ISSN 2278-3091
Volume 9, No.1.1, 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse7491.12020.pdf
https://doi.org/10.30534/ijatcse/2020/7491.12020

Near-Duplicate Images Detection and Clustering in Large-Scale Image Dataset: A Systematic Literature Review

Nadiah Yusof[1], Amirah Ismail[2], Nazatul Aini Abd Majid[3]
[1]Universiti Tun Hussein Onn Malaysia, nadiahyusof@gmail.com

ABSTRACT

Near-duplicate images detection (NDID) is a different image but have similarity in scenery, object and content. Research in NDID always related with the database either conventional or cloud computing. The content inside the database is in high quantities. Various techniques are used to help manage, detect and clustering ND images contained in the database, but the problem is, how accurately detecting NDID in features extraction and this study are still ongoing and facing several issue and problems. The issue and problem in NDID related to detec-tion and clustering the similar images. Therefore, this study was conducted to study the techniques that have been used previously for de-fined the suitable technique that can be used and help in images features extraction because features extraction is the most important part in NDID research, in order to allow the system to understand what requirements and unique structure for different images. The methodology is implemented in five phases which are review paper, collecting data, compare data, find out strength and weaknesses and come out with innovative ideas. The results of this study are to discuss the framework for the use of existing techniques to see gap for improvement that can be done in subsequent studies..

Key words : : Near-duplicate images detection

1. INTRODUCTION

Daily life activity in uploading images by users through various online media applications increase the number of images in the online database [1], [2]. This action triggers a dumping of identical a few images of an object's position or a scene's similar in the image and is identified as ND images. According to the [3] study, the percentages of ND for images should almost exceed 85% of the equation then the image is categorized as ND images. ND Images are also similar in terms of the structural arrangement of objects and movement in images but differ in colour mapping, scaling, rotation, size, and format conversion [4], [5].

Research in near-duplicate images detection (NDID) has two different points of view, the first point of view is; more images near to the user query in the image detection, will increase NDID image similarity and help users to have more options to retrieve a similar image that meets the user's requirement. While, the second point of view is, wasting of database storage in placing ND images. However, despite these two points of view, most of researchers agree that the bigger problem faced by online images is how the system will help to detect and filter roughly similar to the image for matching comparison purposes and then clustering the images according to similar features and structure of the images [1][6]–[11].

'How the system will help to detect and filter roughly similar image?' is the most important research question in NDID Research field. The flow and process in NDID displaying that this research is definitely important to proceed and need to be fulfilled by researcher because a system has to be discovered and recognize relevant image following by the user query image. In logical declaration when the system cannot retrieve relevant result the system will be considered as no longer applicable and the user will not be engaging with the system in the near-future. The aimed of this paper to discuss all important and point of view are related in NDID field in a critical thinking way and we call it as systematic literature review. Finding for this paper is theoretical framework for NDID research.

The arrangement of this paper is structured as follows. The first is the introduction, following by elaborate discussion in CBIR component in section II. The aim of this paper and research questions will be defined and described in section III. Briefly, the word in NDID challenges and characteristics will be discussed in section IV. Next section is about NDID characteristic. Last part in this paper is the discussion area will be discussed in section VI.

1.1 Research Background in Content-Based Image Retrieval

Fundamental in content-based image retrieval will be described as discovering similar image for query image accordingly features structure inside the images. Figure 1 shows a traditional CBIR technique process. The flows starting with user query by uploading either text or image into query section then the system will be described the query by extract the features, in order to proceed in similarity comparison part with images insides database, continues by a system need to index and retrieve the result following by specific features and component in the images.

According to [12][13] traditional CBIR technique have five basic elements to perform near-duplicate image

detection process. The five elements in traditional CBIR is text, color, texture, shape and sketch. First element is text, matching in text are performed by a representation of images or information contained in the database using binary code. Hence, texts used by users when searching for information via search engines are processed and matched by the text that has been adjusted for images or information contained in the database [14]. Text-based is using a traditional database to manage images. Through text descriptions, images can be organized by topical or semantic hierarchies to facilitate easy navigation and browsing based on standard Boolean queries. However, since automatically generating descriptive texts for a wide spectrum of images is not feasible, most text-based image retrieval systems require manual annotation of images [15], [16]. Obviously, annotating images manually is a cumbersome and expensive task for large-scale image datasets, and is often subjective, context-sensitive and incomplete. As a result, it is difficult for the traditional text-based methods to support a variety of task-dependent queries [17] due to the fact of that reason, more researcher is trying to solve this problem by way of proposed a new technique like the color and so on.

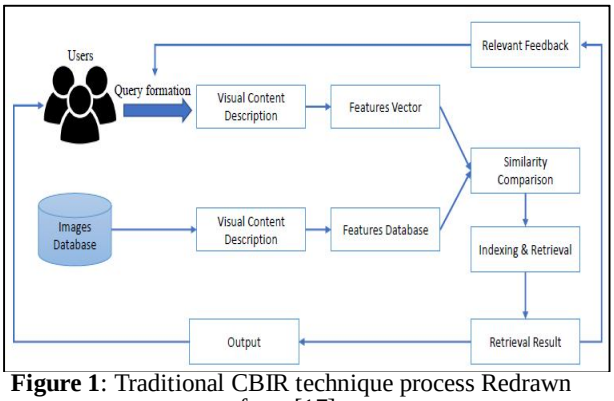

Figure 1: Traditional CBIR technique process Redrawn from [17].

Second element in CBIR is a colour, Colour is the first and one of the most widely used visual features in image retrieval and indexing [13]. The most important advantages of colour element are power of representing visual content of images, simple extracting colour information of images and high efficiency, relatively power in separating images from each other [18], relatively robust to background complication and independent of image size and orientation [19][12][20][21]. The process carried out in this element through the decomposition of the colour histogram found in the query image and the corresponding match executed in colour similar to the image of the query image [22], [23].

The colour histogram method introduced in [24][21] has shown to be very effective and simple to implement. Use of colour histogram is the most common way of representing colour feature [25][21]. Despite some drawbacks, the colour histogram had been used in many kinds of research and significant efforts were done for overcoming its weakness [26][27][28][21]. One the disadvantage of the colour histogram method is that it is not robust to significant appearance changes because it does not include any spatial or texture information [27][21].

Third element are available in CBIR are Texture. Various texture representations have been investigated in pattern recognition and computer vision. Basically, texture representation methods can be classified into two categories: structural and statistical [29], [30]. Structural methods, including morphological operator and adjacency graph, describe texture by identifying structural primitives and their placement rules. They tend to be most effective when applied to textures that are very regular. Statistical methods, including Fourier power spectra [31], co-occurrence matrices[28], [32], [33], shift-invariant principal component analysis (SPCA) [34], [35], Tamura feature[36], Wold decomposition [37], Markov random field [38], fractal model [39], and multi-resolution filtering techniques such as Gabor and wavelet transform [40], characterize texture by the statistical distribution of the image intensity [28], [32], [36], [40], [41]. However, to differentiate between two or extra object in the image need to apprehend the shape first, in order to retrieve an excellent result.

Forth element is shape-based technique is formed very popular technique compared to other techniques found in CBIR. This is because the forming technique is parallel to the needs of the user as it helps to elaborate on the characteristics of the shapes found in the image, the results obtained more precisely than the above three techniques [29]. Compared with colour and texture features, shape features are usually described after images have been segmented into regions or objects [41]–[44]. Since robust and accurate image segmentation is difficult to achieve, the use of shape features for image retrieval has been limited to special applications where objects or regions are readily available. The state-of-art methods for shape description can be categorized into either boundary-based (e.g. rectilinear shapes [45], polygonal approximation [46], finite element models [47], and Fourier-based shape descriptors [48][49][50] or region-based methods[44][51] (e.g. statistical moments [52][53]). A good shape representation feature for an object should be invariant to translation, rotation and scaling [17].

Fifth element in CBIR is sketching techniques. This element implementation and motivated by children who have weaknesses in reading and writing but basically children can sketch their imagined image and this study is implemented to help children experiment the context of their query in the query space through the sketch technique obtained the query image results [54]. Sketch techniques are implemented through local image matching methods by distinguishing query images through spaces and matching image features with each space contained in the storage images in the database [55]. Early 1990s research in Sketch-Based Image Retrieval (SBIR) studies have been started [26][56]. Based on the cartoon image domain, using sketching technique is far more superior than keyword search and more user-friendly [57][58] especially children who don't have knowledge in reading and writing. Generally, to helping the user find the desired

image easily and effectively [59][56] this technique have been approached. However, this technique facing limitation according to sketching by the user cannot be matched with the images in a repository and another limitation is sketch features extraction need the right bone to be matching and it was not an easy task [60][56]. According to [10] traditional CBIR approach have a top-notch process in small-scale image dataset. However, when the system dealing with a massive or large-scale image dataset the system will be going through a little bit slow to process image features extraction. This is the strong reason why many other researchers proposed a various technique in order to resolve this problem.

