GRAPH CUTS BASED RELEVANCE FEEDBACK IN IMAGE RETRIEVAL

Lelin Zhang, Sidong Liu, Zhiyong Wang, Weidong Cai, Yang Song, David Dagan Feng

School of Information Technologies, The University of Sydney, NSW 2006, Australia

ABSTRACT

Relevance feedback (RF) allows users to be actively involved in the information retrieval process and has been widely used in various information retrieval tasks. While most existing RF methods in content-based image retrieval (CBIR) focus on visual features of individual images only, in this paper we formulate the relevance feedback process as an energy minimization problem. The energy function takes into account both the feature aspect of each image and the manifold structure among individual images. The solution of labelling images as relevant or irrelevant is obtained with the graph cuts method. As a result, our method enables flexibly partitioning the feature space and labelling of images and is capable of handling challenging scenarios (or queries). Experimental results demonstrate that our proposed method outperforms the popular RF methods.

Index Terms— Content-based image retrieval, relevance feedback, graph cuts, energy minimization, interactive retrieval

1. INTRODUCTION

With the explosive growth in web and personal image collections, image retrieval has become an active research direction in the past decades [1, 2]. Many existing systems retrieve images based on the metadata such as keywords, date/geolocation tags or descriptions associated with the images. However, relying only on metadata can be problematic, since it is often incomplete and noisy in large-scale image databases.

Content-Based Image Retrieval (CBIR) systems retrieve images based on the actual contents of images rather than their matadata, where an image is represented by visual features such as color, texture, or shape. The images are retrieved based on the similarity metrics between their visual features and that of a query image. However, the "semantic gap" between these low-level features and high-level concepts is a challenging issue. In addition, different users may have different preferences even for the same query image [2]. Fixed retreival mechanisms may not capture such dynamics.

The work presented in this paper was partially supported by ARC (Australian Research Council) grants.

Relevance feedback (RF) is a powerful tool to fill in the semantic gap and meet the demand of user's preference, thus to improve the performance of CBIR systems [3, 4]. In a RF process, after the initial results are displayed, a user is able to label some of the images as relevant (positive) or irrelevant (negative). The CBIR system then refines the retrieval results based on these feedback samples. The process is carried out iteratively, and the results are improved by gradually learning the query.

In general, existing RF techniques can be classified into two categories: query reformulation based and classification based methods. Query reformulation based methods assume that a specific user's need can be modeled as an optimal similarity function. This can be achieved by learning a similarity metric [3, 5, 6, 7, 8, 9, 10], or adjusting the the query in the feature space [11, 12].

However, the gap between low-level features and the user's search semantic is so wide that any such linear (or even non-linear) function may not describe the user's need in an optimal way. Classification based methods aim to model the user's preference into a two-class learning problem, where a classifier is employed to take the user's feedback as learning samples, and to predict candidate images as relevant/irrelevant. The early works include decision trees [13] and Bayesian [14]. Recently, Support Vector Machines (SVM) based methods [15, 16, 17, 18] have shown promising results. However, most of the classification based methods make use of only the labeled samples as individual features and ignore the manifold structure of the features between both labeled and unlabeled samples. Generally, if the neighbors of an image are relevant to user's preference, the image itself is assumed to be relevant as well. We believe such manifold can be used to facilitate the classification.

Therefore, in this paper we propose a novel classification based method by taking both the features and the manifold structures into consideration. We define two cost functions to model the two types of information, and aim to find optimal classification with minimum global energy. Finally, graph cuts [19] is employed to obtain the solution efficiently. As a result, our proposed method provides more flexible partitioning of the feature space that describes user's need more accurately. Also, our method is able to incorporate more advanced features and similarity functions.

