Relevance Feedback For Image Retrieval Using Transfer Learning and Improved MQHOA

Huaqiu Wang\textsuperscript{1,a} and Qian Liu\textsuperscript{1,*,b}

\textsuperscript{1}Liangjiang School of Artificial Intelligence, Chongqing University of Technology, 400045, China

E-mail: \textsuperscript{a}wanghuaqiu@cqut.edu.cn, \textsuperscript{*,b}LiuQian2020@2019.cqut.edu.cn

Abstract. Image retrieval is a challenging technology in multimedia applications where meeting the users’ subjective retrieval needs while achieving high retrieval performance is insufficient for existing methods. In this work, a related feedback image retrieval algorithm based on deep learning and optimization algorithm (CAMQHOA-RF) is proposed. Transfer learning based on the deep convolutional neural network is applied to extract deeper image features to reduce the semantic gap. The multi-scale quantum harmonic oscillator algorithm improved by the idea of “aggregation” is introduced to search the feature space effectively. The covariance matrix is used to strengthen the relationship between feature points at different scales to guide feature points to approach ideal query points faster. Moreover, the query point is reselected based on the feedback information to explore more potential users’ interest areas. Experiments have shown that compared with other algorithms, the proposed algorithm has fewer parameters that need to be set, but higher retrieval accuracy, faster retrieval speed, and stronger robustness are obtained, which can meet users better.

1. Introduction

With the development of computer and multimedia technology, the type and scale of information are increasing rapidly. As an important carrier of information, images are research objects of many fields. How to retrieve images that match the user’s query purpose from the massive picture library quickly and accurately has become a research hotspot. After the 1990s, content-based image retrieval technology has gradually developed, but traditional content-based image retrieval technology cannot avoid the ”semantic gap” problem. In response to this problem, many researchers have proposed various feature representation methods. Especially in recent years, deep learning technology has been widely used in CBIR tasks to extract high-level semantic features of images [1–3].

Relevance Feedback (RF) has been introduced into a large number of CBIR tasks because those technologies proposed are still unable to meet users’ purpose on retrieval. For this reason, an adaptive weight retrieval system was designed to verify the effectiveness of feature weight estimation [4], but it focuses on distinguishing good features from bad features so that good features have higher weights, which make the system more likely to fall into local optimums. Many researchers regard the CBIR task based on RF technology as a binary classification problem. To solve the imbalance and small sample problems, the expectation-maximization parameter was applied to the relevant feedback based on SVM classification [5]. Li proposed a semi-supervised active learning algorithm, which integrates unlabeled images into learning
to build a better classification model \[6\]. The kernel empirical orthogonal complementary component analysis based on the covariance matrix was applied to implement a support vector machine correlation feedback method based on feature reconstruction \[7\]. Tzelepi et al. retrained the fully connected layer by combining the convolutional neural network with users’ feedback \[8\]. However, those methods enumerated above cannot make full use of unknown feature regions in the feature space to further explore users’ needs.

In view of the above-mentioned, many researchers have applied optimization algorithms in combination with RF technology in CBIR tasks in recent years. \[9, 10\] separately employed particle swarm and genetic algorithm optimizers to track users’ retrieval preferences, but they are both easy to fall into local optimum. \[11\] integrated the Firefly algorithm with the SVM algorithm and RF technology to achieve better retrieval performance, which was Inefficient and complicated. In summary, the methods above have many problems, such as numerous and complex parameters, not satisfying the requirements of real-time retrieval, and being unable to utilize the potential feature space of users. Prompted by these approaches, this paper we improves the multi-scale quantum harmonic oscillator algorithm (MQHOA) and proposes a relevant feedback image retrieval method (CAMQHOA-RF) based on the multi-scale ant colony quantum harmonic oscillator algorithm, which models the movement of query points as an optimization problem and then searches for better feature points in the feature space through the optimization algorithm. Finally, the feature points are reselected depends on user feedback at each round to maximize the correlation between user feedback information and feature space points. Experiments have shown that in the process of user feedback, the CAMQHOA-RF algorithm which requires fewer parameters, is difficult to fall into a local optimal solution, can quickly converge, and increase the capacity to explore unknown related regions, and can be used to effectively search the feature space, especially when most images have been marked as relevant, the user’s interest area can still be explored, which improves the retrieval performance while reduces the number of user interactions, and make the search results more in line with users’ needs.

