Multimodal Kernel Learning for Image Retrieval

Yen-Yu Lin[1,2]
[1]Institute of Information Science
Academia Sinica
Taipei, Taiwan
yylin@iis.sinica.edu.tw

Chiou-Shann Fuh[2]
[2]Dept. of Computer Science and Information Engineering
National Taiwan University
Taipei, Taiwan
fuh@csie.ntu.edu.tw

***Abstract*—We propose a semi-supervised learning technique to address the problem of fusing multimodal information sources for CBIR. In our approach, user's preferences in the form of reference feedback are treated as labeled data, and the key idea is to devise an on-line scheme to effectively transform the abstract semantics into useful training data for improving the query performance. Specifically, our method can be characterized with the following three advantages: 1) Kernel matrices are used to encode each modality of information so that the fusion can be conveniently carried out via boosting; 2) The base kernel matrices are derived from eigendecomposing the graph Laplacian, and further refined to satisfy a pivotal *monotone* property that ensures intrinsic structure will be reasonably maintained for each modality; 3) The adopted optimization criterion in boosting is to align with a target kernel matrix accounting for relevance feedback, and the learned multimodal kernel matrix can be used for training, and then for testing with those unlabeled ones in the database. To demonstrate the efficiency of the proposed framework, experimental results on CBIR are provided to illustrate several practical considerations.**

***Keywords*—image retrieval, boosting, kernel fusion**

I. INTRODUCTION

Exploiting relevance feedback for retrieving multimedia information such as webpages, images, and video data is an interesting and challenging research topic. Such problems typically deal with at least two different sources of information: the intrinsic data relations entailed by the underlying multimedia data, and the abstract semantics given by on-line users. One important aspect of these tasks is that the emphasis may not be on learning a *general* classifier for inferencing, but finding a reliable way to transfer user's preferences to locate relevant data from those already included in a database. In this work, we consider kernel matrices as the information bottleneck for *semi-supervised learning* so that the system can respond more properly to user's queries. Furthermore, to accommodate the rich characteristics embodied in multimedia data, we design a wide variety of *base* kernel matrices, each of which is to account for a particular modality of feature representation, and its corresponding distance measure. Our formulation uses boosting to fuse the multimodal information, and then applies SVMs with the learned ensemble kernel for finding those relevant to a query.

A. Previous Work

As we shall experiment the efficiency of our method with *content-based image retrieval*, our discussion will mostly focus on those related to CBIR. In particular, we investigate several trends in CBIR, and their connection to our approach.

Supervised versus semi-supervised learning: Since labeled data provided in relevance feedback are few, they generally can not faithfully represent the underlying joint distribution of the label and the feature. Current mainstream classification techniques, say AdaBoost and SVMs, consequently may not be stable enough for learning a reliable decision boundary. This phenomenon is the so-called *small sample problem* in retrieval [26]. Alternatively, semi-supervised learning, e.g., [2], [11], appears to be more promising in that both the labeled and unlabeled data are considered in the construction of a decision boundary. The use of the unlabeled data usually builds upon the assumption that two nearby points are likely to have the same class label. It follows that the learned decision boundary has a tendency not passing through regions with high sample density. Another recent trend of designing semi-supervised algorithms for retrieval is based on *manifold learning* [9], [10], [15]. Such techniques assume that the images of interest spread as manifolds embedded in the feature space, and then consider this particular image structure as well as the information of labeled data for addressing queries.

Single-modal versus multimodal information sources: Fusing multimodal information sources is one of the feasible ways to bridge the semantic gap. One main reason for the effectiveness is that through the information fusion, the respective irrelevant factors in each source may be smoothed out or reduced (though this is not always the case). Indeed, multimodal information fusion has been shown to be successful in, e.g., image and web retrieval [25], and video retrieval [29]. The most widely-used fusion strategies include averaging, linear combination, min-max aggregation or voting, in which each information source is pre-associated with a weight, and the outcome is the weighted combination of the evidence found in each modality. Behind the simple and intuitive idea, fusing multimodal information often involves two difficult issues: 1) how to find the optimal weighting in each source; and 2) how to perform the fusion from user to user, or query to query. Wu et al. [27] propose *super kernel* for fusing evidence from all sources to form a feature vector, and learn an SVM-based classifier to output the final results. In this case, the multiple information sources can be nonlinearly combined.