2. OUR PROPOSED METHOD

2.1. System Overview

The CBIR system that integrates the proposed RF method is illustrated in Fig. 1. For a given query image in the first iteration, the visual features are extracted and compared with the features of the candidate images using Euclidean metric, which produces a list of results ranked by similarity. The user can provide feedback to indicate his/her preference by labeling the images as relevant (positive) or irrelevant (negative). The system constructs a graph that encodes the user's feedback, the features, and the relationships between features with two cost functions. Graph cuts is employed to find the optimal cuts of the graph, which separate the images into relevant/irrelevant subgraphs. In order to produce a ranked result list from the boolean prediction, constrained similarity measurement (CSM) [16] is employed, where images predicted as relevant will be ranked above the irrelevant ones, and the images with the same label will be ranked by the original (Euclidean) metric. As the retrieval process iterates and more feedback is gathered, the system can find better cuts of the graph, thus obtain improved prediction for the user's preference.

2.2. Problem Formulation

Given a set of candidate images V, we want to label each image $v \in V$ as relevant (P) or irrelevant (N). The predicted label $L_v \in \{P, N\}$ is determined by the contents of v itself and its neighborhood $N_k(v)$, where $N_k(v)$ is k most similar images of v.

We model these two types of information using a graph structure $G = \langle V, E \rangle$. We consider each image as a vertex $v \in V$, and it is connected to its k nearest neighbors by edges $e \in E$. Given a set of labels L on V, we define the global energy $E(L|G)$ as:

$$E(L|G) = \overbrace{\sum_{v \in V} D_v(L_v)}^{\text{Data Cost}} + \overbrace{\sum_{uv \in E} S_{uv}(L_u, L_v)}^{\text{Smooth Cost}} \quad (1)$$

where $D_v(L_v)$ encodes the property of v, $S_{uv}(L_u, L_v)$ encodes the neighborhood relationships of v, and $u \in N_k(v)$.

We regularize our problem in terms of energy minimization. The optimal prediction of labels $\hat{L}$ can be found by minimizing the global energy, i.e.:

$$\hat{L} = \arg\min_{L} E(L|G) \quad (2)$$

2.3. Label Prediction by Energy Minimization

Our current implementation actually makes a double use of the labeled images. Firstly, they provide the hard constraints that a solution L must satisfy. Secondly, they are used to derive the soft constraint of the data cost function, which is defined as the mean of Gaussian between the image v and all relevant (or irrelevant) samples:

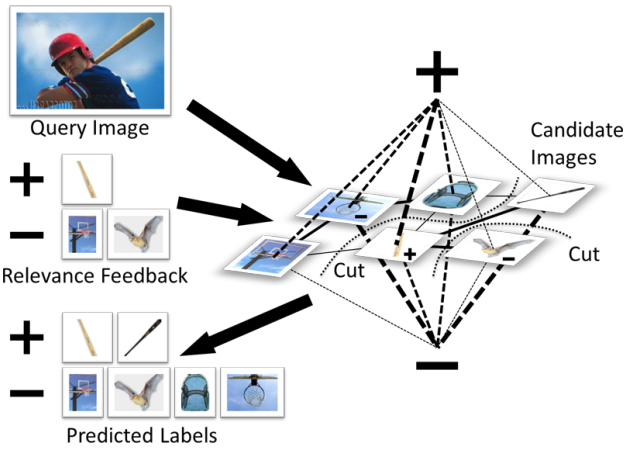

Fig. 1. An illustration of label prediction with graph cuts for 6 candidate images. The labeled images are marked with a "+" (relevant) or "-" (irrelevant). The data-cost is represented by dashed edges and the smooth-cost is represented by solid edges. The thickness of each edge reflects the cost. The graph cuts algorithm finds the solution with minimum global cost and outputs the corresponding labels as predicted results.

$$D_v(L_v) = \begin{cases} 1 + \max\limits_{u \in V} \sum\limits_{uw \in E} S_{uw}(L_u, L_w) & \text{if } L_v = \mathcal{L} \\ 0 & \text{if } L_v \neq \mathcal{L} \\ \alpha \sum_{u \in L_v} e^{-\gamma \|v-u\|^2 / \|L_v\|} & \text{if } v \text{ is unlabeled} \end{cases} \quad (3)$$

where $\mathcal{L}$ is the label of v when v is labeled, α is the trade-off parameter between data cost and smooth cost, γ is the variance parameter of the Gaussian kernel. The first and second lines effectively associate a large cost (more than the maximum smooth cost associated to any image) to the given label of v.