2. Relevant feedback image retrieval based on CAMQHOA algorithm

2.1. Feature description and distance model

In recent years, many researchers have successfully applied convolutional neural networks to image feature extraction, which can extract highly abstract deep features in images. Su et al. proved that the VGG16 model has a stronger transferable learning ability than other common models \[12\]. This paper uses the pre-trained VGG16 model to extract the 1024 dimensional image features from the fc7 layer. To reduce the complexity of retrieval without loss of image feature quality, principal component analysis (PCA) is introduced, which is realized through singular value decomposition (SVD) to achieve dimensionality reduction of high-dimensional features. Thus 128-dimensional high-quality image features are obtained. Considering that different feature components in the feature vector have different physical meanings, to ensure that the weight of each feature component is at the same low level when the similarity is matched, the features are standardized by z-score. The similarity model between the query image and the image set is based on the cosine distance, and the similarity matching method is (1):

$$\text{Dist}(Q, q_i) = \frac{Q \cdot q_i}{||Q|| ||q_i||}$$

(1)
where $F = \{q_1, q_2, ..., q_n\}$ is the feature set of the image set, $Q$ and $q_i$ are the feature vector of the query image and the image in image sets respectively.

### 2.2. Population initialization and fitness function determination

This paper models image retrieval as an optimization problem, which is different from the classic optimization problem as the object to be optimized is collected from each user feedback. If only the images marked as relevant by the user are considered, plus the number of relevant images displayed to the user is small, the retrieval result is likely to stagnate. Therefore, the relevant and irrelevant image information from the user feedback are taken into account, then the kth round of query vector is defined as:

$$Q_k = \alpha * Q_0 + \beta * relf_k - \gamma * irrf_k$$  

where $relf$ and $irrf$ are the centers of related image sets and irrelevant image sets, respectively, and they can be expressed as:

$$relf = \frac{1}{N_{REL}} \sum_{v_j \in X_{REL}} v_j$$

$$irrf = \frac{1}{N_{IRR}} \sum_{v_j \in X_{IRR}} v_j$$

$$v_j = \frac{q}{|q|}$$

The smaller fitness value indicates that the feature point is farther away from the area of the irrelevant image in the feature space, and closer to the area of the related image in the feature space. The fitness function is as follows:

$$F(Q_k) = \frac{1}{N_{REL}} \sum_{i=1}^{N_{REL}} Dist(Q_1, q_i) + \frac{1}{N_{IRR}} \sum_{j=1}^{N_{IRR}} Dist(Q_1, q_j)$$

where $q_i \in X_{REL}$, $q_j \in X_{IRR}$, $N_{REL}$ and $N_{IRR}$ are the size of $X_{REL}$ and $X_{IRR}$, respectively, and $Q_k$ is query the vector.

A random uniformly distributed initial population is generated in the feature space to ensure the ability to explore the entire feature space in the early stages. As the number of user interactions increases, the set of irrelevant images and related images will continue to expand, and the query vector is optimized by the optimization algorithm, which makes the retrieval result closer to the objective of the user’s query. Finally, the optimal feature points can be obtained.

### 2.3. Improved evolutionary stage

The sampling centers of the MQHOA algorithm do not affect each other, and the aggregation process in each area is independent. In this way, the sampling is only determined by the sampling currently selected and is unrelated to the previous sampling, which may cause the optimization process to fall into a local optimal solution and be difficult to search the feature space efficiently. To prevent the situation above, a sampling fitness record sequence for each sampling area to record all sampling conditions in that area is established, which is inspired by the ant colony pheromone model. Then the idea of "aggregation" is introduced in MQHOA, which can be defined as follows:

$$a = \frac{k}{\sum_{i=1}^{k} \text{FuncV}(i) - \text{AveFuncV}}$$
where $F$ is the adjustment factor, which can be defined as:

$$
F = \begin{cases} 
\max(FuncV(i) - AveFuncV) & \text{else} \\
1 \max(FuncV(i) - AveFuncV) & \leq 1 
\end{cases}
$$

(8)

where $FuncV(i)$ denotes the fitness value of the sampling point, and denotes the average fitness value of all sampling points in the sampling area.

