The Imaging Science Journal

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/yims20

Seagull optimization-based near-duplicate image detection in large image databases

Srinidhi Sundaram, S. Kamalakkannan & Sasikala Jayaraman

To cite this article: Srinidhi Sundaram, S. Kamalakkannan & Sasikala Jayaraman (2023): Seagull optimization-based near-duplicate image detection in large image databases, The Imaging Science Journal, DOI: 10.1080/13682199.2023.2190944

To link to this article: https://doi.org/10.1080/13682199.2023.2190944

Published online: 29 Mar 2023.

THE IMAGING SCIENCE JOURNAL
https://doi.org/10.1080/13682199.2023.2190944

THE RPS ROYAL PHOTOGRAPHIC SOCIETY
Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

Seagull optimization-based near-duplicate image detection in large image databases

Srinidhi Sundaram[a], S. Kamalakkannan[b] and Sasikala Jayaraman[a]

[a]Department of Information Technology, Annamalai University, Tamil Nadu, India; [b]Department of Information Technology, School of Computing Sciences, Vels Institute of Science, Technology & Advanced Studies (VISTAS), Chennai, India

ABSTRACT

Many of the near-duplicate (ND) image detection methods involve a greater number of interest points (IPs) and large dimensions of the feature descriptors requiring huge computations and are unsuitable for large image databases. They may fail to detect NDs if the query and images in the database contain sparse IPs due to low entropy. Besides, the k-means algorithm used for the quantization of visual words may land at a sub-optimal minimum for descriptors because of their distance distribution in feature space. This article presents a new ND image detection method, which uniformly distributes the IPs over low and high entropy regions, reduces the dimension of feature descriptors using discrete wavelet transform (DWT) and employs Seagull Optimization algorithm (SOA) for optimally forming the visual words. It examines proposed method performs on image databases of various sizes and shows that the developed method is more reliable and computationally efficient than the alternatives.

ARTICLE HISTORY
Received 29 November 2022
Accepted 9 March 2023

KEYWORDS
Near-duplicate images; seagull optimization; K-means clustering; scale invariant features transform; bag of visual words; discrete wavelet transform; visual words; computationally efficient

Introduction

Near-Duplicate (ND) image detection is an important problem with several applications such as the detection of copyright infringement and the reduction of storage space. A duplicate is a digital image that is visually identical and differs only on the scale, colour schemes or storage format of the original image, while an ND is further varied by contrast, luminance, rotation, translation, a slight change of the background of the original document but is nearly identical from the perception of users. Additional alterations like the insertion of a caption and/or a logo may further alter their appearance. Duplicate detection is a binary decision problem resulting in an answer of 'yes' or 'no' to confirm the ND of a given pair of images. The majority of the images are used unlawfully violating the copyrights. Any image uploaded to a website violates copyrights unless it is a personal photo or a transformed image created by the owner of the personal photo. Automatic online detection of duplicates enables the finding of copyright infringements and allows the rights holders to enforce their rights. The existing ND detection methods aim at detecting the NDs of a given query image among a collection of digital images and differ in making significant innovations in various modules of ND detection systems such as feature descriptors, dimensionality reduction, indexing, quantization and geometric consistency verification of matched features [1].

Several methods for ND detection have been suggested in the literature [2]. These methods are generally classified into watermarking-based strategies and content-based approaches. The former strategies embed a digital signature within the original image before distribution for ascertaining the ownership and subsequently perform checking the presence of embedded signature in website images for NDs. The later approaches analyse the contents of digital images by extracting relevant visual features. They compare the query image features with those of the digital images in the website and identify the images, whose features are closer to those of the query image, as ND images. Many of the content-based duplicate detection algorithms identify interest points (IPs), extract local feature descriptors by using algorithms like Scale Invariant Features Transform (SIFT), Speeded Up Robust Features (SURF), etc., and perform indexing or quantization of the evaluated feature database. The number of IPs and the length of the descriptors are so large and become a limiting factor for application to large image databases (IDBs). Though these approaches are robust, they may not detect NDs if the query and images in the database may contain sparse IPs due to low entropy regions indicating smooth regions like plains, sea and so on. Besides, the k-means algorithm used for the quantization of features may land at a sub-optimal trap for descriptors given their distance distribution in feature space. There is thus a need for dimensionality

CONTACT Srinidhi Sundaram srinidhisristar@gmail.com Department of Information Technology, Annamalai University, Tamil Nadu, India

reduction of the features before hashing or quantization and avoiding local minima during quantization into visual words.

Recently, a Seagull Optimization algorithm (SOA), inspired from the natural seagulls' migratory and assault patterns, has been suggested for handling optimization problems [3] and portrayed to be superior to other evolutionary algorithms like genetic algorithm, bacterial foraging, biogeography based optimization and so on. In SOA, seagulls representing probable solutions migrate from one location to another in search of foods like fish, earthworms, insects, amphibians, reptiles and other small animals. During migration, seagulls frequently attack other birds in a spiral natural shape at sea. The SOA was applied to solving a variety of real-world optimization problems such as classification [4] and detection of NLOS nodes in VANET [5].

This paper proposes a new ND image detection method for improving the computational efficiency and robustness through blending the IPs obtained by two different techniques for uniformly distributing the IPs over low and high entropy regions, reducing the dimension of feature descriptors using discrete wavelet transform (DWT) and employing SOA for optimally forming the visual vocabulary. The proposed method has been studied on large IDBs and their performances have been discussed in the paper.

There are five sections in this paper. The associated work is surveyed in Section 2, the ND detection methodologies are outlined in Section 3, the developed ND detection scheme is described in Section 4, the results are shown in Section 5 and the article is concluded in Section 6.

Related work

Several methods were suggested for detecting NDs in large IDBs in recent decades. This section reviews a few of these techniques.

Ke et al. performed ND detection using sparse descriptors, obtained from each image and related with a locality sensitivity hashing (LSH) for quick search on the individual features [6]. The methodology used SIFT for evaluating features and employed PCA for dimensionality reduction. The computational overhead was found to be prohibitive especially for larger IDBs. Wang et al. presented an ND detection algorithm, which first calculates the k-bit hash code for each image and then performs the ND detection using the hash codes [7]. Chen and Stentiford developed an ND detection method involving colour and texture-based signatures. The approach did a heuristic matching based on unrestricted competition [8]. Foo et al. outlined a clustering method for ND image detection employing invariant image local descriptors and adopting ND text-document clustering techniques [9]. Zhao et al. employed SIFT, PCA and DoG for image representation and adopted LIP-IS for better nearest neighbour search (NNS) for developing an ND identification system. The scheme also eliminated false matches and located the exact nearest neighbour [10]. Chum et al. developed two schemes for ND image and video-shot detection. The former one used hierarchical colour histograms and LSH for quick detection, while the latter one used SIFT descriptors and min-hash algorithm. Both methods required less data storage and yielded good results [11].