We further model the smooth cost of a neighborhood of v with the Cosine similarity:

$$S_{uv}(L_v, L_u) = \sum_{u \in N_k(v)} \frac{u \cdot v}{\|u\|^2 \|v\|^2} \quad (4)$$

The Gaussian is used for the data cost to make it comparable with the Gaussian Kernel used in SVM. The Cosine similarity is used for the smooth cost because it provides a complementary view besides the relationships defined by Gaussian.

Note that the above choice of cost functions may not be optimal. In fact, they are chosen to demonstrate the properties of our method and to compare with other methods. The ability of our method to handle different energy functions left us the possibility to incorporate more advanced similarity metrics, thus further improve the performance.

Now that both data cost and smooth cost are defined, we employ graph cuts [20] to solve the energy minimization in

Eq. (2) by cutting the graph into relevant/irrelevant subgraphs to obtain the optimal label prediction.

3. EVALUATION

3.1. Experimental Settings

For the experiments we use the first 30 object categories of the Caltech-256 data set [21] comprising 3,646 images. Full list of selected categories can be found in Fig. 3. Since the images are collected from Google and Picsearch, the images exhibit great variety. The query is performed in a query by example manner, where both query and candidate images come from the data set. For a given query image, we treat the images from the same category as relevant, and others as irrelevant.

We extract three different features from the images: a 64-dimension HSV color histogram [22] with 4 bins for hue, saturation, and value, a 128-dimension color coherence vector (CCV) [23], and a 48-dimension texture feature extracted by pyramidal wavelet transform (PWT). The value of each dimension of all three features are normalized to the range of 0 to 1.

In the experiment, we follow the greedy user model [4] which assumes users are impatient, such that: 1) they expect the best possible results after each RF iteration; 2) they would only label a small number of images in the top results. In the experiment, only top 5 relevant and top 5 irrelevant images in top 50 results will be used as feedback. The images already labeled by the user will be excluded from the results of the following iterations. Consequently, the samples of different iterations do not overlap.

We compare the proposed graph cuts (GC) based method with the classification based Support Vector Machines (SVM) method [16] and the query reformulation based weight updating (WU) method [3]. For GC and SVM, we combine the features to form a 240-dimension feature vector and use Euclidean distance to compute the ranking. For WU, three Euclidean distances are computed from the three features and normalized to the range of 0 to 1, then combined with different weights to form a final distance. For all the methods, we report the parameters that give the best performance. For GC, we use the implementation of [20] to solve the problem in Eq. (2). We set $\alpha = 50,000$ and $\gamma = 0.3$ for Eq. (3), and $k = 20$ for Eq. (4). For SVM, we conduct the classification using LIBSVM library [24]. We use Gaussian kernel $K(x, y) = \exp(-\gamma\|x - y\|^2)$ with $\gamma = 1$ and the penalty term $c = 10$.

3.2. Results

We conducted 10 RF iterations for each query and calculated the average precision for the top 10, 20, 30, 100 results of all 3,646 queries.

As shown in Fig. 2, our proposed method consistently outperforms SVM in all ranges, which indicates that it is able to predict relevant/irrelevant images more accurately.

Table 1. Performance after 10 iterations for different groups of queries in top 30 results.

	Initial	**WU**	**SVM**	**GC**
All	0.119	0.344(+189%)	0.467(+292%)	0.500(+317%)
Hard	0.085	0.257(+201%)	0.373(+338%)	0.398(+367%)
Easy	0.163	0.458(+181%)	0.587(+260%)	0.625(+284%)

The overall performance of WU (top20, 30 and 100) is worse than GC and SVM. Also, the performance growth of WU slows down at a much faster rate compared to GC and SVM. After the 4th iteration, WU achieved little further improvement no matter in top 10, 20, 30, or 100 results. This suggests that the ability of WU to model the query is weaker than GC and SVM, since WU only learns a linear combination of features and similarity functions. However, we note that WU greatly improves the topmost results (top 10) during the first few iterations (1 to 4), beating both GC and SVM at the same stage. The reason lies on the CSM re-ranking mechanism used in classification based method, which ranks the topmost results (all predicted to be relevant) by the ineffective Euclidean metric, and it is obviously sub-optimal. This issue is more distinct in top results and early iterations, where few feedback samples are available and the classifiers are yet to achieve their full power. This issue could be rescued if a better similarity metric is used originally.

