Internet Cross-media Retrieval Based on Deep Learning

Bin Jiang[a], Jiachen Yang[a,d], Zhihan Lv[b,*], Kun Tian[c], Qinggang Meng[d], Yan Yan[e]

[a] *School of Electrical Automation and Information Engineering, Tianjin University, Tianjin, P.R. China*
[b] *Dept. of Computer Science, University College London, London, WC1E 6EA, UK*
[c] *National Key Laboratory of Science and Technology on Aerospace Intelligence Control,Beijing,P.R. China*
[d] *Department of Computer Science, School of Science at Loughborough University, UK*
[e] *Department of Information Engineering and Computer Science ,University of Trento, Italy*

Abstract

With the development of Internet, multimedia information such as image and video is widely used. Therefore, how to find the required multimedia data quickly and accurately in a large number of resources , has become a research focus in the field of information process. In this paper, we propose a real time internet cross-media retrieval method based on deep learning. As an innovation, we have made full improvement in feature extracting and distance detection. After getting a large amount of image feature vectors, we sort the elements in the vector according to their contribution and then eliminate unnecessary features. Experiments show that our method can achieve high precision in image-text cross media retrieval, using less retrieval time. This method has a great application space in the field of cross media retrieval.

Keywords: Cross-media retrieval, Deep learning, Feature extracting, Multimedia information

1. Introduction

With the growth of Internet services, multimedia technology has a continuous development , which brings a tremendous increase in the amount of Internet

[*] Corresponding author
Email address: Z.Lu@cs.ucl.ac.uk (Zhihan Lv)

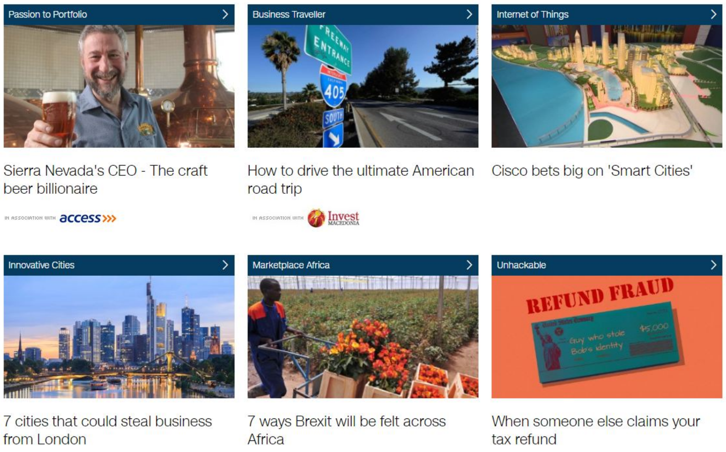
data. At the same time, the number of digital images per day is also growing exponentially at an alarming rate,with the popularity of digital cameras or other devices,as shown in Fig.1. How to make effective content retrieval and analysis from such a large number of images or texts, become the focus of scientific research[1].

Image retrieval technology is divided into the text based image retrieval(TBIR) technology and content based image retrieval(CBIR) technology[2, 3]. The traditional image retrieval system is based on the images with manual text annotation, and it is keyword search. This kind of system consumes huge manpower and material resources on the massive image data sets,so the retrieval precision is greatly disturbed by the subjective factors[4, 5]. The content based image retrieval technology is different, most of the content based systems use low-level visual features of the image itself, such as color, texture and shape. Some similarity evaluation methods are used to match the images using visual features,then the query image and database image are matched. Without manual intervention, speed and precision of the whole process have been greatly improved. The application field is very wide and can be used for the search of a particular picture on the network for common users, and can also be used for various types of professional organizations[6].

In this paper, we explore the cross media retrieval technology, texts are used to search related image information, images are used to search related text information,as shown in Fig.2.

Cross media retrieval technology is a new technology field, it is related to the multidisciplinary cross and comprehensive, some fields are considered: pattern recognition, machine learning, image processing, video processing, speech recognition, data mining technology, agent of artificial intelligence and natural language processing.

However, unlike single media learning, cross media learning is still a new research direction of salary. At present, the relevant research is still relatively preliminary, there are still a lot of challenges and difficulties in the following areas need further research.



Sierra Nevada's CEO - The craft beer billionaire

How to drive the ultimate American road trip

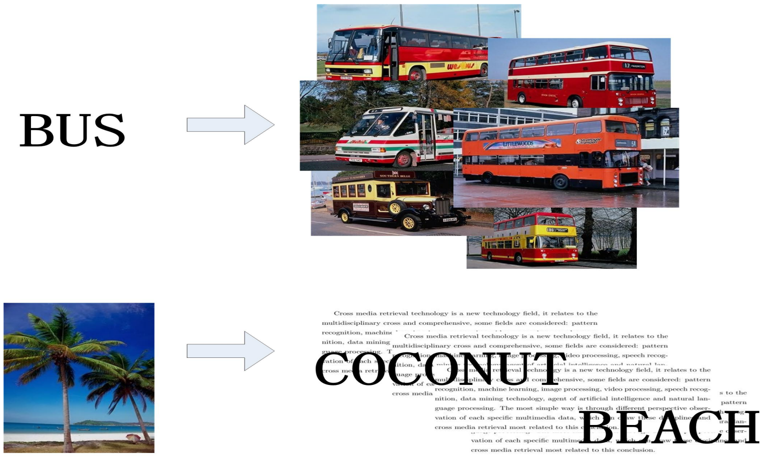
Cisco bets big on 'Smart Cities'

7 cities that could steal business from London

7 ways Brexit will be felt across Africa

When someone else claims your tax refund

Figure 1: Internet Cross-media: we can see that there will be a lot of texts around the images.



Cross media retrieval technology is a new technology field, it relates to the multidisciplinary cross and comprehensive, some fields are considered: pattern recognition, machine learning, image processing, video processing, speech recognition, data mining technology, agent of artificial intelligence and natural language processing. The most simple way is through different perspective observation of each specific multimedia data, which can draw these disciplines and cross media retrieval most related to this conclusion.

Figure 2: Text query images and image query texts

- Consistency description of cross media data

  Due to the different modes of cross media data, expression with different dimensions and different attributes cannot directly be the computing correlations. In addition to the heterogeneous characteristics, between the bottom of the content and high-level semantic semantic gap, characteristics by the traditional method can not be applied to the study of cross media for learning. Therefore, it is very necessary to excavate the consistency of cross media data and to realize the flexibility of different modes.

- Incremental learning across media

  The vast majority of cross media learning methods are involved in the singular value decomposition for the input matrix. Therefore, once to insert the new data to the original data set, the algorithm will spend more computation time is recalculated augmented the input data matrix singular value decomposition. Therefore, in order to avoid the high computational complexity, it is very important to propose an efficient incremental updating algorithm for singular value decomposition[7, 8].

- The lack of modal complement

  With the rapid development of information technology, more and more cross media data appear in many applications, such as medical diagnosis, web page classification and cross media analysis. However, due to the high cost of data acquisition, the lack of authenticity and rejection and other reasons, these applications are faced with the problem of lack of mode. It is inevitable that the application of the full modality of these requirements can be handled with a very limited amount of data. So, the cross media applications need a cross media data deletion modal complement method[9, 10].

- Correspondence between different modal descriptions

In practice, there is noise in the description of some modes, which can destroy the corresponding relationship. The vast majority of applications in the real world (such as cross media retrieval and pattern recognition) need complementary descriptions from different models to obtain more accurate and robust estimation. So there is an urgent need to find a cross media denoising method, to re-establish the correspondence between the cross media description.

