IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 12, Issue 3, March 2023

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.423

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 3, March 2023 ||

| DOI:10.15680/IJIRSET.2023.1203126 |

Sensing of Near Duplicates in Large Image Database

M.Parvathi, K.Kanishka, R.Dhivya, M.Anna Poorani, R.Abirami

Department of Computer Science and Engineering, Mahindra Institute of Technology, Namakkal, India

ABSTRACT: Traditional techniques for image retrieval are not supported for the ever expansive image database. These downsides can be removed by utilizing contents of the image for image retrieval. This sort of image retrieval is called as Content Based Image Retrieval (CBIR). D-SIFT is works with CBIR is focused around the visual features like shape, color and texture. The Density- Scale Invariant Feature Transform (D-SIFT) is a stand out amongst the most locally feature detector and descriptors which is utilized as a part of the majority of the vision programming. We focus texutre ,color,shape,size,string based image matching with better accuracy. These features include Texture, Color, Shape and Region. It is a hot research area and researchers have developed many techniques to use these feature for accurate retrieval of required images from the databases. In this work we present a literature survey of the Content Based Image Retrieval (CBIR) techniques based on Texture, Color, Shape and Region. We also review some of the state of the art tools developed for CBIR.

I. INTRODUCTION

Current image search engines heavily rely on image-totext relevance whereas the surrounding texts extracted from the Web are often inadequate or inaccurate. Therefore, image-to-text techniques, i.e. annotating images with semantic and informative text descriptions, are very useful for improving text-to-image search. One key idea of image-totext is to propagate textual annotations from similar images in a large database to a query image. However, such approaches can hardly scale up to billions of Internet images as web-scale similar image retrieval still remains a challenging problem.
As many images tend to have duplicates on the Web, and each instance of a duplicate may be annotated differently, it is straightforward to aggregate annotations from the duplicates of an input image to obtain accurate, comprehensive textual annotations for the input. Such an idea has been successfully explored in [24]. Although obtaining accurate text tagging for images with duplicates may seem insufficient in the computer vision community, the amount of duplicates is a good indication for the importance of an image. Therefore, it is of great value to obtain accurate tagging for frequently appeared, duplicate images on the Web.

It is very inefficient to retrieve duplicates for each image on the web since in this case, to annotate all images with duplicates would require to take each image in the database and query the database with it, which is quadratic in the size of the database and hence not feasible for large databases. A natural next step is then to discover the duplicate image clusters from all the web images, to aggregate text tags for each duplicate image cluster, and to assign the tags to all the instances of a cluster. In this paper, we study this duplicate discovery problem for billions of images on the Web. the input is billions of images, and the output is duplicate image clusters. Based on the discovered clusters, image annotation can thus be both reliable and efficient through aggregation and assignment within each cluster.

Although duplicate discovery may seem trivial at a glance, it is indeed a very challenging problem. We are not only interested in clustering exact duplicate, but also global duplicate and near duplicate. Exact duplicates are images having exactly the same appearance. Global duplicates tolerate small alterations on the content, whereas near duplicates allows rotating, cropping, transforming, adding, deleting and altering image contents. It is challenging to define a good distance metric such that within a threshold all samples are duplicates (high precision), and all the duplicates fall into the hyper-sphere defined by the threshold (high recall). Furthermore, it is nontrivial to discover duplicate image clusters in web-scale images.

II. IMAGE PROCESSING

Image processing involves changing the nature of an image either improve its pictorial information for human interpretation or render it more suitable for autonomous machine perception. The digital image processing, which involves using a computer to change the nature of a digital image.The digital image define as a two-dimensional

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 3, March 2023 ||

| DOI:10.15680/IJIRSET.2023.1203126 |

function, f(x, y), where x and y are spatial (plane) coordinates, and the amplitude of f at any pair of coordinates (x, y) is called the intensity or gray level of the image at that point. When x, y, and the amplitude values of f are all finite, discrete quantities. The field of digital image processing refers to processing digital images by means of a digital computer. Note that a digital image is composed of a finite number of elements, each of which has a particular location and value and the elements are referred to as picture elements, image elements, pels, and pixels. Pixel is the term most widely used to denote the elements of a digital image.

Large scale duplicate image discovery was often investigated in the context of finding spatially related images, where s-tuple min-Hash values based on bag-of-words (BOW) representations are used to discover seed images, which are further used to retrieve near duplicates. The minhash function was then improved by applying geometric constraints on image features using multiple independent min-hash function. This idea was extended by [9], where s-tuples of min-hash values are extracted within image partitions on grids. Although these local feature-based approaches were shown to be feasible on millions of images, it is still challenging to directly apply them to billion-scale.

