A Unified Optimization Based Learning Method for Image Retrieval*

Hanghang Tong[1,†], Jingrui He[1,†], Mingjing Li[2], Wei-Ying Ma[2],
Changshui Zhang[3], Hong-Jiang Zhang[2]
[1,3]*Department of Automation, Tsinghua University, Beijing 100084, China*
[2]*Microsoft Research Asia, 49 Zhichun Road, Beijing 100080, China*
[1]*{walkstar98, hejingrui98}@mails.tsinghua.edu.cn,* [2]*{mjli, wyma, hjzhang}@microsoft.com*
[3]*zcs@tsinghua.edu.cn*

Abstract

In this paper, an optimization based learning method is proposed for image retrieval from graph model point of view. Firstly, image retrieval is formulated as a regularized optimization problem, which simultaneously considers the constraints from low-level feature, online relevance feedback and offline semantic information. Then, the global optimal solution is developed in both closed form and iterative form, providing that the latter converges to the former. The proposed method is unified in the senses that 1) it makes use of the information from various aspects in a global optimization manner so that the retrieval performance might be maximally improved; 2) it provides a natural way to support two typical query scenarios in image retrieval. The proposed method has a solid mathematical ground. Systematic experimental results on a general-purpose image database demonstrate that it achieves significant improvements over existing methods.

1. Introduction

Initial image retrieval is based on keyword annotation, which is a natural extension of text retrieval [2, 14]. The typical query scenario in such image retrieval systems is query by keyword (QBK). However, it suffers from several main difficulties, e.g., the large amount of manual labor required to annotate the whole database, and the inconsistency among different annotators in perceiving the same image [11].

To overcome these difficulties, an alternative scheme, content-based image retrieval (CBIR) is proposed in the early 1990's. The typical query scenario in such image retrieval systems is query by example (QBE). Its advantage over keyword based image retrieval lies in the fact that feature extraction can be performed automatically and the image's own content is always consistent. Despite the great deal of research work, its performance is far from satisfactory due to the well-known gap between low-level feature and high-level semantic concepts [3].

To narrow or bridge the gap, a great deal of work has been performed. From a learning point of view, these works can be categorized into two major groups: one is to search for appropriate metrics so that similarity measured from low-level feature can best approximates that from high-level semantics; the other is to incorporate high-level semantic information to learn better representation of images as well as the query concept.

To pursue an optimal similarity metric for low-level feature, many distance functions have been used, including perceptual distance function (DPF) [8], Earth Mover's Distance (EMD) [10], Manhattan ($L1$) distance, Euclidean ($L2$) [7], etc. However, these metrics are based on pair-wise distance calculation and might oversimplify the relationship among all the images in the database.

In terms of high-level semantic information incorporation, it makes use of the additional high-level semantic information provided by users to improve the performance of retrieval systems. According to the source of such information and corresponding learning technology, it can be categorized into online short-term learning and offline long-term learning.

The semantic information for on line learning comes from relevance feedback. State-of-the-art learning techniques can be classified into inductive and transductive ones according to whether unlabeled data is utilized in the training stage or not [3]. The goal of

*This work was performed at Microsoft Research Asia.
† The first two authors contribute equally to this paper.

Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)
1063-6919/05 $20.00 © 2005 **IEEE**

an inductive method is to create a classifier which separates the relevant and irrelevant images and generalizes well on unseen examples. One of the most effective inductive learning techniques is support vector machines (SVM) [16]. However, one major problem with inductive methods is the insufficiency of labeled examples, which might bring great degradation to the performance of the trained classifier. On the other hand, transductive methods aim to accurately predict the relevance of unlabeled images which are attainable during the training stage. A representative work belonging to this category is D-EM [15]. However, if the components of data distribution are mixed up, which is often the case in CBIR, the performance of D-EM will be compromised [15]. Despite the immaturity of transductive methods, we see with them great potential since they provide a way to solve the small sample size problem.

In contrast to online learning, fewer efforts have resolved offline learning issues. In both QBE and QBK, the semantic information accumulated in the log can be used for offline learning. Moreover, the initial manual annotation stage in QBK might also provide such information. The learning technology differs according to the specific query scenario.