- Sample redundancy in the same mode

  Within the same mode of an object to describe possible multiple, how to use the description of the other modes, elimination of redundant description, save storage space, is also very worthy of study.

2. Background and motivation

Rasiwasia et.al [11] proposed a cross media retrieval method based on canonical correlation analysis(CCA) to obtain the shared description between different modes. In addition, a general method of using CAA to study the semantic description of web images and related texts was presented[12, 13].

The cross media clustering uses the similarity of the samples in different modes, and the clustering of multiple modes is used to reveal the potential sharing structure among different modes, as shown in Fig.3. At present, relevant researchers have put forward some effective cross media clustering method [12]. A cross-media clustering method based on CCA is proposed to project data onto a low-dimensional space [12]. By using non negative matrix factorization, Liu[7] search compatible clustering method for different modes. Cai et.al [14] proposed a robust large scale multi modal clustering method to deal with the heterogeneous description of large scale data. In [8] , the proposed model is used to update the clustering model by using alternate propagation constraints among different modes.

Cross media classification[15, 16] is a method to construct a classifier based on the shared features of the objects found in the multiple modes, which can be

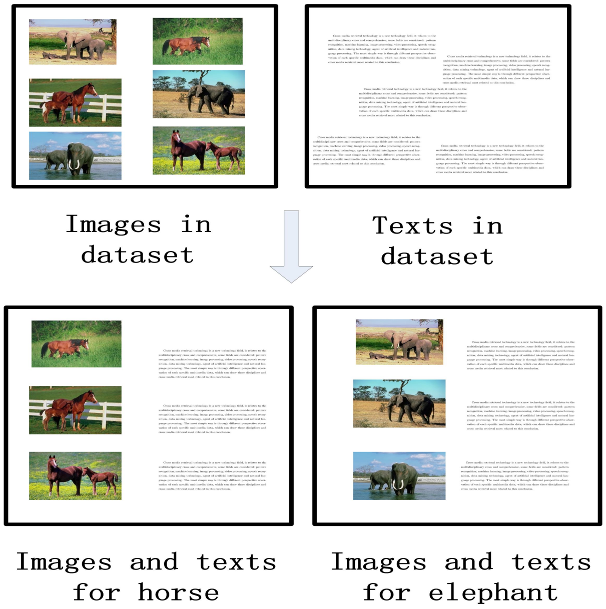


Figure 3: Crossmedia clustering

used to classify the multimedia. In recent years, some progress has been made in the field of cross media classification. A cross media feature extraction method based on CCA is proposed to realize cross media classification [15]. In [16], the paper proposed an 3D object class model, which was based on the estimation of pose and the recognition of invisible modes. An active transfer learning method is proposed to correlate the features of different scenes[17]. Based on boosting, a weak classifier is learned in each mode and the integrated classifier is produced by weighted combination[18].

Cross media essence lies in high-level information carrier (text, image, audio, video), the existence of a relatively low-level primitives and the basic element theory of multimedia. These characteristics are massive, heterogeneous and in high dimension .Multi-level characteristics will undoubtedly increase the difficulty of retrieval, literature and the association between the bottom and top is called semantic gap.

Based on the traditional content of cross media learning research, we can not solve the cross media correlation metric problem. However, data mining has been many related research work. In order to solve the learning process in the characteristics of heterogeneous cross media to reduce the semantic gap between low-level content and high-level semantics. In recent years, researchers proposed a variety of cross media learning method to explore cross the potential relationship between media data, improving the efficiency of cross media learning. These studies mainly include: collaborative training, multi core learning, subspace learning.

Blum et.al [19] proposed co-training method as one of the early schemes to deal with the problem of cross media learning. The algorithm is based on semi supervised learning, using the redundancy mode. Each modality is enough to describe the problem, if there are enough training samples, each of the modes is sufficient to learn a strong network. In the given marker, each mode are independent of another mode.

Collaborative training algorithm is iterative,as shown in Fig.4. The algorithm firstly uses the labeled samples to train each other's related classifiers separately, which makes the output of the same verification samples have to be similar to each other. Then in coordination in the process of training, each classifier with unlabeled samples selected several markers of high credibility samples . The labeled samples are added to another classifier of the labeled training set to each other, using these newly labeled samples are updated until it reaches a stop condition. In the principle of consistency, the purpose of each iteration is to maximize the consistency between the classifiers on the validation set. Although there are certain differences in the prediction results of each classifier in the validation set, these differences can be propagated back to the training set to train a more accurate classifier. In the next iteration, the difference between the predicted results of each classifier on the validation set is minimized.

However, in the usual case for the vast majority of cross media data sets, making the two modal fully redundant conditions are difficult to meet.

Although kernel method is a effective method to solve the nonlinear model , the number of feature samples containing heterogeneous information is very

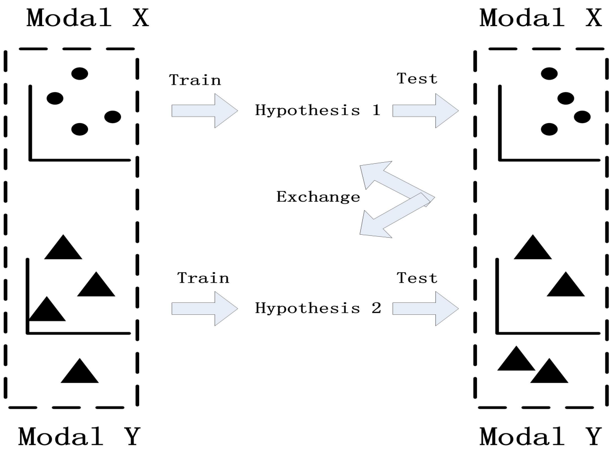


Figure 4: The process of co-training style algorithms

large . Therefore, in recent years, there has been a lot of research on the combination of nuclear, which is the multi core learning. This method is becoming a new hot spot in the field of nuclear machine learning. The process of multi core learning is given in this paper. In the course of study, it is not to select a single kernel function to carry out the multi core learning, but to select a set of kernel function, which is a combination of different nuclei. Because the different verification should be different from the different modes of the concept of similarity, so the nuclear combination may be a better integration of multiple information sources, the way to find the optimal solution [20, 21, 22].

In the multi core framework, the representation of the sample in the feature space is transformed into the basic core and the combination weight coefficient,as shown in Fig.5. The data in the new feature space to obtain a better expression, and greatly improve the precision of prediction or classification accuracy. However, the most important issue here is how to get the feature space of the combination, which is how to get the combination weight coefficient through learning.

As is shown in the graph, subspace learning aims at obtaining the potential subspace of multiple modes to capture the complementary information among

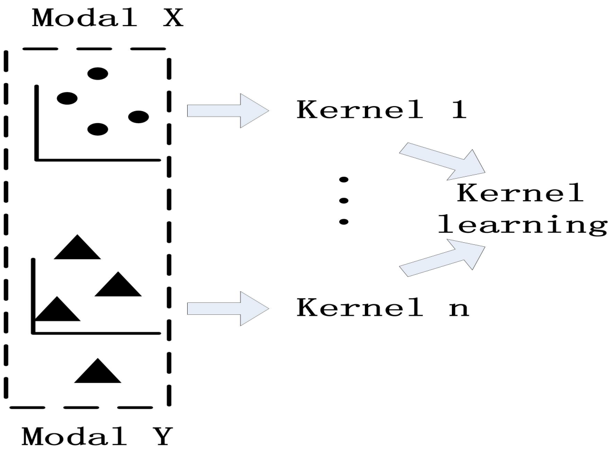


Figure 5: The process of Multiple Kernel Learning(MKL) algorithms

different modes,as shown in Fig.6. The current subspace learning methods are divided into four categories, which are based on projection, matrix decomposition, task and measure.