Query-class independent versus query-class dependent: Retrieval methods that adopt the same information fusion strategy for all queries are considered as *query-class independent*. As the best fusion policy may vary from query to query, a *query-class dependent* scheme is preferred. The concept of query-class dependent is proposed by Yan et al. [29]. Their algorithm has two stages. An input query is first classified into one of the predefined classes. Then the retrieval

result is obtained according to the fusion strategy associated with the class. However, the effectiveness of a query-class dependent method is based on two assumptions: 1) Samples in the same class are assumed to prefer the same fusion strategy; 2) The classifier for separating the classes is required to achieve high accuracy.

B. Our Approach

Concerning multiple modalities, since the feature representations and their associated similarity measures are different, fusion in the domains of representations and distance outputs are difficult. For example, an image can be represented by a feature vector, a tensor, or a bag of feature vectors. And their corresponding distances could be a metric, or non-metric. Thus, instead of directly considering the representations as well as the distance outputs, we propose to perform multimodal fusion in the domain of the kernel matrices, such a fusion strategy would generally give a more flexible and unified view for handling multimodal information sources.

In designing kernel matrices for semi-supervised learning, two viewpoints are often adopted: the first relies on *kernel alignment* [7], [13], [16], and the second has to do with integrating the concepts of *cluster assumption* [4] or *manifold assumption* [24] into kernel matrix generation. We propose to use boosting to link the two aspects of considerations. By our design of base kernel matrices, the scheme will nicely preserve intrinsic structure of each information modality. Furthermore, a boosting algorithm itself will conveniently select good base kernels from the many modalities such that their combination/fusion would best explain the query and relevance feedback. Since the ensemble kernel matrix is dynamically learned from multiple modalities, our method is *query dependent*.

To achieve high fusion performance in CBIR, good and diversified image representations and their corresponding similarity measures are required. In our system, we implement six representation-distance pairs in three levels, i.e., global-based level, region-based level and patch-based level. Some of them are suggested in the retrieval literature, and some are proposed by us. Each of these combinations may achieve good retrieval results in some image categories, but none is always the best. Our experimental results show that the performance could be significantly improved through the fusion of six modalities by the proposed scheme.

II. Kernel Design and Principles

Kernel-based methods have attracted considerable attention in recent years, owing to their effectiveness in classification, regression, and ranking. For multimedia retrieval, it is nature to use a kernel matrix to record a particular modality of information source. Under this setting, the problem of multimodal fusion can be transformed into combining all the corresponding kernel matrices into an informative one.

A. Kernel Matrix by Alignment

Consider a set of training samples $S = \{(\mathbf{x}_i, y_i)\}$ where $\mathbf{x}_i \in \mathcal{X}$ (the input space) and $y_i \in \{+1, -1\}$ is the binary label. In practice, a kernel matrix K associated with S can be obtained by specifying a kernel function.

As suggested by several researchers [6], [7], [13], [17], [19], model selection can also be done in terms of kernel matrix. Cristianini et al. [7] suggest to use *kernel alignment* $\hat{A}(S, K_1, K_2)$ to estimate the degree of agreement between two kernel matrices K_1 and K_2 with respect to S, where

$$\hat{A}(S, K_1, K_2) = \frac{\langle K_1, K_2 \rangle_F}{\sqrt{\langle K_1, K_1 \rangle_F \langle K_2, K_2 \rangle_F}}, \quad (1)$$

$$\text{with } \langle K_p, K_q \rangle_F = \sum_{i,j} K_p(\mathbf{x}_i, \mathbf{x}_j) K_q(\mathbf{x}_i, \mathbf{x}_j). \quad (2)$$

Let K be the learned kernel matrix. To connect K with a given classification task, we define the *target kernel* to be $\mathbf{y}\mathbf{y}^T$, where $\mathbf{y} = [y_1, ..., y_{|S|}]^T$. Then a *well-learned* K would have a large value in the following equation,

$$\hat{A}(S, K, \mathbf{y}\mathbf{y}^T) = \frac{\langle K, \mathbf{y}\mathbf{y}^T \rangle_F}{|S| \sqrt{\langle K, K \rangle_F}}. \quad (3)$$

In optimizing with equation (3), Lanckriet et al. [13] use *semi-definite programming* (SDP) to search the optimal kernel by linearly combining several predefined kernel matrices.