First, false positives resulted from min-Hash is a serious problem on a large-scale dataset, which we also reinforce in our study. Second, the complexity of seed growing. Estimated the computational complexity of the seed growing step to be O(NL), where L is the number of seeds and N the dataset size. We observed that L ≈ 0.2B when N = 2B by our approach, resulting in O(NL)≈O(N2). On the other hand, global feature-based compact codes have been widely used for image retrieval or clustering on large-scale datasets. For example, semantic hashing is used to convert the GIST descriptor to a few hundred bits so that real-time search performed on millions of images achieves comparable recognition results to full GIST descriptors. Even raw pixels of tiny 32×32 images were shown to be useful in identifying semantically similar images from a set of 80M images used Haar wavelets of image thumbnails to represent an image, and conducted image clustering on 1.5B images via tree-based nearest neighbor search.

III. LITERATURE SERVEY

1. High-Confidence Near-Duplicate Image Detection

Author – Wei Dong, Zhe Wang, Moses Charikar

Year – 2019

In this paper, we propose two techniques for near-duplicate image detection at high confidence and large scale. First, we show that entropy-based filtering eliminates ambiguous SIFT features that cause most of the false positives, and enables claiming nearduplicity with a single match of the retained high-quality features. Second, we show that graph cut can be used for query expansion with a duplicity graph computed offline to substantially improve search quality. Evaluation with web images show that when combined with sketch embedding, our methods achieve false positive rate orders of magnitude lower than the standard visual word approach. We demonstrate the proposed techniques with a largescale image search engine which, using indexing data structure offline computed with a Hadoop cluster, is capable of serving more than 50 million web images with a single commodity server.

2. A Review on Near-Duplicate Detection of Images using Computer Vision Techniques

Author – K. K. Thyagharajan, G. Kalaiarasi

Year – 2020

Nowadays, digital content is widespread and simply redistributable, either lawfully or unlawfully. For example, after images are posted on the internet, other web users can modify them and then repost their versions, thereby generating near-duplicate images. The presence of near-duplicates affects the performance of the search engines critically. Computer vision is concerned with the automatic extraction, analysis and understanding of useful information from digital images. The main application of computer vision is image understanding. There are several tasks in image understanding such as feature extraction, object detection, object recognition, image cleaning, image transformation, etc. There is no proper survey in literature related to near duplicate detection of images. In this paper, we review the state-of-the-art computer vision-based approaches and feature extraction methods for the detection of near duplicate images. We also discuss the main challenges in this field and how other researchers addressed those challenges. This review provides research directions to the fellow researchers who are interested to work in this field.

3. Efficient Near-duplicate Detection and Sub-image Retrieval

Author - Yan Ke, Rahul Sukthanka, Larry Huston

Year – 2018

We introduce a system for near-duplicate detection and sub-image retrieval. Such a system is useful for finding copyright violations and detecting forged images. We define near-duplicates as images altered with common transformations such as changing contrast, saturation, scaling, cropping, framing, etc. Our system builds a parts-based

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 3, March 2023 ||

| DOI:10.15680/IJIRSET.2023.1203126 |

representation of images using distinctive local descriptors which give high quality matches even under severe transformations. To cope with the large number of features extracted from the images, we employ locality-sensitive hashing to index the local descriptors. This allows us to make approximate similarity queries that only examine a small fraction of the database. Although locality-sensitive hashing has excellent theoretical performance properties, a standard implementation would still be unacceptably slow for this application. We show that, by optimizing layout and access to the index data on disk, we can efficiently query indices containing millions of keypoints. Our system achieves nearperfect accuracy (100% precision at 99.85% recall) on the tests presented and consistently strong results on our own, significantly more challenging experiments. Query times are interactive even for collections of thousands of images.

4. Duplicate Discovery on 2 Billion Internet Images

Author - Xin-Jing Wang, Lei Zhang, Ce Liu

Year – 2021

Duplicate image discovery, or discovering duplicate image clusters, is a challenging problem for billions of Internet images due to the lack of good distance metric which both covers the large variation within a duplicate image cluster and eliminates false alarms. After carefully investigating existing local and global features that have been widely used for large-scale image search and indexing, we propose a two-step approach that combines both local and global features: global descriptors are used to discover seed clusters with high precision, whereas local descriptors are used to grow the seeds to cover good recall. Using efficient hashing techniques for both features and the MapReduce framework, our system is able to discover about 553.8 million duplicate images from 2 billion Internet images within 13 hours on a 2,000 core cluster.