Xu et al. suggested a two-stage scheme for the detection of NDs and retrieval. The distances among two rectangle blocks of two images were evaluated using SIFT descriptors in the first stage, while in the next stage, several hypotheses were performed for detecting scale variations. The method may fail if the image underwent image rotations [12]. Wang et al. proposed an ND detection approach combining both local and global features, and using an efficient hashing technique and map-reduce framework [13]. Hsieh et al. outlined an ND image detection technique adopting hash tables for fast image matching and ND detection. It initially extracted the image features and stored them in the slots of multiple hash tables. It then hashed the descriptors and assessed whether the query image was an ND one with a low computational burden [14]. Yao et al. presented an ND image detection method adopting a contextual descriptor for measuring the similarity and discarding the mismatches and reducing the number of images [15].

Liu et al. outlined an adjustable-length signature for ND detection, wherein an image was denoted by a signature, whose length depends on the number of patches. The method utilized the earth mover's distance to handle adjustable-length signatures [16]. Pawar and Mankar proposed a method that extracted patches of variable length signatures through clustering adjacent and visually similar pixels of the images for ND detection. Probabilistic binary pattern and distance were employed to calculate the similarity between the two images [2]. Deshmukh and Lambhate developed a MapReduce-based ND image identification technique for improving the efficiency and reliability of the search. MapReduce was simple and parallel computing techniques normally used for analysing huge data with minimum storage space [17]. Layek et al. developed a hybrid methodology combining global moment-invariant features and local feature vectors for identifying and grouping NDs from images posted on social media [18]. Landge and Mane surveyed the literature about ND image-matching techniques and discussed several schemes for representing images, extracting features and evaluating the similarity between two images [19]. Zhang et al. presented a Bitwise LSH method employing a

bit per hash, thereby significantly reducing the memory for storing hash values, and performing the ND detection of images, videos and web documents much faster [20]. Jayshree and Bhale suggested an ND detection method involving Earth Move's Distance Algorithm. The method evaluated patches that represented changeable length signatures by clustering the adjacent and visually same pixels of the image [21]. Thaiyalnayaki et al. proposed a method for indexing ND Images in the Web Search. The method employed SURF for extracting the features and hashing for sorting ND images based on the query image [1]. Fellah proposed a generalized domain-independent clustering for detecting ND records. The method worked recursively for arranging NDs as hierarchical clusters with a view of reducing the search space [22]. Albayrak et al. developed a duplicate record detection scheme for e-commerce applications using a real-world dataset. The method generated potential duplicate product pairs for training using text similarity and domain-specific distance metrics. It was shown that the scheme could detect duplicates with good accuracy [23]. Chevallier et al. introduced the idea of an ND dataset involving data exploration, data integration and data quality. It employed a methodology for artificially creating the training data [24]. Gusev and Xu presented a method for detecting ND images from a large image database involving candidate production, selection and clustering. It employed visual embedding to lower the computational burden [25]. Wang et al. studied the ND text alignment search problem and suggested leveraging the bottom-k-sketch. It identifies groups of sections with comparable sketches as NDs [26]. Mehta and Tripathi suggested an ND detection technique by analysing the edge profile obtained by the edge histogram descriptor and SVM classifier [27]. Outlined an ND detection method involving a sectional MinHash algorithm, which predicts the similarity between two documents. The method attempted to lower the hashing time while retaining the detection accuracy [28].

Proposed SOA-based ND detection method

The conventional scheme of ND detection of a query image is given in Figure 1. It initially pre-processes all the images in a large IDB by removing their noises and converting them into grey scale. It then identifies the IPs either by using the Difference of Gaussian (DoG) or fast Hessian matrix (FHM) or FAST corner points, and evaluates local descriptors employing SIFT or SURF. It forms a visual vocabulary as depicted in Figure 2 using K-means clustering. It then searches the features of a query image in the clustered visual vocabulary by the concept of inverted file indexing as explained in Figure 3. The geometrical consistence of matched points is verified by applying RANSAC [29] and the false matches are eliminated. It then ranks the images based on the number of matched feature points. A few of the top-ranked images are considered NDs.

Instead of using a single technique for identifying the IPs, the proposed methodology uses DoG and minimum eigenvalue (MEV) based approaches for finding the IPs at both low and high entropy regions [30]. Then 50% of the strongest IPs of each approach are used for computing the local descriptors using SIFT [31]. This process is repeated for all images in the entire IDB. The dimension of SIFT descriptor is 128, which is large enough to make the ND detection process highly inefficient. DWT, a powerful technique using dyadic positions and scales, is applied for reducing the dimension of each descriptor. The 128-dimensional SIFT descriptors at each IP are rearranged into a (16×8) sized matrix, which is then passed through a DWT as shown in Figure 4 to form 4 sub-bands (LL, LH, HH and HL). The approximation component (LL) with a size of (8×4) is rearranged into a (1×32) sized vector and considered as the reduced descriptors, while discarding the detailed components of LH, HL and HH [32].

