MINISTRY OF EDUCATION AND TRAINING

VIETNAM ACADEMY OF SCIENCE AND TECHNOLOGY

GRADUATE UNIVERSITY OF SCIENCE AND TECHNOLOGY

AN HONG SON

CONTENT-BASED IMAGE RETRIEVAL WITH REPRESENTATIONS LEARNING AND DATA DIMENSIONALITY REDUCTION

Major: Computer Science

Major code: 9 48 01 01

SUMMARY OF COMPUTER SCIENCE DOCTORAL THESIS

Ha Noi - 2023

The thesis has been completed at: Graduate University of Science and Technology - Vietnam Academy of Science and Technology

Scientific Instructor: Assoc.Prof. Phd. Nguyen Huu Quynh

Reviewer 1: ……..

Reviewer 2: ……..

Reviewer 3: ……..

The thesis shall be defended in front of the Academy level doctoral thesis grading committee, meeting at the Graduate University of Sciences and Technology - Viet Nam Academy of Sciences and Technology on ..... hour, date ..... month …. year 2023.

This thesis could be found at:

- The National Library of Vietnam
- The Library of Graduate University of Science and Technology

INTRODUCTION

1. The necessity of the thesis

In recent years, with the rapid increase of social networks along with the strong development of 4.0 technology and smart mobile devices, multimedia applications have generated a massive digital image database. Digital images play an important role in many different fields of life such as remote sensing, fashion, medicine, education, architecture, crime prevention, ... Therefore, the fast and accurate retrieval of an image in a large and diverse digital image database is a challenge and an urgent task in the current field of computer vision.

In the field of computer vision, Content-Based Image Retrieval (CBIR) is currently one of the actively researched directions. The goal of CBIR is to search for images based on the analysis of the visual content of the query image [3]. However, this method faces the challenge of the "semantic gap" between low-level image features and high-level concepts that humans perceive [4], which can lead to irrelevant images being returned. To overcome this, various methods have been proposed to bridge the semantic gap by transforming high-level concepts in images into low-level features. These features are categorized into global features (including color, shape, texture, and spatial information) and local features depending on the feature extraction method [4]. The representation of these features is the foundation for CBIR.

Machine learning is an important tool for mining data structures, obtaining better data representations, and uncovering hidden data patterns so that relevant information can be extracted. In machine learning, there are three main approaches, including:

supervised learning, unsupervised learning, and semi-supervised learning. The difference between these approaches is the use of labeled samples during the learning process.

In recent years in Viet Nam, there have been many graduate students, research teams have effectively applied machine learning techniques for the CBIR with relevance feedback, to narrowing the "semantic gap" and improving the retrieval accuracy of image retrieval systems. However, these studies have not focused on addressing the issue of small class size and have not exploited the row-sparse attribute of the projection matrix. In addition, the superiority of deep learning techniques for image retrieval on large, unlabeled and high-dimensional data sets has also not been exploited. This is a research orientation in accordance with the common research trend of the world, highly urgent and effective applicable in practice and this is also the research direction that graduate student are pursuing. Therefore, graduate student have chosen the topic "***Content - Based image retrieval with representations learning and data dimensionality reduction***" as their thesis topic.

2. Research objectives of the thesis

The research, proposes several methods to improve the accuracy and retrieval time for with problems have small class size, small sample size, and high-dimensional data by incorporating machine learning techniques into the CBIR with relevance feedback.

3. The main research contents of the thesis

The thesis focuses on researching and exploring the following main contents: (1) CBIR and image features representation; (2) Semantic gap in CBIR; (3) Relevance feedback, techniques, and challenges in relevance feedback; (4) Machine learning, deep learning, Autoencoder networks; (5) Experimental environment, experimental image dataset, and performance evaluation methods.

CHAPTER 1. OVERVIEW OF CONTENT-BASED IMAGE RETRIEVAL WITH RELEVANT FEEDBACK

1.1. Content-Based Image Retrieval

Content-based image retrieval is an application of computer vision techniques to image retrieval problems [12].

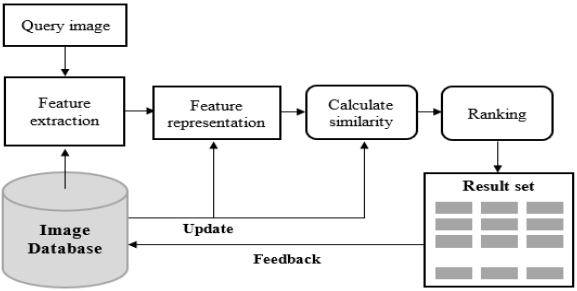

Figure 1.1. Model of the CBIR system

The goal of the CBIR is to use the visual content of an image to find images relate from a large image database (the content here is understood as color, shape, texture or any other information can be extracted from the image itself).

1.2. Low level features

Image features can be divided into global features and local features. Global features, including: color features, texture features, shape features and spatial information, in which color features are considered as one of the most important features in image retrieval. The local features include: Scale Invariant Feature Transform (SIFT), Strong and Fast Features (SURF), Local Binary Pattern (LBP).

1.3. Feature select

Feature selection is the process of selecting the most relevant subset of features that most efficiently representasion data object. These features are selected from the original data features and sorted in descending order of importance. Several approaches have been

proposed in recent years such as: Fisher weight [33], Relief [34], Relief-F [35], Mutual information [36], Hilbert Schmidt (HSIC) [37], Laplace [38]. In which, Fisher weighting technique, Relief and Relief-F algorithm are commonly used.

