LEARNING A WEIGHTED SEMANTIC MANIFOLD FOR CONTENT-BASED IMAGE RETRIEVAL

Ran Chang[a], Zhongmiao Xiao[a], KokSheik Wong[b], and Xiaojun Qi[a]

[a]ran.c@aggiemail.usu.edu, [a]zhongmiao.xiao@aggiemail.usu.edu and [a]Xiaojun.Qi@usu.edu
Department of Computer Science, Utah State University, Logan, UT 84322-4205
[b]koksheik@um.edu.my
Faculty of Computer Science & Information Technology, University of Malaya, Malaysia

ABSTRACT

We propose a novel weighted semantic manifold ranking system for content-based image retrieval. This manifold builds a more accurate intrinsic structure for the proper image space by combining visual and semantic relevance relations. Specifically, we apply the learning mechanism to capture users' semantic concepts in clusters and extract high-level semantic features for each database image. We then incorporate the reliability score, the fuzzy membership, and the composite low-level and high-level relation into the traditional affinity matrix to construct a weighted semantic manifold structure. We finally create an asymmetric relevance vector to propagate positive and negative labels via the proposed manifold structure to images with high similarities. Extensive experiments demonstrate our system outperforms other manifold systems and learning systems in the context of both correct and erroneous feedback.

Index Terms— CBIR, semantic clusters, semantic features, weighted semantic manifold

1. INTRODUCTION

Current content-based image retrieval (CBIR) techniques focus on utilizing users' interactions to bridge the semantic gap. Retrieval techniques based on relevance feedback (RF) [1] are generally considered as promising interactive approaches to formulate user query and improve retrieval performance. However, most RF techniques ignore the manifold structure of image features. As a result, the latest trend has been shifted to RF-based transductive learning, which recovers the intrinsic structure for a proper image space, spreads labeled and unlabeled images as a manifold, and derives the similarity measure based on the learned manifold. Here, we review several representative transductive learning techniques in CBIR.

He *et al.* [2] propose the generalized manifold ranking based image retrieval algorithm (gMRBIR) to represent image relationships as a graph and propagate labeled information among images using the graph structure. They exploit the distribution of unlabeled images to improve ranking and retrieval accuracy. He *et al.* [3] use Laplacian eigenmaps to preserve geodesic distances between image pairs along the manifold. They then use neural networks in online retrieval to map low-level features to high-level semantics for inferring the semantics of a new image. Cai *et al.* [4] incorporate a locality preserving regularizer and user's RF into the manifold structure to learn a better classification function. Wang *et al.* [5] apply the affinity propagation clustering algorithm to reduce the manifold graph while preserving its structure. However, its retrieval performance may not be optimal since clustering results cannot match users' high-level semantic concepts. Bian and Tao [6] combine the biased discriminative Euclidean embedding with the manifold regularization-based item to discover the manifold structure for better classification. Chang and Qi [7] propose a semantic clusters-based manifold structure to represent image relationships for better retrieval performance. All these manifold learning techniques improve retrieval performance after each iterative step. However, they are sensitive to noise inherent in the natural images. They are also sensitive to users' erroneous RF due to the propagation of wrong labels. Moreover, they do not study the accumulated feedback from multiple query sessions to improve the manifold structure.

In this paper, we propose a novel weighted semantic manifold that builds a more accurate intrinsic structure for the proper image space by combining low-level visual similarity and high-level semantic relevance relations. First, we apply the learning mechanism to capture users' semantic concepts in clusters and apply a minimum-distance-based strategy to assign each non-clustered image into an estimated cluster. These clusters approximately divide database images into meaningful semantic categories to facilitate future learning. Second, we extract high-level semantic features of each database image based on users' historical retrieval experiences. These features are used to estimate the high-level semantic relations among images. Third, we incorporate the reliability score, the fuzzy membership, and the composite low-level and high-level relation into the traditional affinity matrix to build the semantic manifold structure. This incorporation suppresses the noisy propagation of erroneous feedback and strengthens the propagation powers of images with more relevant semantic information. Fourth, we construct the asymmetric relevance vector and propagate ranking scores of its labeled images to unlabeled images via the manifold structure. This assignment ensures the propagation on positive images is dominated and helps unlabeled images to obtain proper ranking scores. The rest of the paper is organized as follows: Section 2 presents our proposed semantic manifold learning approach. Section 3 compares our CBIR system with four systems. Section 4 draws conclusions and presents future directions.

