ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 5, Issue 3, March 2018

A Study on Near-Duplicate Image Detection System

[1] Dr.A.Mercy Rani, [2] B.Anitha, [3] K.Anukeerthi
[1] Asst.Professor, [2][3] Scholar,Sri SRNM College, Tamilnadu, India

Abstract: **- Due to the abundant increase of imaging technologies, manipulation of digital images create a severe problem in various fields such as medical imaging, journalism, scientific publications, digital forensics etc. This gives the challenges in a matching of slightly modified images to their original ones which are called as Near-Duplicate image detection. The images are altered using some features such as cropping, changing its shape, contrast, saturation, framing etc. Digital Image Processing plays a vital role in finding Near-duplicate images in various applications. The near-duplication image detection process is used to find the duplicate image by comparing the slightly altered images to the original one to assist in the detection of forged images. This paper presents the overview of near duplication images, near duplicate image detection system, algorithms and it gives the analysis of the various researchers held in this field.**

Keywords**: Image Processing, Near-Duplicate Image, Similarity matching, Image Matching Algorithms.**

I. INTRODUCTION

Nowadays the Internet usage is increasing rapidly by the social network users. They are accessing the web for viewing and sharing images, videos, documents etc. Due to this reason the web is occupied with large number of duplicate contents. These duplicate contents must be removed to increase the efficiency of web in terms of its speed, storage, resource sharing etc. Moreover, Near-Duplicate contents also occupy the web heavily. Near-Duplicate contents are created by altering the original content slightly with addition or deletion of some features. Images play a vital role in duplicate contents since they can be modified easily than videos. The original images are altered using some transformations such as scaling, rotation, colouring, compression, brightness changes, cropping etc. to create Near-Duplicate images. These operations perform simple modifications such as bit-level hashing on the original image. Hence it is necessary to detect the Near-Duplicate images to make the web efficient for their users. The method which has the capability to detect variants in images with acceptable degree of reliability and accuracy is used to detect the copyright violations and reduces the redundancy in the collection of images [1].

The digital images are two-dimensional like a photograph or screen display or three-dimensional like statue or hologram. The optical devices cameras, mirrors, lenses, telescopes, microscopes are used to capture the images. Recently a large amount of optical data such as digital images and videos has been captured by visual sensor nodes and they are distributed on web. These visual data creates a lot of Near-Duplicate images.

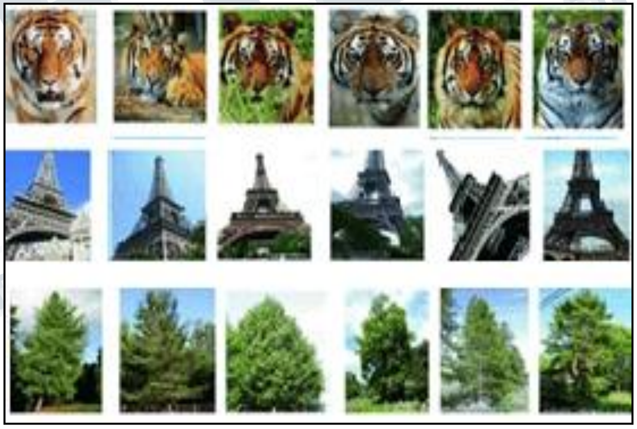

Figure 1. Types of Near-Duplicate images

The images which are in the form of left rotation, right rotation, flipping, cropping, resizing or changing the resolution of an image is known as Near-Duplicate Image. These Near-duplicate images waste the limited storage of memory. The Near-duplicate images can be classified as major duplicate, partial duplicate and scene-object duplicate.

The Figure 1 shows the types of Near-Duplicate images. First row represents the Major duplicate. Second row represents the Partial duplicate. Third row denotes the Scene-object duplicate. The major duplicate images are images which are exactly the original image same or images which have smaller difference in scales, colors, file formats, luminance intensities, and so forth. The parts of the images are fully duplicate then they are known as partial duplicates. The Scene-object duplicates are images they sharing the same 3D scene or the same object[2].

II. NEAR-DUPLICATE IMAGE DETECTION

Duplicate image detection is needed for reducing the storage space, providing user with unique image and for copyrights. In traditional duplicate image detection system, initially the images are converted into a particular image representation and then it is stored in image indexing structure. When a query image is received by the system, it calculates the similarities of the images in the indexing structure by assigning score to each image based on the received query image. Then the Near-Duplicate images are found by applying threshold value to the similarity values of the image [3].

A. Image Matching Algorithms

The Near-Duplicate images are detected using image matching algorithms. Some of the Image matching algorithms are discussed below.