1.4. Feature extraction

Feature extraction is an important method for generating new features based on some combination or transformation of the original features. Feature extraction methods also help to obtain more discriminant data representations. Feature extraction is done by projecting the original data into the embedding spaces. Typical methods include: Linear Discriminant Analysis (LDA) [44], Robust Sparse Linear Discriminant Analysis (RSLDA) [41], and Feature Extraction using Gradient Descent (FE_GD) [43], Principal Component Analysis (PCA) [45].

1.5. Machine learning for CBIR

Machine learning techniques commonly used in CBIR include: (1) Unsupervised learning (including: Clustering K-means and K-means++ [48]); (2) Supervised Learning (including: Support Vector Machine SVM [51] and Artificial Neural Network [55]); (3) Deep learning (including: Autoencoder and ResNet Network [68]); (4) Associative learning [69].

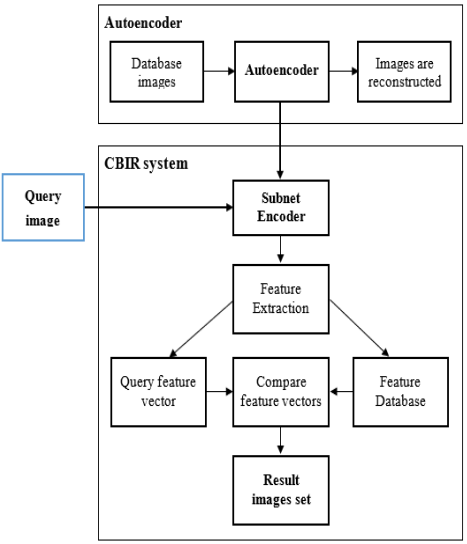

Figure 1.3. Integrate autoencoder with CBIR model

1.6. Relevance Feedback

Relevance Feedback (RF) is a powerful tool commonly used in CBIR systems [76]. It was introduced in the early 1990, with the aim of involving users in the image retrieval process to reduce the semantic gap between what is described by queries (low-level features) and what the user thinks. By continuously learning through user interaction, RF has significantly improved the performance of CBIR systems [77].

1.7. Measure similarity between images

Measure similarity to determines which image is the most relevant to the query image. Therefore, measuring similarity directly affects the accuracy and computational complexity of the CBIR system. Some measures are widely used in CBIR such as: Minkowski distance; Manhattan distance; Chessboard distance; Hamming distance; intersection schema distance; Mahalanobis distance; Canberra distance; cosine distance; Chi-square statistics; Squared Chord. Choosing the right similarity measure is a difficult task, and many research have done this experimentally.

1.8. Some research on CBIR

1.8.1. International research

In 2016, Ponomarev et al., in [90] presented a CBIR system based on the integration of color, texture and shape. The main limitation of the system is the increased computational complexity due to the integration of many features. In 2017, Srivastava & Khare in [91] developed a new multi-resolution analysis algorithm that analyzes images at multiple levels, with other levels capturing information that one level missed. This approach is based on extracting texture and shape features using a local binary pattern descriptor (LBP). A new CBIR approach is presented by combining color, shape and texture features proposed by Z.Zhao et al., in [99]. Although the proposed system

obtains high accuracy, the performance of the system is affected when the query image contains many complex objects.

In 2018, Sajjad et al., in [92] proposed a CBIR system that is invariant to rotation and color change. The proposed system is based on combining color and texture features to form a common feature vector. To reduce the semantic gap, Ashraf et al., in [94] proposed a CBIR system that combines color and edge features to form a feature descriptor. However, it still suffers from a lack of spatial information and no computational cost-effectiveness information. Phadikar et al, in [100] proposed a CBIR system in the discrete cosine domain. Although the use of a genetic algorithm has a positive effect on the accuracy of the system, it increases the time it is used.

In 2019, Pavithra & Sharmila in [93] proposed a new method to select seed points for dominant color-based image retrieval. However, the proposed method needs to be merged with other feature extraction methods (shape, texture and spatial information) to reduce the semantic gap, since the same color information can be assigned for images in different semantic classes. A new CBIR system was presented by Bani & Ershad in [98], based on the extraction of local and global texture features in both the frequency and spatial domains as well as the color features in the spatial domain. The proposed system shows values with high accuracy and is compared with other modern methods. In addition, it is reported to be rotation-invariant and less sensitive to noise, but it has a high runtime due to the use of different features.

In 2020, Ashraf et al., in [96] developed a methodology for CBIR systems based on combining low-level features (texture and color). However, the proposed model lacks structural and spatial information, like many other researchs; Alsmadi et al., in [97] introduced a new content-based image retrieval technique that takes advantage of color, shape, and texture. The proposed technique has applied genetic algorithm, thus improving the quality of the solution.

However, it suffers from the importance of the process and needs to be repeated many times, which slows down the computation time.

1.8.2. Viet Nam research

In Vietnam, in recent years, there have been many research works and doctoral theses related to CBIR problem published, especially research works by the research team of Assoc.Prof.Phd. Nguyen Huu Quynh, Assoc.Prof.Phd. Ngo Quoc Tao, PhD student and associates published in doctoral theses:

- In 2017, Vu Van Hieu successfully defended his doctoral thesis "Research on some classification techniques in content-based image retrieval" [101]. The limitation is that the accuracy of the result set in the thesis is still low because the thesis's approach is to consider a single region containing related points, ignoring the fact that the images are scattered throughout the entire feature space. The point of note here is that although the thesis collects training samples through the related feedback mechanism, the thesis's approach is not in the direction of learning the projection matrix.