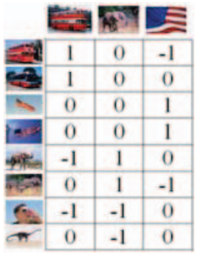
2. WEIGTHED SEMANTIC MANIFOLD APPROACH

Unlike the traditional manifold systems that exclusively use the low-level visual similarity to build the image space, our semantic manifold system builds a more accurate intrinsic structure by adding high-level semantic relevance relations. Low-level relations are evaluated by applying the Minkowski-based distance on low-level visual features. High-level relations are evaluated by applying the semantic correlation-based distance on high-level semantic features, which are learned through past retrieval experiences.

In our system, each image is represented by both low-level and high-level features. Low-level features contain color moments, edge direction histogram, and wavelet domain entropy-based textures. Initially, high-level features of each image are empty since no knowledge has been learned. They are updated after each query session. As a result, the dimensionality of the high-level features is N×p, where N is the total number of images in the database and p is the number of queries submitted so far. For each query session, we construct a new semantic cluster using all images marked as relevant to the query by the user in all iterations. At the end of the query session, the dimension of high-level semantic features is incremented by 1 and all values in the newly added dimension are initially set to 0's. Based on the user's feedback, all relevant images have their respective slots incremented by 1 and all irrelevant images have their respective slots decremented by 1. Each newly created semantic cluster may be combined with any existing semantic clusters until all clusters have no significant overlapping and are distinct to each other.

In order to maximize the semantic information inherited in high-level semantic features, we deliberately ensure retrieved images will not be returned in the following RF iterations in the off-line training process. We also return 25 images per iteration to allow users to conveniently label the relevancy of returned images in a screen.

Fig. 1 shows high-level semantic features of each of eight database images after using three images (bus, elephant, and flag) as queries to perform the retrieval process at the training stage. In this example, the user labels two images as positive (marked by 1's) and two images as negative (marked by -1's) for each query. Each column stores the query's semantic information learned from the user's RF. All positively labeled images in a column form a semantic cluster. Each row corresponds to high-level semantic features of an image. It clearly shows that the dimensionality of high-level semantic features grows as more queries are performed. For example, the high-level semantic features of the first database image are (1, 0, -1) after performing three queries.



	1	0	-1
	1	0	0
	0	0	1
	0	0	1
	-1	1	0
	0	1	-1
	-1	-1	0
	0	-1	0

Fig. 1: Illustration of semantic clusters and semantic features

We combine two semantic clusters if the number of images coexisting in both clusters is more than half of the total number of images in the smaller cluster. These users' feedback-based clusters achieve better clustering results since humans tend to classify objects into semantic categories and remember how well each object belongs to each category [8]. However, some database images are not assigned to any semantic cluster since they are not returned or not positively labeled in any query session. To this end, we apply a minimum-distance-based strategy to assign each non-clustered image to its estimated semantic cluster. Specifically, we find the Euclidean distance-based nearest neighbor of the non-clustered image among all the images within the semantic clusters. We then assign the non-clustered image to the cluster that contains this nearest neighbor. In this way, each non-clustered database image is assigned to a known semantic cluster.