When the sampling points are more scattered, the clustering degree is larger. At this time, the global search ability is stronger, and new sampling points are continuously generated iteratively. When the clustering degree shrinks and gradually converges. At this time, the local search ability is stronger. To expand the search area of the feature space and reduce the number of iterations, the CAMQHOA algorithm selects one of the top $s$ of the fitness of each sampling area as the optimal fitness value of the current area. The selection probability is as follows:

$$
P_i(j) = \frac{\text{Fitness}(AllFunc_{V_i}(j))}{\sum_{l=1}^{s} \text{Fitness}(AllFunc_{V_i}(l))}
$$

(9)

where $i = 1, 2, 3, ..., k$, and there are total of $k$ sampling areas. $AllFuncV$ is the fitness record sequence of all sampling points. $AllFunc_{V_i}(j) \in \{ AllFunc_{V_i}(1), AllFunc_{V_i}(2), AllFunc_{V_i}(3), ..., AllFunc_{V_i}(q) \}$, $q$ is the number of samples. $s$ changes dynamically according to the degree of aggregation, which is defined as:

$$
s = \begin{pmatrix}
1 & a > 1.05 \\
... & ...
4 & a < 0.95
\end{pmatrix}
$$

(10)

When the quantum harmonic oscillator converges, it reaches the energy ground state at the current scale. At this time, the sampling point is located near the global optimal solution. The sampling point of the covariance matrix is applied to strengthen the relationship between each sampling, and there is a clearer direction in the feature space for the next round of sampling points. The scale is dynamically changed according to the sampling in each region, which further improves the sampling accuracy, and the feature points can be more effectively approached to the optimal feature point. The scale of CAMQHOA changes according to the following:

$$
\sigma_{sq+1} = \frac{\|X_{i:q+1} - m_q\|}{\sqrt{C_q}}
$$

(11)

where $X_{i:q+1}$ is the adopted coordinates of $k$ regions from the present age are arranged according to their fitness value. $m_q$ and $C_q$ are the average vector of the fitness of $k$ sampling regions of the previous generation and the covariance matrix of the sampling points of the previous generation respectively, and can be directly calculated as:

$$
m_q = \sum_{i=1}^{u} w_i X_{i:k}^q
$$

(12)

$$
C_q+1 = (1 - \frac{u}{d^2})C_q + \frac{u}{d^2\sigma_{sq}} \sum_{i=1}^{u} w_i (X_{i:k}^q - m_q)(X_{i:k}^q - m_q)'
$$

(13)

where $u = \frac{k}{2}$, $w_i = \frac{1}{n}$, $C^0 = I$, $d$ denotes the function dimension.

Combined with the nature of image retrieval, the purpose of each optimization is to maximize a set of users’ potential points of interest. To be distinguished from traditional optimization
problem the feature point with the smallest fitness value may not be the closest to the ideal point. That means the discovery of the user’s interest feature area may be limited if the point with optimal fitness value is selected. After fitness evaluation of all generated feature points is completed, the first 10 feature point set $B$ with the best fitness value is re-elected. When the $N_{IRR}$ is large, the feature point farthest from the center of $X_{IRR}$ is selected. When the $N_{REL}$ is large, the feature point closest to the center of $X_{REL}$ is selected. When $N_{REL}$ increases further, the feature point with the best fitness value is selected. The query described point above can be defined as follows:

$$Q = \arg \min_{q_i} \{ \text{Dist}(q, irr f) \}, q \in B$$  \hspace{1cm} (14)

$$Q = \arg \min_{q_i} \{ \text{Dist}(q, rel f) \}, q \in B$$  \hspace{1cm} (15)

2.4. The processes of Image retrieval based on CAMQHOA-RF

The CAMQHOA-RF algorithm mainly includes two processes, which are optimizing query points by CAMQHOA and user feedback. The overall flow chart of the image retrieval system is shown in Figure 1.

