www.nature.com/scientificreports

SCIENTIFIC REPORTS

OPEN

Received: 28 November 2016
Accepted: 6 June 2017
Published online: 30 June 2017

Automatic Detection of Galaxy Type From Datasets of Galaxies Image Based on Image Retrieval Approach

Mohamed Abd El Aziz[1,3,5], I. M. Selim[2,4] & Shengwu Xiong[1]

This paper presents a new approach for the automatic detection of galaxy morphology from datasets based on an image-retrieval approach. Currently, there are several classification methods proposed to detect galaxy types within an image. However, in some situations, the aim is not only to determine the type of galaxy within the queried image, but also to determine the most similar images for query image. Therefore, this paper proposes an image-retrieval method to detect the type of galaxies within an image and return with the most similar image. The proposed method consists of two stages, in the first stage, a set of features is extracted based on shape, color and texture descriptors, then a binary sine cosine algorithm selects the most relevant features. In the second stage, the similarity between the features of the queried galaxy image and the features of other galaxy images is computed. Our experiments were performed using the EFIGI catalogue, which contains about 5000 galaxies images with different types (edge-on spiral, spiral, elliptical and irregular). We demonstrate that our proposed approach has better performance compared with the particle swarm optimization (PSO) and genetic algorithm (GA) methods.

Astronomy has become an immensely data-rich field. For example, the Sloan Digital Sky Survey (SDSS) will produce more than 50,000,000 images of galaxies in the near future[1]. In turn, galaxy morphology can be used to provide an independent test of the two proposed scenarios for galaxy formation. Elliptical galaxies, for example, are believed to be formed through major mergers[2], whereas disk-dominated galaxies cannot have undergone recent major mergers, as such mergers would have severely disrupted their shape[3]. Thus, the class of quenching models is sufficient to explain the full range of morphological types observed for quenched galaxies. For example, bars can be found in all types of disk galaxies, from the earliest to the latest stages of the Hubble sequence. Barred galaxies constitute a major fraction of all disk galaxies. A small number of galaxies that appear unbarred at visual wavelengths have actually been found to be barred when observed in the near infra-red. The three clearest cases are NGC 1566[4], NGC 1068[5,6] and NGC 309[7]. De Zeeuw and Franx[8] surveyed the literature for the dynamics of these objects. We are still far from a complete understanding of the dynamical structure of galaxies. Here, we will be able to do no more than scratch the surface of the majority of these problems.

The development of galaxy morphological schemes can be used to successfully determine galaxy morphology via classification or image-retrieval methods. For example, the Deep Neural Network (DNN) algorithm has been used to classify the Galaxy Zoo (e.g. ref. 9). This method minimizes the sensitivity to changes in the scaling, rotation, translation and sampling of an image by using a rotation-invariant convolution. The results of this method is better than 99% with respect to human classification; however, as human classification has several associated errors, in turn the DNN approach also suffers from the same errors[10]. In ref. 11, the random forest method was used to classify an *HST*/WFC3 image containing 1639 galaxies, which identified disturbed morphologies using

[1]School of Computer Science and Technology, Wuhan University of Technology, Wuhan, China. [2]National Research Institute of Astronomy and Geophysics, Astronomy Department, Cairo, Egypt. [3]Department of Mathematics, Faculty of Science, Zagazig University, Zagazig, Egypt. [4]Higher Institute of Technology, Department of Computer Science, Tenth of Ramadan City, Egypt. [5]Scientific Research Group in Egypt (SRGE), Cairo, Egypt. Mohamed Abd El Aziz, I. M. Selim and Shengwu Xiong contributed equally to this work. Correspondence and requests for materials should be addressed to M.A. (email: abd_el_aziz_m@yahoo.com) or S.X. (email: xiongsw@whut.edu.cn)

multimode, intensity and deviation statistics. Additionally[12], proposed a method that consists of two stages: first, feature extraction (shape, color and concentration) of galaxy images from the SDSS DR7 spectroscopic sample, followed by the classification of these features using a support vector machine.

The authors in ref. 10, proposed a different approach called MORFOMETRYKA, which used the Linear Discriminant Analysis (LDA) algorithm to classify various features (concentration, asymmetry, smoothness, entropy and spirality) extracted from the galaxy images. The results of their approach were better than 90% based on 10-fold cross validation to classify a galaxy as either an elliptical or a spiral.

These galaxy classification methods have provided powerful results. However, there is another trend to deal with galaxy images, i.e. to determine the most similar images to query image, not classify them into groups only, therefore, the image retrieval techniques are needed[13].