At the end of the training phase, we build the weighted semantic manifold by incorporating three additional pieces of information into the affinity matrix $W = [w_{ij}]_{N\times N}$ as used in the traditional manifold ranking system. Here, each element w_{ij} in W represents the similarity between the i^{th} and j^{th} images in the database and N is the total number of images in the database. The resultant W will be directly used in the online retrieval phase without any further update. The three additional pieces of information are:

1) The Reliability Score: It measures the importance of each image in the manifold structure. The higher the score, the more important the image is, and the more propagation power of the image. Since non-positively labeled images are assigned to their clusters based on the low-level Euclidean distance, we are not sure about their importance in the manifold structure and empirically set their reliability scores to equal to a small value, e.g., 0.05. For each positively labeled image i, its reliability score r_i is computed as follows:

$$r_i = \exp\left(1 - \frac{d_{i,Rep(i)}}{A}\right) \tag{1}$$

where $d_{i,Rep(i)}$ is the Euclidean distance between low-level features of image i and $Rep(i)$, which is the representative image of the cluster that image *i* belongs to. This representative image has the closest Euclidean distance to the centroid of all positively labeled images within image *i*'s cluster. *A* is the average of the Euclidean distances between all positively labeled images and their respective representative image. This computation ensures that a positively labeled image that is closer to the representative image of its cluster has a higher reliability score. For each cluster, its representative image has the highest reliability score and therefore has the highest importance and propagation power.

2) The Fuzzy Membership: It measures the membership of each image in its assigned semantic cluster. The higher the membership, the more important the image is, and the more propagation power of the image. Since the jointly, positively labeled images in a search session likely contain similar semantic content, we group them in a cluster and assign them a fuzzy membership of 1's. For each non-positively labeled image *i*, its fuzzy membership μ_i is computed as follows:

$$\mu_i = \frac{\frac{1}{n}\sum_{k=1}^{n} d_{P_k,Rep(i)}}{d_{i_i,Rep(i)}} \qquad (2)$$

where P_k denotes a positively labeled image that is in image *i*'s cluster and *n* is the total number of positively labeled images in image *i*'s cluster. This computation ensures that a non-positively labeled image that is closer to the representative image of its cluster has a larger fuzzy membership. A linear normalization function is applied on μ_i's for all non-positively labeled images to ensure that its values fall in the small range of [0, 0.25].

3) The Composite Relation: It measures the relation between two images by combining the low-level feature-based visual similarity and the high-level feature-based semantic relevance. The composite relation between images *i* and *j* is computed as follows:

$$Relation_{i,j} = (1-\eta)\times d_{i,j} + \eta\times(1-S_{i,j}) \qquad (3)$$

Here, η is the contribution factor of high-level semantic features; $d_{i,j}$ denotes the Euclidean distance between normalized low-level features of images *i* and *j*; and $S_{i,j}$ denotes the high-level semantic relevance relation between images *i* and *j*, which is computed by the semantic-correlation-based distance:

$$S_{i,j} = HSF_i \bullet HSF_j = \sum_{k=1}^{m} HSF_i(k)\times HSF_j(k) \qquad (4)$$

where HSF_i and HSF_j respectively represent the semantic features of images *i* and *j*, $HSF_i(k)$ and $HSF_j(k)$ respectively are the k^{th} element of the semantic features of images *i* and *j*, *m* is the dimensionality of semantic features of each image, and the × operation is defined as follows:

$$HSF_i(k)\times HSF_j(k) = \begin{cases} 1 & if\ HSF_i(k)=1, HSF_j(k)=1 \\ -1 & if\ HSF_i(k)\times HSF_j(k)=-1 \\ 0 & otherwise \end{cases} \qquad (5)$$

This operation yields 1's when two images share the same semantics, -1's when two images have different semantics, and 0's when no semantics is learned for either image.

Our semantics-based affinity matrix is computed below by incorporating the three additional pieces of information:

$$w_{ij} = r_i \times r_j \times \mu_i \times \mu_j \times \exp\left(-\frac{Relation_{i,j}{}^2}{2\sigma^2}\right) \qquad (6)$$

where σ is the overall variance of image features.