Shared subspace learning method based on projection uses feature mapping to extract the latent subspace of multiple modes. Such methods can be divided into linear projection and nonlinear projection. Classical linear projection methods include: canonical correlation analysis, partial least squares, and the mainstream of nonlinear projection methods mainly related to: the kernel canonical correlation analysis method, and the depth of canonical correlation analysis[23].

Subspace learning method based on matrix decomposition using matrix decomposition to extract the basis vectors of shared subspace between different modes. This kind of methods mainly include two kinds: one kind is based on non negative matrix factorization method, such as the joint sharing non negative matrix factorization algorithm; the other is based on the feature decomposition method, such as multiple output regular projection algorithm[24, 25].

The task based shared subspace learning method can improve the overall generalization performance of each task by learning multiple related tasks at

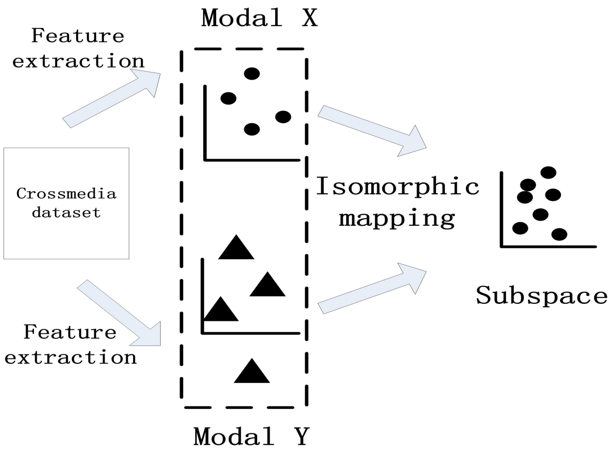


Figure 6: The process of Subspace Learning(SL) algorithms

the same time. This kind of method contains multi task learning, multi label learning and multi class learning. Multi task learning compared to the classical algorithms including: alternating structure optimization method and convex multi task feature learning algorithm also, efficient multi label learning method for multi label classification of shared subspace learning algorithm for; and the new multi class learning method is for multi class classification of share structure called method[26].

The subspace learning method based on metric is designed to learn the good measurement between different modes to realize the comparison of the similarity between the modes. The methods are divided into two categories: a class using the Euclidean distance metric, such as multi modal distance metric learning framework; another based on the Mahalanobis distance metric, the representative method for shared subspace multi metric learning algorithm[27].

Text retrieval processing is based on the characteristics of words, phrases paragraph and text structure statistics to achieve internal search. Image retrieval is based on the pixels within the image, such as gray value and texture, skeleton structure, boundary feature or some transformation by the wavelet analysis. For frequency-domain characteristics, video and audio have similar spectral characteristics. If these features are extracted as the composition of a high-dimensional feature vector, the internal data will be redundancy. However, we found some interesting phenomenons , kernel function and kernel model is widely used. This model is the first classification model from the classifier design. SVM has been successfully used, using the principle of the essence [28].

Each multimedia low-level feature structures may be no rules , we can maximize various multimedia data with class separability criterion through the classifier design. In this way, we can ensure retrieval process from a cross media integrated search cable expression model based on data classification properties . Based on the characteristics of multimedia data, classification process also has dimension reduction function. Automatic matching for certain characteristics is needed, which will be matched up to the cross media integrated model. This means that we don't bother the various media for manual annotation. The template structure still need us to carry on the bottom of the cross media integrated features extraction. Therefore, the establishment of cross media extraction is finished in the data separability criterion and expression[29].

3. Proposed method

In this part, we will introduce in detail the retrieval algorithm, which is based on the similarity theory, and then exit the machine learning algorithm based on DBN. Finally, the results of the most matching image retrieval can be obtained.In image feature extraction, although a large number of features lead to slow calculation, but we added a dimension reduction algorithm, used to solve this problem. The retrieval process shown in Fig.7.

3.1. Feature extraction

3.1.1. CCM and DBPSP

As a kind of commonly methods, which used to describe the texture feature of image, a 3×3 window for color co-occurrence matrix (CCM) with four directions respectively can be obtained in an image. As shown in Fig.8 and 9,CCM reflects

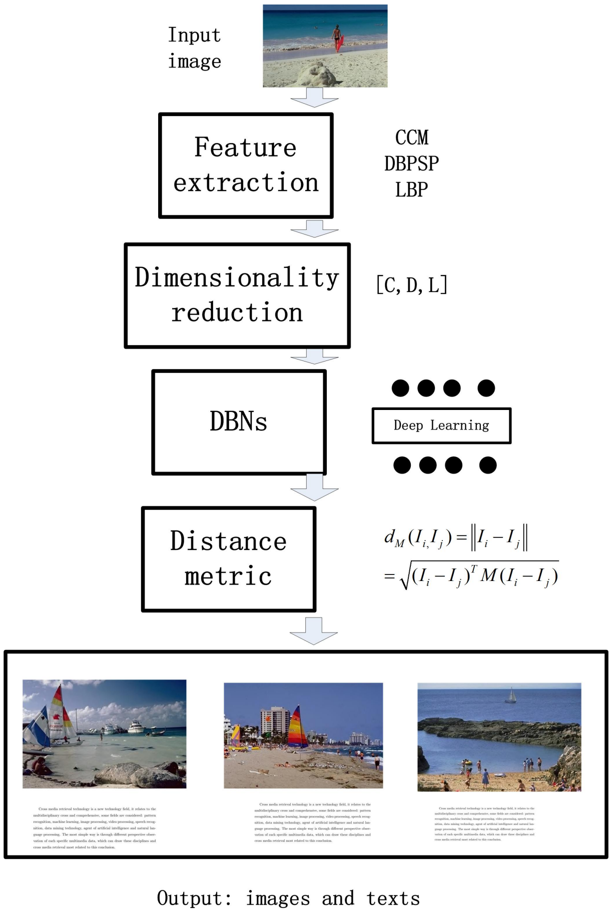


Figure 7: Proposed cross media retrieval method,this method is divided into four steps. Specific image features are extracted and dimension reduction is necessary. DBN is used to classified and distance metric is used to rank.