The reduced descriptors, organized in the form of a matrix [$\$^R$] of size $(N \times L)$, are clustered into K-number of visual words $[vw_1, vw_2, \ldots, vw_K]$ by using the K-means clustering technique [33], which lowers the sum of squares of Euclidean distance (ED) between the visual words and the feature descriptors.

$$\text{Minimize } \Omega = \sum_{m=1}^{K} \sum_{n=1}^{N} \|\$_n^R - vw_m^2\| \tag{1}$$

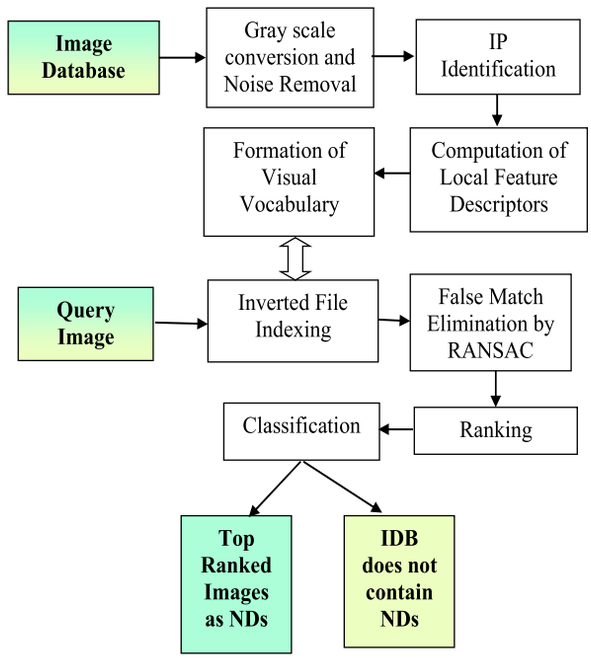
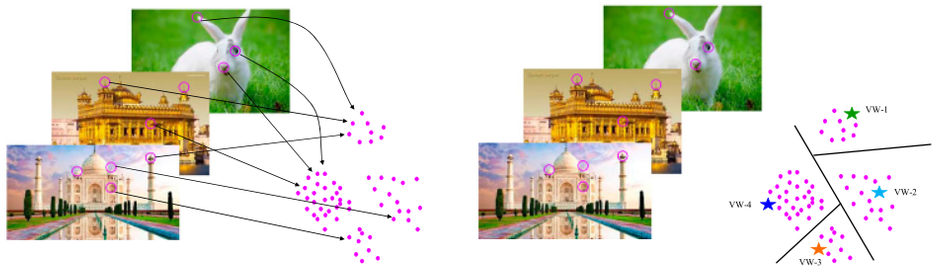


Figure 1. Generalized ND detection.



(a) feature points mapped in feature space (b) voronoi cell based partition

Figure 2. Construction of visual vocabulary.

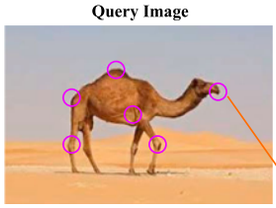
where

$$\|\$_n^R - vw_m\| = \sqrt{\sum_{i=1}^{D} (\$_{n,i}^R - vw_{m,i})^2}, \quad (2)$$

represents Euclidean distance

$\$_n^R$ denotes the *n*th reduced feature

Visual Word No	Image Nos
1	2, 3, 20, 37
2	1, 78, 100, 250
....	
50	4, 50, 75, 226, 750
...	
101	2, 6, 28, 76, 350
...	
150	76, 92, 102



Visual Word No	Image Nos
1	2, 3, 20, 37
2	1, 78, 100, 250
....	
50	4, 50, 75, 226, 750
...	
101	2, 6, 28, 76, 350
...	
150	76, 92, 102

(a) IDB to index mapping (b) Query image mapped to index of I

Figure 3. Principle of Inverted File Indexing.

$\$_{n,i}^R$ represents the *i*th value of the *n*th reduced feature

vw_m indicates the *m*th visual word

$vw_{m,i}$ denotes the *i*th value of *m*th visual word

N represents the number of features

L denotes the length of each feature vector.

The K-means technique may land at a sub-optimal trap for image descriptors given their distance distribution in feature space. The proposed methodology adopts SOA for optimally forming the visual words. Each seagull's position in any flock denotes a solution to the clustering problem as

$$\overrightarrow{PSG} = [vw_1, vw_2, \ldots, vw_K] = [(vw_{11}, vw_{12}, \ldots, vw_{1L}), (vw_{21}, vw_{22}, \ldots, vw_{2L}), \cdots, (vw_{K1}, vw_{K2}, \ldots, vw_{KL})] \quad (3)$$

The optimization process uses the seagulls' behaviour of migration and attack during the search for foods like fish, earthworms, insects, amphibians, reptiles and other small animals. The fitness of a trial solution for the problem variable vector can be assessed by a fitness function formed from the problem's objective function of Equation (1). The fitness value is calculated by the following fitness function (F) adopting Equation (1) for each candidate solution during the iterative

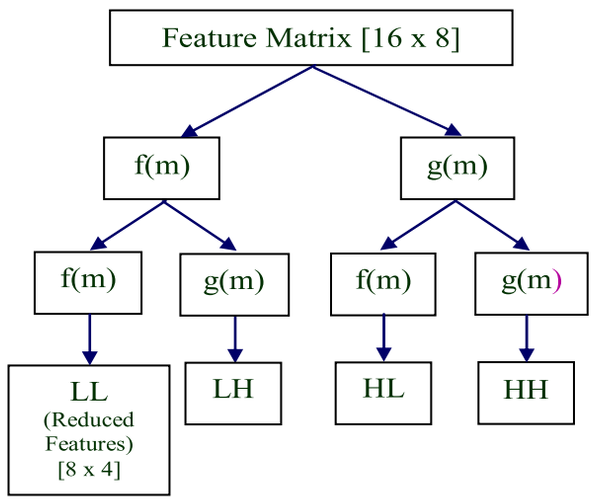


Figure 4. Two-dimensional DWT.

process.

$$\text{Maximize } F = \frac{1}{1 + \sum_{m=1}^{K} \sum_{n=1}^{N} \$_n^R - vw_m^2} \quad (4)$$

Seagulls' migration is considered a global search as such search represents large-scale flights, while their attack is represented as a local search.

Migration: During migration, seagulls must fly in the path of the best location. A supplementary variable 'S' is introduced in between neighbouring seagulls to evade collisions. The change in the position required to evade collisions can be expressed by Equation (5).

$$\overrightarrow{\Delta PSG}^{*}(t+1) = S \times \overrightarrow{PSG}(t) \quad (5)$$

where

$\overrightarrow{\Delta PSG}^{*}$ reflects the change in the position of seagulls required for evading collisions

$\overrightarrow{PSG}(t)$ denotes the current position of seagulls at instant-t.

S reflects the seagulls' movement factor, which is linearly varied from S^o to 0 during iterations by Equation (6).

$$S = S^o - \frac{S^o\, t}{MNI} \quad (6)$$

where

t is the iteration counter

S^o is the initial S parameter

MNI denotes maximum number of iterations.