For each online query request, we propagate ranking scores of positively and negatively labeled images collected during RF iterations to unlabeled images via the proposed semantic manifold structure. The propagated ranking scores are then used as the similarity scores between query and database images. We initially encode a relevance vector $Y=[y_i]_{N\times1}$by setting the row corresponding to query as 1's and setting the remaining elements as 0's. If query is a positively labeled image in a cluster, we also set the rows corresponding to all the other positive images in this cluster as 1's. The relevance score of each image is determined by the propagation of *Y* through $M=(1-\alpha D^{-1/2}WD^{-1/2})^{-1}$, where *D* is a diagonal matrix with the (*i*,*i*)-element being the sum of the i^{th} row of *W* and α is a parameter in [0, 1). That is, we compute the relevance scores for all images $P=[p_i]_{N\times1}$by multiplying *M* with *Y* and return *n* images with the highest scores. Based on the user's RF, we update $Y=[y_i]_{N\times1}$by:

$$y_i = \begin{cases} 1 & \text{if the ith image is judged as relevant} \\ -0.25 & \text{if the ith image is judged as irrelevant} \end{cases} \qquad (7)$$

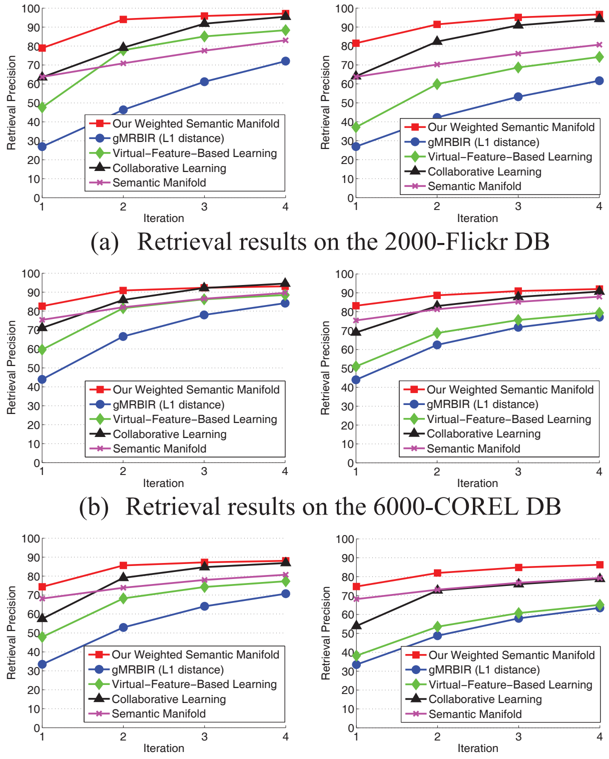
The proposed manifold structure *M* is then multiplied with this updated *Y* to compute relevance scores for the next round. This process continues for a few iterations or until the user is satisfied with retrieval results. This asymmetric assignment ensures the propagation on negatives will not be dominated since negative images do not provide sufficient information as positive images.

3. EXPERIMENTAL RESULTS

We tested our proposed system on the 2000-Flickr DB, the 6000-COREL DB, and the combined 2000-Flickr and 6000-COREL DB. Flickr and COREL DBs contain 20 and 60 categories with 100 images per category, respectively.

In our system, each image is represented by a 100 dimensional low-level feature vector. All pertinent parameters are set as follows: η = 0.7, σ = 0.05, α = 0.99.

To facilitate the evaluation process, we designed an automatic feedback scheme to construct the semantic manifold structure by performing query sessions using 10% unique, randomly chosen database images in the training process. For each query session, our system performed 4 iterations and returned top 25 images per iteration. A retrieved image is considered to be relevant if it belongs to the same category as the query. The retrieval precision is computed as the ratio of the relevant images to the total returned images. We compared our system with L_1-based gMRBIR [2], long-term virtual-feature-based [9], long-term collaborative learning-based [10], and our semantic manifold-based CBIR systems [7] on three DBs. Fig. 2 shows the average retrieval precision of these systems in the context of having no erroneous feedback and having a level of 5% erroneous feedback. To introduce the noise, we let the simulated "user" misclassify some relevant images as irrelevant and irrelevant images as relevant. The remaining 90% of the database images are used as queries for all experiments. It clearly shows that our proposed system outperforms two peer systems and the other two long-term systems and is more resilient to erroneous feedback since the other four systems significantly drop the precision in all iterations. Specifically, at the last iteration, our system respectively achieves the average retrieval accuracy of 98.5%, 94.8%, and 89.8% on 2000, 6000, and 8000 DBs when correct RF is involved; it respectively achieves the average retrieval accuracy of 97.8%, 92.9%, and 87.8% on 2000, 6000, and 8000 DBs when erroneous RF is involved. This noise resilient feature results from robust, meaningful semantic clusters and rich semantic features learned in the training process.