IV. EXISTING SYSTEM

This is the most common form of text search on the Web. Most search engines do their text query and retrieval using keywords. The keywords based searches they usually provide results from blogs or other discussion boards. The user cannot have a satisfaction with these results due to lack of trusts on blogs etc. low precision and high recall rate. In early search engine that offered disambiguation to search terms. User intention identification plays an important role in the intelligent semantic search engine. The similarity assessment is fundamentally important to many multimedia information processing systems and applications such as compression, restoration, enhancement and copy detection etc. The image similarity assessment is to design algorithms for repeated and objective evaluation of similarity in a consistent manner with individual human judgment.

DISADVANTAGE

- Loss of Global Weighting. Predefined fixed weights are adopted to fuse the distances of different low-level visual features.
- Loss of Adaptive Weighting. adaptive weights for query images to fuse the distances of different low-level visual features. It is adopted by Bing Image Search.
- For our new approaches, two different ways of computing semantic signatures are compared.
- Not Visual : Query-specific visual semantic space using Reciprocal Hash Maps. For an image, a single semantic signature is computed from one SVM classifier trained by combining all types of visual features.

V. PROPOSED SYSTEM

The proposed system Content-Based Image Retrieval (CBIR) uses D-SIFT algorithm the visual contents of an image such as color, shape, texture, and spatial layout to represent and index the image. Active research in CBIR is geared towards the development of methodologies for analyzing, interpreting cataloging and indexing image databases. In addition to their development, efforts are also being made to evaluate the performance of image retrieval systems. The quality of response is heavily dependent on the choice of the method used to generate feature vectors and similarity measure for comparison of features. In this work we proposed an algorithm which incorporates the advantages of various other algorithms to improve the accuracy and performance of retrieval. Halftone is the reprographic system that re-enacts nonstop tone symbolism through the utilization of specks shifting either in size or in separating subsequently creating a slope like effect. Halftone can likewise be utilized to allude particularly to the picture that is delivered by this process.

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 3, March 2023 ||

| DOI:10.15680/IJIRSET.2023.1203126 |

ADVANTAGE

- It Represent all of the descriptors of an image via sparse representation and assess the similarity between two images via sparse coding technique.
- The main advantage is, a feature descriptor is sparsely represented in terms of a Dictionary Score or transferred as a linear combination of Dictionary Score atoms, so as to achieve efficient feature representation and robust image similarity assessment.
- Best Results.
- High accuracy.
- High performance in search of related image re-ranking.

MODULE DESCRIPTION

Image Recognition

Image recognition, SIFT features are first extracted from a set of reference images and stored in a database. A new image is matched by individually comparing each feature from the new image to this previous database and finding candidate matching features based on Euclidean distance of their feature vectors. The fast nearest-neighbor algorithms that can perform this computation rapidly against large databases. The keypoint descriptors are highly distinctive, which allows a single feature to find its correct match with good probability in a large database of features .A cluttered image, many features from the background will not have any correct match in the database, giving rise to many false matches in addition to the correct ones.

Image Copy Detection

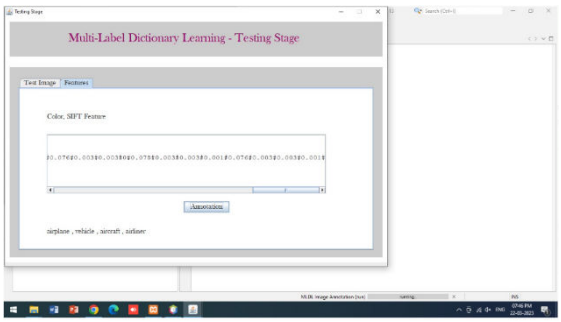
The increasing availability of digital multimedia data, the integrity verification of image data becomes more and more important. Digital images distributed through the Internet may suffer from several possible manipulations. To ensure trustworthiness, image copy detection techniques have emerged to search duplicates and forgeries.Image copy detection can be achieved via image hashing or watermarking techniques.

Image Retrieval

The most popular image retrieval approach is Content-based approach image retrieval (CBIR). A query image, extract its Dictionary Score feature (with atoms) and transmits the features into an image database, the each image is stored together with its Dictionary Score feature and original SIFT feature vectors . The most common technique is to measure the similarity between two images by comparing the extracted image features.