The seagulls follow the fittest neighbour after evading the collisions of neighbouring seagulls. Such movement is so randomized to achieve a better balance between exploration and exploitation during

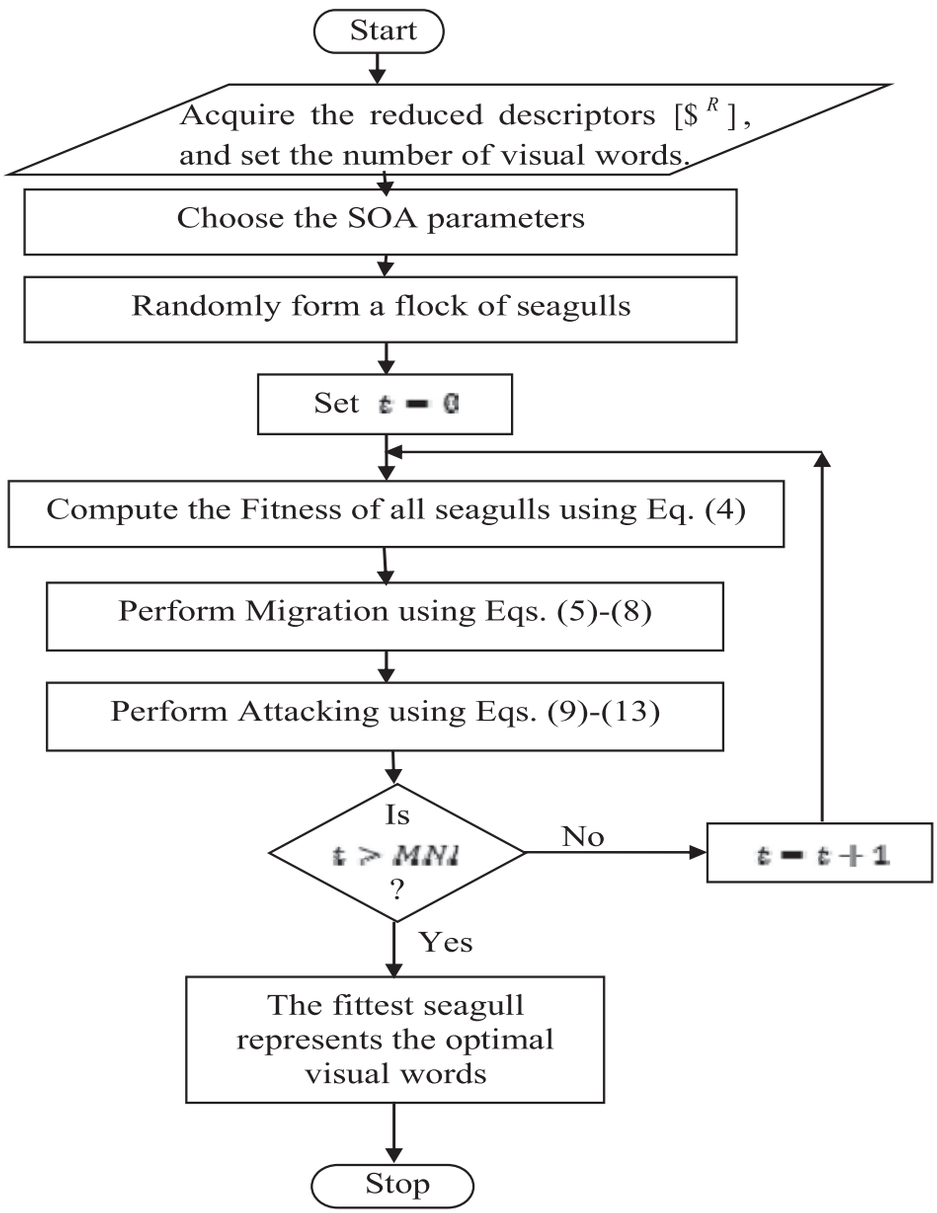


Figure 5. Flowchart of SOA.

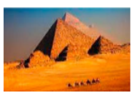
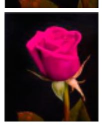
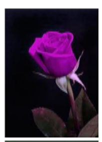
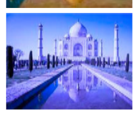
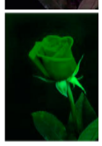
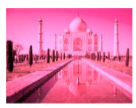
Table 1. Modules of proposed and existing methods.

	IPs			Descriptors		Feature reduction	Clustering		False matches elimination
Method	DoG	FHM	MEV	SIFT	SLCBD	Wavelet	K-means	SOA	RANSAC
EM-1	✓			✓			✓		✓
EM-2		✓		✓			✓		✓
EM-3					✓		✓		
SNDD	✓		✓	✓		✓		✓	✓

Table 2. Query images.

No.	Image	Size	No.	Image	Size
1		277* 182*3	4		201* 251*3
2		1200* 900*3	5		275* 183*3
3		259* 194*3	6		276* 183*3

Table 3. NDs in the IDB.

QI-1	QI-2	QI-3	QI-4	QI-5	QI-6
					
					
					
					

the search as given by Equation (7).

$$\overrightarrow{\Delta PSG}^{\#}(t+1) = 2 \times S^2 \times \text{rand}(0, 1) \times (\overrightarrow{PSG}^{Fit}(t) - \overrightarrow{PSG}(t)) \quad (7)$$

where $\overrightarrow{\Delta PSG}^{\#}$ reflects the change in position of seagulls required to follow the fittest seagull

$\overrightarrow{PSG}^{Fit}(t)$ denotes the fittest seagull in the flock at tth iteration

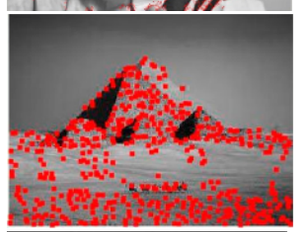
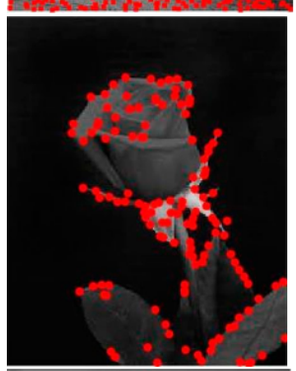
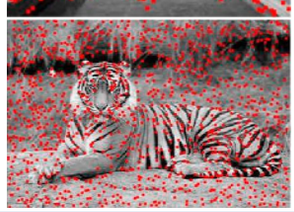
rand(0, 1) denotes a random number that falls in between 0 and 1.