B. Graph Laplacian and Kernel Matrix

The previous section describes how to learn a kernel matrix by alignment; however, it assumes the training samples are all labeled. Concerning CBIR, only a small portion data are labeled, and the remaining are unlabeled. Thus techniques other than kernel alignment are required.

Let the set of data points be $S = S_L \cup S_U$, where $S_L = \{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_l, y_\ell)\}$ for data with labels, and those without labels are included in $S_U = \{\mathbf{x}_{\ell+1}, ..., \mathbf{x}_{\ell+u}\}$. We define a graph G with vertices over S, and an affinity matrix W to record the non-negative weights over edges of G. Typically, the value of W_{ij} is set to 1 if samples $\mathbf{x}_i$ and $\mathbf{x}_j$ are neighbors or close enough, and otherwise 0. Then the *graph Laplacian* of G can be defined as $L = D - W$, where D is a diagonal matrix with $D_{ii} = \sum_{j=1}^{\ell+u} W_{ij}$.

From the *spectral graph theory* [5], we know the following equation holds for any given vector $\mathbf{v} = [v_1, ..., v_{\ell+u}]^T$:

$$\mathbf{v}^T L \mathbf{v} = \sum_{i,j=1}^{\ell+u} W_{ij} (v_i - v_j)^2. \quad (4)$$

In other words, the degree of smoothness of vector $\mathbf{v}$ varying along the intrinsic structure can be measured by the Laplace operator L. We next eigendecompose L as follows:

$$L = \sum_{i=1}^{n} \lambda_i \mathbf{v}_i \mathbf{v}_i^T, \quad (5)$$

where $\lambda_1, ..., \lambda_n$ are the eigenvalues of L in non-decreasing order, and $\mathbf{v}_1, ..., \mathbf{v}_n$ are the corresponding eigenvectors with unit length. The eigenvector corresponding to a smaller eigenvalue will have a smoother change in the intrinsic structure. Note that the *Laplacian eigenmap* [1] for manifold learning builds upon the above property.

The kernel matrices yielded from the graph Laplacian, as are proposed in [24], have the following form,

$$K = \sum_{i=1}^{n} g(\lambda_i)\mathbf{v}_i\mathbf{v}_i^T, \qquad (6)$$

where g is a real-valued function: 1) being monotonically decreasing in λ_i; and 2) $g(\lambda_i) \geq 0, \text{for all } i$. The first property ensures that the smoother eigenvectors are emphasized more, and the second guarantees K is a kernel matrix.

III. Learning Multimodal Fusion

Having described useful principles for kernel design, we are now in a position for learning multimodal kernel matrices.

A. Modal-wise CBIR Information

Assume that each data instance in the dataset S has F feature representations, i.e., $\mathbf{x}_i = [\mathbf{x}_i^1, ..., \mathbf{x}_i^F]$. To leave room for extending our approach to handling other multimedia information, no specific form of a feature representation is enforced. The space of the fth representation is denoted as $\mathcal{X}^f$, for $f = 1, \ldots, F$. The distance function with respect to the fth representation is denoted as $d_f : \mathcal{X}^f \times \mathcal{X}^f \longrightarrow \mathbf{R}$. Again no assumptions have been made on the distance functions. They could be either a metric or a non-metric.