- In 2019, Dao Thi Thuy Quynh successfully defended her doctoral thesis "Improving the accuracy of content-based image retrieval using the distance function weight adjustment technique" [102]. The limitation is that the method does not consider the heterogeneity of the feature space and does not solve the problem of approximate access on non-metric spaces. Although the thesis collects training samples via RF, the approach of the thesis is a projection matrix based on taking advantage of the locality of each feature point region.

- Most recently, in 2022, PhD student Cu Viet Dung carried out his doctoral thesis "Improving the precision of content-based image retrieval through on a manifold learning approach from user feedback" [103]. Although the approach of the thesis is to learn the projection matrix with the training samples obtained from the related feedback mechanism, the image retrieval is performed on the projection space.

In general, these works have effectively approached and exploited machine learning chniques for CBIR and experimented on popular and professional image data sets. However, these works have not exploited the sparse property of the projection matrix and learned image representation by deep learning approach. This is a practical and highly feasible research direction that the PhD student aims at in of this thesis.

1.9. Experimental and evaluate performance

1.9.1. Experimental image database

The experimental data used in this thesis are professional image databases, which have been widely used to evaluate the performance of the CBIR system [104], including COREL (Figure 1.7) and CIFAR -100 (Figure 1.8) image database.

Figure 1.7. Some representative images in the COREL image dataset

Figure 1.8. Some representative images in the CIFAR-100 image dataset

1.9.2. Methods of performance evaluation

In this thesis, the measure used to evaluate the performance of the proposed methods are: *AP* and *mAP*.

$$\mathrm{AP} = \frac{\sum_{k=1}^{N} P(k).rel(k)}{R} \quad \textbf{(1.25)}$$

$$mAP = \frac{1}{Q}\sum_{q=1}^{Q} AP(q) \quad \textbf{(1.26)}$$

1.10. Conclusion of chapter 1

In this chapter, the thesis has systematized the basic theoretical knowledge and research related to CBIR, and analyzed the research related to the stages in CBIR to see the advantages and limitations of current researchs, as a basis to confirm the feasibility of the research direction and determine the contents to be solved in the next chapters of the thesis.

CHAPTER 2. IMAGE RETRIEVAL METHOD WITH SPARSE DISCRIMINANT ANALYSIS

2.1. Introduction

The image retrieval with related feedback uses a classification approach that only includes two classes is negative and positive, so it has several problems: (1) The number of samples is often too small compared to the dimension of feature space [115], (2) The number of negative samples is often much more than the number of positive samples [115], and (3) The number of classes is too small, resulting in number projection directions is small, because the number projection directions is closely related to the number of classes. To solve these three problems, the thesis proposes a new supervised image retrieval method, combining an important feature extraction model based on the RSLDA method with a classification model in the CBIR system in order to improve accuracy and query time. The method is name **SDAIR** (Sparse Discriminant Analysis for Image Retrieval).

SDAIR has the following characteristics: (1) The model is very flexible, can be applied to any image similarity measure, any feature selection learning model, and any classification learning model; (2) Not affected by the small class size problem, while it still eliminates redundant and irrelevant features, and takes advantage of discriminant information; (3) The number of positive samples is not required to be large enough because it can provide a mechanism to automatically add positive samples to the training sample set (no need to re-train the projection learning model); (4) Simultaneous support for these two tasks is the selection of the important feature set and the addition of a positive training sample.

2.2. Proposed method

2.2.1. Model of the method

The image retrieval model is proposed in Figure 2.1. The retrieval process begins with feature extraction of the query image. Use these feature vectors together with a predefined similarity

measure to measure the similarity between the query image and database images. Then, a set of images relate to the query image is selected and sorted in descending order of similarity to obtain the retrieval result set.

User feedback on the retrieval result set to get the feedback set, which is also the training set. On the basis of this training set, the projection learning algorithm is performed to obtain the projection learning model A. Use the projection learning model A on the feedback set and arrange the features in descending order of important to obtain an important set of features. To solve the problem of small sample size and numbe positive samples less than negative samples in the CBIR with feedback, the model automatically adds positive samples by applying projection A has been learned on the feature set to obtain the incremental set. Collect important features on both feedback and incremental sets to generate a training set for classifier learning, thus obtaining a classifier learning model C. The ranking of images will be performed according to the classifier learning model C to get the retrieval result set. This process will be repeated if the user is not satisfied with the retrieval result, otherwise the final result set.

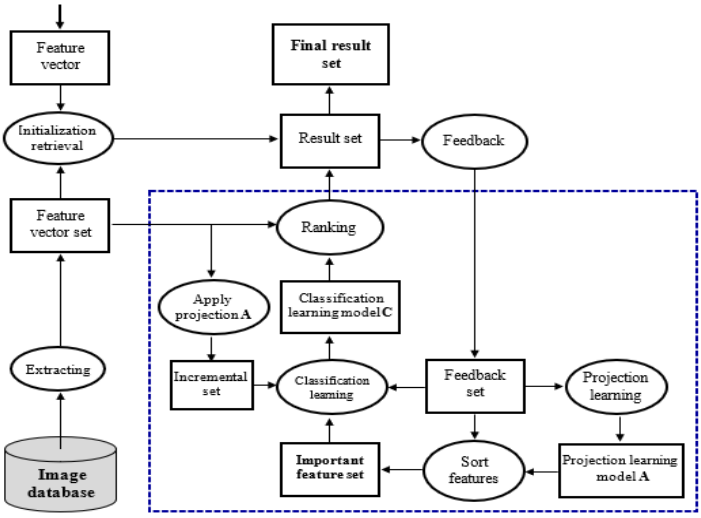

Figure 2.1. Model of the proposed image retrieval method

2.2.2. Projection learning model for selecting an important feature set

Robust Sparse Linear Discriminant Analysis (RSLDA) [41] is a feature extraction method based on LDA. It minimizes the $\ell_{2,1}$ norm of the linear projection matrix Q. RSLDA can recover original data from low dimensional projected data