The seagull can update its location following the best search agent. The net distance between the seagulls and the fittest seagull can be written by Equation (8).

$$\overrightarrow{\Delta PSG}^{net}(t+1) = (\overrightarrow{\Delta PSG}^{*}(t+1) - \overrightarrow{\Delta PSG^{\#}}(t+1)) \quad (8)$$

Attack: The purpose of exploitation is to exploit the experience and history of the search process. At the time of migration, seagulls change their velocity and attack angle. They use their wings and weight to keep their altitude and do spiral movements in the sky while attacking the prey. The behaviours in the *x*, *y* and *z*

Table 4. Query images with detected IPs.

No.	Image with DoG IPs	No. of IPs	Image with MEV IPs	No. of IPs
1		242		561
2		202		2690
3		45		466
4		39		130
5		173		519
6		262		1088

planes are given in Equations (9)–(12).

$$x^* = \check{A} \times \cos(\theta) \tag{9}$$

$$y^* = \check{A} \times \sin(\theta) \tag{10}$$

$$z^* = \check{A} \times \theta \tag{11}$$

$$\check{A} = \beta \times \tag{12}$$

where

$\check{A}$ is the spiral's radius for each rotation

θ is a random number

β and α are constant factors representing spiral rotation.

The seagulls' position can be modified using Equations (8)–(12) as

$$\overrightarrow{PSG}(t+1) = (\overrightarrow{\Delta PSG}^{net}(t+1) \times x^* \times y^* \times z^*) + \overrightarrow{PSG}^{Fit}(t) \tag{13}$$

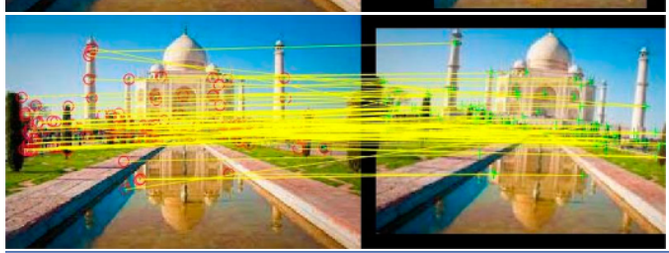
where $\overrightarrow{PSG}(t+1)$ denotes the position of seagulls in the next iteration.

The suggested SOA begins with a flock of seagulls that is randomly formed. During the iterative process, the seagulls modify their positions following the fittest seagull. The S is linearly varied from S^o to 0 during iterations for realizing a linear transition among exploration and exploitation, and lands at the global best position. The flow of the SOA's solution process is explained in Figure 5. The method also adopts the conventional modules of inverted file indexing, false match elimination by RANSAC, ranking the NDs based on the number of matches and classification. The top-ranked ones are considered NDs.

Results and discussion

The proposed SOA-based ND Detection (SNDD) was studied on five IDBs containing 5000, 10,000, 15,000, 20,000 and 25,000 images. In this regard, 4976 images were collected from websites through browsing. Among the collected images, six sample query images (QI-1 to QI-6) that classically contain photographs of Church, Mother Theresa, Pyramids, Rose, Tajmahal and Tiger, were selected and four NDs were artificially created for each of the chosen six images through colour change, rotation, intensity variation, cropping, etc. and included in the IDB, thereby making the size of 5000. Large IDBs comprising

Table 5. ND detection of SNDD for QI-5 before RANSAC.

Querry and ND images with matched IPs	No. of matched IPs
	208
	174
	128
	115

10,000, 15,000, 20,000 and 25,000 images were obtained by duplicating the IDB containing 5000 images. These IDBs (denoted as IDB-1- IDB-5) were used for studying the performances of the PMs through indexing and searching for NDs of query images. In addition, the results of the PM were compared with those of two EMs (EM-1 and EM-2) involving DoG/FHM, SIFT, K-means clustering and RANSAC, and another existing method (EM-3) adopting Spatial Layout of Colour Based Descriptor (SLCBD) and K-means clustering. The combinations of various modules employed in the proposed and EMs are given in Table 1.

The sample six QIs with their sizes and NDs are given in Tables 2 and 3 respectively. Table 4 displays the images with detected DoG and MEV-based IPs. All the detected IPs either by DoG or MEV were used in the existing methods, while in the PM, 50% of the strongest DoG-based IPs and 50% of the MEV-based IPs were used in the subsequent processes.

It is obvious from Table 4 that the DoG-based scheme identifies lesser number of IPs at abruptly changing regions with large gradients, while the MEV finds a larger number of IPs distributed throughout the image. Combining the strongest IPs of each approach gives a balance between the low entropy and high entropy regions.

The query image and the NDs marked with matched IPs before and after applying RANSAC are shown in Tables 5 and 6 respectively for QI-5. The tables also include the number of matched pairs in each detected ND image. The number of matched IPs after the removal of inconsistent IPs is slightly smaller than the number of matches before removal. However, they contain only the true matches belonging to the query and ND images.

The successfully detected and undetected NDs for all six query images are given in Table 7. It is very clear from the table that the developed method identified all the NDs of the given query images, while the existing methods failed to detect one or more NDs, especially the cropped and rotated images. The correct or wrong classification by the proposed and existing methods was also studied for all the image databases, and are represented by

- true positive (T+), the ND image correctly classified as ND;

Table 6. ND detection of SNDD for QI -5 after RANSAC.

Querry and ND images with matched IPs	No. of matched IPs
	203
	166
	122
	99

Table 7. Success (✓) and Failure (×) in detecting NDs.

Query image	Near duplicates	EM-1	EM-2	EM-3	SNDD
QI-1	ND-1	✓	✓	×	✓
	ND-2	×	×	✓	✓
	ND-3	✓	✓	×	✓
	ND-4	✓	✓	×	✓
QI-2	ND-1	✓	✓	✓	✓
	ND-2	×	✓	✓	✓
	ND-3	✓	✓	×	✓
	ND-4	✓	✓	×	✓
QI-3	ND-1	✓	✓	✓	✓
	ND-2	✓	✓	✓	✓
	ND-3	✓	✓	×	✓
	ND-4	✓	✓	×	✓
QI-4	ND-1	✓	✓	×	✓
	ND-2	✓	✓	✓	✓
	ND-3	✓	✓	×	✓
	ND-4	✓	✓	×	✓
QI-5	ND-1	✓	✓	✓	✓
	ND-2	✓	✓	×	✓
	ND-3	✓	✓	×	✓
	ND-4	✓	✓	✓	✓
QI-6	ND-1	✓	✓	✓	✓
	ND-2	✓	✓	×	✓
	ND-3	✓	✓	×	✓
	ND-4	✓	✓	✓	✓

- true negative (T–), the original image correctly classified as original;
- false positive (F+), the original image wrongly classified as ND; and
- False negative (F–), the ND image wrongly classified as original.