The performance of individual categories in top 30 results is shown in Fig. 3. Among the 30 test categories, our proposed method performs best in 20 categories while SVM performs best in 10 categories, WU performs worst in all categories. If we separate the categories into hard (initial precision < 0.1) and easy (initial precision ≥ 0.1) groups, we can see that GC performs best in both groups, and the performance gain ratio on hard queries is even larger than that on easy queries, as shown in Table 1. This validates the superior ability of our proposed GC method in learning complex query models.

4. CONCLUSION AND FUTURE WORK

In this paper we present a novel RF algorithm by taking both the features and the manifold structures into account and formulate the problem in terms of energy minimization. Two cost functions are defined to encode the two types of information. Therefore, a more comprehensive query model can be learnt. We employ graph cuts to solve the minimization problem and obtained the optimal labels. Experimental results based on a subset of Caltech-256 data set show a significant improvement in terms of precision as compared to SVM and WU. The proposed framework is flexible to incorporate more advanced cost functions. In the future, we will further investigate how different cost functions interact with each other.

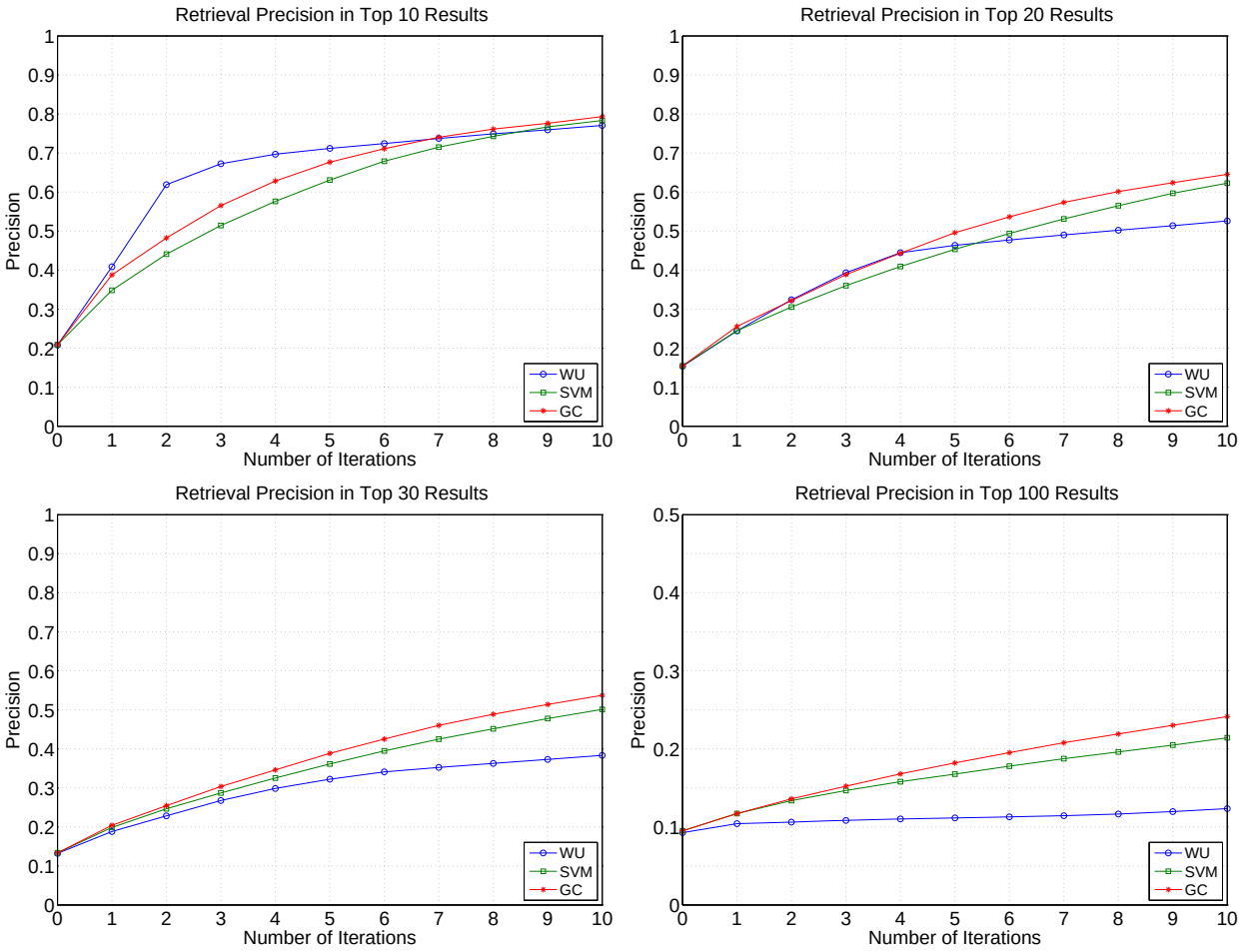

Fig. 2. Performance of the proposed GC method compared with SVM and WU.

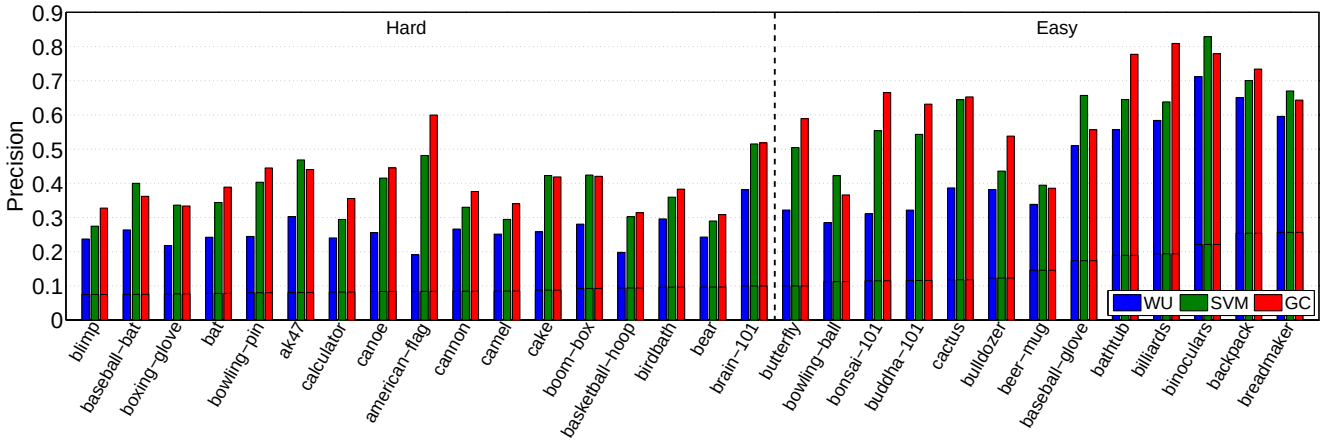

Fig. 3. Performance improvement of individual categories in top 30 results. The final results after 10 iterations are overlaid against the initial results. The categories are sorted by their initial precisions such that categories on the right are "easier" than the ones on the left.

5. REFERENCES

[1] D. D. Feng, W. C. Siu, and H. J. Zhang, *Multimedia Information Retrieval and Management: Technological Fundamentals and Applications*, Springer, Berlin; New York, 2003.

[2] D. Datta, R.and Joshi, J. Li, and J. Z. Wang, "Image retrieval: Ideas, influences, and trends of the new age," *ACM Computing Surveys*, vol. 40, no. 2, pp. 5:1–5:60, May 2008.

[3] Y. Rui, T. S. Huang, M. Ortega, and S. Mehrotra, "Relevance feedback: A power tool for interactive content-based image retrieval," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 8, no. 5, pp. 644–655, Sept. 1998.

[4] X. S. Zhou and T. S. Huang, "Relevance feedback in image retrieval: A comprehensive review," *Multimedia Systems*, vol. 8, pp. 536–544, 2003.