The image-retrieval method is a computer system for browsing, searching, detecting and retrieving images from a large database of digital images[14]. The content-based image retrieval (CBIR) approach is one of the most commonly used image retrieval methods[15], which aims to avoid the use of textual descriptions and instead retrieves images based on similarities in their content. Relevant content can be information related to image patterns, colors, textures, shape and location[16].

Such image content is obtained by using feature-extraction methods, which is then saved in a database. To answer a queried image, the similarity between stored features and the features of a queried image (extracted using the same method) is computed and used to determine the closest between the images. However, the CBIR approach is a challenging problem for galaxy images, because there is a large number of galaxy images and determining the most relevant images from a large database becomes a non-trivial task.

Several methods have been applied to improve the quality of CBIR for galaxy images. Ref. 17 introduced a CBIR method for astronomical images which used a multi-resolution approach to compress the original images in sketches. These sketches (features) were compared with the features of the queried image through the use of correlation and symmetry functions[18]. Next, ref. 19 proposed a CBIR method which summarized and indexed the Zurich archive of solar radio spectrograms. The summarized step was performed by clustering the content of an image into groups (regions) by using the same texture feature, which were represented by a set of parameters (location, a texture roughness and region extensions). The indexing step was then performed by quantizing these regions.

In general, the previous methods consider either the shape, the texture features or the color, or both of them (color/texture, color/shape and shape/texture), but not all of them. Moreover, not all of the extracted features are important: some may be redundant/irrelevant, which in turn reduce the quality of the classification or image-retrieval results. To address this, the aim of this paper is to introduce a new machine-learning approach for the retrieval of galaxy images. Our approach avoids the limitations of previous methods by extracting the shape, color and texture features from galaxy images, and then determining the most relevant features and ignoring other features by using the *K*-NN classifier as measure of the quality of the features which selected by Sine Cosine algorithm (SCA).

The proposed approach consists of two stages: training and image retrieval. In the training stage there are two steps: the first is feature extraction, where the color, shape and texture features are extracted from a dataset of galaxy images. The second step is feature selection, which is performed based on the modified sine cosine algorithm[20] that selects the most relevant features using the classification accuracy as a fitness function. In the second stage, similar images to the queried image are returned by using the Euclidean distance as a measure.

Feature extraction

In this section, visual features such as color, texture and shape are introduced[15].

Color Feature Extraction. The color of an image is one of the most widely used features in image retrieval and several other image-processing applications. It is a very important feature since it is invariant with respect to scaling, translation and rotation[21]. Therefore, the aim of any color feature extraction method is to represent the main colors of the image content (red, green, and blue, i.e. RGB) and then use these color features to describe the image and distinguish it from other images. RGB colors used in this study were obtained by converting from the SDSS color system using the Maxim DL astronomical software[22].

The color histogram is one of the most well-known color features used for image feature extraction[23, 34], which denotes the joint probability of the intensity of an image. From probability theory, a probability distribution can be uniquely characterized by its moments. Thus, if we interpret the color distribution of an image as a probability distribution, moments can be used to characterize the color distribution. The moments of the color distribution are the features extracted from the images; if we denote the value of the *i*th color channel at the *j*th image pixel as P_{ij}, then the color moments can be defined as refs 23 and 24:

- The first-order moment (the mean):

$$E_i = \frac{1}{N}\sum_{j=1}^{N} P_{ij} \tag{1}$$

- The second-order moment (the standard deviation):

$$\sigma_i = \sqrt{\frac{1}{(N-1)}\sum_{j=1}^{N} (P_{ij} - E_i)^2} \tag{2}$$

- The third-order moment (skewedness):

$$s_i = \sqrt[3]{\frac{1}{N}\sum_{j=1}^{N}(P_{ij} - E_i)^3} \tag{3}$$

Texture Feature Extraction. The texture descriptor is an important feature that provides properties such as smoothness, coarseness and regularity[25]. Textures can be rough or smooth, vertical or horizontal. Generally, they capture patterns in the image data, such as repetitiveness and granularity.