Thus, in solving the problems faced by traditional CBIR technique, various techniques and new models introduced to help expand the features of the ND image in the implementation of decomposition and matching tasks, basically, other techniques are also the expansion and extension of the original ideas from the techniques are available in the CBIR technique.

Numerous studies and techniques involving in NDID have been proposed, below is continues with new expansion technique an example is a bag of visual word technique and Min-Hash technique [20], this technique focuses on ND image clustering through space and position matching in the image. While the Singular Value Decomposition-Scale Invariant Feature Transform (SVD-SIFT) technique [21] emphasizes research using catalytic methods to accelerate the process of detecting ND images.

Continuously by Visual Salient Riemannian technique [22] to identify the prominent key space in the processed image and this technique of optimizing the use of databases by reducing the ND image detection process. Following by Others technique in order to help image features extraction is deduplication technique [23]–[25][2]. Deduplication technique thereby helping to reduce energy consumption that can increase heat production. The Similarity Join Operator technique [26][27] detect ND images based on the absolute ratio of the absolute equation. While Fourier-Mellin Transform [28] techniques to detect ND images through rotation of image rotation, image scale measurements and invariant changes contained in the image.

Then the Haar wavelet technique [28], [29] extracts the vector inside the image to determine the distance Manhattan objects in the image aim to evaluate the precision of the image. While Kernel Hashing techniques [30] detects ND images through the diversity of features in the images to define the differences between each image and converts them into binary images into the kernel space. However, most of the proposed techniques more focus on image feature extraction for a small-scale image dataset.

Hence, to solve the NDID features extraction problem, various other techniques have been introduced. Other techniques have been introduced are locality-constrained linear coding (LLC) & max-IDF techniques [10]. This technique is intended for clustering of ND images through the distribution of images into different cluster bucket and each space divided by unique key features contained in the image. The next action for this technique is to match the space ND images to allow cluster the similar image based on the unique key features of ND of images. However, this research emphasizes more on detection of local features extraction image representation.

Previous researchers have concentrated on amending the similarity measurement of NDID features extraction. Bag of words model was proposed by [31] to utilize the advantages of a min-hash technique by extract local features. Jaccard coefficient is a part of Min-Hash technique and has been used to detect the similarity [31], [32]. Although these approaches have shown that they are possible to calculate the similarity of NDID, the weaknesses are how to directly apply Min-Hash technique on large-scale image clustering in order to solve the problem [10].

A few researchers have focused on one part in NDID by enhancing of ND image clustering part. They developed an interesting algorithm to combine both the local and global features to discover the ND of images by approach Locality-Constrained Linear coding and max-idf technique. The local features are to discover growth of seed image and global features is to discover seed cluster [33]. However, this technique has a limitation inherent by global features [10].

This is followed by [34] who said if we focus just on one side either local or global in features extraction, then the similarity will increase up to 54% of similarity, while hybrid the two types of features extraction area the result approximately increases to 65.5%. Hence, a systematic literature review on NDID is required to be done because of it easy for future researcher find a compiled research in NDID topic. An investigation into the systematic literature review in which the most important aspects need to be accomplished by researchers and functional information and statistics must be pulled out [35].

Based on Scimago Journal & Country Rank research in NDID are related and included inside Computer Vision and Pattern Recognition (CVPR) category. Before we continue in the depth into NDID discussion we clarify all the publication are interested in this area. This is because we will see how important NDID research.

2. METHODS

The steps in the systematic literature review in this paper are documented as below;

2.1 Survey Objective and Research Questions

Survey objective and research question are described in section A and statistic on published paper and presented papers in different journals and conferences are presented in section B.

The authors of the current paper have profited from "Cloud computing service composition: A systematic literature Review"[61], “Guidelines for performing Systematic Literature Reviews in Software Engineering” [62] to conduct and perform this research.

2.2. Publication statistics

The present research objective at collecting and investigating all the credible and effective studies that have to examine NDID field. More specifically, the technique of features extraction and methods of paper will be considered, and their characteristic will be described in this paper.

To fulfil the above-mentioned objective and identify the techniques have been selected by the researcher for their studies are covered for which new techniques are proposed, dataset and benchmarks have been used. Most researchers considered NDID research parameter and objective function and user requirement that are important in designing these NDID techniques. The following statement is the research questions (RQs) are raised.

RQs 1: What are the main objective of the researcher?
RQs 2: What is the proposed approach and what are the techniques have been used?
RQs 3: What datasets or benchmark are used for features extraction calculation?
RQs 4: What evaluating procedures have been used to compare the better result in NDID area?
RQs 5: What others research has been considered in each paper to compare the result?
RQs 6: How the SLR discussion will be give an extra-or dinary view to convince another researcher this re search is important?

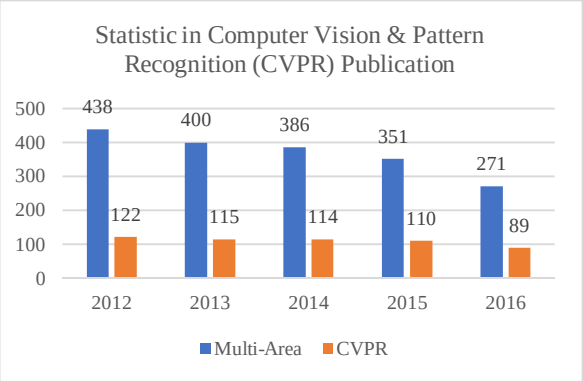

Figure 2: Statistic graph of journal and conference publication for two types data are related in CVPR. (Data: Scimago Journal & Country Rank)

In this study, attempts have been made to examine all off-journal and conference publication are related to image processing under computer vision and pattern recognition specific in image retrieval processing field. To achieve this goal and answer the research questions in section A.

Based on Scimago Journal & Country Rank (SJR), journal and proceeding statistics shows the reduction of a journal and proceeding related CVPR listed in SJR ranking from 2012 to 2016. Research in NDID usually placing under CVPR categories and this is the main reason in this paper we are choosing CVPR as an important keyword for searching NDID research area in journal and proceeding publication. The fields involved in the publication of the Journal and the proceedings displayed in the graph shows in Figure 2. The blue blocks in this graph involving the field multi-area such as artificial intelligence, software engineering, applied mathematics, computational theory and mathematics, human-computer interaction, computer graphic, computer-aided design and so on. While orange blocks are specific for CVPR area.

Ranking of journal and proceeding publication in CVPR area decreased can be caused by several factors. The first factor is relating to the specialization of the field in CVPR has been restructured according to the suitability of the scope in the field conducted by SJR. This can be seen with the specialization of the Human-Computer Interaction which was previously a field listing with CVPR. Furthermore, the second factor is the publication that involves the diversity of fields in computer science and is no longer specific to any specific area leading to a decline in CVPR publishing ranking. However, this decrease does not affect the expansion of the study conducted in CVPR, especially studies related to CBIR and particularly ND image detection.

3. NEAR-DUPLICATE IMAGE DETECTION

A large amount of data in multimedia such as images, videos, audio and animation has been exploding into websites such as YouTube, Facebook, Google video, flicks, Instagram, Twitter and many others [63]. This activity will give negative impact into database management system and cause repetition of the same image and we call it as an ND image. Nowadays shows, more than 80 percent of the web images sources are not original [64][11]. This issue makes it this research become important because the massive of repetition data need to cluster according to their category.

3.1. Near-Duplicate Image Detection Problem and Issue

In this section, we will be discussing problem and issue in NDID from the previous researcher. The widespread problem in NDID is how to extract image features and cluster the large-scale image dataset and it's very challenging for effective indexing and retrieval [10], [66][1][77]. Below are several discussions of the specific issue and problem in NDID and clustering from the previous researcher.