For QBE, He et al proposed in [4] using singular value decomposition (SVD) method to infer a hidden semantic space from the log. They also proposed an online learning strategy based on SVM. However, in their method, the information from low-level feature is ignored. Therefore, the performance might be largely degraded, especially when the log data is limited.

For QBK, a representative work is based on classification [1]. Jing et al in [6] further extended this work by introducing labeling vector to online collect training samples and offline update the keyword models. However, the ratio of initial manually labeled images is relatively high to achieve a satisfactory result (ten percent in their experiments), which is still a heavy burden especially when the database is large. On the other hand, Jing et al also proposed combining online learning to refine the retrieval result. However, the combination scheme is somewhat heuristic. Moreover, their method will not work if only positive examples are provided in relevance feedback.

To deal with the limitations of existing methods in terms of similarity measurement and online learning, we applied a recently developed manifold ranking algorithm for the scenario of QBE in [3]. The proposed method evaluates the relevance between two images by exploring the relationship of all the data points in the low-level feature space. It also provides a natural way to perform online learning in a transductive manner. Experimental results demonstrated that it outperforms existing methods by a large margin. However, its application in QBK as well as incorporation with offline learning is not investigated.

In order to address all the drawbacks mentioned above in image retrieval, in this paper, we propose an optimization based learning method to integrate similarity measurement, online learning and offline learning in a unified manner. Different from most of previous work which is based on vector model, the proposed method is based on graph model, that is, the information from various aspects is firstly expressed as the relevance between two images or between an image and the query concept. Then, image retrieval is formulated as a regularized optimization problem which simultaneously considers the constraints from low-level feature, online feedback and offline information. Finally, the global optimal solution is developed in both closed form and iterative form, providing that the latter converges to the former. The proposed learning method is unified in the senses that 1) it makes use of the information from all aspects in a global optimization manner so that the retrieval performance might be maximally improved; 2) it provides a natural way to support both QBE and QBK.

The main contribution of this paper can be summarized as follows:

1. An optimization based learning method is proposed. It unifies low-level feature and high-level semantic concept learning in a global optimization manner. It supports both QBE and QBK.
2. Significant improvement in image retrieval performance is achieved.

The organization of the paper is as follows: the proposed optimization based learning method is presented in Section 2, we address the implementation details in Section 3; systematic experimental results are provided in Section 4; finally, we conclude in Section 5.

2. Optimization based learning method

2.1. Notation

Suppose we have totally n image in the database: $\{I_i, i=1,2,\cdots,n\}$ and q denotes the query. The proposed method is based on graph model, that is, the information from low-level feature and high-level semantics is denoted as the relevance between two images or between an image and the query concept:

Let $W^{low}=\left(W_{i,j}^{low}, i,j=1,2,\cdots,n\right)$ be an $n\times n$ affinity matrix constructed from low-level feature, where $W_{i,j}^{low}$ denotes the relevance between I_i and I_j measured

from low-level feature. Normalize W^{low} by $S^{low} = (D^{low})^{-1/2} W^{low} (D^{low})^{-1/2}$, where D^{low} is the diagonal matrix with (i,i)-element equal to the sum of the ith row of W^{low};

Let W^{off}, D^{off}, and S^{off} be defined similarly as above, except that they are constructed from offline high-level semantics;

Let $y^{on} = [y_i^{on}, i = 1,2,\cdots,n]^T$ be an $n \times 1$ vector, where y_i^{on} denotes the relevance between I_i and q measured from online relevance feedback;

Let $f = [f_i, i = 1,2,\cdots,n]^T$ be an $n \times 1$ ranking vector, where f_i denotes the total relevance between I_i and q measured simultaneously from low-leve feature, offline high-level semantics and online relevance feedback.