In our formulation, a *modality* of CBIR information is decided by a pair of feature representation and its distance function. Under this setting, we can now compute the graph Laplacian for each modality. Specifically, we use a *simple-minded* scheme to define the fth affinity matrix by

$$W_{ij}^f = \begin{cases} 1, & \text{if } \mathbf{x}_i^f \in k\text{-NN of } \mathbf{x}_j^f \\ & \text{or } \mathbf{x}_j^f \in k\text{-NN of } \mathbf{x}_i^f, \\ 0, & \text{otherwise,} \end{cases} \qquad (7)$$

where $1 \leq i, j \leq |S|$. With (7), the Laplacian matrix L^f of each modality f is computed as follows:

$$L^f = D^f - W^f, D^f = diag\left(\sum_{j=1}^{|S|} W_{1j}^f, ..., \sum_{j=1}^{|S|} W_{|S|j}^f\right). \quad (8)$$

Analogous to equations (5) and (6), we can now perform eigendecomposition on L^f, and conveniently construct the corresponding kernel matrix M^f for the fth modality, i.e.,

$$M^f = \sum_{i=1}^{|S|} g(\lambda_i^f)\mathbf{v}_i^f{\mathbf{v}_i^f}^T \qquad (9)$$

$$= \sum_{i=1}^{|S|} \alpha_i^f \mathbf{v}_i^f{\mathbf{v}_i^f}^T, \quad \alpha_1^f \geq \cdots \geq \alpha_{|S|}^f \geq 0 \quad (10)$$

$$\simeq \sum_{i=1}^{N} \alpha_i^f M_i^f, \quad \alpha_1^f \geq \cdots \geq \alpha_N^f \geq 0, \qquad (11)$$

where $\lambda_1^f, ..., \lambda_{|S|}^f$, arranged in a *non-decreasing* order, are the eigenvalues of L^f, and $\mathbf{v}_1^f, ..., \mathbf{v}_{|S|}^f$ are the corresponding normalized eigenvectors. From (9) to (10), we impose the *order constraint* on $\alpha_1^f, ..., \alpha_{|S|}^f$ to avoid the use of a parametric form due to the function g. Empirically we set $N = 30 \ll |S|$, and let $M_i^f = \mathbf{v}_i^f{\mathbf{v}_i^f}^T$. The concise form in (11) gives a unified view for addressing multiple information sources.

B. Monotone Base Kernel (MBK)

Suppose we treat the set of relevance feedback examples as labeled data. Following our notations in the previous section, the dataset is $S = S_L \cup S_U$, and $|S_L| = \ell, |S_U| = u$. Clearly, a labeled data point with $y = 1$ implies relevance to a query and vice versa. In practice, we have $\ell \ll u$.

To learn reasonable values of $\alpha_1^m, ..., \alpha_N^m$, Zhu et al. [30] design an optimization scheme to connect kernel alignment with graph Laplacian. Specifically, aligning the submatrix M_{LL} (related to only the labeled data) to a target kernel serves as the objective function, and the order constraint in (11) is enforced as the constraint. They use *quadratically constrained quadratic program* (QCQP) to obtain the optimal kernel for a single modality. While QCQP may not be efficient enough for on-line applications such as CBIR, more critical is that the fusion of multiple information modalities remains unsolved.

We instead consider *boosting* to accomplish the kernel learning with S. It implies we need to construct a set of *base kernel matrices* such that the outcome after boosting is still a kernel, and more challengingly, satisfactorily resolves the two critical issues discussed earlier in this section. To this end, we define modal-wise *monotone base kernel* (MBK) matrices by

$$B_i^f = B_{i-1}^f + {M_{LL}}_i^f \quad \text{for } i = 2, \ldots, N, \qquad (12)$$

where $B_1^f = {M_{LL}}_1^f$. Let $\Psi_f = \{B_i^f\}_{i=1}^N$ and $\Psi = \{\Psi_f\}_{f=1}^F$. From the definition in (12), we know each base kernel matrix is of size $\ell \times \ell$, and is derived solely based on labeled data (i.e., relevance feedback). It can be readily verified boosting with Ψ, the set of base kernel matrices, will automatically satisfy the order constraint, and the mechanism of boosting provides a systematic way to fuse base kernels across multiple modalities.