The true and estimated classifications of the proposed method are studied through a confusion matrix. It is expected that the proposed method should have ideally zero F^+ and F^-. The classifications by the developed and existing methods are obtained for all the IDBs and consolidated in Table 8 in addition to including the per cent of T^+, T^-, F^+ and F^-values concerning a total number of images. The table indicates that the T^+andT^-per cent values of the proposed SNDD are (99.92% & 100%), (99.9% & 100%), (99.9% & 100%), (99.88% & 95.83%) and (99.86% & 95.83%), which are relatively greater than or equal to

Table 8. Confusion matrices of developed methods.

			Detected class	
		IDB-1 (5000 images)		
			Original (4976)	ND (24)
	EM-1	Original (4976)	4970 (99.88%)	6 (0.12%)
		ND (24)	2 (8.33%)	22 (91.67%)
True Class	EM-2	Original (4976)	4962 (99.72%)	14 (0.28%)
		ND (24)	1 (4.17%)	23 (95.83%)
	EM-3	Original (4976)	2472 (49.68%)	2504 (50.32%)
		ND (24)	14 (58.33%)	10 (41.67%)
	SNDD	Original (4976)	4972 (99.92%)	4 (0.08%)
		ND (24)	0 (0%)	24 (100%)
		IDB-2 (10,000 images)		
			Original (9952)	ND (48)
	EM-1	Original (9952)	9934 (99.82%)	18 (0.18)
		ND (48)	6 (12.5%)	42 (87.5)
True Class	EM-2	Original (9952)	9914 (99.62%)	38 (0.38%)
		ND (48)	4 (8.33%)	44 (91.67%)
	EM-3	Original (9952)	4944 (49.68%)	5008 (50.32%)
		ND (48)	28 (58.33%)	20 (41.67%)
	SNDD	Original (9952)	9942 (99.90%)	10 (0.10%)
		ND (48)	0 (0%)	48 (100%)
		IDB-3 (15,000 images)		
			Original (14928)	ND (72)
	EM-1	Original (14928)	14898 (99.80%)	30 (0.20%)
		ND (72)	9 (12.5%)	63 (87.5%)
True Class	EM-2	Original (14928)	14868 (99.60%)	60 (0.40%)
		ND (72)	6 (8.33%)	66 (91.67%)
	EM-3	Original (14928)	7416 (49.68%)	7512 (50.32%)
		ND (72)	42 (58.33%)	30 (41.67%)
	SNDD	Original (14928)	14913 (99.90%)	15 (0.10%)
		ND (72)	0 (0%)	72 (100%)
		IDB-4 (20,000 images)		
			Original (19904)	ND (96)
	EM-1	Original (19904)	19856 (99.76%)	48 (0.24%)
		ND (96)	12 (12.5%)	84 (87.5%)
True Class	EM-2	Original (19904)	19820 (99.58%)	84 (0.42%)
		ND (96)	8 (8.33%)	88 (91.67%)
	EM-3	Original (19904)	9888 (49.68%)	10016 (50.32%)
		ND (96)	56 (58.33%)	40 (41.67%)
	SNDD	Original (19904)	19880 (99.88%)	24 (0.12%)
		ND (96)	4 (4.17%)	92 (95.83%)
		IDB-5 (25,000 images)		
			Original (24880)	ND (120)
	EM-1	Original (24880)	24815 (99.74%)	65 (0.26%)
		ND (120)	15 (12.5%)	105 (87.5%)
True Class	EM-2	Original (24880)	24770 (99.56%)	110 (0.44%)
		ND (120)	10 (8.33%)	110 (91.67%)
	EM-3	Original (24880)	12360 (49.68%)	12520 (50.32%)
		ND (120)	70 (58.33%)	50 (41.67%)
	SNDD	Original (24880)	24845 (99.86%)	35 (0.14%)
		ND (120)	5 (4.17%)	115 (95.83%)

Table 9. Comparison of performances.

	Method	Accuracy	Sensitivity	Specificity	Precision	F1
IDB-1	EM-1	99.84	91.67	99.88	78.57	84.62
	EM-2	99.70	95.83	99.72	62.16	75.41
	EM-3	49.64	41.67	49.68	0.40	0.79
	SNDD	99.92	100.00	99.92	85.71	92.31
IDB-2	EM-1	99.76	87.50	99.82	70.00	77.78
	EM-2	99.58	91.67	99.62	53.66	67.69
	EM-3	49.64	41.67	49.68	0.40	0.79
	SNDD	99.90	100.00	99.90	82.76	90.57
IDB-3	EM-1	99.74	87.50	99.80	67.74	76.36
	EM-2	99.56	91.67	99.60	52.38	66.67
	EM-3	49.64	41.67	49.68	0.40	0.79
	SNDD	99.90	100.00	99.90	82.76	90.57
IDB-4	EM-1	99.70	87.50	99.76	63.64	73.68
	EM-2	99.54	91.67	99.58	51.16	65.67
	EM-3	49.64	41.67	49.68	0.40	0.79
	SNDD	99.86	95.83	99.88	79.31	86.79
IDB-5	EM-1	99.68	87.50	99.74	61.76	72.41
	EM-2	99.52	91.67	99.56	50.00	64.71
	EM-3	49.64	41.67	49.68	0.40	0.79
	SNDD	99.84	95.83	99.86	76.67	85.19

the existing methods (EM-1, EM-2 and EM-3), thereby portraying the superiority of the proposed methods.