2.2. Optimization problem formulation

With the above notation, the learning task is to infer the ranking vector f from W^{low}, W^{off} and y^{on} as Eq.1. Once f is obtained, it can be used to rank all the images in the database (largest ranked first).

$$\{(W^{low}, D^{low}, S^{low}); (W^{off}, D^{off}, S^{off}); y^{on}\} \to f \qquad (1)$$

To maximally make use of S^{low}, S^{off} and y^{on} to improve retrieval performance, a 'good' ranking vector should be as consistent as possible with these information, that is to say, if two images are measured as relevant by S^{low} or S^{off}, they should receive similar ranking scores in f and vice versa. On the other hand, if an image is marked as highly relevant with the query by y^{on}, it should receive a high ranking score in f and vice versa. We consider all these constraints in the a regularized optimization framework by defining the following cost function with f:

$$Q(f) = \left\{ \mu \sum_{i,j=1}^{n} W_{i,j}^{low} \left\| \frac{1}{\sqrt{D_{i,i}^{low}}} \cdot f_i - \frac{1}{\sqrt{D_{j,j}^{low}}} \cdot f_j \right\|^2 + \eta \sum_{i,j=1}^{n} W_{i,j}^{off} \left\| \frac{1}{\sqrt{D_{i,i}^{off}}} \cdot f_i - \frac{1}{\sqrt{D_{j,j}^{off}}} \cdot f_j \right\|^2 + \varepsilon \sum_{i=1}^{n} \left\| f_i - y_i^{on} \right\|^2 \right\} \qquad (2)$$

The first, second and third items on the right hand of Eq.2 correspond to the constraints from S^{low}, S^{off} and y^{on}, respectively. The trade-off among these constraints is captured by the regularization parameters μ, η and ε, where $0 < \mu, \eta, \varepsilon < 1$ and $\mu + \eta + \varepsilon = 1$.

With the above optimization criterion, the optimal ranking vector f^* is achieved when $Q(f)$ is minimized:

$$f^* = \arg\min_f Q(f) \qquad (3)$$

2.3. Optimization problem solving

Differentiating $Q(f)$ with respect to f leads to the following optimal ranking score f^* in closed form:

$$f^* = (1 - \mu - \eta)(I - \mu S^{low} - \eta S^{off})^{-1} \cdot y^{on} \qquad (4)$$

Although the closed form for f^* is achieved, in some practical cases, the iterative form might be more preferable. We also develop an iterative solution for solving the optimization problem defined in Eq.2 and Eq.3:

$$f(t+1) = \mu S^{low} f(t) + \eta S^{off} f(t) + (1 - \mu - \eta) y^{on}, \quad where\ f(0) = y^{on} \qquad (5)$$

The relationship between the above two versions of optimal solutions can be given as[1]:

$$f^* = \lim_{t \to \infty} f(t) \qquad (6)$$

3. Image retrieval process: implementation details

To apply the proposed method to image retrieval, there are two graphs (one from low-level feature W^{low}; and the other from offline high-level semantic information W^{off}) and one vector y^{on} from online semantic information. Once constructed, W^{low} is fixed; while W^{off} and y^{on} are updated according to the additional semantic information obtained offline and online, respectively. We should also determine the regularized parameters in Eq.2.

3.1. Graph construction and update

The construction of W^{low} is similar with that in [3]:

1. Take each image as a vertex; calculate the K nearest neighbors for each point; and connect two points with an edge if they are neighbors.
2. Since $L1$ distance can better approximate the perceptual difference between two images than other popular Minkowski distances when using either color or texture representation or both [3], it is adopted to define the edge weights in W^{low}:

[1] The proof in this subsection is similar with that in [17, 18]. For the limited space, we will not provide details here.

$$W_{i,j}^{low} = \prod_{l=1}^{m} \exp\left(-\left|x_{il} - x_{jl}\right| / \sigma_l\right) \quad (7)$$

where x_{il} and x_{jl} are the lth dimension of x_i and x_j respectively; m is the dimensionality of the feature space; and σ_l is a positive parameter that reflects the scope of different dimensions; set $W_{i,i}^{low} = 0\ (i = 1,2,\cdots,n)$.

On the other hand, the construction and update of W^{off} can be performed as follows:

1. Initialize W^{off} as an $n \times n$ matrix with $W_{i,j}^{off} = 0$ $(i, j = 1,2,\cdots,n)$;
2. For every two image I_i and I_j $(i \neq j)$, if they are labeled with the same keyword (in the initial manual annotation stage in QBK) or marked as relevant simultaneously in the same query session, update $W_{i,j}^{off} \leftarrow W_{i,j}^{off} + 1$.