To extract features while preserving the main energy of the data, RSLDA solves the following optimization problem:

$$\min_{P,Q,E} Tr(Q^T(S_w - \lambda S_b)Q) + \lambda_1\|Q\|_{2,1} + \lambda_2\|E\|_1 \qquad \textbf{(2.6)}$$
$$s.t.\ X = PQ^TX + E,\ P^TP = I$$

Taking the motivation to overcome the limitation of LDA, and inheriting the advantages of the RSLDA method, thesis propose an improved learning model by adding a term to fit the class label (samples with the same label in projection space will be closer together while samples with different labels will be further apart) to increase the classification property of the resulting projection matrix. Minimizing the objective function (2.7) below.

$$\min_{P,A,E} Tr(A^T(S_w - \lambda S_b)A) + \lambda_1\|A\|_{2,1} + \lambda_2\|E\|_1 + \frac{1}{2}\|Y - AX\|_F^2 \quad \textbf{(2.7)}$$
$$s.t.\ X = PA^TX + E,\ P^TP = I$$

Algorithm 2.1: Select the set of important features

Input:	- Training sample matrix X, label matrix Y - Parameters λ_1, λ_2, number of important features k
Output:	- Projection Matrix A - Important features matrix X_k
Step 1:	Calculate S_b according to formula (2.2); Calculate S_w according to formula (2.3)
Step 2:	Solve the optimization problem (2.7) according to [132] get theo projection matrix A
Step 3:	Calculate $\|a_i\|_2, i = 1,2,\dots,m$ of the projection matrix A
Step 4:	Sort m rows of X in descending order of $\|a_i\|_2$. Construct X_k consisting of k rows on the top of X.
Step 5:	Return A and X_k

2.2.3. Learning model for classification

This section inherits the solution of the small sample size problem in Algorithm 2.1 and focuses on solving the classification phase of the image retrieval problem with relevant feedback.

To solve the above small class size problem, the thesis proposes a learning model classification but it is performed on the original feature space. When performing classification on the original feature space, face the problem of high dimensionality of the feature space, therefore propose to remove redundant features (see Algorithm 2.1).

The classification algorithm is summarized in Algorithm 2.2:

Algorithm 2.2: Building a classification model

Input: - Training sample matrix X, label matrix L
- Projection matrix A ;
- Important feature matrix X_k
- Set of feature vectors F

Output: Classification learning model R

Step 1: Apply the projection learning model A to the feature vector set F.

Step 2: Construct an incremental matrix $X^{(e)}$ consisting of e points x_i corresponding to e points y_i which are neighbors of $y_i^{(q)}$. Construct a label matrix $L^{(e)}$ consisting of e positive labels of $x_i \in X^{(e)}$.

Step 3: Merge the matrix $X^{(e)}$ into the matrix X according to the principle that the first column of $X^{(e)}$ is placed to the right of the last column of X. Similar to merging the matrix $L^{(e)}$ into L.

Step 4: Applying the classification learning model to X and L.

Step 5: Return R.

2.2.4. Proposed image retrieval algorithm

The proposed algorithm calls Algorithm 2.1 in Step 2 to reduce the number of dimensions and obtain the important feature set. This step helps to solve the highdimensional data problem and helps

to solve the small class size problem (in Algorithm 2.2) of the image retrieval with relevance feedback, which uses the classification technique. Step 3 solves the problem of small class size, small sample size, and unbalanced sample set by calling Algorithm 2.2.

The proposed algorithm is summarized in Algorithm 2.3:

Algorithm 2.3: SDAIR

Input: **F**: feature set of database image, **q**: query image vector, ***N***: Number of images returned at each iteration.

Output: **S**: Result set.

Step 1: Query image q to get the initial result set. On this set, construct the result set *I* by taking the top *N* image vectors.

Step 2: **Repeat**

Step 2.1: User responds on set *I* to obtain the feedback set RF

Step 2.2: Implement Algorithm 2.1 to get the important feature set X_k

Step 2.3: Implement Algorithm 2.2 to get the classification learnin model C

Step 2.4: Ranking the feature set F according to the classificatic learning model C to get the list of results.

Step 2.5: Take the top N image of the list in Step 2.4 as the resulting image set S.

Until (User stops responding)

Step 3: Return S.

2.3. Experimental Results

The first experiment is to compare the proposed method with typical image retrieval methods, to show that the proposed method has a higher overall precision than the remaining methods. The second experiment is to test the effect of removing redundant and irrelevant features, while solving the small class size problem on the CIFAR-100 database. The mAP measure (in 1.9.3) is also used to evaluate the precision of the proposed method.

The DLRPIR and RDA_FSIS method [42] are used to compare with the proposed method because it uses the same similarity precision and feedback mechanism as the proposed method, and it uses discriminant low-class projection to project the original data into a projection space, and then perform classification on this projection space to rating the images.