The most important performance metrics of accuracy, sensitivity, specificity, precision and F1 were also computed for the proposed and existing methods and compared for all the five IDBs in Table 9. The performance metrics of the proposed SNDD are much better than the existing methods of EM-1, EM-2 and EM-3. The average performance metrics of the proposed SNDD method are graphically compared with EM-1, EM-2 and EM-3 in Figure 6, which exhibits the superior performances of the proposed SNDD for different-sized IDBs.

$$Accuracy = \frac{(T^+ + T^-)}{(T^+ + T^- + F^+ + F^-)} \tag{14}$$

$$Sensitivity = \frac{T^+}{(T^+ + F^-)} \tag{15}$$

$$Specificity = \frac{T^-}{(T^- + F^+)} \tag{16}$$

$$Precision = \frac{T^+}{(T^+ + F^+)} \tag{17}$$

$$F_1 = \frac{2 \times Precision \times Sensitivity}{Precision + Sensitivity} \tag{18}$$

With a view of studying the computational efficiency of the proposed SNDD, the time for evaluating the features and creating the Bag of Visual Words by EM-1, EM-2, EM-3 and SNDD for all the five IDBs were measured, and given in Table 10. Analysing the computation times, it is generally clear that the proposed SNDD is

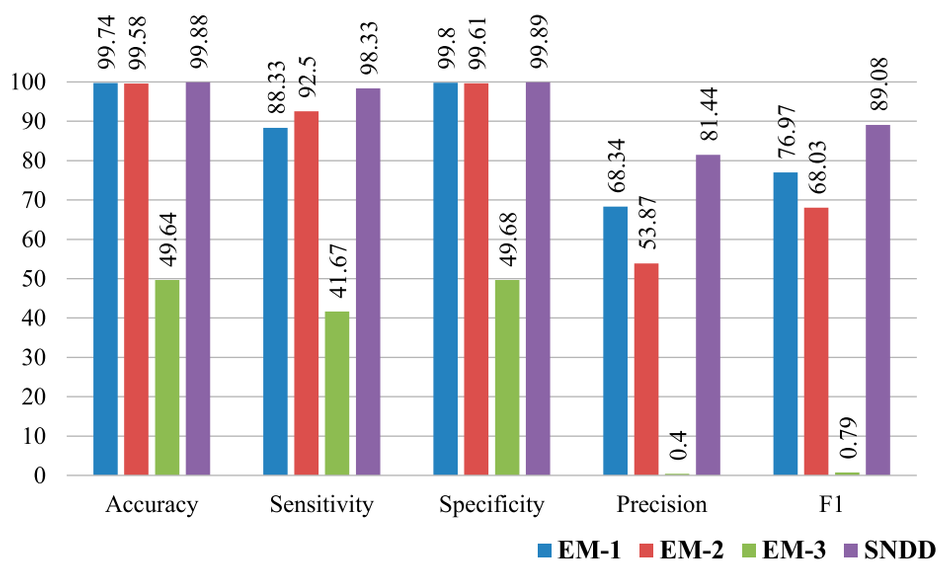


Figure 6. Comparison of average performance metrics.

Table 10. Computation time for creating bag of visual words.

Database size	EM-1	EM-2	EM-3	PM
IDB-1 (5000)	320	340	256	273
IDB-2 (10,000)	580	680	483	556
IDB-3 (15,000)	860	980	698	740
IDB-4 (20,000)	1150	1340	967	1110
IDB-5 (25,000)	1650	1900	1184	1538

Table 11. Querying time (seconds).

Database size	EM-1	EM-2	EM-3	SNDD
IDB-1 (5000)	2.1	2.4	1.7	1.9
IDB-2 (10,000)	2.7	2.9	2.2	2.3
IDB-3 (15,000)	3.4	3.1	2.4	2.6
IDB-4 (20,000)	3.7	3.3	2.9	3.1
IDB-5 (25,000)	4.2	3.7	3.1	3.4

faster than the EM-1 and EM-2, slower than EM-3 at the cost of losing other performances. It is to be pointed out that the computation time exponentially increases with the size of the IDB. The large computation times are acceptable, as they are to be formed only once, before searching for NDs. The querying time for searching NDs for a query image is very important and hence measured for all the six query images at all the five IDBs, and compared in Table 11. It is seen from the table that the querying time is much smaller than the existing methods of EM-1 and EM-2, and slightly larger than EM-3. The proposed SNDD is computationally superior to existing methods.

Conclusion

A new ND image detection method involving DoG, MEV, DWT and SOA was developed in this paper. The strongest DoG and MEV-based IPs were used for distributing the IPs over low and high entropy regions of images, and the DWT was used for reducing the dimensionality of each feature vector. The SOA was employed for optimally forming the visual words. The developed method was studied on IDBs comprising 5000–25000 digital images and found that the developed method was more robust and computationally efficient, thereby making it suitable for online applications. Moreover, the blending of the DoG and MEV-based IPs makes the detection process more robust, even on images with low entropy regions. The method can further be modified to detect image forgeries. Moreover, other types of IP detection methods and other metaheuristic algorithms can be applied to the developed ND detection method for future work. Intelligent algorithms like fuzzy logic and deep learning can also be applied for enhancing the ND detection process.

Acknowledgements

The authors thankfully acknowledge the administrative officers of Annamalai University for the computing and internet facilities provided to perform this work.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Data availability statement

Data sharing does not apply to this article as no new data has been created or analysed in this study.

Notes on contributors

Srinidhi Sundaram, Assistant Professor, Department Computer Science and Engineering, Agni College of Technology, Chennai, Tamil Nadu, received the B.Sc Degree in Computer Science from Pandit Ravishankar Shukla University, M.C.A from Madurai Kamaraj University and M. Tech Degree in Information Technology from Sathyabama University, in 1999, 2003 and 2010 respectively. She is presently pursuing part-time Ph.D, Department of Information Technology, Annamalai University, Tamil Nadu, India. She has 15 years of teaching experience and is specialized in the area of image forensics, soft computing and metaheuristic optimization.

S. Kamalakkannan received his Ph.D. Degree in Computer Science from Vels Institute of Science, Technology & Advanced Studies (VISTAS), Chennai, Tamil Nadu, India. He is currently working as Associate Professor, Department of Information Technology, School of Computing Sciences, Vels Institute of Science, Technology & Advanced Studies (VISTAS), Chennai, Tamil Nadu, India, which is a well-known university. He has 18 years of teaching experience in both UG and PG level. His research interest includes Data Mining, Big Data Analytics, Cloud Computing and Block Chain Technology. He has produced one Ph.D Research scholar. He has published more than 50 research articles in various International journals such as Sci, WOS, Scopus and UGC referred journals. He serves as an Examiner in various Universities and Colleges. He received Best Young Scientist award and Best Faculty award.