First issue inside features extraction in image management system is image clustering and this part are very important because will help image management system structure their image according to the same image features similarity. Without thinking about the sizes of images sets,

the ND images clustering issue can be dealt with as a confined CBIR issue, and it isn't exceptionally hard to support the identification of ND images from small-scale images datasets. But, when the images dataset scale increments to a huge level, a few significant issues may arise. Firstly, the computation complexity for most clustering algorithm is O(n2) or significantly more, in this way they can't specifically be applying for distinguishing ND images from large-scale images datasets. Secondly, most clustering algorithms are huge data dependent, which may not ready to help parallel computing. Thirdly, the highlights utilized as a part of the customary CBIR algorithm are generally in high dimension, which may bring about high computational complexity calculation [10].

Second issue is the greater part of the spots of interests has a tremendous number of images shared by the worldwide user. Meanwhile, some of the images are upload, already adjusted and duplicated by the different user previously being partaken in social networks. Likewise, there exists a huge number of ND images in the website. The current ND images processing approaches mainly focused on finding the ND images, where an inquiry image is required. Be that as it may, how to process the ND images clustering automatically from the web-scale social images is very difficult [66].

Third issue in NDID is computational proficiency, the vectorially representations were first implanted into binary codes in a few works [78], [79]. In this specific circumstance, a key issue was to guarantee that the images that were comparable in the original vector space ought to be low dimension [73]. In the binary code space. The ND images comparability can be effectively figured out by the Hamming distance between the binary codes. In spite of the simple way, representing an image by a single vector, as a rule, neglects to adapt to the varieties among the ND pictures as shown in [80]. Moreover, the dimension of the images features must be resolved from the earlier, before solving the images characteristic. What's more, the vectors are bad at displaying the connections among a few sections of the image [73].

Fourth issue in NDID is multimedia data indexing is a challenging process because of calculation dimensionality and it's much debated among researcher. Thus, the design of special indexes for multimedia has been a currently functioning of research [82][68] and recognizing an image that has been conceivably altered in NDID is a vital request of numerous applications, frequently including large-scale image datasets, extending from several thousand to a huge number of images. The undertaking is made complex in light of the fact that the query image is only here and there an perfect similarity of the reference images in the database, having endured numerous conceivable changes, including cropping and occlusions, changes of scale, rotation, non-affine geometric transformations, photometric and colorimetric changes, compressions and noise, and other arranged transformation, such as dithering and fancy artistic side effect [68].

The last issue in NDID is false positive is a complicated issue for large-scale image clustering since encoding the original features into binary codes isn't a simple issue. next, the learning-based hash techniques require various iteration for preparing the models, which may set aside a ton of time for a large-scale image dataset. these techniques are altogether intended to keep running on a single machine. for the issue of large-scale image clustering, huge quantities of an image cannot fit on one single machine and can't basically apply the conventional calculations to help NDID clustering [10].

Based on a previous research problem, a various hybrid technique has been approach, all the approach technique will be including with algorithm for features image detection and clustering. In section 4.3 comparing and discussion NDID algorithm has been proposed by previous researcher.

3.2 Comparing NDID algorithm

In this section, we discuss about and evaluate NDID present- day algorithm that has been approach and practice by the previous researcher. According to the ordinary lookup and finding, the new researcher found many strategies how to solve the bother are associated to NDID that can purpose the hassle on NDID.

Table 1: compile of algorithm in NDID research area

Reference	Techniques	Algorithm	Justification
Zheng et. al. 2012 [63]	Rotation, Scale, Translation (RST) Invariance features + Salient Covariance Matrix (SCOV) + ICA Independent component analysis	Original $p(F(X)) = \prod_i p(f_i(X))$ Enhancement Algorithm $p(F(X)) = \prod_i p(f_{\Delta i}(X))$	"Note that because H1...H2 consists of only binary coefficients (1and 1), the projections therefore only contain addition operations which is a great deal less difficult than doing real-value multi- plication ".

Bueno et. al. 2012 [68]	Scale in Variance Features Transform (SIFT)	$L_i = \frac{P(D_i\|X)+\epsilon}{P(D_i\|\overline{X})+\epsilon}$ + $P_j(X\|D_1 \cap D_2 \ldots \cap D_n) = \frac{\prod_{i=1}^{n} L_i \times P(X)}{\prod_{i=1}^{n} L_i \times P(X) + P(\overline{X})}$ = $LQ_i = \frac{P(D_i\|\overline{X})+\epsilon}{P(X)\times P(D_i\|X) + P(\overline{X})\times P(D_i\|\overline{X})+\epsilon}$	Those that correspond to features between "correct" images (between near-duplicates) and those between "incorrect" (unrelated) images.
Dong et. al. 2012 [4]	Scale in Variance Features Transform (SIFT)	$N_d(f_k) = \{\langle f_{ij}, i\rangle \in \mathbf{DB} \mid \|f_k - f_{ij}\|_2 \le d\}.$ Merge $N_d(Q) = \cup_k N_d(f_k)$	Can be applied to improve recall
Z. Li & Feng 2013 [67]	Locality Sensitive Hashing + K-Nearest Neighbor SIFT + K means + BoVW	$L_i = \frac{P(D_i\|X)+\epsilon}{P(D_i\|\overline{X})+\epsilon}$ + $P_j(X\|D_1 \cap D_2 \ldots \cap D_n) = \frac{\prod_{i=1}^{n} L_i \times P(X)}{\prod_{i=1}^{n} L_i \times P(X) + P(\overline{X})}$ = $LQ_i = \frac{P(D_i\|\overline{X})+\epsilon}{P(X)\times P(D_i\|X) + P(\overline{X})\times P(D_i\|\overline{X})+\epsilon}$	In some features the question images may additionally moreover now not have a correspondent near-duplicate in database (absent queries). In these situations, the Bayesian search proposed till now cannot study an absent query, and many, or even all the sides of a query images would be processed to get hold of no reply or a wrong retrieved image.
Kalaiarasi & Thyagharajan 2013[69]	Colour Texture Moment (CTM)	Not mentioned	
J. Wang 2013 [70]	Strong Geometry Consistency (SGC) + Scale Weighting	$s = s_q/s_p, \theta = \theta_q - \theta_p$ Enhancement $SW = (s_1 \times s_2)^n$	The error of spatial areas of the detected feature points is (empirically) assumed to be inside 5 pixels.
L. Li et. al. 2013 [71]	Bag of Visual Word (BoV)	$S(r_k) = \sum_{r_i \neq r_k} \exp(D_s(r_k, r_i)/\sigma_s^2) w(r_i) D_r(r_k, r_i)$ Combination $D_s(r_k, r_i) = \sum_{m=1}^{N_k}\sum_{n=1}^{N_i} f(c_k, m) f(c_i, n) D(c_{k,m}, c_{i,n})$	The discriminative power of the traditional BOV model is limited by quantization errors
J. Li et. al. 2014 [66]	Color Moment + Wavelet Transform + SIFT	$Merge(\Omega_i, \Omega_j) = \begin{cases} 1, & \text{if } M_{ij} \ge P_i/2 \,\&\, M_{ij} \ge P_j/2 \\ 0, & \text{otherwise} \end{cases}$	Two images have sufficient matched SIFT point pairs
Battiato et. al. 2014 [72]	Bag of Visual Word (BoVW)	$\tau(\Psi_I, \Psi_J) = \sum_{p=1}^{P} \min(\Psi_p(I), \Psi_p(J))$	Search for the same scene with possibly different photometric and/or geometric variations, given an image as query.
L. Liu et. al. 2015 [73]	K-Nearest Neighbour	– Euclidean distance $\mathfrak{D}(\mathbf{H}, \hat{\mathbf{H}}) = \sqrt{\sum_{i=0}^{b-1} (\mathbf{H}[i] - \hat{\mathbf{H}}[i])^2}$ – Histogram intersection [38] $\mathfrak{D}(\mathbf{H}, \hat{\mathbf{H}}) = 1 - \sum_{i=0}^{b-1} \min(\mathbf{H}[i], \hat{\mathbf{H}}[i])$ – Chi-square statistic [38] $\mathfrak{D}(\mathbf{H}, \hat{\mathbf{H}}) = \sum_{i=0}^{b-1} \frac{(\mathbf{H}[i] - \hat{\mathbf{H}}[i])^2}{\mathbf{H}[i] + \hat{\mathbf{H}}[i]}$ – Canberra distance [39] $\mathfrak{D}(\mathbf{H}, \hat{\mathbf{H}}) = \sum_{i=0}^{b-1} \frac{\lvert\mathbf{H}[i] - \hat{\mathbf{H}}[i]\rvert}{\mathbf{H}[i] + \hat{\mathbf{H}}[i]}$ – Bhattacharyya distance [40] $\mathfrak{D}(\mathbf{H}, \hat{\mathbf{H}}) = -\log \sum_{i=0}^{b-1} \sqrt{\mathbf{H}[i] * \hat{\mathbf{H}}[i]}.$	With each image represented by a signature, calculating the similarity between two images transforms to the similarity computation between two signatures.
S. Kim et. al. 2015[74]	Min-Hashing + Jaccard Similarity	$P(survival) = (1 - (1 - p(\boldsymbol{c}_i, \boldsymbol{x}))^M)^S \triangleq (P_{MHF})^S,$	Efficiently partition images into buckets, where the images having the same sketch.
F. Nian et. al 2015 [9]	Bag of Word	$S = \frac{D_h}{3 * N},$	A distance of zero indicates that they are very similar or have little variation
H. Wang et. al	K-Mean Clustering + Bag of	Not mentioned	