2.3.1. Experiment on the overall performance of the proposed method

Figure 2.8 shows the average precision of the three methods at the top 100 images for the first three iterations. With these results, it is shown that the precision of the RDA_FSIS method is higher than that of DLRPIR because it learns a sparse discriminant projection matrix according to the structure of each class and reduces the small class size problem. The precision of the proposed method is the highest among the three methods because it eliminates redundant and irrelevant features. Besides, it also effectively solves the small class size problem.

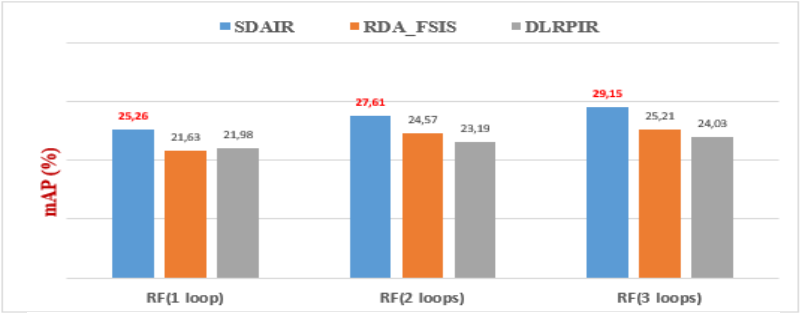

Figure 2.8. mAP of three methods on top 100

2.3.2. Experiment on efficiency when removing redundant features and solving small-class size problem

The thesis designs three experimental scenarios as follows:

Scenario (**1**): Compare the retrieval efficiency without using feedback (using only Euclidean) on a space of 1305 dimensions and original space but eliminating redundant and unimportant dimensions.

Scenario (**2**): Compare the retrieval efficiency without using feedback (using only Euclidean) on the original space (but removing redundant and unimportant dimensions) and on the projection space.

Scenario (**3**): Compare the retrieval efficiency using feedback on spaces including: (1) original original space (with 1305 dimensions); (2) original space (but remove redundant and unimportant dimensions); and (3) projection space. In this scenario, the SVM model is used to rating the images and obtain a retrieval result set.

The number of dimensions that the thesis experiments in all three scenarios above includes: 30 original dimensions (type 1275), 20 original dimensions (type 1285), and 10 original dimensions (type 1295). Tables 2.2, 2.3 and 2.4 are the results for respectively scenarios (1), (2), and (3).

Table 2.2. Image retrieval results based on scenario (1)

Method	**OIR i**				
Dimensions	1305	128	30	20	10
mAP(%)	16,07	**18,27**	16,63	16,15	15,6

Table 2.3. Image retrieval results based on scenario (2)

Method	**OIR i**				**PIR i**			
Dimensions	128	30	20	10	128	30	20	10
mAP(%)	**18,27**	**16,63**	**16,15**	**15,6**	17,21	15,68	15,3	15,05

Table 2.4. Image retrieval results based on scenario (3)

Method	**OIRRF i**					**PIRRF i**			
Dimensions	1305	128	30	20	10	128	30	20	10
mAP(%)	20,3	**25,26**	20,56	19,16	18,63	20,9	19,76	18,96	18,63

Looking at Table 2.2, we see that, the precision when choosing 128 dimensions is the highest among the dimensions including 128, 30, 20, and 10. This is evidence to confirm the effectiveness when removing redundant and irrelevant features of the proposed method.

Table 2.3, the precision of the proposed method on the original space is higher than the precision on the projection space in all dimensions including 128, 30, 20, and 10. The reason for this is because on the original space, it is possible to determine which features are most important to keep, while on the projection space, it is not known which features are important to keep, leading to features that are less important can to keep, and removing important features.

The data in Table 2.4 show that, in dimensions 128, 30, 20, and 10, the precision of the proposed method on the original space is always higher than on the projection space. The reason for this is that in addition to eliminating redundant and irrelevant features, it also reduces the impact of the small class size problem.

Table 2.5 below shows the query time of image retrieval method on original space and projection space.

Table 2.5. Image query time by dimensionality on Original space and Projection space

Method	Runtime of OIR i					Runtime of PIR i			
Dimensions	1305	128	30	20	10	128	30	20	10
Time (s)	0.5531	**0.35**	0.20	0.19	0.18	0.44	0.49	0.42	0.34

2.4. Conclusion of chapter 2

In this chapter, the thesis has proposed a flexible model, by automatically adding positive samples to the training set, which does not require the number of positive samples to be large enough. In addition, it can simultaneously serve two tasks: selecting important feature sets and addition positive training sample. Experimental results have shown that the proposed method can improve the performance of the image retrieval problem with related feedback, where the sample size is small, the class size is small, and high-dimensional data.

The main contributions of this chapter have been published in the works [**CT4, CT2**].

CHAPTER 3. LEARNING IMAGE REPRESENTATIONS WITH DEEP CONVOLUTIONAL NEURAL NETWORK AUTOENCODER FOR IMAGE RETRIEVAL WITH RELEVANCE FEEDBACK

3.1. Introduction

The performance of any CBIR method depends mainly on the descriptive representation of the image and is also expectation to be discriminatory, strong, and low-dimensional. Manually designed feature for image retrieval is an area of very active research, however its performance is limited because manual design cannot represent image features in an accurate way. 135].

To solve the limitations mentioned above, the thesis proposes a semi-supervised method the name is AIR, based on three components (autoencoder convolutional neural network, image feature extraction and SVM classification in related feedback). The AIR method overcomes two problems: (1) the ability to distinguish the poor features of the previous methods due to the integrated RF mechanism and classifier via the SVM support vector machine, and (2) mitigate the problem vanishing/exploding gradients and computational complexity through the use of shortcut connections in the autoencoder architecture and resulting in the possible use of deep autoencoder.