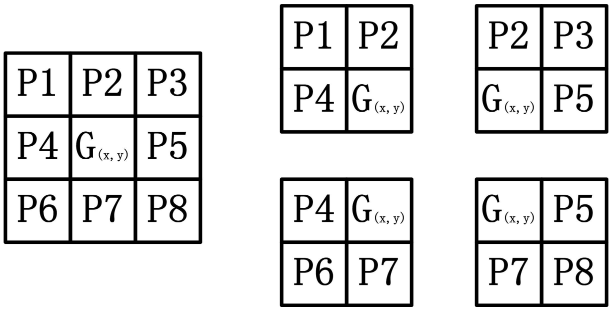


Figure 8: 3 × 3 window for color co-occurrence matrix (CCM)

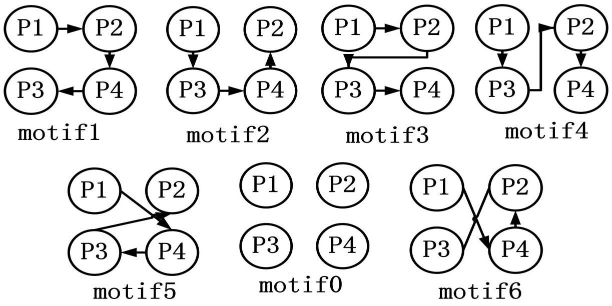


Figure 9: 7 motifs color co-occurrence matrix (CCM)

the relationship between the pixels of the neighborhood. CCM features can be divided into seven categories, as used by the previous[30]. The final matrix descriptor is a 7 × 7 form.

$$\begin{aligned} M_i(u,v) &= M_i(u,v|\delta_x,\delta_y) \\ &= M_i(P_i[x,y],P_i[x+\delta_x,y+\delta_y]) \end{aligned} \tag{1}$$

$$m_i(u,v) = \frac{M_i(u,v)}{N_i} \tag{2}$$

$$N_i = \sum_{u=0}^{6}\sum_{v=0}^{6} M_i(u,v) \tag{3}$$

Color co-occurrence matrix (CCM) is designed to describe the direction of local texture , which can not be used to describe the internal complexity of the texture. Difference between pixels of scan pattern(DBPSP) is put forward.

$$\begin{aligned}
d^1(x,y) &= |P_1 - P_2| + |P_2 - P_4| + |P_4 - P_3| \\
d^2(x,y) &= |P_1 - P_3| + |P_3 - P_4| + |P_4 - P_2| \\
d^3(x,y) &= |P_1 - P_2| + |P_2 - P_3| + |P_3 - P_4| \\
d^4(x,y) &= |P_1 - P_3| + |P_3 - P_2| + |P_2 - P_4| \\
d^5(x,y) &= |P_1 - P_4| + |P_4 - P_3| + |P_3 - P_2| \\
d^6(x,y) &= |P_1 - P_4| + |P_4 - P_2| + |P_2 - P_3|
\end{aligned} \tag{4}$$

$$AR_i = \frac{1}{N_i} \sum_{N_i}^{j} d_i^j(x,y) \tag{5}$$

3.1.2. Local binary pattern (LBP)

The initial function of Local binary pattern (LBP) is the auxiliary local contrast of image ,but not a complete feature descriptor[31].In the field of digital image processing and pattern recognition, and then upgrade to a kind of effective texture description operator, measure and extract the texture information of the image, which is invariant to illumination. LBP has many variants, or improvements.

in the 3×3 window, original LBP Operator is defined, adjacent pixels are compared with gray value center pixel, if the center pixel value is greater than the surrounding pixel value , the pixel is labeled as 0, otherwise 1. So,3×3 neighborhood of eight points compared with center produce 8-bit binary number (a total of 256 codes), that is to be the window center pixel LBP value, and the value can be used to reflect the texture information of the region.

In this paper, we use the famous rotation invariant LBP as an operator [32, 33]. The continuous rotation of the circular neighborhood to get a series of initial LBP values, the minimum value as the LBP value of the neighborhood,as shown in Fig.10.

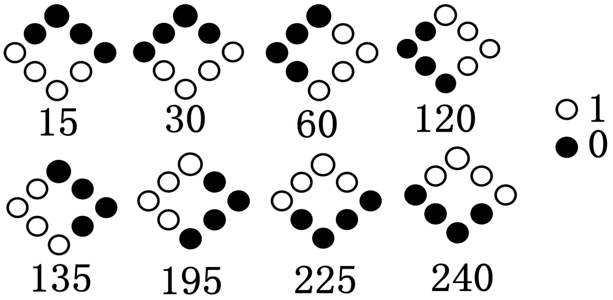


Figure 10: Rotation invariant local binary pattern

3.2. Dimensionality reduction

In image feature extraction, although a large number of features lead to slow calculation, but we added a dimension reduction algorithm, used to solve this problem.We can get the most efficient retrieval algorithm according to the characteristic of the degree of contribution.

$$I(x,y) = \sum_{i,j} p(x_i, y_j) log \frac{p(x_i, y_j)}{p(x_i)p(y_j)} \tag{6}$$

$$maxV = \frac{1}{|S|^2} \sum_{i \in S} I(c, x_i) \tag{7}$$

$$minW = \frac{1}{|S|^2} \sum_{i,j \in S} I(x_i, y_j) \tag{8}$$

$$max\phi = (V - W) \tag{9}$$

By this method, we can sort the image features according to the importance degree.

3.3. Deep belief networks

Deep learning refers to a variety of machine learning algorithms to solve the problem of image, text and other issues on the multi-layer neural network. Depth study can be classified into neural network from the large class, but there

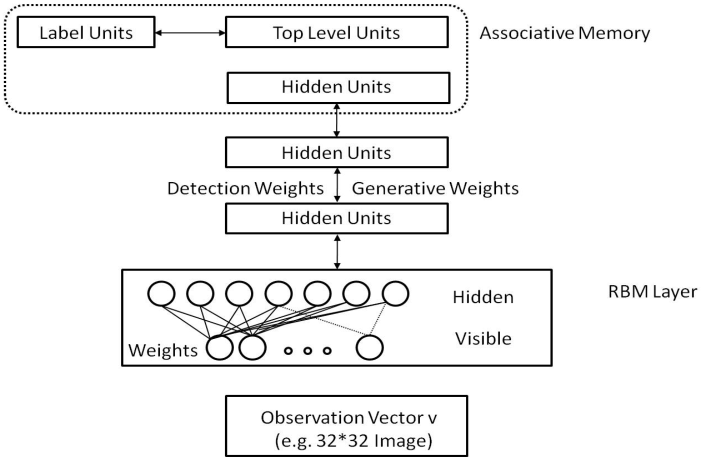


Figure 11: Deep belief network framework

are many changes in the specific implementation. The core of deep learning is the feature learning, which aims to get the characteristic information of the hierarchical network , so as to solve the important problems that need to be used in the past. Deep learning is a framework that contains several important algorithms: Convolutional Neural Networks(CNN),Sparse AutoEncoder(SAE),Restricted Boltzmann Machine(RBM),Deep Belief Networks(DBN)and Recurrent neural Network(RNN).For different issues (image, voice, text), it is necessary to use different network models to achieve better results [34, 35].

In this paper ,we choose deep belief network as our machine learning algorithm,as shown in Fig.11. DBNs was proposed in 2006 by Geoffrey Hinton and it is a generative model. By training the weights between the neurons, we can make the whole neural network according to the maximum probability to generate the training data. We can not only use the DBN recognition feature, classification data, but also can use it to generate data. For deep neural network, if the traditional BP algorithm is applied, DBNs encountered the following problems: A big set of labeled samples for training should be provided; The learning process is slow; Inappropriate choice of parameters lead learning to converge to local optimal solutions.

DBN is composed of a plurality of neurons, which are divided into explicit and hidden neurons. The explicit element is used to accept the input, and the hidden element is used to extract the feature. So the hidden element also can be called the detectors . The top two layers of the connection is no direction, the composition of the joint memory is made. The lower layer is connected between the upper and the lower. The lowest level represents the data vector , each neuron represents one dimension of the data vector. The components of DBN is Restricted Boltzmann Machine (RBM). The process of training DBN is carried out layer by layer. In each layer, the data vector is used to infer the hidden layer, and then the hidden layer is used as the data vector of the next layer [36].