Scale Invariant Feature Transform(SIFT) :

The SIFT [4] was developed by David Lowe in 2004 and it presents a method for detecting distinctive invariant features from images. It can be later used to perform reliable matching between different views of an object or scene. The two key concepts such as distinctive invariant features and reliable matching are used here. In this method, the cascade filtering approach is used to detect the features which transform image data into scale-invariant coordinates relative to local features. This approach has four major computational stages: First stage as Scale-Space extrema detection, second stage as Keypoint localization, third stage as orientation assignment and fourth stage as Keypoint descriptor. According to the name cascade approach, these stages are executed in the descending order. In each stage, a filtering process is applied so that only the key points which are robust enough are allowed to move to the next stage. The researches who tested the SIFT algorithm stated that although SIFT seemed to be the more appealing descriptor; the 128-dimensions of the descriptor vector turn the feature detection into a relatively expensive process[5].

Principal Component Analysis for SIFT(PCA-SIFT) :

The new algorithm which is emerged to improve SIFT and eliminate the computational costs carried with Lowe's implementations is PCA-SIFT[6]. Initially, the PCA-SIFT was developed by Ke and Sukthankar in 2004. After that an evaluation was conducted by Mikolajczyk and Schmid in 2004, and identified that the SIFT algorithm as being the most resistant to common image transformation of the stable feature detection algorithms. Hence, Ke and Sukthankar decided to improve the local image descriptor used by SIFT. This approach uses a Principal Component Analysis (PCA) to detect the local features instead of the SIFT's smoothed weighted histograms. The Principal Component Analysis is a standard technique for dimensionality reduction and it has been applied to a broad call of computer vision problem, including feature selection, object recognition and face recognition. Though the PCA has several limitations, due to its simplicity it remains popular. The PCA-SIFT achieved the ability to speed up the SIFT's matching process by an order of magnitude, but it was proved to be less distinctive than SIFT.

Min-Hash Algorithm :

Min-hash is a Locality Sensitive Hashing scheme [7] that estimates similarity between set of visual words. The similarity between two images can be defined as the Jaccard similarity between the two corresponding sets of visual words I1 and I2: sim(I1, I2) = |I1 ∩ I2|/ |I1 ∪ I2| , which is simply the ratio of the intersection to the union of the two sets. Min-hash is a hash function h : I → v, which maps a set I to some value v. The minimum hash value is found by applying hash function to each visual word in the set I, and it returns the hashed value which is the visual word that has minimum hashed value : min-hash h(I). The hash function is implemented by using a look-up table, it contains visual word and with a random floating-point value followed by a min operator. The computation of the min-hash of a set I is performed by computing a hash of every element in the set. The time taken for its computations is therefore linear in the size of the set|I|. Min-hash has the property that the probability of hashing collision of two sets is equal to their Jaccard similarity: P(h(I1) = h(I2)) = sim(I1, I2). The min-Hash method stores only a small constant amount of data per image[8].

III. A STUDY ON NEAR-DUPLICATE IMAGE DETECTION SYSTEM

SaehoonKim et al [9] presented a scalable solution of Near-Duplicate image discovery on billions of images. This proposed method, initially generate some cluster seeds with min-hashing to divide-and-conquer the images. Then remove the false positive images by growing the cluster seed value using the carefully designed growing function. The bottom–k min-hash is a basis component of the seed growing step and it is used to generate different signatures for removing all candidate images which has only one common visual word with a cluster seed. This method can discover Near-Duplicate clusters with high precision and recall.

JagtapAnkita K et al [3] provides the features used in content based image retrieval and it also discussed the various methods used to reduce semantic gap. It also proposed a new system for retrieval of relevant images with duplicate detection framework. The proposed system performs five steps to retrieve relevant images based on user's query. The proposed method uses a visual vocabulary of vector quantized local feature descriptors to find similarity measures to evaluate Near-Duplicate image detection. Using this duplicate image detection technique with existing Intent search system improves precision of top ranked images. Yue Wang et al [10] presented a new keypoint-based approach to Near-Duplicate images detection. This approach combines the advantages of appearance-based method and keypoint-based method for affine Near-Duplicate image detection. The proposed approach consists of three steps. First, the keypoints of images are extracted and then matched. Second, the matched keypoints are on an affine invariant ratio of normalized lengths. Finally, the matching is confirmed by using the color histograms of areas formed by matched keypoints in two compared images. The proposed algorithm has been tested on Columbia dataset and conducted the quantitative comparison with RANdomSAmple Consensus (RANSAC) algorithm and Scale-Rotation Invariant Pattern Entropy (SR-PE) algorithm.

Ondrej Chum et al [11] proposed and compared two novel schemes for Near-Duplicate image and video-shot detection. They mainly focused on scalability to very large image and video databases, where fast query processing is necessary. The first approach uses Locality Sensitive Hashing for fast retrieval which is based on global hierarchical colour histograms. The second approach uses local feature descriptors and it computes approximate set intersections between documents using a min-Hash algorithm. In both methods, each image is stored only with the small amount of data. The proposed system provides a very time-efficient approximation and it locates duplicate video clips in large corpora automatically.