2015 [1]	Word		
L. Carvalho et. al. 2015 [75]	K-Mean Clustering	$\otimes_{(d(r_i,s_i)\leq\xi),s_i\in} S \equiv \mathcal{F}_{(cord\leq\infty)}\left(\overline{\pi}_{\{s_i,s_j,\mathcal{F}(d(s_i,s_j))\to cord\}}\left(S \bowtie_{(d(s_i,s_j)\leq\xi)} S\right)\right)$	The distances between elements si,sj and returns the ordinal in (1), we employ F as an aggregate function that receives classification of the dissimilarity values.
Zhao et. al. 2016 [10]	Locality Linear Coding + MaxIDF-cut + K-Means clustering	$maxIDF-cut(s_i)=\{s_i^k \cdot mIDF \to s_i^K \cdot mIDF\}$ $if\left(s_i^j \cdot mIDF < i \cdot \sum_k^K s_i^k \cdot mIDF\right)$ $s_i^j \cdot mIDF=0$ ✚ $Sim_s(A_1,A_2)=\frac{\lvert A_1\cap A_2\rvert}{\lvert A_1\cup A_2\rvert}$ $Sim_h(I_i,I_j)=\frac{\sum_w \min\left(s_i^w,s_j^w\right)}{\sum_w \max\left(s_i^w,s_j^w\right)}$	Using maxIDF cut can achieve better performance for duplicate clustering and LLC may lead to an image representation that is not discriminative enough for near-duplicate images clustering
M. Alsmadi 2017 [76]	Canny Edge Detection+ Great Deluge Algorithm	$\text{Similarity Difference}=\sum_{j=1}^{n} f_j(I_1)-\sum_{j=1}^{n} f_j(I_2)\leqslant 0.005$	Comparison chromosome structure
R. Hassanian & M.K. Javad 2018 [11]	Min-Hash Locality Sensitive Hashing	$\Pr[x_A^\pi = x_B^\pi]=\Pr[x_{\tilde{A}}^\pi = x_{\tilde{B}}^\pi]$	constructs signatures that are more compact by generating single bits using the parity of the minHash values that result in the reduction of required storage and the search time.

Based on finding which state in Table 1. We can see the large problem in NDID discipline with the aid of preceding researchers are focusing on how to differentiate between an incorrect and right images that is close to the query image both in local or global shape in the images [67], [68]. They proposed many methods such as impartial component analysis (ICA) and scale invariant facets transform (SIFT), this is allowed how to extract seed structure from the photographs or in a different word is the matrix that is filling in the images structure contain.
Besides that, another researcher is focusing on the distances structures (Bag of Word; K-mean clustering) that are contained in the images. From this lookup they are attempting to calculate and evaluate the distances every image and if the distances in images are close to and similar to the different images in database they classify as near-duplicate. However, this algorithm will be going through a huge complex image structure because, if the images are now not similar but have a small distance they still expect that image are similar.

A few researchers involved how to speed-up the time similarity detection in NDID research [11]. Speed-up the time in detection similarity will make consumer fulfil because they will locate the images barring need ready for the NDID process the images. But they nevertheless want to focus on features extraction in order to make sure the image has higher similarity detection.

4. DISCUSSION

4.1 Objectives of Research

To reach the first research question RQs1 and accomplish a complete perspective of the theme, it is basic to classify the objective of the researches. The objective of the considered papers demonstrates the presence of ten different focus objective, RO 1 to RO 10, in which each paper can be set in at least one classifications, as portrayed in Table 2. In light of Table 2, the biggest measure of specialists' consideration has been centred around RO 1 and there is a huge contrast as far as consideration paid in the writing

RO 3 and RO 2. This distinction might be claiming many researchers agree clustering in ND images are important but the most important things in NDID is how we can represent and extract the images features structure, because the system needs to acknowledge the unique ND image features structure before proceeding with clustering part in order to place the ND images following their unique categories. The research objective from the previous researcher is described as below

Table 2: Entire research objective related in NDID area investigate by researcher

Reference	Techniques	Local /Global	O 1	O 2	O 3	O 4	O 5	O 6	O 7	O 8	O 9	O 10
Zheng et. al. 2012 [63]	Rotation, Scale, Translation (RST) Invariance features + Salient Covariance Matrix (SCOV) + ICA Independent component analysis	Global	√	√		√	√					√
Bueno et. al. 2012 [68]	Scale In variance Features Transform (SIFT)	Local	√		√							
Dong et. al. 2012 [4]	Scale In variance Features Transform (SIFT)	Local	√		√							
H. Wang et. al. 2012 [81]	Scale In variance Features Transform (SIFT) + Histograms of oriented Gradient (HoG) + BoF + K-Mean Clustering	Local	√	√	√				√			

Z. Li & Feng 2013 [67]	Locality Sensitive Hashing + K-Nearest Neighbor SIFT + K means + BoVW	Local	√	√	√		√				√
Kalaiarasi & Thyagharajan 2013[69]	Colour Texture Moment (CTM)	Local	√		√		√			√	
J. Wang 2013 [70]	Strong Geometry Consistency (SGC) + Scale Weighting	Local	√		√						
L. Li et. al. 2013 [71]	Bag of Visual Word (BoV)	Local	√		√			√	√		
J. Li et. al. 2014 [66]	Color Moment + Wavelet Transform + SIFT	Local & Global	√		√	√					√
Battiato et. al. 2014 [72]	Bag of Visual Word (BoVW)	Local	√		√					√	
L. Liu et. al. 2015 [73]	K-Nearest Neighbor	Local	√		√						
S. Kim et. al. 2015[74]	Min-Hashing + Jaccard Similarity	Global	√			√			√		
F. Nian et. al 2015 [9]	Bag of Word	Local	√		√						
H. Wang et. al 2015 [1]	K-Mean Clustering + Bag of Word	Global	√	√		√					
L. Carvalho et. al. 2015 [75]	K-Mean Clustering	Local		√	√				√		
Zhao et. al. 2016 [10]	Locality Linear Coding + MaxIDF-cut + K-Means clustering	Local		√	√						
M. Alsmadi 2017 [76]	Canny Edge Detection+ Great Deluge Algorithm	Local	√		√						
R. Hassanian & M.K. Javad 2018 [11]	Min-Hash Locality Sensitive Hashing	Local	√	√	√				√		

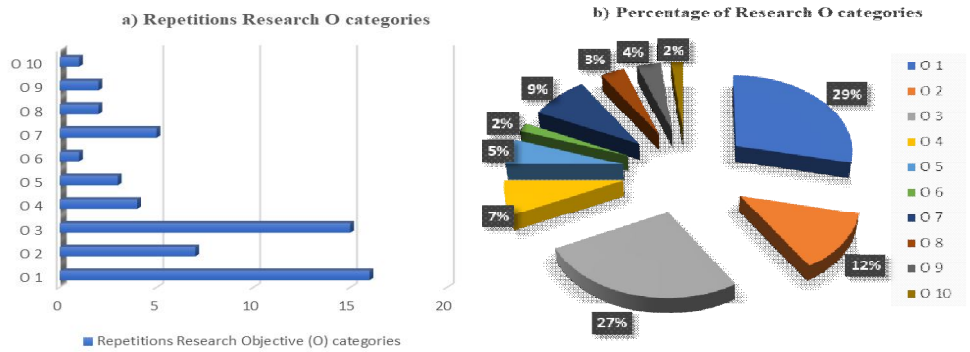

Figure 4: Number of two column a) Repetitions RO categories b) Percentage of RO categories

Finding the ND images for given input images, where a query image is required and Enhance productivity of NDID while maintaining precision: how to discover and deal with the ND image consequently from the web-scale social images is extremely difficult [66].

1. To detect and cluster ND images into their specific category in Large-scale image dataset [10].
2. To encode an image as a binary vector, which is called Local-based Binary Representation (LBR).