3.2. Proposed method

The proposed method include of three components. The first component is the unsupervised training of a deep autoencoder neural network on a subset of the image set. The second component is to apply the learning model from the first component to extract low-dimensional features from the database image set (both the first and second components are taken offline). The third component is to retrieval images that are similar to the query image based on the related feedback. The autoencoder model is trained on a subset of the CIFAR-100 image database.

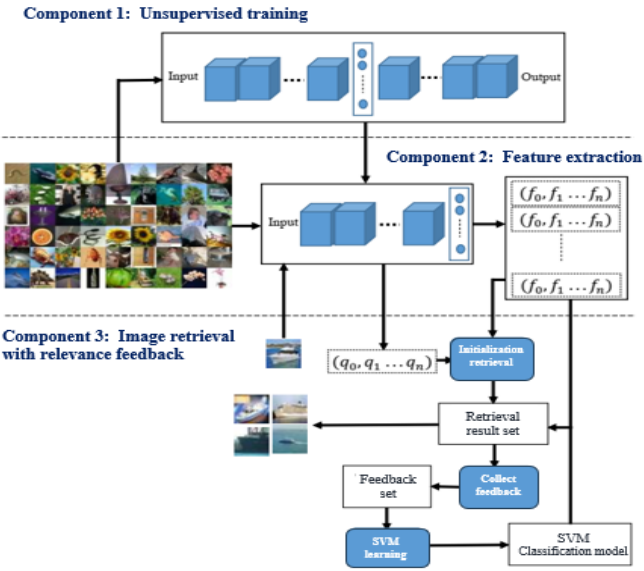

Figure 3. 1. Model of the proposed image retrieval method

3.2.1. Learning image representations with autoencoder

3.2.1.1. Convolutional neural network autoencoder

First, the input image is encoded so that each time a patch of $d \times d$ pixel $p_i, i = 1,2, \dots k$, is selected from the input image, and then the weight w_j of the convolutional j used for convolution calculations. Finally, the neuron value $a_{ij}, j = 1,2, \dots, m$ is calculated from the output layer.

$$a_{ij} = f(p_i) = \sigma(w_j . p_i + b) \qquad \textbf{(3.1)}$$

$$RElu(p) = \begin{cases} p \text{ nếu } p \geq 0 \\ 0 \text{ nếu } p < 0 \end{cases} \qquad \textbf{(3.2)}$$

Then the o_{ij} output from the convolution decoder is encoded that p_i is reconstructed through a_{ij} to produce $\hat{p}_i$.

$$\hat{p}_i = f'(a_{ij}) = \emptyset(w_i . a_{ij} + \hat{b}) \qquad \textbf{(3.3)}$$

p*i*is generated after each convolution encoding and decoding. We get the patch P obtained from the reconstruction operator. We use the mean square error between the original patch of the input image pi, i=1,2,…k and the reconstructed patch of the image pi, i=1,2,..,k.

The cost function is described in equation (3.4), and the reconstruction error is described in equation (3.5)

$$L(\theta) = \frac{1}{k}\sum_{i=1}^{k} E(p_i, \hat{p}_i) \quad \textbf{(3.4)}$$

$$E(p_i, \hat{p}_i) = \|p_i - \hat{p}_i\|^2 = \|p_i - \emptyset(\sigma(p_i))\|^2 \quad \textbf{(3.5)}$$

3.2.1.2. Pooling layer

Similar to in CNN, the convolution layer is connected to the pooling layer [92]. In the convolutional neural network architecture autoencoder, the max pooling layer is placed after the convolution layer:

$$a_j^i = \max(p_j^i) \quad \textbf{(3.6)}$$

In equation (3.6), p_j^i represents the i region of the j feature map, and a_j^i represents the i neuron of the j feature map.

3.2.1.3. Convolutional network architecture autoencoder

The deep neural networks suffer from vanishing/exploding gradients problems and computational complexity. Because autoencoders have many convolutional and deconvolutional layers, there is information loss and performance degradation when reconstructing images. Inspired by from ResNet networks, which include shortcut connections [75], we additional shortcut connections into the autoencoder network as shown in Figure 3.2. These connections make it possible to directly send feature maps from the first layer of the encoder to several later layers.

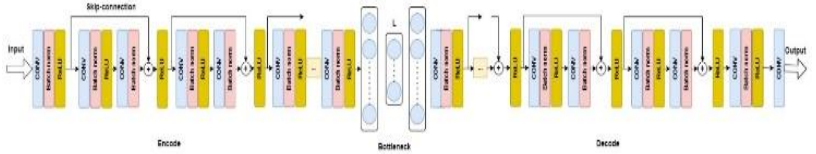

Figure 3.2. Proposed autoencoder network architecture for feature extraction

3.2.2. Retrieval images with relevance feedback using SVM

3.2.2.1. Support vector machine (SVM)

In this section, the thesis chooses Support Vector Machine (SVM) [39] for image classification and ranking because: Firstly, it is a powerful classifier, especially for binary classification, which is the image retrieval with related feedback is a two-class problem.

Secondly, through the optimization of the hyperplane, the distance from each sample to the optimized hyperplane can be used as a value for ranking the images.

3.2.2.2. Image retrieval

As the method model in Figure 3.1, after training the autoencoder convolutional neural network model in Component 1, we proceed to remove the decoder part and keep the encoder part to have the learning model as in Component 2. Using the learning model in Component 2 of the model to extract low-dimensional feature vectors to obtain a set of n feature vectors $(f_0, f_1 \ ... \ f_n)$.