Note that in the case that there is no log data, or in QBK, only one image is manually labeled in the initial manual annotation stage, W^{off} is empty and the proposed method is simplified into the work in [3].

3.2. y^{on} setup: initial query

In QBE, if the query image is in the database, the element of y^{on} corresponding to the query image is set 1, while all the other elements are set 0. On the other hand, if the query image is not in the database, in order to apply Eq.4 or Eq.5, W^{low} and W^{off} should be firstly expanded by adding one row and one column corresponding to the query image. However, it might be time-consuming. For simplicity, we can only use low-level feature by $L1$ distance for the initial retrieval and all the element of y^{on} is set 0.

In QBK, y^{on} is constructed from the initial manual annotation stage: if an image is not labeled in this stage, the corresponding element in y^{on} is set 0. On the other hand, the labeled images are treated differently: if the keywords of an image cover the query, it is considered a relevant image and the corresponding element in y^{on} is set 1; otherwise, it is considered an irrelevant one and the corresponding element is set $-\gamma(0 \le \gamma \le 1)$. In this way, its influence is suppressed. The reason can be ascribed to the asymmetry between relevant and irrelevant images [3]: generally speaking, relevant images should make more contribution to the overall ranking score than irrelevant ones. Here the parameter γ controls the suppression extent: the smaller γ is; the less impact irrelevant images will have on the overall ranking score. If $\gamma = 1$, there is no suppression for irrelevant image; if $\gamma = 0$, the effect of irrelevant images is ignored.

3.3. y^{on} update: relevance feedback

In relevance feedback, the additional online semantic information can be used to update y^{on} for both QBE and QBK: for a positive image, the corresponding element in y^{on} is set 1; while for a negative image; the corresponding element is set $-\gamma(0 \le \gamma \le 1)$ for the same reason as discussed above.

As mentioned in the introduction section, if negative examples are unavailable or we only consider the positive examples, the method proposed in [6] will not work. However, it is not the problem for the proposed method.

Another important issue in relevance feedback is how to select unlabeled images for users' feedback so that the convergence to the query concept can be maximally speeded up. In [3], we proposed three active learning schemes. Namely, 1) to select the most positive images; 2) to select the most informative images; and 3) to select the most positive and inconsistent images. All of these schemes can be combined into the proposed method. Here, we simply adopt the first scheme since active learning is not the main focus of this paper.

3.4. Regularization parameter selection

Since $\mu + \eta + \varepsilon = 1$, there is actually two independent parameters needed to be set. Note that in Eq.4, the final ranking result will not be influenced by $(1 - \mu - \eta)$, therefore it is fixed to be 0.01. Thus, we only need to determine η ($0 < \eta < 0.99$). η reflects the trade-off of the constraints between low-level feature and offline high-level semantic information. Ideally, it should be adaptively set according to the relative contribution of W^{off} to the final ranking vector compared with W^{low}. Currently, it is roughly set according to the amount of offline data: the more offline information, the higher η is. We will pursue the more principled way to determine η in future work.

4. Experimental results

4.1. Experiment design

We have evaluated the performance of the proposed method using a general-purpose image database consisting of 5,000 Corel images. The images are

categorized into 50 groups, each having 100 images. Images belonging to the same group are considered to be relevant. The precision vs. scope curve is used to evaluate the performance of various methods.

Low-level feature has an important influence on the retrieval performance. However, we do not perform careful feature selection in this paper since what we want to propose is a general learning method which can be applied with any kind of feature or feature combination. In our current implementation, the features that we use to represent each image include color histogram [12], color correlogram [5], Tamura feature [13], and pyramid wavelet texture feature [9].

Besides the regularization parameters discussed in Section 4, there are four parameters need to be set: K, σ_l, γ and the iteration steps. The number of iteration steps is set to be 50 since we observe no improvement with more iterations. To determine the other parameters, a parametric study has been performed and the final parameters adopted are: $K = 100$; $\sigma_l = 0.05$ and $\gamma = 0.1$.

Relevance feedback (RF) is simulated as follows. For a query, 5 iterations of RF are carried out. At each iteration, the system examines top 5 images.