Table 3: Utilizes tools; Dataset; Dataset sources and Image dataset parameter considered

References	Tools	Dataset	Dataset source	Image Dataset Parameter Considered
L. Zheng et. al. (2012) [63]	Not Mentioned	General Images	collected from 15 viewers on 1003 images, including 779 landscape images and 228 portrait images. Flickr	resizing plus JPEG compression cropping image surface strong transformations including print & scan, contrast change, blur, etc.
Bueno et. Al 2012 [68]	Not Mentioned	General images	containing 110,000 unrelated images obtained from the Web (Random).	Strong editing by cropping, rotation, scaling, shearing, gamma correction, and dithering.
H. Wang et. al. 2012 [81]	Java 1.7 Hadoop 1.0.1	Hollywood action & Human Action	Not mentioned sources (200 to 400 Images) + (12790 videos)	To detect interest points for both images such as the Harris-Laplace detector
Z. Li & X. Feng 2013 [67]	Not Mentioned	includes 157 original	INRIA CopyDays dataset (229 transformed images for	Image resizing followed by JPEG compression ranging from JPEG3 to JPEG75,

		images which containing a variety of scene. Trademark & Landmark	the strong transformation) Flickr 1000 images collected Google Images 1,000 images	Cropping ranging from 5% to 80% of the image surface, Strong transformations: print and scan, paint, change in contrast, perspective effect, blur, very strong crop and so on.
J. Li et. al 2014 [66]	C/C++ Server with 2.0GHz CPU and 24GB memory	General Images & Landmark and landscape	COREL5k 5000 Images The Oxford Buildings Dataset 5000 Images Flickr 5,200,000 images taken from 1,447 different places all over the world. WIKI.com	Modified and copied by other users before being shared in social communities
F. Nian et. al. (2015)[9]	Not mentioned	General images	Google 1260 images Several TV programs Video 24 hours data.	Focusing in ND images dataset and defined with those variations: resizing, cropping, changing luma and chroma, adding text or watermark, changing layout slightly
H. Wang et. al 2015 [81]	Java virtual machine C/C++ GPU – Cuda C environment. HDFS	13,137 Videos	Google Video Yahoo Video YouTube	Hamming distance image dataset
W. Zhao et. al. 2016 [10]	Hadoop 2.4.0 for images clustering	100M General Images Dataset	Yahoo! Webscope Yahoo Flickr Creative Commons 100M (YFCC-100M) dataset	Discriminative representations of similarity images.
M. Alsmadi 2017 [76]	Not Mentioned	Chromosomes	Coral Dataset consists of 10,908 different images with the size of 256 * 384 or 384 * 256 for each image.	crossover, mutation (genetic operators) and great deluge algorithm local search in order to generate new chromosome.
R. Hassanian & M.K. Javad 2018 [11]	Not Mentioned	General data and Articles	GoldSet corpus Stanford University 2,160 manually labelled news articles in 68 directories	to detect clustering of document

3. Local regions are extracted densely from the image, and each region is converted to a simple and effective feature describing its texture [9].
4. To encode an image as a binary vector, which is considering as Global-based Binary Representation (GBR). Global regions are separated thickly from the images, and each region is changed over to a straightforward and successful part depicting on texture images [83].
5. Qualitative or Quantitative transformation method.
6. General or specific dataset for testing part.
7. To productively detect and speed up searching them is imperative to applications, for example, copyright infringement identification and discovering interchange variants of existing images [4].
8. To reduce redundancy of database spacing for ND images.
9. To protect the intellectual property rights of the content originators and creative artists.
10. To filter out the non-salient regions from an image, which also helps to eliminate some background noises in the images.
11. To improve user experience in search engine.
12. To increase the efficiency of tagging the document by reducing the need for manual inspection of the document.

Figure 4 shows two type of graph; a) is repetitions RO categories and b) is a percentage of RO categories. Based on previous RO have been achieve by researcher in NDID are more focus on RO 1; finding the ND images for a given input image, where a query image is needed and Improve efficiency of NDID while maintaining accuracy: how to find and manage the near-duplicate image automatically from the web-scale social images is very challenging. From RO 1 the system will recognize different features in a different image and similar features in the different image.

The second higher RO in NDID research is RO 3. Occasionally the different image has similar scenery but the different object or people in the images so that this the strong reasons why many researchers more focus in RO 3 compare to RO 4. Otherwise if image similarity detection more focus on global encodes binary will disturbing user activity by uploading images that have similar scenery but the different person in the images. Another objective that approaches in NDID is the main contribution for every researchers and novelty for their research.

Based on numerous objective in RQs 1 from previous researcher, research question number 2 (RQs 2) has been develop. Literature review result and discussion in section 2 and section 4.3 indicate all related technique and algorithm has been approaching to solve and justify image features extraction. Those technique facing the different problem and because of that multi-hybrid technique are approaches to solve the NDID features extraction. The different dataset has a different calculation on image features extraction, not all approaches technique can solve similarity on unique image dataset (e.g.: such as heritage,

scientific, architecture and so on) from this situation choosing the right technique will help increase the similarity detection in NDID and this answering the RQs 3.

This response for RQs 4 in evaluation for comparing a significant result. The calculation technique for a better result in order to get precision and recall value has been applied by many previous researchers in NDID area. Many other researchers agree with the following three criteria to measure the performances of NDID. They are precision (PR), recall (RC) and F- Measure, which are expressed as follows [66][10], [74], [94], [95];

$$PR = AS/AC * 100\% \quad (1)$$

$$RC = AS/TC * 100\% \quad (2)$$

$$F - measure = 2\frac{* PR * RC}{PR+RC} \quad (3)$$

where AC is the number of detected NDID, AS is the number of correctly detected NDID, and TC is the number of NDID in ground truth [66].

To address RQs 5 and RQs 6 by developing Theoretical Framework (TF) are shown in Figure 5. This TF are helping to summarize all the related technique has been applying in NDID research area and helping another researcher to recognize suitable technique will be applied in different group images and dataset. Various studies and techniques involving in ND image detection have been carried out as a bag of visual word technique and Min-Hash [84], this technique focuses on similarity image through space and position matching in the different image. While the Singular Value Decomposition-Scale Invariant Feature Transform (SVD-SIFT) technique [85] emphasizes features extraction using catalytic methods to accelerate the process of detecting the similarity of images. Next, the Visual Salient Riemannian technique [63] seeks to identify the prominent key space in the processed image. The technique of optimizing the use of databases by reducing the ND images same as the technique of data deduplication [86]–[88]. Otherwise, the data deduplication technique thereby helping to reduce energy consumption that can increase heat production. Next, the Similarity Join Operator technique [75][89] assessed the ND images based on the absolute ratio of the absolute equation. Fourier-Mellin Transform [90] techniques measure similarity through .