(a) Retrieval results on the 2000-Flickr DB

(b) Retrieval results on the 6000-COREL DB

(c) Retrieval results on the 8000-image DB

Fig. 2: Comparison of five CBIR systems on three DBs with correct (left) and 5% erroneous (right) RF.

In our experiments, our system generates 39, 106, and 213 semantic clusters for 2000-, 6000-, and 8000-image databases when using the randomly chosen 10% of the database images as training queries to perform the query task, respectively. These cluster numbers are in accordance with the number of semantic categories in their corresponding image database. As a result, the merged semantic clusters approximately divide database images into meaningful semantic categories, which represent the intrinsic structure of the database by capturing the intention of multiple users.

4. CONCLUSIONS AND FUTURE WORK

We propose a novel weighted semantic manifold ranking system for CBIR. Our semantic manifold builds a more accurate intrinsic structure for the proper image space by combining low-level and high-level relations. Major contributions are: 1) Apply the learning mechanism to create semantic clusters and approximately divide database images into meaningful semantic categories. 2) Extract high-level semantic features of each image based on users' retrieval experiences. 3) Incorporate the reliability score, the fuzzy membership, and the composite relation into the affinity matrix to build the weighted semantic manifold structure. 4) Construct the asymmetric relevance vector to propagate ranking scores of its labeled images via the manifold to images with high similarities. Extensive experiments show our system outperforms two manifold systems and two long-term CBIR systems.

We will investigate other strategies to incorporate three additional pieces of information into the affinity matrix. We will also investigate other strategies to reduce the manifold structure to be applicable in a large-scale DB.

5. REFERENCES

[1] Y. Liu, D. Zhang, G. Lu, and W. Y. Ma, "A Survey of Content-Based Image Retrieval with High-Level Semantics," *Pattern Recognition*, Vol. 40, No. 1, pp. 262-282, 2007.

[2] J. He, M. Li, H. Zhang, H. Tong, and C. Zhang, "Generalized Manifold-Ranking-Based Image Retrieval," *IEEE Trans. Image Processing*, Vol. 15, No. 10, pp. 3170-3177, 2006.

[3] X. He, W. Ma, and H. Zhang, "Learning an Image Manifold for Retrieval," *Proc. of Int. Conf. on Multimedia*, pp. 17-23, 2004.

[4] D. Cai, X. He, and J. Han, "Regularized Regression on Image Manifold for Retrieval," *Proc. of Int. Workshop on Multimedia Information Retrieval*, pp. 11-20, 2007.

[5] F. Wang, G. Er, and Q. Dai, "Inequivalent Manifold Ranking for CBIR," *Proc. of ICIP*, pp. 173-176, 2008.

[6] W. Bian and D. Tao, "Biased Discriminant Euclidean Embedding for Content-Based Image Retrieval," *IEEE Trans. on Image Processing*, Vol. 19, No. 2, pp. 545-554, 2010.

[7] R. Chang and X. Qi, "Semantic Clusters Based Manifold Ranking for Image Retrieval," *Proc. of ICIP*, pp. 2473-2476, 2011.

[8] S. Pinker, How the Mind Works, W. W. Norton & Company, New York, New York, 1997.

[9] P. Y. Yin, B. Bhanu, K. Chang, and A. Dong, "Long-Term Cross-Session Relevance Feedback Using Virtual Features," *IEEE Trans. KDE*, Vol. 20, No. 3, pp. 352-368, 2008.

[10] X. Qi, S. Barrett, and R. Chang, "A Noise-Resilient Collaborative Learning Approach to CBIR," *Int. J. of Intelligent Systems*, Vol. 00, pp. 1-23, 2011.