Jun Jie Foo et al [1] reported two related investigations, first it gathered the results which are returned by a web search engine for popular image queries and it manually analyzed to identify instances and types of near-duplication. Second, it used the combination of existing image matching and mining techniques to verify whether the Near-Duplicate images are removed from the answer sets. The image-matching algorithms such as DPF, PCA-SIFT, and HBC are evaluated on the common alterations images. The query based approaches DPF and PCA-SIFT demonstrated that the effectiveness for Near-Duplicate image detection is more efficient in the former and it is more accurate in later. The non-query-based HBC method is truly effective for automatic detection of Near-Duplicate instances. The proposed method achieves the best efficiency amongst all the considered methods with a trade-off in accuracy compared to the query-based PCA-SIFT method.

Yan Ke Rahul Sukthankar et al [12] proposed a system for near-duplicate detection and sub-image retrieval. This proposed system constructs a parts-based representation of images using distinctive local descriptors since it provides good quality matches though the severe transformations are occurred in the image. It uses locality-sensitive hashing to index the local descriptors to extract the large number of features from the images. The high recall and precision is achieved using distinctive local descriptors and locality-sensitive hashing helps to provide efficient data layout for interactive operation.

Li Chen et al[13] presented an attention based similarity measure in which only very weak assumptions are imposed on the nature of the features employed. This approach determined the similarity by the amount of matching structure detected in the pairs of images. The proposed method is compared with Colour histogram LDF intersection and Gabor based signature matching method. The proposed method evaluation is conducted on the BBC open news archives, which
contain a great diversity in the image database. The observed results demonstrated that the Cognitive Visual Attention CVA based similarity measurement achieves the best precision and recall performance, followed by Gabor-based similarity measurement and color histogram intersection.

IV. CONCLUSION

Recently, the usage of social network is increased rapidly and they share the data such as text, images, videos etc in web. They share the original data or they slightly modify the received data and share it in the web. This causes the storage of duplicate data excessively in memory. Due to this process, the memory space may be reduced abundantly and it may also slow down the network process. Hence, the Near-duplication detection process is necessary to detect the duplicate data from the web. In this paper, the Near-Duplicate image detection system is studied for analyzing the various methods available for near-duplicate detection. From the analysis, it revealed that min-hash, visual vocabulary of vector quantized local feature descriptors, appearance-based method, keypoint-based method, Locality Sensitive Hashing, SIFT and SIFT(PCA-SIFT) methods are widely used method for Near-Duplicate detection system.

REFERENCES

1) Foo, Jun Jie, et al. "Detection of near-duplicate images for web search." Proceedings of the 6th ACM international conference on Image and video retrieval.ACM, 2007.

2) Qiao, Fengcai, et al. "Large scale near-duplicate celebrity web images retrieval using visual and textual features." The Scientific World Journal 2013 (2013).

3) JagtapAnkita K., Tidke B. A on “Review on Content Based Duplicate Image Detection”. International Journal of Science and research (IJSR) -2013.

4) Lowe, David G. "Object recognition from local scale-invariant features." Computer vision, 1999. The proceedings of the seventh IEEE international conference on. Vol. 2. Ieee, 1999.

5) M. Guerrero, "A Comparative Study of Three Image Matcing Algorithms: Sift, Surf, and Fast," Utah State University, Utah, 2011.

6) Ke, Yan, and Rahul Sukthankar. "PCA-SIFT: A more distinctive representation for local image descriptors." Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on. Vol. 2. IEEE, 2004.

7) Indyk, P., Motwani, R.: Approximate nearest neighbors: towards removing the curse of dimensionality. In: Proc. of ACM symposium on Theory of computing. (1998).

8) Lee, David C., Qifa Ke, and Michael Isard. "Partition min-hash for partial duplicate image discovery." European Conference on Computer Vision. Springer, Berlin, Heidelberg, 2010.

9) Kim, Saehoon, et al. "Near duplicate image discovery on one billion images." Applications of Computer Vision (WACV), 2015 IEEE Winter Conference on.IEEE, 2015.

10) Wang, Yue, ZuJunHou, and Karianto Leman. "Keypoint-based near-duplicate images detection using affine invariant feature and color matching." Acoustics,Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on. IEEE, 2011.

11) Chum, Ondřej, et al. "Scalable near identical image and shot detection." Proceedings of the 6th ACM international conference on Image and video retrieval.ACM, 2007.

12) Ke, Yan, et al. "Efficient near-duplicate detection and sub-image retrieval." Acm Multimedia.Vol. 4.No. 1. 2004.

13) Chen, Li, and Fred Stentiford. "Comparison of near-duplicate image matching." (2006): 38-42.