Sasikala Jayaraman received the B.E. Degree in Electronics and Communication Engineering from Madras University, India in 1993, and the M.E and Ph. D degrees in Computer Science and Engineering from Annamalai University in 2005 and 2011 respectively. She has been working as an Associate Professor, Department of Information Technology, Annamalai University, Tamil Nadu, India since 1999. Her research interests are in the area of optimization, evolutionary algorithms and image processing.

References

[1] Thaiyalnayaki S, Sasikala J, Ponraj R. Indexing near-duplicate images in web search using minhash algorithm. Mater Today Proc. 2018;5(1):1943–1949.

[2] Pawar VB, Mankar JR. A survey on matching and retrieval of near duplicate images. Int J Sci Res Develop. 2016;3(11):580–581.

[3] Dhiman G, Kumar V. Seagull optimization algorithm: theory and its applications for large-scale industrial engineering problems. Knowl Based Syst. 2019;165:169–196.

[4] Jiang H, Yang Y, Ping W, et al. A novel hybrid classification method based on the opposition-based seagull optimization algorithm. IEEE Access. 2020;8:100778–90.

[5] Mani R, Jayaraman S, Ellappan M. Hybrid seagull and thermal exchange optimization algorithm-based NLOS nodes detection technique for enhancing reliability under data dissemination in VANETs. J Int Commun Syst. 2020;33(14):e4519.

[6] Ke Y, Sukthankar R, Huston L, et al. Efficient near-duplicate detection and sub-image retrieval. ACM multimedia 2004 (Vol. 4, No. 1, p. 5).

[7] Wang XJ, Zhang L, Ma WY. Duplicate-search-based image annotation using web-scale data. Proc IEEE. 2012;100(9):2705–2721.

[8] Chen L, Stentiford F. Comparison of near-duplicate image matching. Proc. 3rd European conference on visual media production, CVMP; 2006.

[9] Foo JJ, Zobel J, Sinha R. Clustering near-duplicate images in large collections. Proceedings of the international workshop on workshop on multimedia information retrieval; 2007 (pp. 21–30).

[10] Zhao WL, Ngo CW, Tan HK, et al. Near-duplicate keyframe identification with interest point matching and pattern learning. IEEE Trans Multimed. 2007;9(5):1037–1048.

[11] Chum O, Philbin J, Zisserman A. Near duplicate image detection: Min-hash and TF-IDF weighting. In Bmvc. 2008; (Vol. 810,: pp. 812-815).

[12] Xu D, Cham TJ, Yan S, et al. Near duplicate identification with spatially aligned pyramid matching. IEEE Trans Circuits Syst Video Technol. 2010;8:1068–1079.

[13] Wang XJ, Zhang L, Liu C. Duplicate discovery on 2 billion internet images. Proceedings of the IEEE conference on computer vision and pattern recognition workshops; 2013, pp. 429–436.

[14] Hsieh SL, Chen CC, Chen CR. A novel approach to detecting duplicate images using multiple hash tables. Multimed Tools Appl. 2015;74(13):4947–4964.

[15] Yao J, Yang B, Zhu Q. Near-duplicate image retrieval based on contextual descriptor. IEEE Signal Process Lett. 2014;22(9):1404–1408.

[16] Liu L, Lu Y, Suen CY. Variable-length signature for near-duplicate image matching. IEEE Trans Image Process. 2015;24(4):1282–1296.

[17] Deshmukh AS, Lambhate PD. A methodological survey on mapreduce for identification of duplicate images. Int J Sci Res (IJSR). 2016;5(1):206–210.

[18] Layek AK, Gupta A, Ghosh S, et al. Fast near-duplicate detection from image streams on online social media during disaster events. 2016 IEEE annual India conference (INDICON); 2016, pp. 1–6. IEEE.

[19] Landge A, Mane P. Near duplicate image matching techniques. (2016). International conference on information communication and embedded systems (ICICES) (pp. 1–5). IEEE.

[20] Zhang W, Ji J, Zhu J, et al. Bithash: an efficient bitwise locality sensitive hashing method with applications. Knowl Based Syst. 2016;97:40–47.

[21] Jayshree B, Bhale NL. A survey on finding image similarity and retrieval of near duplicate images.

[22] Fellah A. All-Three: near-optimal and domain-independent algorithms for near-duplicate detection. Array. 2021;11:100070.

[23] Albayrak OS, Aytekin T, Kalaycı TA. Duplicate product record detection engine for e-commerce platforms. Expert Syst Appl. 2022;193:116420.

[24] Chevallier M, Rogovschi N, Boufarès F, et al. Detecting near duplicate dataset with machine learning. Int J Comp Inf Syst Indust Manag Appl. 2022;14:374–395.

[25] Gusev A, Xu J. Evolution of a web-scale near duplicate image detection system. https://deepai.org/publication/evolution-of-a-web-scale-near-duplicate-image-detection-system. 2022.

[26] Wang Z, Zuo C, Dend D. Txtalign: efficient near-duplicate text alignment search via bottom-k sketches for plagiarism detection. SIGMOD '22: proceedings of the 2022 international conference on management of data; 2022; p. 1146–1159.

[27] Mehta P, Tripathi RK. Near-duplicate detection for LCD screen acquired images using edge histogram descriptor. Multimed Tools Appl. 2022;81:30977–30995.

[28] Shayegan MJ, Faizollahi-Samarin M. An extended version of sectional MinHash method for near-duplicate detection. J Supercomput. 2022;78:15638–15662.

[29] Fischler MA, Bolles RC. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Commun ACM. 1981;24(6):381–395.

[30] Kok KY, Rajendran P. Validation of Harris detector and eigen features detector. IOP conference series: materials science and engineering 2018 (Vol. 370, No. 1, p. 012013). IOP Publishing.

[31] Lowe DG. Distinctive image features from scale-invariant keypoints. Int J Comput Vision. 2004;60 (2):91–110.

[32] Grzegorzek M, Sav S, O'Connor NE, et al. Local wavelet features for statistical object classification and localization. IEEE MultiMedia Magazine. 2010;17(1):118.

[33] Morissette L, Chartier S. The k-means clustering technique: general considerations and implementation in mathematica. Tutor Quant Methods Psychol. 2013;9 (1):15–24.