4.2. Experimental results

For limited space, we only present the result of QBE in this paper. In this scenario, to generate the log data, a small portion of the images in the database are used as queries. In each query session, the system examines the first top 20 images. After the log is generated, we use each image in the whole database as a query, and average the results over the 5,000 queries. The proposed method is compared with SVD-based method and 'SVD+SVM' [4].

First, the initial retrieval result is evaluated. In order to perform a systematic evaluation, we vary the percentage of training data, i.e. images used to generate the log data, and compare the average precision of top 20 retrieved images (P20) with that by SVD-based method [4]. The precision vs. the percentage of the training data curve is shown in Figure 1. From the figure, it can be seen that the proposed method outperforms SVD-based method by a large margin. Then, we fix the percentage of the training data to 5% and evaluate the effect of simultaneous learning from W^{low} and W^{off}. To this end, we compare the retrieval result with that by 1) setting W^{off} empty and using W^{low} only (LOW); and 2) setting W^{low} empty and using W^{off} only (OFF). The average precision vs scope is shown in Figure 2. From the figure, it can be seen that 1) the proposed method takes the advantage of both low-level feature and high-level semantic information so that it achieves a high performance; 2) even the curve of 'OFF Only ' outperforms that of 'SVD', indicating that in terms of utilizing the log data alone, the proposed scheme is more effective.

In relevance feedback, we fix the percentage of training data to be 5% and evaluate two situations: both positive and negative examples are available (PN); only positive examples are considered (OP). The average precision of top 20 retrieved images (P20) vs. iteration number is shown in Figure 3. The proposed method outperforms 'SVD+SVM' by a large margin. The reason might be that, in 'SVD+SVM', 1) the low-level information is totally ignored; 2) according to [4], the images which are not in the log will not receive the hidden semantic feature so that when the amount of the log is small, there is actually not any high-level information about many images in the database. On the other hand, if we compare Figure 1 and Figure 3, 'SVD+SVM' actually causes degradation in performance. Only after the system has accumulated enough labeled examples, can 'SVD+SVM' refine the retrieval result. This observation is consistent with the experimental results in [3]. On the other hand, the proposed method consistently increases the precision and outperforms 'SVD+SVM'.

5. Conclusions

In this paper, we have investigated image retrieval under a regularized optimization framework to make use of the information from both low-level feature and high-level semantics in a global optimization manner. Different from most of the existing methods, the proposed one is based on graph model in which the information from various aspects is expressed as the relevance between two images or between an image and the query concept. The proposed optimization criterion as well as optimization objective consider simultaneously the constraints from low-level feature, online feedback and offline semantic information. The global optimal solution is developed in both closed form and iterative form. Systematic experimental results demonstrate the effectiveness of the proposed method.

6. Acknowledgements

This work is supported by the project (60475001) of the National Natural Science Foundation of China. The authors would give thanks to Xing Zheng for valuable discussions.

7. References

[1] Chang, E., et al., "CBSA: Content-Based Soft Annotation for Multimodal Image Retrieval Using Bayes Point Machines", IEEE Trans on Circuits and Systems for Video Technology, January 2003, Volume 13, No. 1, pp.26-38.

[2] Chang, S.K. and Hsu, A., "Image Information Systems: Where do We Go from Here?", IEEE Trans. on Knowledge and Data Engineering, Oct. 1992, 4(5).

[3] He J., Li M., Zhang H.J., Tong H., and Zhang C., "Manifold-Ranking Based Image Retrieval", Proc. ACM International Multimedia Conference, 2004.

[4] He X., King O., Ma W.Y., Li M., and Zhang H.J., "Learning a Sementic Space from User's Relevance Feedback for Image Retrieval", IEEE Tran. on Circuits and Systems for Video Technology, January 2003, Vol. 13, No. 1.

[5] Huang, J., et al., "Image Indexing Using Color Correlograms", Proc. IEEE Conf. on Computer Vision and Pattern Recognition, 1997, pp. 762-768.