There are several texture extraction methods, such as the discrete cosine transform (DCT), the discrete Fourier transform (DFT), discrete wavelet transform (DWT) and the Gabor filter[26, 27]. The Gray Level Co-Occurrence Matrix (GLCM) and Color Co-Occurrence Matrix (CCM) are the most commonly used statistical approaches used to extract the texture of an image[28]. These features include the contrast, correlation, entropy, energy and homogeneity, which are defined as:

- The contrast represents the amount of local variation in an image. This concept refers to pixel variance, and it is defined as:

$$CN = \frac{1}{(G-1)^2}\sum_{u=0}^{G-1}\sum_{v=0}^{G-1}|u - v|^2 p(u, v) \tag{4}$$

- The correlation represents the relation between pixels in an image, which determines the linear dependency between two pixels and is defined as:

$$CR = \frac{1}{2}\sum_{u=0}^{G-1}\sum_{v=0}^{G-1}\frac{(u - \mu_u)(v - \mu_v)}{\sigma_u^2\sigma_v^2}p(u, v) + 1 \tag{5}$$

- The energy (*En*) represents the textural uniformity, where large values of *En* indicate a completely homogeneous image.

$$En = \sum_{u=0}^{G-1}\sum_{v=0}^{G-1}p(u, v)^2 \tag{6}$$

- The entropy (*ET*) measures the randomness of the intensity distribution. It is inversely correlated to *En*, and is defined as:

$$ET = \frac{1}{2log(G)}\sum_{u=0}^{G-1}\sum_{v=0}^{G-1}p(u, v)\,log_2 p(u, v) \tag{7}$$

- The homogeneity (*H*) is used to measure the closeness of the distribution, where large values *H* indicate that the image contrast is low. The definition of *H* is given in the following equation:

$$H = \sum_{u=0}^{G-1}\sum_{v=0}^{G-1}\frac{p(u, v)}{1 + |u - v|} \tag{8}$$

where u, v are the coordinates of the co-occurrence matrix, G is the number of grey levels, and μ_u, μ_v, σ_u, and σ_v are the mean values and the standard deviations of the uth row of the vth column of the co-occurrence matrix, respectively.

Shape Feature Extraction. Shape features were extracted by using the contour moments defined mathematically as follows. Let $z(i)$ be an ordered sequence that represents the Euclidean distance between the centroid and all N boundary pixels of the object. The rth contour sequence moment m_r[14] is defined as:

$$m_r = \frac{1}{N} \times \sum_{i=1}^{N}[z(i)]^r \tag{9}$$

Sine Cosine Algorithm

In this section, the sine cosine algorithm (SCA) is illustrated[20], this algorithms is a new meta-heuristic algorithm which used either the sine or cosine function to search about the best solution. Consider the current solution X_i, $(i = 1, 2, \ldots, pop_{size})$ from the population of solutions is updated as in the following equation[20]:

$$X_i = X_i + r_1 \times \sin(r_2) \times |r_3 P - X_i| \tag{10}$$

$$X_i = X_i + r_1 \times \cos(r_2) \times |r_3 P - X_i| \tag{11}$$

The previous two equations were combined to update the solution that can be simultaneously by switching between the sine or cosine function[20]:

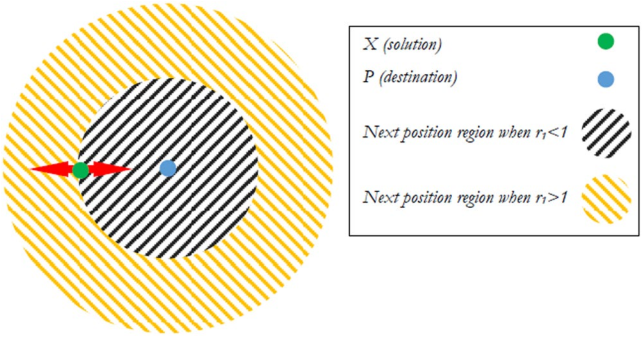
X (solution) ●
P (destination) ●
Next position region when $r_1<1$ ▨
Next position region when $r_1>1$ ▨

Figure 1. The Sine and Cosine functions effects on the next solution[20].

Algorithm	Parameters	Value
BSCA	a	2
PSO	Inertia weight	0.5
	Maximum velocity	1.0
	Minimum velocity	−1.0
	Cognitive coefficient	1
	Cognitive coefficient	2
GA	cross probability of	0.7
	Mutation Percentage	0.4
	Mutation Rate	0.1

Table 1. The parameter settings of each algorithm.

$$X_i = \begin{cases} X_i + r_1 \times \sin(r_2) \times |r_3 P - X_i| & if\ r_4 < 0.5 \\ X_i + r_1 \times \cos(r_2) \times |r_3 P - X_i| & if\ r_4 \geq 0.5 \end{cases} \tag{12}$$

where r_1, r_2, r_3 and r_4 are random variables, P is the best solution, and $|\cdot|$ represents the absolute value[20].