RBM is the component of DBN [37]. In fact, each RBM can be used alone as a cluster. RBM only has two layers of neurons, one is called visible layer for input training data. Another layer is called the hidden layer , corresponding to the implicit element , used as a feature detector .In fact, RBM training process is to find a probability distribution of the most able to produce a training sample. That is, the requirements of a distribution is made for the maximum probability of training samples. Because the distribution of the decisive factor is the weight of W, so the goal of RBM is to find the best weight. In order to maintain the interests of readers, here we do not give the maximum likelihood function of the derivation process, a direct description of how to train RBM.

DBN is composed of multilayer RBM for a neural network, which can be regarded as a generative model. Here is the training process: First full training of the RBM; Fix the first RBM's weight and offset, and then use the state of the hidden neuron as the input vector of the second RBM; Full training after second RBM, the second RBM stacked on the top of the first RBM; Repeat the above three steps any number of times; If the training set of data is labeled, the RBM in the display layer need to have the classification of labeled neurons, together with the training.

3.4. Distance and matching strategy

Manhattan distance and Euclidian distance are often used. This method is simple and easy to understand.The relationship between the two feature vectors can be computed.

$$d(q, x_j) = \sum_{i=1}^{i=m} |x_{ij} - q_i| \tag{10}$$

$$d(q, x_j) = \sqrt{\sum_{i=1}^{i=m} (x_{ij} - q_i)^2} \tag{11}$$

Weighted Euclidian distance is often used considering the weight factor.The relationship between the two feature vectors can be computed.

$$d(q, x_j) = \sqrt{\sum_{i=1}^{i=m} \rho_i (x_{ij} - q_i)^2} \tag{12}$$

$$\rho_i = \frac{n}{\sum_{j=1}^{j=n} (x_{ij} - \bar{z}_i)^2} \tag{13}$$

$$\bar{z}_i = \frac{\sum_{j=1}^{j=n} x_{ij}}{n} \tag{14}$$

Image similarity learning(ISL) can compute the dissimilarity extent of two vectors extracted from two images. The famous learned manhalanobis distance metric $d_M(I_i, I_i)$ can be caculated in the form:

$$d_M(I_i, I_i) = \|I_i - I_j\| = \sqrt{(I_i - I_j)^T M (I_i - I_j)} \tag{15}$$

This linear equation can also be converted to get

$$d_M(I_i, I_i) = \|I_i - I_j\|_M = \sqrt{(WI_i - WI_j)^T (WI_i - WI_j)} \tag{16}$$

$$M = W^T W \tag{17}$$

4. Evaluation metric and databases

In this part, we will introduce the evaluation criteria of the retrieval algorithm and the experimental database. In the cross media retrieval evaluation algorithm, we use three parameters: MAP, Percentage and MRR.In the selection of the experimental database, we choose the Wiki Text-Image dataset [11] and NUS-WIDE dataset [38].

4.1. Performance evaluation

For a cross media retrieval system, the evaluation index is very important. Recently, many parameters are proposed to evaluate the advantages and disadvantages of the algorithm. In this paper, we adopt precision and recall as the evaluation index to carry out our experiments[39]. Precision is the ratio between TP and TP + FP ,in which TP is the retrieval similar images number,and TP + FP is the retrieval total images number. Precision is used to measure the Probability of success for an image retrieval system. Recall is the ratio between TP and TP + FN, in which TP is the retrieval similarity image number and TP + FN is the total number in the database. Recall is used to measure the percentage of the retrieved for image retrieval system. The definition of precision and recall is shown in Table.1. The following formula is the expression of mathematical language, for the two evaluation indicators.

$$precision = \frac{TP}{TP + FP} \tag{18}$$

$$recall = \frac{TP}{TP + FN} \tag{19}$$

In previous studies , if there are more relevant information in feedback of retrieved images, the system is better.On the contrary, the system is worse.

However, it is mutually restricted between precision and recall. Specifically speaking,if there are 1000 images in database, but only 100 of them are related to the query. If the former 10 retrieved images are related, precision is considered to be 1, but recall is considered to be only 0.1; Instead, if the 100 images that

Table 1: Definition of precision and recall

	Relevant	Irrelevant	Total
Retrieved	True Positive	False Positive	Predicted Positive
Not retrieved	False Negative	True Negative	Predicted Negative
Total	Actual Positive	Actual Negative	TP+FP+FN+TN

has just been described is used as a search result, recall must be 1, but precision is considered to be only 0.1. In the following experiment, the precision-recall curve analysis is used to assess the various retrieval methods. On the other hand, calculation of comprehensive evaluation index is used to evaluate the performance of the analysis.

In order to take care of the difference between the sample in the process of user experience, the mean precision of the retrieval results is indicated by the average precision of the former N images. In search results, users usually focus only on the image or text that is in the front row, the former N image contains a lot of relevant image , we can think that the system can meet the requirements of the search. In fact, the N here is equivalent to the sum of TP and FP, TP is the number of relevant images retrieved and FP is the number of irrelevant images the formula is:

$$meanprecision = \frac{TP}{N} \tag{20}$$

Comprehensive evaluation index , $F - measure$ is one of the common evaluation criteria in the field of information retrieval. When we consider the effects of both the precision and the recall, the $F - measure$ is obtained, $F - measure$ is calculated as follows.

$$F - measure = \frac{(\varphi^2 + 1) \times precision \times recall}{\varphi^2 \times (precision + recall)} \tag{21}$$

In the above formula, φ It is used to adjust the weighted proportion of both the precision and the recall .If $\varphi = 1$,F-measure is $F1$,

$$F - measure = \frac{2 \times precision \times recall}{precision + recall} \tag{22}$$

Obviously, $F1$ is the perfect combination of both the precision and the recall . If $F1$ is greater, the ability of the retrieval system is stronger.

4.2. Evaluation metric

If the result of the search is the same category as the query, we think the result is relevant. In other words, the relevant retrieval results refer to exactly match the content of the query in the evaluation process.

Different from the simple judgment of retrieval results, category relevance is more widely used in the evaluation of the types of the algorithm.

In our experiments, we evaluate the retrieval performance for different methods using the retrieved results and category relevance.

Three evaluation metrics are used in two aspects are as follows.

4.2.1. MAP

MAP(mean average precision) is mainly used to determine whether the results of retrieval content is the same category with query, which is related .Or the contents of the retrieval information does not belong to the same category, which is not related.

If an image information or a text message is given, the corresponding result in the retrieval process is R, mean average precision(MAP) can be defined as

$$AP = \frac{1}{L} \sum_{r=1}^{r=R} prec(r)\delta(r) \tag{23}$$

L is the amount of data that is retrieved from the database. $prec(r)$ is the reaction of the accuracy of the top r retrieval datas. Here, we use the ratio of the top r in the relevant retrieval results and the full results to represent. If a retrieval result and query are the same category , we set $\delta(r) = 1$, and if a retrieval result and query are not the same category , we set $\delta(r) = 0$.

In this paper, we set $R = 50$. We obtain an average value of the AP calculated from the retrieval results, as MAP(mean average precision) .