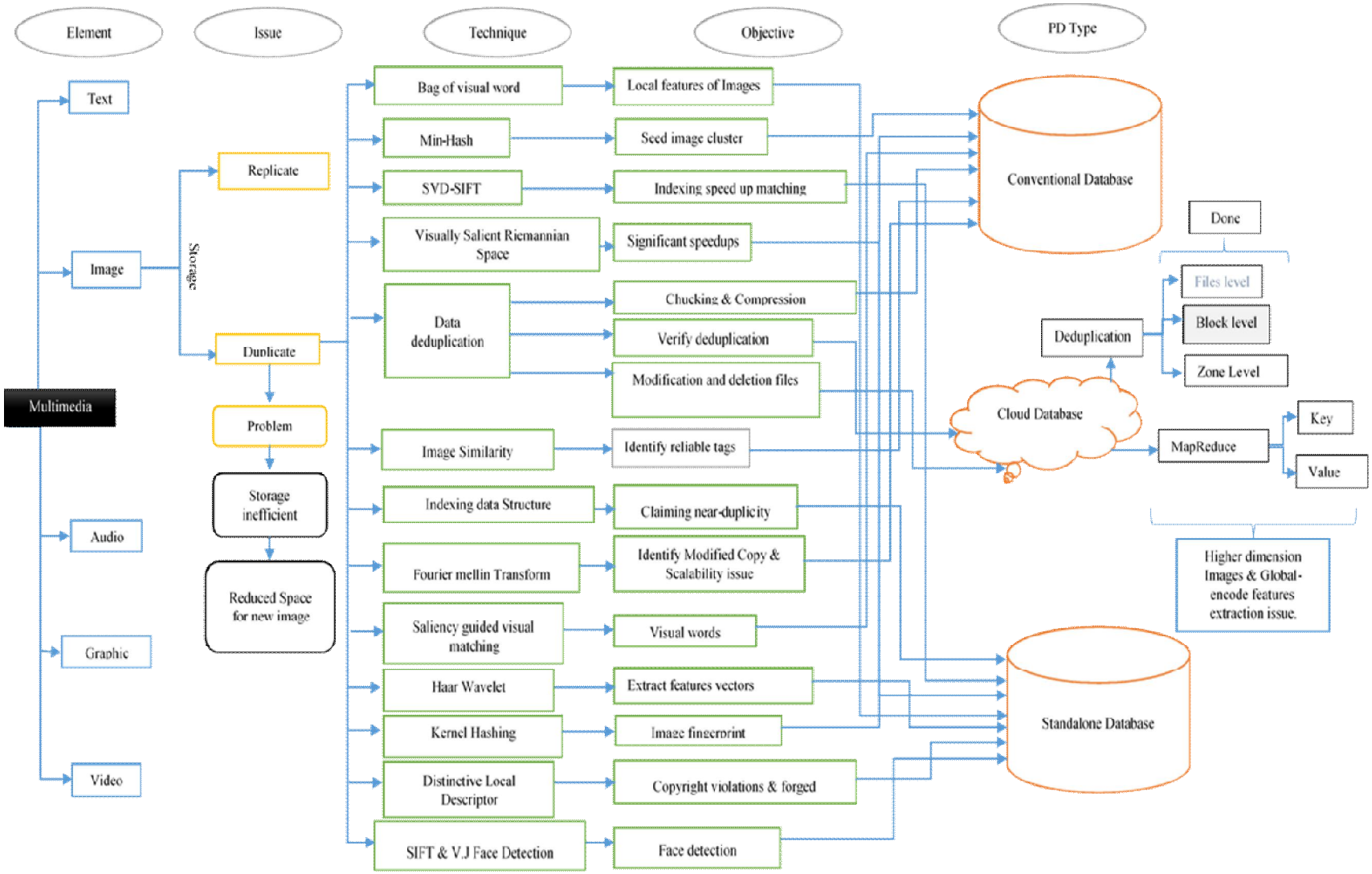

Figure 5: Theoretical Framework of existing Technique in NDID

rotation of images, image scale and the invariant change contained in the image. Thus, the Haar-wavelet technique [91][92] extracts the vector found in the image to determine the distance of the Manhattan object in the image to assess the accuracy images. According to [93] Kernel Hashing technique is allowed detects ND images through the diversity research of the features contained in the image in order to define the differences of each image and converts to the binary image and place into kernel space. However, most of the techniques described are more focused on image extraction for small-sized databases.

5. CONCLUSION

This investigation demonstrates that all NDID research can be group into 5 groups: a) Query of images, b) Features extraction, c) Similarity measurement, d) data/image clustering, e) output either indexing or retrieving ND image. Based on Table 2 and Table 3 research in NDID more focus on how to enable the system to understand what the user needs by calculating the features representation in the data or images according to multi-technique have been proposed in the literature review.

ACKNOWLEDGMENT

This research was supported by GUP-2017-077. We thank our associates who gave understanding and skill that enormously helped in the exploration.

REFERENCES

1. H. Wang, F. Zhu, B. Xiao, L. Wang, and Y. Jiang, "***GPU-based MapReduce for large-scale near-duplicate video retrieval***," *Multimed. Tools Appl. J.*, vol. 74, no. 23, pp. 10515–10534, 2015 https://doi.org/10.1007/s11042-014-2185-x
2. J. Zhang *et al.*, ***"IM-Dedup: An image management system based on deduplication applied in DWSNs,"*** *Int. J. Distrib. Sens. Networks*, vol. 2013, 2013
3. N. Soferman, ***"How-to automatically identify similar images using pHash Image de-duplication,"*** Cloudinary Blog, 2015. [Online]. Available: https://cloudinary.com/blog/how_to_automatically_identify_similar_images_using_phash.
4. W. Dong et al., ***"High-Confidence Near-Duplicate Image Detection," in Proceeding ICMR*** '12 Proceedings of the 2nd ACM International Conference on Multimedia Retrieval, 2012, pp. 3304--3311.
5. Y. Maret, F. Dufaux, and T. Ebrahimi***, "Image replica detection based on support vector classifier,"*** Proc. SPIE - Int. Soc. Opt. Eng., vol. 5909, pp. 1–9, 2005.
6. X. Dai and S. Khorram***, "A feature-based image registration algorithm using improved chain-code representation combined with invariant moments,"*** IEEE Trans. Geosci. Remote Sens., vol. 37, no. 5 II, pp. 2351–2362, 1999. https://doi.org/10.1109/36.789634
7. J. J. F. J. J. Foo, R. Sinha, and J. Zobel, *"SICO: A System for Detection of Near-Duplicate Images During Search,"* Multimed. Expo, 2007 IEEE Int. Conf., pp. 595–598, 2007.
8. H. Kim, H. Chang, J. Lee, and D. Lee, "Basil: Effective near-duplicate image detection using gene sequence alignment," Adv. Inf. Retr., no. 1, 2010.
9. [9]F. Nian, T. Li, X. Wu, Q. Gao, and F. Li, "Efficient near-duplicate image detection with a local-based binary representation," Multimed. Tools Appl., vol. 75, no. 5, pp. 2435–2452, 2015.
10. [10] W. Zhao, H. Luo, J. Peng, and J. Fan, "MapReduce-based clustering for near-duplicate image identification," Multimed. Tools Appl., vol. 76, no. 22, pp. 23291–23307, 2016.
11. [11] R. Hassanian-esfahani and M. javad Kargar, "Sectional MinHash for near-duplicate detection," Expert Syst. Appl., vol. 99, no. 1 June 2018, pp. 203–212, 2018. https://doi.org/10.1016/j.eswa.2018.01.014
12. [12] Y. Rui, T. S.Huang, and S.-F. Chang, "Image Retrieval : Current Techniques , Promising Directions , and Open Issues," J. Vis. Commun. Image Represent., vol. 10, no. 1, pp. 39–62, 1999.
13. [13] S. Torres and A. X. Falcão, "Content-Based Image Retrieval: Theory and Applications.," Rev. Informática Teórica e Apl., vol. 13, no. 2, pp. 161–185, 2006.
14. [14] P. Wilkins, P. Ferguson, A. F. Smeaton, and C. Gurrin, "Text based approaches for content-based image retrieval on large image collections," in 2nd European Workshop on the Integration of Knowledge, Semantics and Digital Media Technology (EWIMT 2005), 2005, pp. 281–288.
15. A. K. JAIN and A. VAILAYA, ***"Shape-Based Retrieval: a Case Study With Trademark Image Databases,"*** Pattern Recognit., vol. 31, no. 9, pp. 1369–1390, 1998.
16. J. S. Dhanoa and A. Garg***, "A Novel Technique for Shape Feature Extraction Using Content Based Image Retrieval,"*** 4th Int. Conf. Adv. Eng. Technol., vol. 57, pp. 1–6, 2016.
17. D. Feng, W. C. Siu, and H. Zhang, Multimedia ***Information Retrieval and Management,*** 1st (2003). New York: Springer, 2003.
18. S. Panchanathan, Y. C. Park, K. S. Kim, P. K. Kim, and F. Golshani, ***"The Role of Color in Content-Based Image Retrieval,"*** in Proceedings 2000 International Conference on Image Processing, 2000, pp. 517–520.
19. T. S. Kumar, V. V. Kumar, and B. E. Reddy***, "Image retrieval based on hybrid features,"*** ARPN J. Eng. Appl. Sci., vol. 12, no. 2, pp. 591–598, 2017. https://doi.org/10.1007/s11760-017-1197-1
20. J. Zhang, W. Hsu, and M. L. Lee***, "An Information-driven Framework for Image Mining,"*** in International Conference on Database and Expert Systems Applications, 2001, pp. 232–242.
21. F. Alamdar and M. Keyvanpour, ***"A New Color Feature Extraction Method Based on QuadHistogram,"*** in 2011 3rd International Conference on Environmental Science and Information Application Technology (ESIAT 2011) ESIAT-2011, 2011, vol. 10, no. Part A, pp. 777–783.
22. K. Jenni, S. Mandala, and M. S. Sunar, ***"Content based image retrieval using colour strings comparison,"*** in