Following ref. 20, each parameter was used to perform a specific task. For example, the r_2 parameter defines the direction of X_i (i.e., towards or away from P), while r_3 gives random weights to P in order to stochastically emphasize ($r_3 > 1$) or deemphasize ($r_3 < 1$) its influence when defining the distance. Next, r_4 is responsible for switching between the sine and cosine functions in equation (12)[20]. Finally, r_1 was used to determine the next position regions (or movement direction), which could be either in the space between X_i and P or outside of this space, and it is also responsible for balancing between the exploration and exploitation to improve the convergence performance by updating its value as ref. 20:

$$r_1 = a - t\frac{a}{t_{max}} \tag{13}$$

where t is the current iteration, t_{max} is the maximum number of iterations, and a is a constant. Figure 1 shows how equation (12) defines a region between two solutions in the searched space.

The Proposed Image Retrieval Approach

In this section, we investigate a new approach to galaxy image retrieval as illustrated in Algorithm 1. Our proposed approach consists of two stages: a training stage and the galaxy image retrieval stage.

In the first stage, the input is the dataset of galaxy images. Then the shape, texture and color features are extracted for each galaxy image I, which are combined into a feature vector FV_I, where I is the current image. The next step in the training stage is to reduce the size of FV through using the Binary SCA (BSCA) algorithm (see Algorithm 2) to select the most relevant features. This process is performed by maximizing the accuracy of the K-NN classifier, which is used as a fitness function.

The BSCA starts by generating a random population of size pop_{size}, and the output is the best solution P that points to the selected features (Sel_{Feat}). The solution in the population of the BSCA algorithm is represented as a binary vector by using the sigmoid function which transforms a real number into a binary number as:

$$X_i = \begin{cases} 1 & if\ S(X_i) > \sigma \\ 0 & otherwise \end{cases}, \quad S(x_i) = \frac{1}{1 + e^{-X_i}} \tag{14}$$

where $\sigma \in [0, 1]$ and X_i is the current solution (for example, the solution $X_i = 001100$ with six features means that the third and fourth features are selected).

After the solutions are converted to binary vectors, the fitness function is computed for each solution. The fitness function is defined according to the classification accuracy rate as:

$$F_i = \frac{N_C}{N_I} \times 100 \tag{15}$$

where N_C is the number of correctly predicted samples, and N_I represents the total number of images. The dataset is divided by using a 10-fold cross validation (CV), and then the K-NN algorithm predicts, using the label of the testing set, where the output from 10-fold CV is the average of accuracy through 10 runs.

The solution X_i is updated using equations (10) or (11) based on the value of r_4. This process is repeated until the maximum number of iterations is reached, or there is only a small difference between F_i^{old} and F_i. The output of this stage is the global best solution P, which represents the optimally selected features Sel_{Feat}.

The second stage starts by extracting the features of a queried image FQ, and then the same features corresponding to Sel_{Feat} are selected. Then the Euclidian distance is used to compute the similarity between FQ and FV, and the closest images to the query image are returned (based on the small difference or the required number of images).

Algorithm 1 The Proposed approach For Galaxy Image Retrieval

```
Input: database of images, queried image.
Output: precision and recall.
Training stage:
  • Compute the feature vectors $FV_I$ for all images in the database.
  • Select features $Sel_{Feat} = \text{BSCA}(FV)$.
  • Update the set of features $FV = FV(Sel_{Feat})$.
Image retrieval stage:
  • Compute the feature vector $FQ$ of the queried image $I_Q$.
  • Update $FQ = FQ(Sel_{Feat})$.
  • For {all $\mathbf{I}_i$ % in parallel techniques}
  • Compute the distance between $FQ$ and $FV_i$ using the Euclidean distance $EDist_i$.
  • end for
Select the smallest distance from $EDist$ and determine the index $S_{index}$ that satisfies $EDist < \epsilon$.
Select from the database any images with index $S_{index}$.
Compute the precision and recall.
```

Algorithm 2 Binary Sine Cosine Algorithm (BSCA)

```
Input: features of each image ($FV$).
Initialize a set of solutions ($X$) with size $pop_{size}$, and set the maximum number of iterations $t_{max}$.
for i = 1: $pop_{size}$ do
    Convert $X_i$ to a binary vector using equation (14).
    Compute the fitness function $F_i$ based on the selected features from $FV$ and using 10-fold cross-validation.
    if $F_i < F_P$ then
        $F_P = F_i$.
        $P = X_i$.
    end if
end for
repeat
    Update $r_1$, $r_2$, $r_3$, and $r_4$.
    Update the position using equation (12).
until ($t < t_{max}$)
Return the best solution $P$ obtained so far as the global optimum $F_P$.
```

Experimental Results

We tested our proposed approach using the EFIGI catalogue, which consists of 4458 galaxy images[29]. We also compared the performance of our method with the particle swarm optimization (PSO)[30] and genetic algorithm (GA)[31] methods. The parameters used in each algorithm is given in Table 1. The common parameters between the three algorithms are the population size, the maximum number of iterations which was set to 20 and 100, respectively, and the maximum number of iterations used as the stopping criteria. The experiments were implemented in Matlab and run in the Windows environment with 64-bit support.