C. Boosting for Semi-Supervised Learning

Equipped with the set of Ψ, we still need to define a target kernel K^* so that boosting by aligning with K^* would imply good CBIR performance. Recall that S_L contains the ℓ label data that are indeed the relevance feedback. Since the irrelevance of a pair of images to a query generally does not imply the two must be similar, the binary classification problem with respect to S_L is not *standard*. Consequently, the target kernel given by the $\ell \times \ell$ matrix $\mathbf{y}\mathbf{y}^T$, [7], [13], [30], may not be adequate. We thus consider the following definition for our target kernel K^* for CBIR:

$$K^*(i,j) = \begin{cases} 1, & \text{if } i = j \text{ or } y_i = y_j = 1, \\ -1, & \text{if } y_i * y_j = -1, \\ 0, & \text{otherwise,} \end{cases} \qquad (13)$$

in which the elements with zero value will have no influence in the boosting algorithm. The exact steps of multimodal kernel boosting are listed in Algorithm 1. We remark that in the algorithm, the final boosted kernel matrix is added with a *prior*

Algorithm 1: *Multimodal Kernel Boosting*

Input : Number of iterations T;
An $\ell \times \ell$ prior kernel matrix $K_p = cI$;
An $\ell \times \ell$ target kernel matrix, K^*;
A set of multimodal MBK matrices, Ψ.

Output: An $\ell \times \ell$ boosted kernel matrix $\hat{K} = K + K_P$.

Initialize: $K \leftarrow \mathbf{0}$

for $t \leftarrow 1, 2, \ldots, T$ **do**

1. Update the weight distribution for $1 \le i, j \le \ell$:

$$w_t(i,j) = \begin{cases} \exp(-K^*(i,j)K(i,j)), & \text{if } |K^*(i,j)| > 0 \\ 0, & \text{otherwise.} \end{cases}$$

2. Select the base kernel $B_t \in \Psi$ maximizing

$$\sum_{i,j=1} w_t(i,j)K^*(i,j)B_t(i,j).$$

3. Find the combinational weight, β_t, of B_t:

$$\mathbf{I}^+ = \{(i,j) \mid K^*(i,j)B_t(i,j) > 0\},$$
$$\mathbf{I}^- = \{(i,j) \mid K^*(i,j)B_t(i,j) < 0\},$$
$$W^+ = \textstyle\sum_{(i,j)\in\mathbf{I}^+} w_t(i,j)|B_t(i,j)|,$$
$$W^- = \textstyle\sum_{(i,j)\in\mathbf{I}^-} w_t(i,j)|B_t(i,j)|,$$
$$\beta_t = \tfrac{1}{2}\log(\tfrac{W^+}{W^-}).$$

4. Update the boosted kernel:

$$K \leftarrow K + \beta_t B_t.$$

kernel matrix $K_P = cI$, where c is a positive scalar between $(0, 1)$. This is equivalent to putting a bit more confidence on the diagonal entries since they correspond to the *autocorrelations* of samples in the relevance feedback.

Let $\hat{K} = K + K_P$ be the learned multimodal kernel matrix by Algorithm 1. Assume that the base kernel selection through boosting implies the following expansions:

$$K = \sum_{t=1}^{T} \beta_t B_t = \sum_{f=1}^{F} \sum_{i=1}^{N} \gamma_i^f M_{LL_i}^f, \qquad (14)$$

where the values of γ_i^f can be easily calculated as we know how K is derived through boosting iterations. It follows that we can use $\hat{K}$ to train an SVM-based classifier and then reference the following rectangular matrix

$$K' = \sum_{f=1}^{F} \sum_{i=1}^{N} \gamma_i^f \left[M_{LL_i}^f, M_{LU_i}^f \right]_{\ell \times (\ell+u)} \qquad (15)$$

for testing with those unlabeled data. The CBIR system then returns the samples with top classification scores as the retrieval results. (We use LIBSVM [3] to implement this step.)

IV. Image Features and Distances

Good image representations and similarity measures are also essential for designing an effective retrieval system. In our system, a coupling of image representation and similarity measure defines its respective graph Laplacian, and is associated with a specific set of weak kernels. We implement various ``feature + distance" pairs. These pairs are expected to complement each other, and some combinations of them will get better performances. We roughly divide the image representations into three levels, namely, global level, region level, and patch level, according to the scale described by a single feature.