[5] Y. Rui and T. S. Huang, "Optimizing learning in image retrieval," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2000, vol. 1, pp. 236–243.

[6] J. Laaksonen, M. Koskela, S. Laakso, and E. Oja, "Self-organising maps as a relevance feedback technique in content-based image retrieval," *Pattern Analysis & Applications*, vol. 4, no. 2-3, pp. 140–152, 2001.

[7] Z. Wang, Z. Chi, D. Feng, and A. C. Tsoi, "Content-based image retrieval with relevance feedback using adaptive processing of tree-structure image representation," *International Journal of Image and Graphics*, vol. 03, no. 01, pp. 119–143, 2003.

[8] X. Tian, D. Tao, X. Hua, and X. Wu, "Active reranking for web image search," *IEEE Transactions on Image Processing*, vol. 19, no. 3, pp. 805–820, 2010.

[9] C. D. Ferreira, J. A. Santos, R. da S. Torres, M.A. Gonçalves, R. C. Rezende, and W. Fan, "Relevance feedback based on genetic programming for image retrieval," *Pattern Recognition Letters*, vol. 32, no. 1, pp. 27–37, 2011.

[10] X. Tian, D. Tao, and Y. Rui, "Sparse transfer learning for interactive video search reranking," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 8, no. 3, pp. 26:1–26:19, Aug. 2012.

[11] D. Tao, X. Tang, X. Li, and Y. Rui, "Direct kernel biased discriminant analysis: A new content-based image retrieval relevance feedback algorithm," *IEEE Transactions on Multimedia*, vol. 8, no. 4, pp. 716–727, Aug. 2006.

[12] W. Bian and D. Tao, "Biased discriminant Euclidean embedding for content-based image retrieval," *IEEE Transactions on Image Processing*, vol. 19, no. 2, pp. 545–554, Feb. 2010.

[13] S. D. MacArthur, C. E. Brodley, and Shyu C. R., "Relevance feedback decision trees in content-based image retrieval," in *IEEE Workshop on Content-based Access of Image and Video Libraries*, 2000, pp. 68–72.

[14] I. J. Cox, M. L. Miller, T. P. Minka, T. V. Papathomas, and P. N. Yianilos, "The bayesian image retrieval system, PicHunter: Theory, implementation, and psychophysical experiments," *IEEE Transactions on Image Processing*, vol. 9, no. 1, pp. 20–37, Jan. 2000.

[15] S. Tong and E. Chang, "Support vector machine active learning for image retrieval," in *ACM International Conference on Multimedia*, New York, NY, USA, 2001, pp. 107–118, ACM.

[16] A. K. Guo, G. D.and Jain, W. Y. Ma, and H. J. Zhang, "Learning similarity measure for natural image retrieval with relevance feedback," *IEEE Transactions on Neural Networks*, vol. 13, no. 4, pp. 811–820, July 2002.

[17] D. Tao, X. Tang, X. Li, and X. Wu, "Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 7, pp. 1088–1099, July 2006.

[18] R. Min and H. D. Cheng, "Effective image retrieval using dominant color descriptor and fuzzy support vector machine," *Pattern Recognition*, vol. 42, no. 1, pp. 147–157, 2009.

[19] Y. Y. Boykov and M. P. Jolly, "Interactive graph cuts for optimal boundary & region segmentation of objects in n-d images," in *IEEE International Conference on Computer Vision*, 2001, vol. 1, pp. 105–112.

[20] A. Delong, A. Osokin, H. N. Isack, and Y. Boykov, "Fast approximate energy minimization with label costs," *International Journal of Computer Vision*, vol. 96, pp. 1–27, 2012.

[21] G. Griffin, A. Holub, and P. Perona, "Caltech-256 object category dataset," Tech. Rep. 7694, California Institute of Technology, 2007.

[22] M. J. Swain and D. H. Ballard, "Color indexing," *International Journal of Computer Vision*, vol. 7, pp. 11–32, 1991.

[23] G. Pass, R. Zabih, and J. Miller, "Comparing images using color coherence vectors," in *ACM International Conference on Multimedia*, New York, NY, USA, 1996, pp. 65–73, ACM.

[24] C. C. Chang and C. J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011.