tools Appl* 80, 13841–13864 (2021).

[25] N. Zhang, Y. Jiang and J. Wang, "The Research of Data Recovery on Windows File Systems," 2020 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS), 2020, pp. 644-647, doi: 10.1109/ICITBS49701.2020.00141.