During the image retrieval as in Component 3 of the model, the user provides a query image q, the vector of the query image will be passed through the encoder learning model to get the feature vector of the query image (q_0, q_1,..... q_n). The initial retrieval process compares (using Euclidean) the query image's vector with the database image's vector to obtain the retrieval result set. On this result set, users feedback to obtain a feedback set (this response set includes samples with negative and positive labels, it is also a training set). SVM learning is applied on the training set to obtain the SVM classification model. Applying the classification model on the feature vector set of the image database: the predicted positively labeled images that have the longest distance from the optimal hyperplane) will be ranked at number one of the results list, the positively labeled images that are the second furthest from the optimal hyperplane will be ranked at the number two position of the result list, ... this process repeats until the user gives feedback until the user responds.

3.3. Experimental assessment

3.3.1. The results on the image dataset CIFAR-100

Figure 3.9 shows that the optimal number of layers of the autoencoder network architecture for image retrieval on the CIFAR-100 set is 40 layers and the network configuration using the pooling layer is effective for the deeper the network architecture. Out of the 5

configurations, two in the proposed network architecture give the best results across the entire 20, 40, and 60 layers. This demonstrates that using asymmetric shortcut connections to the autoencoder to generate autoencoder deep networks is efficient on the CIFAR-100 set.

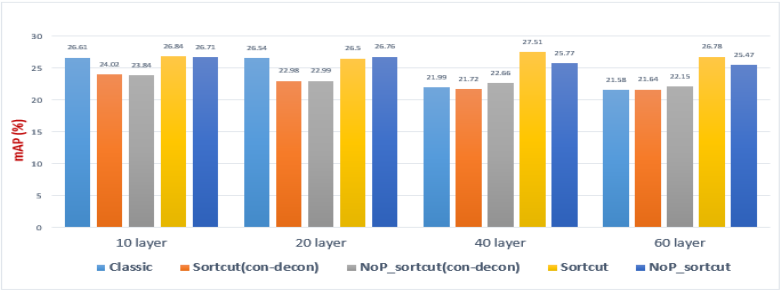

Figure 3.9. Image retrieval results at different depths of autoencoder network on the CIFAR-100 dataset

Figure 3.10 shows the mAP of the four methods Baseline (Non-RF), AIR, EDSSCIR, and SSCAIR for the first three response iterations. In which, Baseline method gives the lowest precision. The reason for this is that the Baseline method has no learning mechanism, it only calculates the Euclidean distance between the feature vector of the query image and the database image. The AIR method performed better than the other two on all loops. The performance of AIR is significantly better than Baseline, which indicates that the relevant feedback provided by the user is very helpful in improving retrieval performance. AIR performs better than EDSSCIR because AIR obtains a good feature representation.

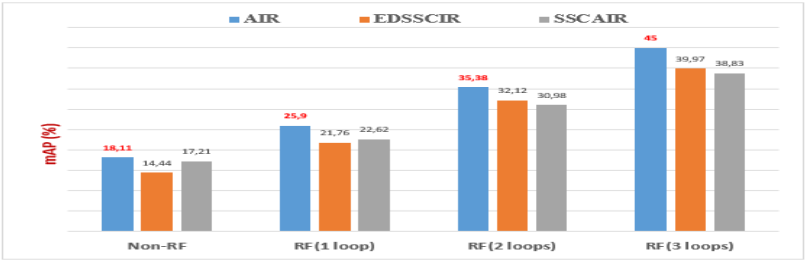

Figure 3.10. Performance comparison (mAP) for the first three loop

Table 3.4. Query execution time of AIR on CIFAR-100 dataset

Feedback Loop	Average time for one query with configuration		
	Shortcut(con-decon) (s)	**Shortcut (s)**	**Classic (s)**
No Feedback	0.2449	0.2650	0.2335
First Loop	25.5623	28.1375	24.0926
Second Loop	26.2186	28.9882	24.4392
Third Loop	27.2913	29.1830	24.5538

Table 3.4 shows that the method of using Shortcut has a higher time than Shortcut(con-decon) (~2s). The reason for this is that it requires additional time to compute the shortcut connections.

3.3.2. The results on the image dataset COREL

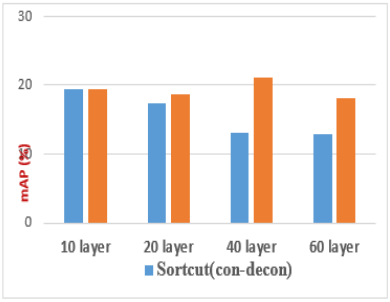

Figure 3.11. Image retrieval results at different depths of autoencoder network on the COREL dataset

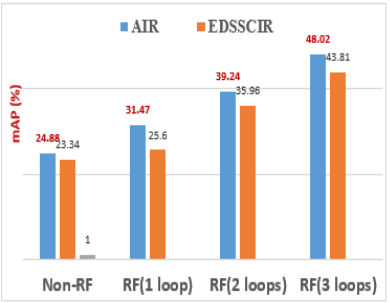

Figure 3.12. Performance comparison (as mAP) of the three methods for the first three loop

Figure 3.11, shows that the proposed network architecture gives the highest results on 40, and 60 layers, while 20 layers result in comparable performance. This proves that using asymmetric shortcut connections in the autoencoder to generate deep autoencoder networks for image matching is effective on the COREL dataset.