Content Based Image Retrieval

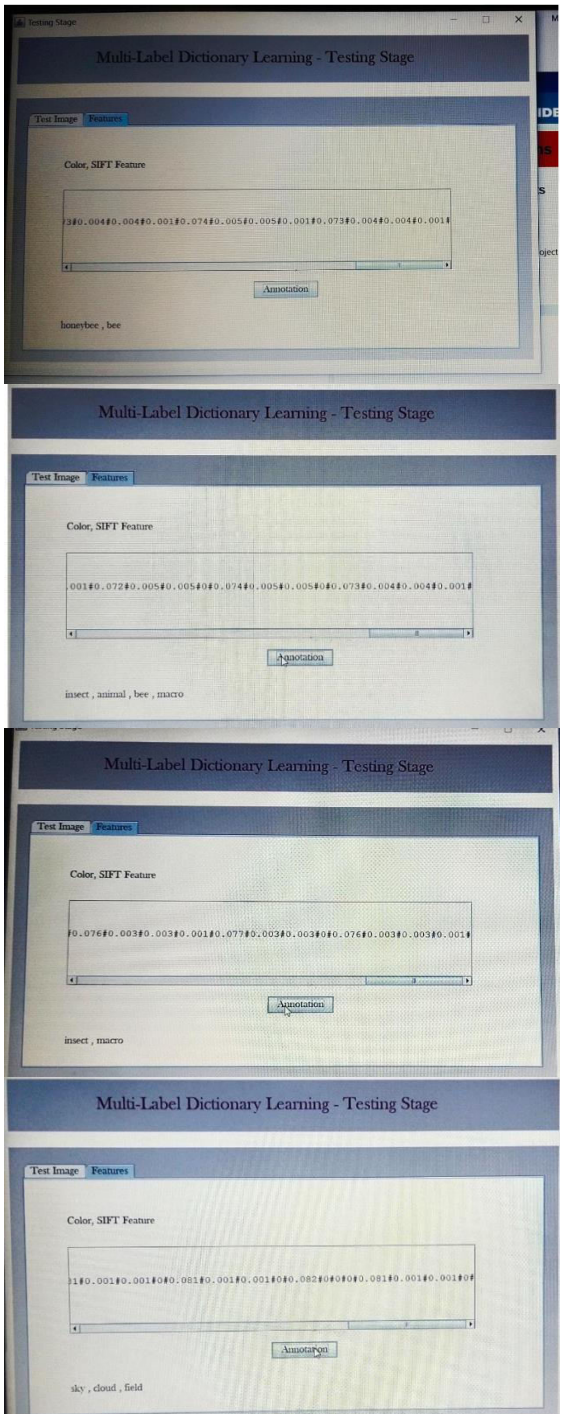
Content-based image retrieval (CBIR) systems needed to effectively and efficiently use large image databases. A CBIR system, users will be able to retrieve relevant images based on their contents. CBIR systems followed two distinct directions.



|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 3, March 2023 ||

| DOI:10.15680/IJIRSET.2023.1203126 |



VI. CONCLUSION

In the D-SIFT feature extraction, D-SIFT transforms image data into scale-invariant coordinates virtual to local features and generates large numbers of features that compactly cover the image over the full range of scales and locations. The low contrast points or poorly localized along an edges are removed by key point localization. A keypoint has been found by comparing a pixel to its neighbors and is to perform a detailed fit to the nearby data for location, scale, and ratio of key curvatures. To make the D-SIFT feature more compact, the bag-of-words (BoW) representation approach quantizes D-SIFT descriptors by vector quantization technique into a collection of visual words based on a pre-defined visual vocabulary or vocabulary tree.

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 3, March 2023 ||

| DOI:10.15680/IJIRSET.2023.1203126 |

REFFERENCES

1. Li-Wei Kang, Member, IEEE, Chao-Yung Hsu, Hung-Wei Chen, Chun- Shien Lu, Member, IEEE, Chih-Yang Lin, Member, IEEE, and Soo-Chang Pei, (2011) "FeatureBased Sparse Representation for Image Similarity Assessment", IEEE Transactions
2. Sivic J and Zisserman A, (2003) "Video Google: A text retrieval approach to object matching in videos," in Proc. IEEE Int. Conf.
3. C. Kim, "Content-based image copy detection," Signal Process.: Image Commun.,
4. Lowe D. G, (2004) "Distinctive image features from scale-invariant keypoints," Int. J. Comput.
5. Ke Y., Sukthankar R and Huston L, (2004) "Efficient near-duplicate detection and subimage retrieval,"
6. H. R. Sheikh H. R and Bovik A. C, (2006) "Image information and visual quality," IEEE Trans.
7. Nistér D and Stewénius H, (2006) "Scalable recognition with a vocabulary tree," in Proc. IEEE Conf.
8. Aharon M., Elad M and Bruckstein A.M, (2006) "The K-SVD: An algorithm for designing of overcomplete dictionaries for sparse representation.
9. Monga V and Evans B. L, (2006) "Perceptual image hashing via feature points: Performance evaluation and tradeoffs," IEEE Trans.
10. Zhang J., Marszalek M., Lazebnik S., and Schmid C, (2007) "Local features and kernels for classification of texture and object categories: A comprehensive study," Int. J. Comput.

INNO SPACE
SJIF Scientific Journal Impact Factor
Impact Factor: 8.118

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details