Procedia Computer Science, 2015, vol. 50.
23. J. Kalpana and R. Krishnamoorthy, ***"Color image retrieval technique with local features based on orthogonal polynomials model and SIFT,"*** Multimed. Tools Appl., vol. 75, no. 1, pp. 49–69, 2016.
24. M. J. Swain and D. H. Ballard, "Color Indexing***," Int. J. Comput. Vis.***, vol. 7, no. 1, pp. 11–32, 1991.
25. L. V. Tran***, "Efficient Image Retrieval with Statistical Color Descriptors,"*** Linkoping University, Sweden, 2003.
26. R. Datta, D. Joshi, J. I. A. Li, and J. Z. Wang, ***"Image Retrieval : Ideas , Influences , and Trends of the New Age,"*** J. ACM Comput. Surv., vol. 40, no. 2, pp. 1–60, 2008.
27. J. Sun, X. Zhang, J. Cui, and L. Zhou***, "Image retrieval based on color distribution entropy,"*** Pattern Recognit. Lett. 27, vol. 27, no. 10, pp. 1122–1126, 2006.
28. J. Yue, Z. Li, L. Liu, and Z. Fu***, "Content-based image retrieval using color and texture fused features,"*** Math. Comput. Model., vol. 54, no. 3–4, pp. 1121–1127, 2011. https://doi.org/10.1016/j.mcm.2010.11.044
29. A. Toshev, B. Taskar, and K. Daniilidis***, "Shape-based object detection via boundary structure segmentation,"*** Int. J. Comput. Vis., vol. 99, no. 2, pp. 123–146, 2012.
30. M. Verma and B. Raman***, "Local tri-directional patterns: A new texture feature descriptor for image retrieval,"*** Digit. Signal Process. A Rev. J., vol. 51, 2016.
31. E. Sokic and S. Konjicija, ***"Phase preserving Fourier descriptor for shape-based image retrieval,"*** Signal Process. Image Commun., vol. 40, pp. 82–96, 2016.
32. V. I. Patil and S. Kotyal, ***"Survey on Content Based Image Retrieval Using Color and Texture Features,"*** Int. J. Adv. Electron. Comput. Sci. ISSN 2393-2835, vol. 2, no. 10, pp. 1424–1429, 2015.
33. J. M. Patel and N. C. Gamit***, "A review on feature extraction techniques in Content Based Image Retrieval,"*** 2016 Int. Conf. Wirel. Commun. Signal Process. Netw., pp. 2259–2263, 2016.
34. P. Sanguansat***, Principal Component Analysis - Multidisciplinary Applications.*** Croatia: InTech, 2012.
35. I. T. Jolliffe, ***Principal Component Analysis, Second Edition***, vol. 98, no. 3. 2002.
36. H. Zhu, Z. Shen, L. Shang, and X. Zhang***, "Parallel Image Texture Feature Extraction under Hadoop Cloud Platform,"*** Springer Int. Publ., pp. 459–465, 2014.
37. J. M. Francos, A. Narasimhan, and J. W. Woods, ***"Maximum Likelihood Parameter Estimation of Textures Using a Wold-Decomposition Based Model,"*** IEEE Trans. Image Process., vol. 4, no. 12, pp. 1655–1666, 1995. https://doi.org/10.1109/83.475515
38. C. Wang, N. Komodakis, and N. Paragios, ***"Markov Random Field Modeling, Inference & Learning in Computer Vision & Image Understanding : A Survey,"*** Comput. Vis. IMAGE Underst., vol. 117, no. 11, pp. 1610–1627, 2013.
39. H. Jiang, T. Feng, D. Zhao, B. Yang, L. Zhang, and Y. Chen, ***"Statistical Fractal Models Based on GND-PCA and Its Application on Classification of Liver Diseases,"*** Biomed Res. Int., vol. 2013, pp. 1–8, 2013.
40. A. Verma***, "Content Based Image Retrieval Using Color , Texture and Shape Features,"*** vol. 4, no. 5, pp. 383–389, 2014.
41. S. Deb***, Multimedia systems and content-based image retrieval.*** 2004.
42. G. Lu and A. Sajjanhar, ***"Region-based shape representation and similarity measure suitable for content-based image retrieval,"*** Multimed. Syst., vol. 7, no. 2, pp. 165–174, 1999.
43. H. Shengjie, ***"Region-based Partial-Duplicate Image Retrieval,"*** 2012 Int. Conf. Ind. Control Electron. Eng., pp. 1521–1524, 2012.
44. M. H. Memon, J. P. Li, I. Memon, and Q. A. Arain, ***"GEO matching regions: multiple regions of interests using content based image retrieval based on relative locations,"*** Multimed. Tools Appl., vol. 76, no. 14, pp. 15377–15411, 2017.
45. H. V. Jagadish, ***"A retrieval technique for similar shapes,"*** in Proceedings of the 1991 ACM SIGMOD international conference on Management of data - SIGMOD '91, 1991, pp. 208–217.
46. E. M. Arkin, L. P. Chew, D. P. Huttenlocher, and K. Kedem, ***"An Efficiently Computable Metric for Comparing Polygonal Shapes,"*** IEEE Trans. Pattern Anal. Mach. Intell., vol. 13, no. 3, pp. 209–216, 1991. https://doi.org/10.1109/34.75509
47. S. Sclaroff and A. P. Pentland***, "Modal Matching for Correspondence and Recognition,"*** IEEE Trans. Pattern Anal. Mach. Intell., vol. 17, no. 6, pp. 545–561, 1995.
48. K. Arbter, W. E. Snyder, H. Burkhardt, and G. Hirzinger, ***"Application of Affine-Invariant Fourier Descriptors to Recognition of 3-D Objects,"*** IEEE Trans. Pattern Anal. Mach. Intell., vol. 12, no. 7, pp. 640–647, 1990.
49. H. Kauppinen, T. Seppanen, and M. Pietikainen, ***"An experimental comparison of autoregressive and Fourier-based descriptors in 2D shape classification,"*** IEEE Trans. Pattern Anal. Mach. Intell., vol. 17, no. 2, pp. 201–207, 1995.
50. E. Persoon and K.-S. Fu, ***"Shape Discrimination Using Fourier Descriptors,"*** IEEE Trans. Syst. Man. Cybern., vol. PAMI-8, no. 3, pp. 388–397, 1986.
51. D. M. Uliyan, H. A. Jalab, A. W. A. Wahab, and S. Sadeghi, ***"Image region duplication forgery detection based on angular radial partitioning and harris key-points,"*** Symmetry (Basel)., vol. 8, no. 7, 2016.
52. M. K. Hu, ***"Visual Pattern Recognition by Moment,"*** IRE Trans. Inf. Theory, vol. 8, no. 2, pp. 66–70, 1962.
53. L. Yang and F. Albregtsen, ***"Fast Computation of Invariant Geometric Moments:,"*** Pattern Recognit., vol. 29, no. 7, pp. 201–204, 1994.
54. T. Bui, L. Ribeiro, M. Ponti, and J. Collomosse, ***"Compact descriptors for sketch-based image retrieval using a triplet loss convolutional neural network,"*** Comput. Vis. Image Underst., 2017.
55. T. Portenier, Q. Hu, P. Favaro, and M. Zwicker, ***"SmartSketcher,"*** in Proceedings of the Symposium on Sketch-Based Interfaces and Modeling - SBIM '17, 2017, pp. 1–12.
56. N. Yusof, T. S. M. T. Wook, and S. F. M. Noor, ***"Songket Motives Retrieval through Sketching Technique,"***