[6] Jing F., Li M., Zhang H.J., and Zhang B., "Keyword Propagation for Image Retrieval", Proc. IEEE International Symposium on Circuits and Systems, 2004

[7] Kokare, M., Chatterji, B.N., and Biswas, P.K., "Comparison of Similarity Metrics for Texture Image Retrieval", IEEE Conf. on Convergent Technologies for Asia-Pacific Region, 2003, vol. 2, pp. 571-575.

[8] Li, B., Chang, E., and Wu, C.T., "DPF-a Perceptual Distance Function for Image Retrieval", Proc. IEEE Int. Conf. on Image Processing, 2002, vol. 2, pp. 597-600.

[9] Mallat, S.G., "A Theory for Multiresolution Signal Decomposition: the Wavelet Representation", IEEE Trans. on Pattern Analysis and Machine Intelligence, 1989, vol. 11, pp. 674-693.

[10] Rubner, Y., Tomasi, C., and Guibas, L., "A metric for distributions with applications to image databases", Proc. IEEE Int. Conf. on Computer Vision, pp. 59-66, 1998.

[11] Shen, H.T., et al., "Giving meanings to WWW images", Proc. 4th ACM Int. Conf. on Multimedia, 2000.

[12] Swain, M., and Ballard, D., "Color indexing", Int. Journal of Computer Vision, 1991, 7(1): 11-32.

[13] Tamura H., Mori S., and Yamawaki T., "Textural Features Corresponding to Visual Perception", IEEE Trans. on Systems., Man and Cybernetics, June 1978, 8(6):460-472.

[14] Tamura, H. and Yokoya, N., "Image database systems: a survey", Pattern Recognition, 1984, Vol. 17, No. 1.

[15] Wu, Y., Tian, Q., and Huang, T., "Discriminant-EM algorithm with application to image retrieval", Proc. IEEE Conf. on Computer Vision and Pattern Recognition, vol. 1, pp. 155-162, 2000.

[16] Zhang, L., Lin, F., and Zhang, B., "Support Vector Machine learning for image retrieval", Proc. IEEE Int. Conf. on Image Processing, 2001, vol. 2, pp. 721-724.

[17] Zhou, D., et al., "Learning with local and global consistency", 18th Annual Conf. on Neural Information Processing Systems, 2003.

[18] Zhou, D., et al., "Ranking on data manifolds", 18th Annual Conf. on Neural Information Processing System, 2003.

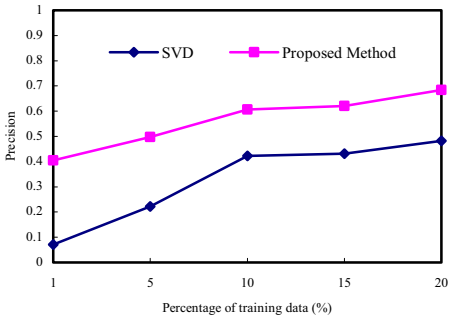

Figure 1. Systematic comparison of P20 under different size of training data.

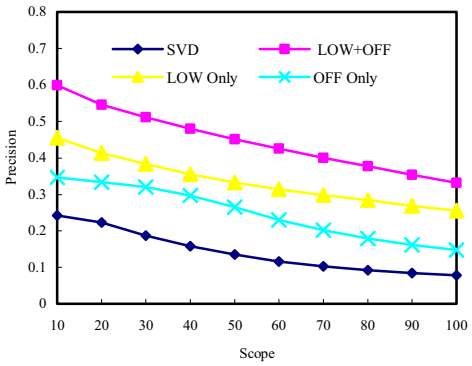

Figure 2. Evaluation simultaneous learning effect

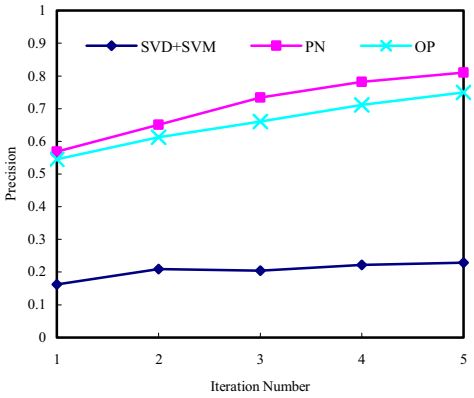

Figure 3. Comparison of relevance feedback