	No. of Features	Name of Selected Features	Accuracy
BSCA	12	Third Color moment (3), Energy(2), Homogenity(3), Entropy(3), Contour moment (1)	94.23
PSO	19	Third Color moment (3), Second Color moment (3), Contrast(2), Energy(3), Homogenity(2), Entropy(2), Contour moment(1)	93.59
GA	20	Third Color moment (3),Second Color moment (3), Contrast(4), Energy(4), Homogenity(4), Entropy(1), Contour moment(1)	92.95

Table 2. The selected features and their accuracy.

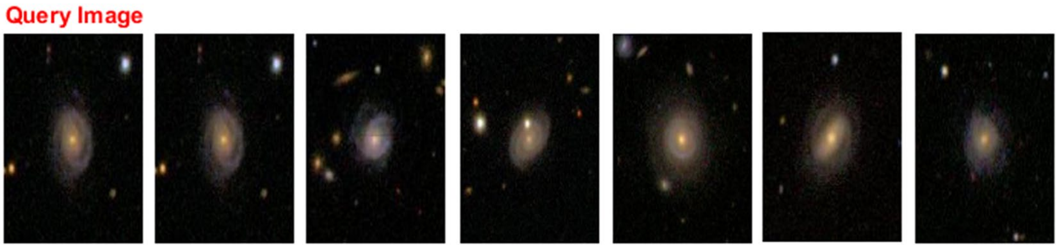

Figure 2. Galaxy image retrieval for a spiral galaxy[29].

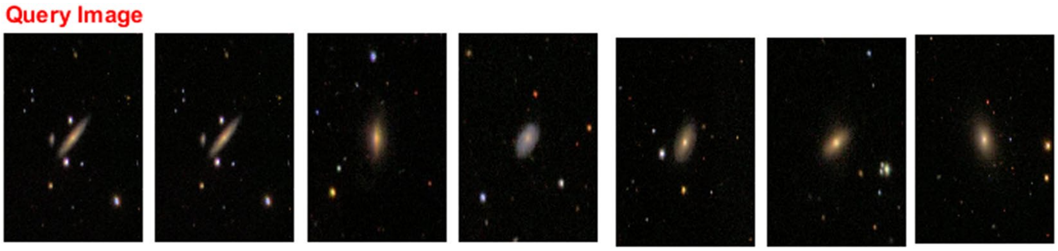

Figure 3. Galaxy image retrieval for an edge-on spiral galaxy[29].

Images Database.

The EFIGI catalogue[29] contains 16 morphological attributes that were measured by visual examination of the composite g, u, r color image of each galaxy, derived from the SDSS FITS images using[29]. The EFIGI catalogue merges data from standard surveys and catalogues (the Principal Galaxy Catalogue, SDSS, the Value-Added Galaxy Catalogue, HyperLeda, and the NASA Extragalactic Database). The bulge-to-disk ratio[32] and the degree of azimuthal variation of the surface brightness were often used as discriminant parameters along the Hubble sequence. This is not surprising since the EFIGI classification scheme is very close to the RC3 system. The final EFIGI database is a large sub-sample of the local universe which densely samples. The EFIGI morphological sequence is based on the RC3 revised Hubble sequence (RHS), which we call the EFIGI morphological sequence (EMS).

Finally, all colors of the original data were used to create composite, "true color", RGB images in PNG format with the Maxim DL astronomical software[22], using the same intensity mapping for all RGB images.

Performance measures.

Two measurements were used to evaluate the performance of the proposed algorithm: the precision rate and the recall rate.

- The precision rate is defined as the ratio of the number of retrieved images similar to the queried image relative to the total number of retrieved images[28].

$$precision = \frac{p}{p + r} \times 100 \tag{16}$$

- The recall rate is defined as the percentage of retrieved images similar to the query image among the total number of images similar to the queried image in the database[28].

$$recall = \frac{p}{p + q} \times 100 \tag{17}$$

where p, q and r are the number of relevant images retrieved, relevant images in the dataset which are not retrieved, and non-relevant images in the dataset which are retrieved, respectively.