4.2.2. Percentage

In the above introduction, MAP(mean average precision) only exists two kinds of absolute classification with the query related and not related. In order to evaluate the accuracy of the retrieval results more flexibly, we first rank the retrieval results according to the correlation . In this evaluation criteria, if a result exists in the percentage of the former t, we can consider that the match is successful.

In this paper, we set $t = 0.2$. Percentage can reflect the ratio of the query correlation degree, and the evaluation is more flexible.

4.2.3. MRR

In this paper,Mean Reciprocal Rank (MRR) is also used to evaluate the retrieval performances for different methods

As similar with Percentage , Mean Reciprocal Rank (MRR) is defined regarding search results in the ranking correlation degree with the query. Mean Reciprocal Rank (MRR) is defined as follows:

$$MRR = \frac{1}{|Q|} \sum_{i=1}^{i=|Q|} \frac{1}{rank_i} \tag{24}$$

4.3. Databases

In the field of cross media retrieval, we have selected two commonly used database to carry our experiments, in order to evaluate the advantages and disadvantages of the retrieval algorithm.

4.3.1. Wiki Text-Image dataset [11]

As the firt experimental Text-Image dataset, Wiki is extracted from the article in Wikipedia, both the image and text information . Total 2866 image-text information pairs form the Wiki Text-Image dataset, which is divided into

Table 2: 10 different categories in Wiki Text-Image dataset

	Art	Bio	Geo	His	Lit	Med	Mus	Roy	Spo	War	Total
Training	138	272	244	248	202	178	186	144	214	347	2173
Testing	34	88	96	85	65	58	51	41	71	104	693
Total	172	360	340	333	267	236	237	185	285	451	2866

10 different categories as shown in Table.2. Each of these images is equipped with the corresponding text interpretation as a description.

In Wiki Text-Image dataset, there are 117.5 words to describe each figure on average. When it comes to the training and testing of the depth of learning, we choose randomly twenty percent of them as a test, and the other eighty percent as a training.

4.3.2. NUS-WIDE dataset [38]

As a supplement, NUS-WIDE dataset is another database we use, and 133,208 images are contained . The NUS-WIDE dataset are with 81-dimensional concepts and 1000-dimensional tags .

The 1000-dimensional tags(annotated tags) can be regarded as the text information, and cach image can be regarded as a image-text pair with the tags. Of course, the image concepts can be treated as the labels for the image-text pairs.In the vast database, we just select the 10 kinds of text-image pairs for our experiments.In this case, we get 26813 sets of information.

Just as the first Wiki dataset, we choose randomly twenty percent of them as a test, and the other eighty percent as a training.In NUS-WIDE Text-Image dataset, there are 7.7 words to describe each figure on average.

4.4. Compared methods

In the field of cross media retrieval, we have selected four commonly used algorithm to be compared with the proposed method, which can demonstrate the effectiveness of our approach

SLDA: SLDA has just been proposed when it is the intrinsic link between the image and the text information. When the image or text is retrieved, the algorithm can be used to get some results around it. According to the symmetric KL-divergence between query and database for images or texts, we can get the ranked retrieved results.

GMA: As a supervised method, Generalized Multiview Analysis (GMA) for cross-media retrieval use pair-wised information and label information. This algorithm has just been proposed as a supervised kernelizable extension for CCA .For each subspace , different modality spaces contain different map data.

SLIM: Just as a supervised dictionary learning method, supervised coupled dictionary learning with group structures for Multi-Modal retrieval (SlIM) is a new approach for crossmedia retrieval. According to multi-model dictionaries and mapping functions ,different modalities for class information can be done using group structures.

MR: Multi-modal Mutual Topic Reinforce Modeling(MR) is proposed for cross-media retrieval by obtaining the common cross-modal semantic information . To explore the internal connection of text or image information, the algorithm also has achieved good results.

5. Experimental results

As mentioned above, whether it is in the field of separate text retrieval, or in the field of cross media retrieval, retrieval accuracy and retrieval time consumption is an important indicator to evaluate the quality of an algorithm. In the experimental part of this paper, we divided into two main content, to detect the performance of the algorithm in two aspects.

Also in the retrieval process, it is different for using image search related text messages and using the text search related image information. The difference between this process resulted in experimental results, so we have a necessary classification to be discussed.

Table 3: MAP, PER and MRR for image-query-texts in both Wiki and NUS-WIDE

	MAP_{Wiki}	PER_{Wiki}	$MRR1_{Wiki}$	MAP_{NUS}	PER_{NUS}	MRR_{NUS}
SLDA [40]	0.2116	0.3037	0.0369	0.1976	0.2369	0.0022
GMA [15]	0.2074	0.2792	0.0153	0.2202	0.3765	0.0043
SLIM[41]	0.2548	0.4048	0.0454	0.3154	0.4639	0.0057
MR[42]	0.2298	0.3735	0.0321	0.2445	0.3896	0.0065
Proposed	0.2576	0.4121	0.0459	0.3373	0.4672	0.0067

5.1. Evaluation of retrieval precision

We evaluate the two tasks in cross-media retrieval:image-query-texts(input one image to search texts) and text-query-images(input one text to search images). For MAP, Percentage and MRR, the performances are list in Table.3 and 4. The results for images query texts are shown in Table.3, and the results for texts query images are shown in Table.4.The experimental results can be seen that our method has a higher accuracy in the two directions of the search.

In order to compare the accuracy of various retrieval algorithms, the results of the first four algorithms in different cross-media datasets are shown in Fig.12 and Fig.13. In Fig.12, image-query-texts results can be compared and the text-query-images results are shown in Fig.13.

Under the support of two kinds of databases, four algorithms are used to be compared. Finally, in order to get a more intuitive understanding of the situation, we will be the two directions on average, get the Table.5 and Fig.14.

In order to further evaluate the accuracy of the algorithm, we consider three other methods(CCA,SM,SCM) to carry out a comparative validation.The performances are list in Table.6.Among them, the experimental results of the two databases are embodied.

Table 4: MAP, PER and MRR for text-query-images in both Wiki and NUS-WIDE

	MAP_{Wiki}	PER_{Wiki}	$MRR1_{Wiki}$	MAP_{NUS}	PER_{NUS}	MRR_{NUS}
SLDA[40]	0.2146	0.2723	0.0241	0.2078	0.2640	0.0035
GMA [15]	0.2542	0.2827	0.0208	0.4199	0.3752	0.0045
SLIM[41]	0.2021	0.3106	0.0261	0.2924	0.3877	0.0045
MR[42]	0.2677	0.4014	0.0400	0.3044	0.4853	0.0071
Proposed	0.2761	0.4097	0.0413	0.4221	0.4943	0.0079

Table 5: MAP, PER and MRR for image-query-texts and text-query-images: average in in both Wiki and NUS-WIDE

	MAP_{Wiki}	PER_{Wiki}	$MRR1_{Wiki}$	MAP_{NUS}	PER_{NUS}	MRR_{NUS}
SLDA [40]	0.2131	0.2880	0.0305	0.2027	0.2518	0.0029
GMA [15]	0.2308	0.2810	0.0181	0.3201	0.3759	0.0044
SLIM[41]	0.2285	0.3595	0.0358	0.3039	0.4258	0.0051
MR[42]	0.2488	0.3875	0.0361	0.2742	0.4375	0.0068
Proposed	0.2669	0.4109	0.0436	0.3797	0.4808	0.0073