![Figure 1. Relevance feedback image retrieval system based on CAMQHOA-RF](image)

3. Experiment and result analysis

In this section, experiments are designed to verify the performance of the method proposed in this paper for CBIR. The following algorithms are selected as comparison algorithms: MQHOA-RF refers to a relevance feedback image retrieval algorithm based on a multi-scale quantum harmonic oscillator algorithm to modify query feature points; PSO-RF refers to a relevance feedback image retrieval algorithm based on particle swarm optimization to modify query feature points; SVM-RF means that the user feedback information is used as the SVM training set to construct a classifier, so that image retrieval is regarded as a two-class classification problem; QPM-RF refers to the relevance feedback image retrieval algorithm that modifies the query feature points without the optimization algorithm.

Five images are randomly selected from each category in the UC Merced Land-Use dataset to form a 105 image query set, and three images are randomly selected from 100 categories in
the Caltech-256 dataset to form a 300 image query set. To ensure the fairness of the experiment, the population size of these optimizing algorithms is 30. For the same reason, the randomness of intelligent algorithms is considered, so the CAMQHOA-RF, MQHOA-RF, and the PSO-RF run 5 times, and the average is compared with the others. To evaluate the effectiveness of the algorithm, the measurements are as follows:

\[
\text{Precision}(i) = \frac{S(i)}{N} \quad (16)
\]

\[
\text{Recall}(i) = \frac{\sum_{i=1}^{S(i)} N_q}{N_q} \quad (17)
\]

where \(S(i)\) is the number of related images in the image displayed during the \(i^{th}\) round of feedback, \(N\) is the number of images displayed to the user, and \(N_q\) is the number of all similar images of this type of image in the image query set.

The average precision when \(N=24\) on the two query sets are shown in Figure 2. The precision of SVM-RF almost no longer to increase in the last few feedbacks because the small and uneven training samples have caused the classifier and features to be relatively stable in the early stages, and the highest fitness corresponding to the later unchanged optimal parameters remain. Figure 2(a) shows QV-RF has fixed parameters, so it is difficult to effectively retrieve the feature space on the Caltech-256 query set (a larger data set). After combining the optimization algorithm, the precision of the two data sets is significantly higher than QV-RF and SVM-RF. In the next 6 rounds of feedback, the accuracy of CAMQHOA-RF is much higher than PSO-RF and MQHOA-RF due to their lack of ability to jump out of local optimal.

![Figure 2](image.png)

**Figure 2.** The average precision of our algorithm compared with other algorithms on the image query set; (a) on Caltech-256 query set; (b) on UC Merced Land-Use query set

To further explore the ability of several algorithms to retrieve related images, especially when most of the existing samples are related, and the recall determines whether it can further explore the space. Figure 3 compares respectively when \(N=24\) and \(N=50\), the average precision at the \(3^{rd}\) round feedback and recall at the \(10^{th}\) round feedback of the CAMQHOA-RF, PSO-RF, and the MQHOA-RF in the seven categories which have high similarity to other categories in the UC
Merced Land-Use query set, like dense residential and medium residential have high similarity. A better retrieval accuracy is achieved by CA-MQHOA on these categories of images, and when \( N=50 \), the recall of the PSO is significantly lower than the others (Figure 3). This is owing to the idea of ”aggregation” of CA-MQHOA, which prevents the optimization from falling into the local optimal space. Moreover, a set of user potential interest feature points that are closer to the ideal point can be found by reselecting the optimal feature points. In summary, the CAMQHOA-RF has the best search ability in the feature space and can retrieve more relevant pictures.