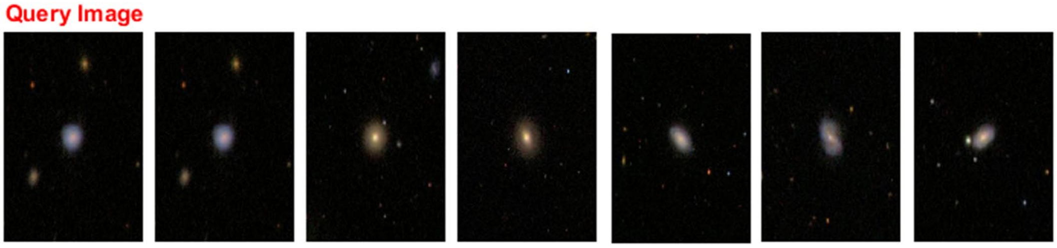

Figure 4. Galaxy image retrieval for an irregular galaxy[29].

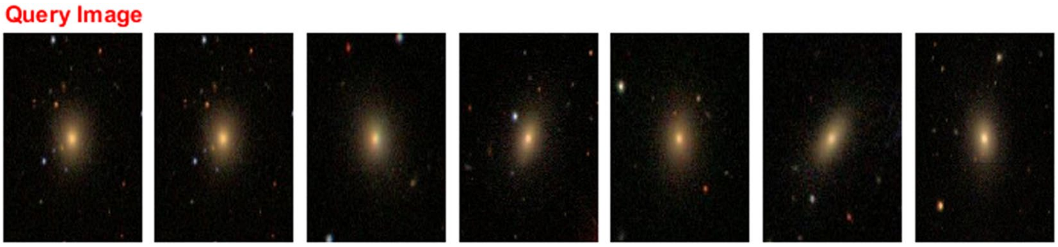

Figure 5. Galaxy image retrieval for an elliptical galaxy[29].

	Proposed approach		PSO		GA	
Dataset	**Recall**	**Precision**	**Recall**	**Precision**	**Recall**	**Precision**
Elliptical	92.68	97.43	90	85.36	82.60	97.44
Spiral Edge	97.50	100	100	100	97.50	100
Spiral	96.87	79.48	91.42	96.96	96.66	74.35
Irregular	90.69	100	92.85	90.69	97.50	100
Avg. Time (s)	292.0		508.1		495.0	

Table 3. A comparison between the proposed approach and the PSO and GA methods for galaxy image retrieval.

	50/50		70/30		85/10	
Dataset	**Recall**	**Precision**	**Recall**	**Precision**	**Recall**	**Precision**
Elliptical	72.67	77.50	86.67	86.33	91.87	94.58
Spiral Edge	80.33	79.65	89.15	88.77	95.93	98.95
Spiral	83.72	68.60	87.60	70.96	93.37	75.28
Irregular	60.61	60.00	82.07	75.76	85.17	94.68
NO. Features/Accuracy	20/81.85		18/88.30		15/92.02	

Table 4. The effect of the size of training set on the performance of the proposed approach for galaxy image retrieval.

Results and Discussion

In order to assess the effectiveness of our approach, we used the leave-one-out cross-validation method, where each image in the dataset was considered as the queried image, and the process was repeated 4458 times. Also, we used the 1-NN method based on 10-fold cross-validation (CV), which was used to evaluate the subset of selected features. This classifier is a parameter-free feature and is easy to implement[33]. As discussed previously, the 10-fold CV works by dividing the dataset into ten groups, and the experiment was performed ten times by selecting one group as the test set and the remaining groups were used as a training set during each run. The output is the average of accuracy of the ten runs.

In general, we used color, texture and shape feature vectors for galaxy image retrieval. The total number of extracted features was 30, where nine features were extracted from the three colors RGB (three moments for each color), 20 texture features (four rotations for each measure) and one shape feature. The extracted feature vectors were applied to the feature selection method (in this study, we compared the BSCA, PSO and GA methods) to determine the relevant features.

The best selected features with their accuracy (the value of fitness function) are given in Table 2. From this table it can be seen that, the BSCA algorithm selects a small number of features with high accuracy followed by the PSO, however, the GA selects a large number of features with low accuracy. In addition, we observed that the more relevant features thatcontain more information and are used to distinguish between the classes are the third color moment, energy, homogeneity, entropy and contour. These features are common between the three algorithms, and all of them are selected by the proposed method.

The comparison results of our proposed method with other methods are illustrated in Figs 2, 3, 4 and 5 and Table 3. From Table 3, we can conclude that the proposed approach is better than PSO and GA in terms of precision and recall measures. The best results were obtained when the spiral-edge type was used as the queried image because they present the most regular structure, while the less accuracy occurs when the spiral type galaxy was tested.