Table 6: Comparision with other three methods in two datasets

	IqT_{Wiki}	TqI_{Wiki}	AVE_{Wiki}	IqT_{NUS}	TqI_{NUS}	AVE_{NUS}
CCA	0.249	0.196	0.223	0.194	0.171	0.183
SM	0.225	0.223	0.224	0.230	0.208	0.219
SCM	0.277	0.226	0.252	0.235	0.212	0.224
Proposed	0.281	0.239	0.257	0.241	0.215	0.228

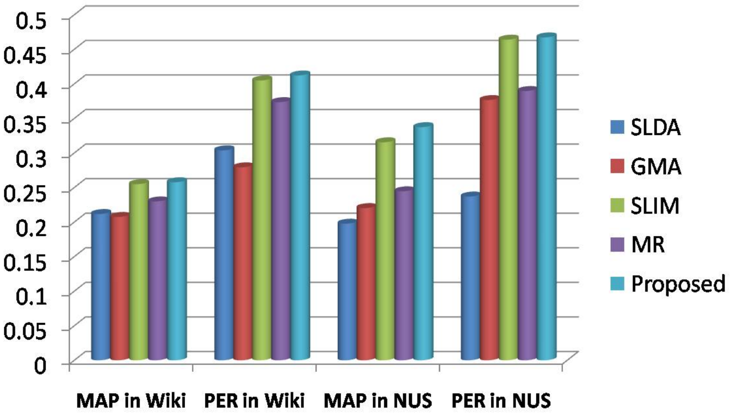


Figure 12: Results of experiments for image-query-texts

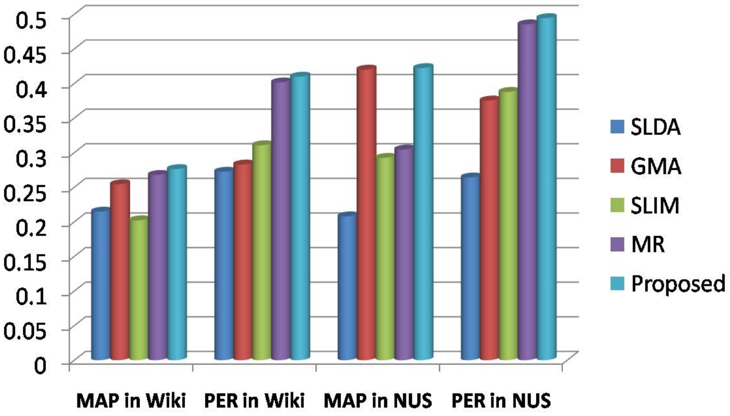


Figure 13: Results of experiments for text-query-images

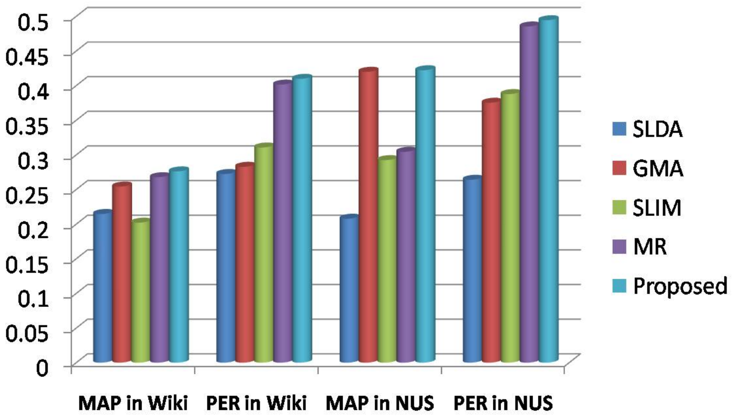


Figure 14: Results of experiments for average value between text-query-images and image-query-texts

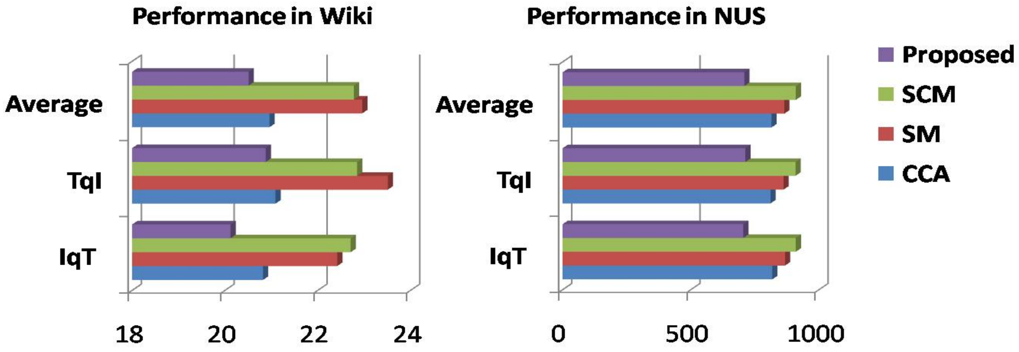


Figure 15: Comparison of retrieval time for different algorithms

Table 7: The comparison of average time consumption for image-query-texts and text-query-images (unit: ms)

	IqT_{Wiki}	TqI_{Wiki}	AVE_{Wiki}	IqT_{NUS}	TqI_{NUS}	AVE_{NUS}
CCA	20.82	21.09	20.96	818	811	815
SM	22.42	23.51	22.96	868	862	865
SCM	22.71	22.85	22.78	910	909	910
Proposed	20.12	20.88	20.50	705	713	709

5.2. Evaluation of retrieval time consumption

Besides the retrieval precision, retrieval time consumption is also a key evaluation result. The performances are list in Table.7.We see that although the number of features extracted in the algorithm is more, our algorithm can ensure that the use of less time consumption. In addition, Comparison of retrieval time consumption can be seen in Fig.15, and our algorithm is able to reduce the retrieval time.

All in all, the method proposed in this paper has some improvement in the accuracy and the retrieval time. In order to meet the needs of cross media retrieval in real time, this kind of method can be used.

6. Conclusion

With the development of the Internet, multimedia information such as image and video, is more convenient and faster. Therefore, how to find the required multimedia data in a large number of resources quickly and accurately, has become a research focus in the field of information retrieval. In this paper, we propose a real time internet cross-media retrieval based on deep learning. As an innovation, We have made full improvement and improvement in extracting feature and distance detection. Experiments show that our method can achieve high precision in image cross media retrieval, using less retrieval time.

The biggest contribution of this paper is that our method can be used for the initial evaluation of image recognition samples to detect whether it meets the requirements of the recognition algorithm. At the same time for a large database of data, our method can be used to detect the error rate of image recognition, which has a very big help for machine learning. Our algorithm has not been verified by more image recognition experiments, so there is still a lot of research space for the quality evaluation of image source, and more work is still needed.

7. Acknowledge

This research is partially supported by National Natural Science Foundation of China (No. 61471260 and No. 61271324), and Natural Science Foundation of Tianjin (No. 16JCYBJC16000).

References

[1] J. Yang, H. Wang, Z. Lv, W. Wei, H. Song, M. Erol-Kantarci, B. Kantarci, S. He, Multimedia recommendation and transmission system based on cloud platform, Future Generation Computer Systems.

[2] J. M. Guo, H. Prasetyo, J. H. Chen, Content-based image retrieval using error diffusion block truncation coding features, IEEE Transactions on Circuits and Systems for Video Technology 25 (3) (2015) 466–481.