A. Global Level

In global-level image representations, each feature describes a specific characteristic about the whole image. These holistic perceptual features will depict an image in a compact form and directly capture the overall properties.

1) Color Histogram + Jeffrey Divergence: Features related to color are adopted mostly for their good performance and human-intuition matching. We apply the 64-bin color histogram defined in the HSV color space. To include the spatial information in the histogram, the color coherence vector (CCV) [20] is concatenated after the color histogram. We use the Jeffrey divergence as the distance function. The Jeffrey divergence of two histograms $P = (p_1, ..., p_n)$ and $Q = (q_1, ..., q_n)$ is defined as

$$d_J(P||Q) = \sum_{i=1}^{n} \left(p_i \log \frac{2p_i}{p_i + q_i} + q_i \log \frac{2q_i}{p_i + q_i} \right). \qquad (16)$$

2) Texture + Euclidean Distance: Texture features refer to the image patterns that display homogeneity. We select and use Tamura *coarseness* and *directionality* features to measure the distribution of the sizes of consistent regions, and the magnitudes and directions of gradients. To cover the texture in the frequency domain, we also apply three-level discrete wavelet transform and calculate the first two moments of coefficients from the nine high-frequency sub-bands. Each feature value is dimension-wise normalized over all images before computing the Euclidean distance.

B. Region Level

Many researchers observe that the reason why images belong to the same class is due to a similar sub-image they commonly share. In other words, the human semantics may only refer to some regions in an image, not the whole image.

1) Normalized Cuts + Principle Angle: We use the *normalized cuts* [22] to segment an image into five regions. Each segmented region is described by a 76-dimensional vector, which comprises a 64-bin color histogram and a 12-bin edge orientation histogram. By stacking the five vectors of the five regions, an image can be represented as a matrix $P \in \mathbf{R}^{76\times5}$. For such a representation, we can measure the dissimilarity between the two column spaces of matrices P and Q by their smallest principle angle(s) [28].

2) k-Means + Integrated Region Matching: We apply the strategy of Li et al. [14] to image segmentation, and compute image distances using *integrated region matching* (IRM): An image is divided into 4-by-4 non-overlapping blocks. From each block we extracted six features, three about color information and three about the wavelet coefficients. The blocks are clustered into five regions via *k*-means on the six features. Then, each region is depicted by nine features, which

respectively characterize color, wavelet coefficients, and spatial information. Finally, the IRM with the *area percentage scheme* is used to measure dissimilarity between image pairs. For more details, please refer to [14].

C. Patch Level

In recent research [8], [12], [23], an image can be represented by a bag of local patches, and using such a representation yields satisfactory results in retrieval.

1) Salient Points + Earth Mover's Distance: We adopt the difference-of-Gaussian (DoG) detector [18] to find the salient points in an image. The 128-dimensional SIFT (Scale Invariant Feature Transform) descriptor [18] and the 64-bin color histogram are used to depict each detected patch. Then images P and Q are represented as $P = \{\mathbf{p}_1, ..., \mathbf{p}_m\}$ and $Q = \{\mathbf{q}_1, ..., \mathbf{q}_n\}$, where m and n are the numbers of detected patches. Because the values of m and n are often large and different, the aforementioned distance functions for the global-based and region-based features are not suitable. We instead apply the earth mover's distance (EMD) [21]. To fit its *weight* and *signature* pair-formulation, images P and Q are expressed as $P = \{(1/m, \mathbf{p}_1), ..., (1/m, \mathbf{p}_m)\}$ and $Q = \{(1/n, \mathbf{q}_1), ..., (1/n, \mathbf{q}_n)\}$ for computing their EMD.