Moreover, from Table 3, it can be seen that the proposed method is faster than the other two algorithms, which takes ~292.0 s (nearly half the time of the other algorithms) to select the best features. We note that the GA method takes less time to complete than the PSO algorithm. In general, the computing time is divided into three parts: the first is the time needed to extract features from the images (~375 s, where each image takes ~0.084). The second part is the time needed to select the most relevant features as in Table 3. The last part is the time needed to compute the matching, which requires ~0.0157 s in addition to the time need to extract the features of the queried image (~0.084).

Figures 2, 3, 4 and 5, show an example of the retrieval images for four galaxy types. In these figures, the five database images that are the closest to the queried image are given as the retrieval results.

In order to investigate the influence of the size of the training set when selecting the best features, the dataset was randomly divided into training and testing sets. The proposed method was then evaluated at three different sizes, i.e. 50%, 70% and 85% of dataset (the remaining is the test set). Our results are shown in Table 4, where it can be seen that the worst accuracy was obtained when the sizes of the training and test sets were equal. The best accuracy was achieved when the training set was 85% of the entire database (as expected: by increasing the size of training set, the accuracy also increases).

Finally, from the previous results, we can conclude on two things: first is that the proposed approach for galaxy image retrieval is better than the PSO and GA algorithms in terms of recall, precision, accuracy and the time complexity. The second is that the most suitable method used to split the dataset (when selecting the best-fitting features) is the 10-fold CV, however, if the dataset is divided randomly then the most suitable size for the training set is in the range 85% to 90%.

Conclusions

In this study, we proposed a machine learning approach for galaxy image retrieval used for the automatic detection of galaxy morphological types from datasets of galaxies images. The automated detection of galaxies types is very important to understand the physical properties of the past, present, and future of the universe, while also offering a means for identifying and analyzing peculiar galaxies that cannot be associated with a defined morphological stage on the Hubble sequence.

Our analysis was performed such that our approach automatically detected specific morphology types from different morphological classes without human guidance. The proposed algorithm was compared with the PSO and GA algorithms, and its performance was evaluated based on recall and precision. The results indicate the superior performance of our proposed approach.

Based on the promising results of the algorithm, our future work will attempt to further investigate its application to other complex problems in astronomy by modifying the proposed method.

References

1. Sloan Digital Sky Survey: Galaxies - Advanced http://skyserver.sdss.org/dr1/en/proj/advanced/galaxies/ (2003).
2. Toomre, A. & Toomre, J. Model of the Encounter Between NGC 5194 and 5195. *Bulletin of the American Astronomical Society* **4**, 2–14 (1972).
3. Toth, G. & Ostriker, J. P. Galactic disks, infall, and the global value of Omega. *Astrophysical Journal* **389**(10) (1992).
4. Hackwell, J. A. & Schweizer, F. Infrared mapping and UBVRi photometry of the spiral galaxy NGC 1566. *Astrophysical Journal*, Part 1 **265**(15), 643–647 (1983).
5. Scoville, N. Z., Matthews, K., Carico, D. P. & Sanders, D. B. The stellar bar in NGC 1068. *Astrophysical Journal*, Part 2 - Letters to the Editor (ISSN 0004-637X) **327**(15), L61–L64 (1988).
6. Thronson, HarleyA. Jr., Bally, John & Hacking, Perry The components of mid- and far-infrared emission from S0 and early-type shell galaxies. *Astronomical Journal* **97**, 363–374 (1989).
7. Block, DavidL. & Wainscoat, RichardJ. Morphological differences between optical and infrared images of the spiral galaxy NGC309. *Nature* **353**, 48–50 (1991).
8. de Zeeuw, Tim & Franx, Marijn Structure and dynamics of elliptical galaxies. *Annu. Rev. Astron. Astrophys.* **29**, 239–274 (1991).
9. Dieleman, Sander, Willett, KyleW. & Dambre, Joni Rotation-invariant convolutional neural networks for galaxy morphology prediction. *MNRAS* **450**, 1441–1459 (2015).
10. Ferrari, F., de Carvalho, R. R. & Trevisan, M. Morfometryka A New Way of Establishing Morphological Classification of Galaxies. *The Astrophysical Journal* **814**(1) (2015).
11. Freeman, P. E. *et al.* New image statistics for detecting disturbed galaxy morphologies at high redshift. *Monthly Notices of the Royal Astronomical Society* **434**(1), 282–295 (2013).
12. Huertas-Company, M., Aguerri, J. A. L., Bernardi, M., Mei, S. & Almeida, J. Sanchez Revisiting the Hubble sequence in the SDSS DR7 spectroscopic sample: a publicly available Bayesian automated classification. *Astronomy& Astrophysics* **A157** (2011).
13. Lintott, C. *et al.* Galaxy Zoo 1: Data release of morphological classifications for nearly 900,000 galaxies. *Mon. Not. R. Astron. Soc.* **410**, 166–178 (2011).
14. Sidhu, S. & Saxena, J. Content based image retrieval a review. *International Journal Of Research In Computer Applications And Robotics* **3**(5), 84–88 (2015).
15. Liu, G.-H. & Yang, J.-Y. Content-based image retrieval using color difference histogram. *Pattern Recognition* **46**(1), 188–198 (2013).