[3] J. A. Piedra-Fernandez, G. Ortega, J. Z. Wang, M. Canton-Garbin, Fuzzy content-based image retrieval for oceanic remote sensing, IEEE Transactions on Geoscience and Remote Sensing 52 (9) (2014) 5422–5431.

[4] W. Li, L. Duan, D. Xu, W. H. Tsang, Text-based image retrieval using progressive multi-instance learning, Proceedings 58 (11) (2011) 2049–2055.

[5] I. Ahamd, T. S. Jang, Old fashion text-based image retrieval using fca, in: International Conference on Image Processing, 2003, pp. III–33–6 vol.2.

[6] Y. Lin, J. Yang, Z. Lv, W. Wei, H. Song, A self-assessment stereo capture model applicable to the internet of things, Sensors 15 (8) (2015) 20925–20944.

[7] J. Liu, C. Wang, J. Gao, J. Han, Multi-View Clustering via Joint Nonnegative Matrix Factorization, 2013.

[8] E. Eaton, M. Desjardins, S. Jacob, Multi-view constrained clustering with an incomplete mapping between views, Knowledge and Information Systems 38 (1) (2014) 231–257.

[9] D. R. Hardoon, S. Szedmak, J. Shawe-Taylor, Canonical correlation analysis: an overview with application to learning methods., Neural Computation 16 (12) (2004) 2639–2664.

[10] L. Sun, S. Ji, J. Ye, Canonical correlation analysis for multilabel classification: A least-squares formulation, extensions, and analysis, IEEE Transactions on Software Engineering 33 (1) (2011) 194–200.

[11] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. R. G. Lanckriet, R. Levy, N. Vasconcelos, A new approach to cross-modal multimedia retrieval, in: International Conference on Multimedia, 2010, pp. 251–260.

[12] K. Chaudhuri, S. M. Kakade, K. Livescu, K. Sridharan, Multi-view clustering via canonical correlation analysis, in: International Conference on Machine Learning, ICML 2009, Montreal, Quebec, Canada, June, 2009, pp. 129–136.

[13] H. Hotelling, Relations between two sets of variates, Biometrika 28 (28) (1935) 321–377.

[14] X. Cai, F. Nie, H. Huang, Multi-view k-means clustering on big data, in: International Joint Conference on Artificial Intelligence, 2013.

[15] A. Sharma, Generalized multiview analysis: A discriminative latent space, in: IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 2160–2167.

[16] S. Hao, S. Min, F. F. Li, S. Savarese, Learning a dense multi-view representation for detection, viewpoint classification and synthesis of object categories, 2009, pp. 213–220.

[17] Z. Zhang, Y. Zhao, Y. Wang, J. Liu, Z. Yao, J. Tang, Transferring training instances for convenient cross-view object classification in surveillance, IEEE Transactions on Information Forensics and Security 8 (10) (2013) 1632–1641.

[18] S. Ko?o, C. Capponi, A boosting approach to multiview classification with cooperation, in: European Conference on Machine Learning and Knowledge Discovery in Databases, 2011, pp. 209–228.

[19] A. Blum, T. M. Mitchell, Combining labeled and unlabeled sata with co-training, in: Eleventh Conference on Computational Learning Theory, COLT 1998, Madison, Wisconsin, Usa, July, 1998, pp. 92–100.

[20] M. Nen, Alpayd, N. Ethem, Multiple kernel learning algorithms, Journal of Machine Learning Research 12 (2011) 2211–2268.

[21] Z. Xu, R. Jin, H. Yang, I. King, M. R. Lyu, Simple and efficient multiple kernel learning by group lasso, in: International Conference on Machine Learning, 2010, pp. 1175–1182.

[22] D. P. Lewis, T. Jebara, W. S. Noble, Nonstationary kernel combination, in: International Conference, 2006, pp. 553–560.

[23] G. Andrew, R. Arora, J. Bilmes, K. Livescu, Deep canonical correlation analysis, in: ICML, 2013, pp. 1247–1255.

[24] S. K. Gupta, D. Phung, B. Adams, T. Tran, S. Venkatesh, Nonnegative shared subspace learning and its application to social media retrieval, in: Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, 2010, pp. 1169–1178.

[25] S. Yu, K. Yu, V. Tresp, H. P. Kriegel, Multi-output regularized feature projection, IEEE Transactions on Knowledge and Data Engineering 18 (12) (2007) 1600–1613.

[26] R. K. Ando, T. Zhang, A framework for learning predictive structures from multiple tasks and unlabeled data, Journal of Machine Learning Research 6 (3) (2005) 1817–1853.

[27] S. Ji, L. Tang, S. Yu, J. Ye, A shared-subspace learning framework for multi-label classification., Acm Transactions on Knowledge Discovery from Data 4 (2) (2010) 890–895.

[28] X. Kong, M. K. Ng, Z. H. Zhou, Transductive multilabel learning via label set propagation, IEEE Transactions on Knowledge and Data Engineering 99 (3) (2013) 704–719.

[29] K. Q. Weinberger, L. K. Saul, Distance metric learning for large margin nearest neighbor classification, Journal of Machine Learning Research 10 (1) (2006) 207–244.

[30] G. Qiu, Color image indexing using btc, IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society 12 (1) (2003) 93–101.

[31] B. Zhang, Y. Gao, S. Zhao, J. Liu, Local derivative pattern versus local binary pattern: face recognition with high-order local pattern descriptor., IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society 19 (2) (2010) 533–544.

[32] Z. Guo, L. Zhang, D. Zhang, Rotation invariant texture classification using lbp variance (lbpv) with global matching, Pattern Recognition 43 (3) (2010) 706–719.

[33] X. Wang, T. X. Han, S. Yan, An hog-lbp human detector with partial occlusion handling, Proceedings 30 (2) (2009) 32–39.

[34] A. Mohamed, G. E. Dahl, G. Hinton, Acoustic modeling using deep belief networks, IEEE Transactions on Audio Speech and Language Processing 20 (1) (2012) 14–22.

[35] A. R. Mohamed, G. Dahl, G. Hinton, Deep belief networks for phone recognition 4.

[36] A. Mohamed, G. Hinton, G. Penn, Understanding how deep belief networks perform acoustic modelling, 2012, pp. 4273–4276.

[37] R. N. Le, Y. Bengio, Representational power of restricted boltzmann machines and deep belief networks., Neural Computation 20 (6) (1989) 1631–49.

[38] T. S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, Y. Zheng, Nus-wide: a real-world web image database from national university of singapore, in: ACM International Conference on Image and Video Retrieval, Civr 2009, Santorini Island, Greece, July, 2009, pp. 1–9.

[39] H. Mller, W. Mller, D. M. Squire, S. Marchand-Maillet, T. Pun, Performance evaluation in content-based image retrieval: overview and proposals, Pattern Recognition Letters 22 (5) (2001) 593–601.

[40] D. M. Blei, J. D. Mcauliffe, Supervised topic models, Advances in Neural Information Processing Systems 3 (2010) 327–332.

[41] Y. Zhuang, Y. Wang, F. Wu, Y. Zhang, W. Lu, Supervised coupled dictionary learning with group structures for multi-modal retrieval, in: AAAI Conference on Artificial Intelligence, 2013.

[42] Y. Wang, F. Wu, J. Song, X. Li, Y. Zhuang, Multi-modal mutual topic reinforce modeling for cross-media retrieval, in: Proceedings of the ACM International Conference on Multimedia, 2014, pp. 307–316.