Figure 3.12 shows the mAP of three methods including Baseline (Non-RF), AIR, EDSSCIR for the first three response iterations. From Figure 3.9, it can be seen that the Baseline method gives the lowest precision. The reason for this is that the Baseline

method has no learning mechanism, it only calculates the Euclidean distance between the feature vector of the query image and the database image. The recommended method AIR performed better than the other two on all iterations. The performance of the AIR method is significantly better than that of Baseline, which indicates that the relevant feedback provided by the user is very helpful in improving the retrieval performance. AIR performs better than EDSSCIR because AIR obtains a good feature representation.

Table 3.5 shows that the method of using Shortcut has a higher time than Shortcut(con-decon) (~0.02s). The reason for this slight increase is due to the additional time required to compute the shortcut connections.

Table 3.5. Query execution time of AIR on COREL dataset

Feedback Loop	**Average time for one query with configuration**		
	Shortcut(con-decon) (s)	**Shortcut (s)**	**Classic (s)**
No Feedback	0.1289	0.1468	0.0457
First Loop	5.5781	5.5734	4.8175
Second Loop	5.6410	5.6508	4.8858
Third Loop	5.8743	5.8919	4.8108

3.4. Conclusion of chapter 3

In this chapter, the thesis presents an image retrieval method consisting of 3 components: (1) semi-supervised training by autoencoder convolutional neural network, (2) image feature extraction and (3) SVM classification in related feedback. This method has taken advantage of the autoencoder network to learn efficient feature representations for image retrieval through the use of Shortcut Connections in the autoencoder architecture.

The main contributions of this chapter have been published in the works [**CT1, CT3**].

CONCLUSION

The thesis has identified the research direction to focus on: approaching using machine learning (especially deep learning) to the image retrieval process with related feedback to shorten the semantic gap, improve the accuracy and Image retrieval speed in CBIR for problems with small class sizes, small sample sizes, large databases, and heights dimensionality.

Some contents of the thesis have been researched and solved such as: (1) using the row-sparsity to remove redundant features, to improve the image retrieval precision even though the class size of the training set may be very small; (2) provide a flexible model, which can select the important feature set, automatically adding positive samples to the training set and does not require a large enough number of positive samples; (3) take advantage of the autoencoder deep convolutional neural network model to learn efficient feature representations for image retrieval through the use of shortcut connections in the autoencoder architecture; (4) design an relevance feedback learning mechanism using a support vector machine SVM to take advantage of labeled samples from user's feedback.

Although the thesis has achieved some important research results on scientific theory and practice in using machine learning techniques in the CBIR process with related feedback, the thesis still has some issues that need to be research, improve and develop further in the future such as: (1) Leveraging the achievements of modern machine learning such as Vision Transformer model, graph convolutional neural network and transmission learning mechanism to improve performance image retrieval; (2) Implement proposed solutions to solve classes of practical problems, using image data with high accuracy, in various fields such as military, medicine, education, ….

NEW CONTRIBUTIONS OF THE THESIS

The thesis has proposed two methods of content-based image retrieval using related feedback, including: method **SDAIR** (Sparse Discriminant Analysis for Image Retrieval) and **AIR** (Autoencoders for Image Retrieval).

1. SDAIR method combines important feature extraction model based on RSLDA method with classification model in content-based image retrieval system to improve accuracy and query time. This method solves three problems: *First*, the number of responses that the user provides is smaller than the dimension of the feature space. *Second*, the number of positive feedback samples is often much lower than the number of negative feedback samples. *Third*, the number of classes is too small, which means that the number of projection directions is limited by the number of classes.

2. AIR method is based on three components: Semi-supervised training by autoencoder convolutional neural network, image feature extraction and SVM classification in related feedback to improve accuracy and time query. This method solves two limitations: First, the poor discriminating ability of the existing methods. Second, mitigate the problem of vanishing/exploding gradients and fast convergence.

LIST OF PUBLISH

1. An Hong Son, Nguyen Huu Quynh, Dao Thi Thuy Quynh, Cu Viet Dung, "Deep Learning of Image Representations with Convolutional Neural Networks Autoencoder for Image Retrieval with Relevance Feedback", *Journal on Information Technologies & Communications*, Vol. 2023, No. 1, pp. 17-24 (ISSN: 1859-3534, DOI: https://https://doi.org/10.32913/mic-ict-research.v2023.n1.1063).
2. Son An Hong, Quynh Dao Thi Thuy, Quynh Nguyen Huu, "Stuck Query Point Processing Of Multi-point Query For Image Retrieval With Relevance Feedback", *Journal of Information Hiding and Multimedia Signal Processing*, Vol. 12, No. 2, pp. 42-55, June 2021. (ISSN:2073-4212/2073-4239; **SCOPUS**).
3. Son An Hong, Quynh Nguyen Huu, Dung Cu Viet, Quynh Dao Thi Thuy, Tao Ngo Quoc (Accepted 20/5/2022), "Learning Binary Codes for Fast Image Retrieval with Sparse Discriminant Analysis and Deep Autoencoders", *Intelligent Data Analysis*, Vol. 27, No. 3, April 2023 (ISSN: 1088-467X/1571-4128; **SCIE**).
4. Son An Hong, Quynh Nguyen Huu, Dung Cu Viet, Quynh Dao Thi Thuy, Tao Ngo Quoc, "Improving image retrieval effectiveness via sparse discriminant analysis", *Multimedia Tools and Applications*, March 2023, pp.1-24 (ISSN: 1380-7501/1573-7721; **SCIE**).