16. Guo, J.-M., Prasetyo, H. & Wang, N.-J. Effective image retrieval system using dot-diffused block truncation coding features. *IEEE Transactions on Multimedia* **17**(9), 1576–1590 (2015).
17. Csillaghy, A., Hinterberger, H. & Benz, A. O. Content-Based Image Retrieval in Astronomy. *Information Retrieval* **3**(3), 229–241 (2000).
18. Ges, VitoDi & Valenti, Cesare Symmetry operators in computer vision. *Vistas in Astronomy* **40**(4), 461–468 (1996).
19. Ardizzone, Edoardo & Maccarone, VitoDi. GesMariaConcetta Suitability of a content-based retrieval method in astronomical image databases. *Vistas in Astronomy* **40**(3), 401–409 (1996).
20. Seyedali Mirjalili, S. C. A. A Sine Cosine Algorithm for Solving Optimization Problems. *Knowledge-Based Systems* (2016).
21. Alamdar, Fatemeh & Keyvanpour, MohammadReza A New Color Feature Extraction Method Based on QuadHistogram. *Procedia Environmental Sciences* **10**(A), 777–783, 9 (2011).
22. Gastaud, R. D., Popoff, F. S. & Starck, J. L. Astronomical Data Analysis Software and Systems XI. *ASP Conf Ser* **281** (2002).
23. Roy, K. & Mukherjee, J. Image similarity measure using color histogram, color coherence vector, and sobel method. *International Journal of Science and Research (IJSR)* **2**(1), 538–543 (2013).
24. Talib, A., Mahmuddin, M., Husni, H. & George, L. E. Efficient, compact, and dominant color correlogram descriptors for content-based image retrieval. In *Proceedings of the Fifth International Conferences on Advances in Multimedia*, *Venice*, *Italy*, 52–61 (2013).
25. Sim, D. G., Kim, H. K. & Park, R. H. Invariant texture retrieval using modified Zernike moments. *Imageand Vision Computing* **22**, 331–342 (2004).
26. Wang, X. Y., Yu, Y. J. & Yang, H. Y. An effective image retrieval scheme using color, texture and shape features. *Computer Standards and Interfaces* **33**, 59–68 (2011).
27. Han, J. & Ma, K. K. Rotation-invariant and scale-invariant Gabor featuresfor texture image retrieval. *Image and Vision Computing* **25**, 1474–1481 (2007).
28. Changa, Bae-Muu, Tsaib, Hung-Hsu & Choub, Wen-Ling Using visual features to design a content-based image retrieval method optimized by particle swarm optimization algorithm. *Engineering Applications of Artificial Intelligence* **26**(10), 2372–2382 (2013).
29. Baillard, A. *et al.* The EFIGI catalogue of 4458 nearby galaxies with detailed morphology. *A&A* **532** (2011).
30. HH, Azar, A. T. & Jothi, G. Supervised hybrid feature selection based on PSO and rough sets for medical diagnosis. *Comput Methods Programs Biomed.* **113**(1), 175–85 (2014).
31. Zhu, Zexuan, Ong, Yew-Soon & Dash, M. Wrapper-Filter Feature Selection Algorithm Using a Memetic Framework. *IEEE Transactions on Systems*, *Man*, *and Cybernetics*, Part B: Cybernetics. **37**(1), 70–76 (2007).
32. de Jong, J. T. A., Kuijken, K. H. & Hraudeau, P. Ground-based variability surveys towards Centaurus A:worthwhile or not? *A&A* **478**, 755–762 (2008).
33. Kudo, M. & Sklansky, J. Comparison of algorithms that select features for pattern classifiers. *Pattern Recog.* **33**(1), 25–41 (2000).
34. Qiu, G. Color image indexing using btc. *IEEE Trans. Image Process* **12**(1), 93–101 (2003).

Acknowledgements

This work was in part supported by National Key Research & Development Program of China (No. 2016YFD0101903). Also, Science and Technology Open Cooperation Program of Henan Province under Grant (#152106000048).

Author Contributions

Selim collected the data from the literature and prepared it; Mohamed Abd El Aziz developed the algorithm; and Shengwu Xiong wrote the manuscript.

Additional Information

Competing Interests: The authors declare that they have no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.