Procedia Technol., vol. 11, no. Iceei, pp. 263–271, 2013.
57. C. Wang, J. Zhang, B. Yang, and L. Zhang, ***"Sketch2Cartoon : Composing Cartoon Images by Sketching,"*** in Proceeding MM '11 Proceedings of the 19th ACM international conference on Multimedia, 2011, pp. 789–790. https://doi.org/10.1145/2072298.2072458
58. N. bt Yusof, T. S. M. T. Wook, and S. F. M. Noor, ***"Comparison Result of Songket Motives Retrieval through Sketching Technique with Keyword Technique,"*** Int. J. Comput. Sci. Netw., vol. 3, no. 2, pp. 70–76, 2014.
59. S. Noah and S. Sabtu, ***"Binding Semantic to a Sketch Based Query Specification Tool,"*** Int. Arab Inf. Technol., vol. 6, no. 2, pp. 116–123, 2006.
60. H. Wang and C. Wang, ***"MindFinder : Interactive Sketch-based Image Search,"*** in Proceeding MM '10 Proceedings of the 18th ACM international conference on Multimedia, 2010, pp. 1605–1608.
61. A. Jula, E. Sundararajan, and Z. Othman, ***"Cloud computing service composition: A systematic literature review,"*** Expert Syst. Appl., vol. 41, no. 8, pp. 3809–3824, 2014.
62. B. Kitchenham, O. Pearl Brereton, D. Budgen, M. Turner, J. Bailey, and S. Linkman, ***"Systematic literature reviews in software engineering - A systematic literature review,"*** Inf. Softw. Technol., vol. 51, no. 1, pp. 7–15, 2009.
63. L. Zheng, Y. Lei, G. Qiu, and J. Huang***, "Near-duplicate image detection in a visually salient riemannian space,"*** IEEE Trans. Inf. Forensics Secur., vol. 7, no. 5, pp. 1578–1593, 2012.
64. W. Youzhong, D. Zeng, Z. Xiaolong, and W. Feiyue, ***"Propagation of online news: Dynamic patterns,"*** in 2009 IEEE International Conference on Intelligence and Security Informatics, ISI 2009, 2009, pp. 257–259.
65. J. Huang, R. Zhang, R. Buyya, J. Chen, and Y. Wu, ***"Heads-Join: Efficient Earth Mover's Distance Similarity Joins on Hadoop,"*** IEEE Trans. Parallel Distrib. Syst., vol. 27, no. 6, pp. 1660–1673, 2016.
66. J. Li, X. Qian, Q. Li, Y. Zhao, L. Wang, and Y. Y. Tang, ***"Mining near duplicate image groups,"*** Multimed. Tools Appl., vol. 74, no. 2, pp. 655–669, 2014.
67. Z. Li and X. Feng, ***"Near duplicate image detecting algorithm based on bag of visual word model,"*** J. Multimed., vol. 8, no. 5, pp. 557–564, 2013.
68. L. Bueno, E. Valle, and R. da S. Torres, ***"Bayesian approach for near-duplicate image detection,"*** in Proceeding ICMR '12 Proceedings of the 2nd ACM International Conference on Multimedia Retrieval, 2012, pp. 1–8.
69. G. Kalaiarasi and K. K. Thyagharajan***, "Visual content based clustering of near duplicate web search images,"*** in Proceedings of the 2013 International Conference on Green Computing, Communication and Conservation of Energy, ICGCE 2013, 2013, pp. 767–771.
70. J. Wang, ***"Strong Geometry Consistency for Large Scale Partial-Duplicate Image Search,"*** in MM '13 Proceedings of the 21st ACM international conference on Multimedia, 2013, pp. 633–636.
71. L. Li, S. Jiang, Z. J. Zha, Z. Wu, and Q. Huang, ***"Partial-Duplicate Image Retrieval via Saliency-Guided Visual Matching,"*** IEEE Multimed., vol. 20, no. 3, pp. 13–23, 2013.
72. S. Battiato, G. M. Farinella, G. Puglisi, and D. Ravì, ***"Aligning codebooks for near duplicate image detection,"*** Multimed. Tools Appl., vol. 72, no. 2, pp. 1483–1506, 2014.
73. L. Liu, Y. Lu, and C. Y. Suen, ***"Variable-length signature for near-duplicate image matching,"*** IEEE Trans. Image Process., vol. 24, no. 4, pp. 1282–1296, 2015.
74. S. Kim, X. J. Wang, L. Zhang, and S. Choi***, "Near duplicate image discovery on one billion images,"*** in 2015 IEEE Winter Conference on Applications of Computer Vision, 2015, pp. 943–950.
75. L. O. Carvalho, L. F. D. Santos, W. D. Oliveira, A. J. M. Traina, and C. Traina***, "Self Similarity Wide-Joins for Near-Duplicate Image Detection,"*** in Proceedings - 2015 IEEE International Symposium on Multimedia, ISM 2015, 2015, pp. 237–240.
76. M. K. Alsmadi, ***"Query-sensitive similarity measure for content-based image retrieval using meta-heuristic algorithm,"*** J. King Saud Univ. - Comput. Inf. Sci., 2017. https://doi.org/10.1016/j.jksuci.2017.05.002
77. L. Wang, H. Wang, and B. Xiao, ***"A GPU-based MapReduce framework for MSR-Bing Image Retrieval Challenge,"*** Proc. 2015 10th Int. Conf. Commun. Netw. China, CHINACOM 2015, pp. 442–447, 2016.
78. B. Wang, Z. Li, M. Li, and W. Y. Ma, ***"Large-scale duplicate detection for web image search,"*** in 2006 IEEE International Conference on Multimedia and Expo, ICME 2006 - Proceedings, 2006, vol. 2006, pp. 353–356.
79. F. Zou et al***., "Nonnegative sparse coding induced hashing for image copy detection,"*** Neurocomputing, vol. 105, pp. 81–89, 2013.
80. O. Chum, J. Philbin, M. Isard, and A. Zisserman, ***"Scalable near identical image and shot detection,"*** in Proceedings of the 6th ACM international conference on Image and video retrieval - CIVR '07, 2007, pp. 549–556.
81. H. Wang, Y. Shen, L. Wang, K. Zhufeng, W. Wang, and C. Cheng***, "Large-Scale Multimedia Data Mining Using MapReduce Framework,"*** in 2012 IEEE 4th International Conference on Cloud Computing Technology and Science, 2012, pp. 287–292.
82. C. Böhm, S. Berchtold, and D. A. Keim***, "Searching in High-dimensional Spaces - Index Structures for Improving the Performance of Multimedia Databases,"*** J. ACM Comput. Surv., vol. 33, no. 3, pp. 322–373, 2001.
83. J. A. R. Serrano and D. Larlus***, "Predicting an Object Location Using a Global Image Representation,"*** 2013 IEEE Int. Conf. Comput. Vis., pp. 1729–1736, 2013.
84. O. Chum, J. Philbin, and A. Zisserman***, "Near Duplicate Image Detection : min-Hash and tf-idf Weighting,"*** in Proceedings of the British Machine Vision Conference, 2008, vol. 810, pp. 812–815.
85. H. Liu, H. Lu, and X. Xue***, "SVD-SIFT for web near-duplicate image detection,"*** Proc. - Int. Conf. Image Process. ICIP, pp. 1445–1448, 2010.

86. Q. He, Z. Li, and X. Zhang***, "Data Deduplication Techniques,"*** in 2010 International Conference on Future Information Technology and Management Engineering, 2010, pp. 430–433.
87. B. Zhou and J. Wen***, "A Data Deduplication Framework of Disk Images with Adaptive Block Skipping,"*** vol. 31, no. 61125102, pp. 820–835, 2016.
88. Z. Wen, J. Luo, H. Chen, J. Meng, X. Li, and J. Li***, "A verifiable data deduplication scheme in cloud computing,"*** Proc. - 2014 Int. Conf. Intell. Netw. Collab. Syst. IEEE INCoS 2014, pp. 85–88, 2014.
89. L. Chen and F. Stentiford***, "Comparison of near-duplicate image matching,"*** Cvmp06, 2006.
90. S. H. Srinivasan and N. Sawant, ***"Finding Near-duplicate Images on the Web using Fingerprints,"*** in Proceeding of the 16th ACM international conference on Multimedia MM 08, 2008, pp. 881–884. https://doi.org/10.1145/1459359.1459512
91. M. Chen, Y. Wang, X. Zou, S. Wang, and G. Wu, ***"A duplicate image deduplication approach via Haar wavelet technology,"*** Proc. - 2012 IEEE 2nd Int. Conf. Cloud Comput. Intell. Syst. IEEE CCIS 2012, vol. 2, pp. 624–628, 2013.
92. S. G. Lakshmi and N. R. Gayathiri, ***"a Framework for Hosting Image Compression in Cloud,"*** Int. J. Comput. Sci. Mob. Comput., vol. 3, no. 3, pp. 845–848, 2014.
93. F. Zou, Y. Chen, J. Song, K. Zhou, Y. Yang, and N. Sebe, ***"Multiple Kernel Hashing,"*** IEEE Trans. Multimed., vol. 17, no. 7, pp. 1006–1018, 2015.
94. F. Cai and H. Chen, ***"A MapReduce scheme for image feature extraction and its application to man-made object detection,"*** in Fifth International Conference on Digital Image Processing, 2013, vol. 8878, pp. 1–7.
95. A. Bala and T. Kaur, ***"Local texton XOR patterns: A new feature descriptor for content-based image retrieval,"*** Eng. Sci. Technol. an Int. J., vol. 19, no. 1, 2016. https://doi.org/10.1016/j.jestch.2015.06.008