![Figure 3](image)

**Figure 3.** The retrieval performance of our algorithm compared with other algorithms on different types of remote sensing images; (a)the average precision when \( N=24 \); (b)the average precision when \( N=50 \); (c)the average recall when \( N=24 \); (d)the average recall when \( N=50 \)

| Algorithm           | SVM-RF | PSO-RF | MQHOA-RF | CAMQHOA-RF |
|---------------------|--------|--------|----------|------------|
| UC Merced Land-Use  | 7.7    | 40.2   | 7.0      | 12.2       |
| Caltech-256         | 115.8  | 48.3   | 8.6      | 13.5       |

Table 1. The average retrieval speed of compared algorithms on UC Merced Land-Use and Caltech-256 (unit: s)

Considering the importance of image retrieval efficiency, Table 1 has shown the average execution time of the algorithms mentioned above on two query sets, which refers to the average time required by each round of feedback in the first 10 rounds. The SVM classifier converges very slowly on large data sets. The retrieval efficiency of the PSO-RF is significantly lower than the others while MQHOA-RF is significantly better than others. In the PSO-RF algorithm each particle needs to iterate continuously to update themselves, whereas, in the MQHOA-RF, only
the fitness assessment is necessary. The efficiency of CAMQHOA-RF is slightly below MQHOA-RF as the calculation of feature points’ clustering degree and reselection, but a better retrieval effect can be achieved through CAMQHOA-RF.

4. Conclusions and Discussion

In this paper, transfer learning is used to extract the deep features of images, and the improved MQHOA algorithm introduced by the idea of “aggregation” is applied to the image retrieval system based on modified query feature points, which can search the feature space effectively and can find the final query point whose semantic features are more in line with the purpose of the user’s queries. The experimental results on Caltech-256 (general data set) and UC Merced Land-Use (remote sensing data set) can prove that this algorithm can effectively improve retrieval performance and reduce the number of user feedback. The next step is to apply the algorithm proposed in this paper to large image sets in more fields. Additionally, we can continue to explore how to construct a better fitness function to evaluate feature points more effectively and further reduce the time complexity.

References

[1] A. Babenko, A. Slesarev, V. Lempitsky and A. Chigorin, "Neural codes for image retrieval," in Springer International Publishing, vol. 8689, pp. 584-599, 2014.
[2] A. Romero, C. Gatta and G. Camps-Valls, "Unsupervised deep feature extraction for remote sensing image classification," IEEE Transactions on Geoscience and Remote Sensing, vol. 54, pp. 1349-1362, 2016.
[3] L. Zheng, Y. Yang and Q. Tian, "SIFT meets CNN: A decade survey of instance retrieval," IEEE transactions on pattern analysis and machine intelligence, vol. 40, pp. 1224-1244, 2018.
[4] X. Lu, J. Wang, X. Li and M. Yang, "An adaptive weight method for image retrieval based multi-feature fusion. Entropy," Entropy, vol. 20, pp. 577, 2018.
[5] Y. Wang, W. Chen and Y. Yang, "A new integrated SVM classifiers for relevance feedback content-based image retrieval using EM, parameter estimation," Applied Soft Computing, vol. 11, pp. 2787-2804, 2011.
[6] G. Li, "Improving Relevance Feedback in Image Retrieval by Incorporating Unlabelled Images," Telkomnika Indonesian Journal of Electrical Engineering, vol. 11, pp. 3634-3640, 2013.
[7] Y. Wang, W. Li and Y. Yang, "An image retrieval scheme with relevance feedback using feature reconstruction and SVM reclassification," Neurocomputing, vol. 127, pp. 214-230, 2014.
[8] M. Tzelapi, A. Tefas, "Relevance Feedback in Deep Convolutional Neural Networks for Content Based Image Retrieval," Neurocomputing, vol. 127, pp. 214-230, 2014.
[9] M. Broilo, F. G. B. D. Natale, "A stochastic approach to image retrieval using relevance feedback and particle swarm optimization," IEEE Transactions on Multimedia, vol. 12, pp. 267-277, 2010.
[10] S. Moreno-Picot, M. Arevalillo-Herraez and F.J. Ferri, "Distance-based relevance feedback using a hybrid interactive genetic algorithm for image retrieval," Applied Soft Computing, vol. 151, pp. 1782-1791, 2011.
[11] T. Kanimozhi, K. Latha and F.J. Ferri, "An integrated approach to region based image retrieval using firefly algorithm and support vector machine," Neurocomputing, vol. 11, pp. 1099-1111, 2015.
[12] D. Su, H. Zhang and H. Chen, "Is Robustness the Cost of Accuracy? ," European Conference on Computer Vision, 2018.