2) Salient Points + EMD with Slack Signatures: Two images regarded as similar may only share common material in the sub-images. By the setting of EMD, it will move *all* the components from one image to another. To integrate the concept of partial similarity, we modify the image representation as follows. First we add an additional *slack* signature $\mathbf{s}$ with a proper weight $0 \leq w < 1$ into the image representation, i.e., $P = \{(w, \mathbf{s}), (1/m, \mathbf{p}_1), ..., (1/m, \mathbf{p}_m)\}$ and $Q = \{(w, \mathbf{s}), (1/n, \mathbf{q}_1), ..., (1/n, \mathbf{q}_n)\}$. Then the moving costs between $\mathbf{s}$ to all other signatures are set as a common and small constant ε. Clearly if $\varepsilon = 0$, only the nearest $(1 - w)$ portion of signature components between the two images will be moved, and thus the concept of partial similarity will be achieved.

V. Experimental Results

In this section, we present several experiments to demonstrate our system and discuss the results.

A. Comparison among Modalities

Before fusing multi-modality information for retrieval, it is reasonable to see first their respective results. The 30-category COREL dataset is used to compare their performances. For each image in the dataset, we search its 20 nearest neighbors (excluding itself) by using the representation and distance function defined in the respective modalities. Then the category-wise average precisions of each modality are shown in Figure 1. For a clearer visualization, the categories are arranged in the order such that the performance of using global color features is monotonically increasing. We highlight some remarks on the results. Because these modalities achieve good results in different kinds of categories, and they seem to complement each other in many categories, this gives us positive evidence to consider multimodal fusion.

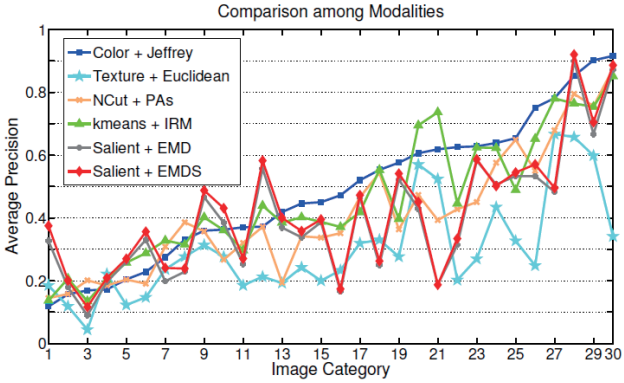

Fig. 1. Accuracy comparisons among modalities

B. Multiple Modalities vs. Single Modality

We use five-fold cross validation to evaluate the proposed algorithm. More precisely, we randomly divide the \$30\$-category COREL dataset into five equal-size subsets. In each run of cross validation, one subset is picked as the query set, and the other four subsets serve as the images in the database. Our system is driven by the query-by-example (QBE) execution, and each image in the query set will be submitted to the system one by one. Once the query and the relevance feedback are acquired, we use the on-line retrieval procedure to handle the query, train the classifier, and return the top-ranking results. For the sake of systematic performance evaluation, the system will automatically feedback some informative samples and insert them into the relevance feedback. In our implementation, five relevant and five irrelevant samples that are closest to the decision boundary but have been misclassified by the current classifier are selected at each iteration. Five iterations are executed for each query. Since the classifier is not trained when a query is given at the first time, the relevance and irrelevance feedback are randomly selected at the first iteration.

Under the mechanism, our system's performance on retrieval with relevance feedback is evaluated. For comparison, we also measure the performance for each of the six modalities. The major difference is that base kernel sets from all modalities can be selected and added into the boosted kernel in our approach. For each of the single modality, only its own base kernels are under consideration. In other words, the difference is that the fusion is performed or not. We highlight some remarks on the experimental results in the following.

Throughout the fusion in boosting the kernel matrix, the performance is significantly improved. In most cases, the outcome with boosted fusion is higher than the best performance among all the modalities. In other words, the proposed method is effective in learning the combination of multi-modality. In the latter feedback iterations, the improvement in precision is more stable (i.e., performance falling behind the one of a single modality rarely happens). That may be due to that the number of feedback examples becomes sufficient for boosting a useful kernel.

The combinational weights roughly match the performance of modalities in each category. Since we learn the combinational weights dynamically, our approach to fusing multiple information sources is query-dependent.

TABLE I

AVERAGE PRECISION RATES EVALUATED UNDER DIFFERENT FEEDBACK ITERATIONS WITH VARIOUS RETRIEVAL MODELS.

Retrieval Model	Feedback Iteration (Top-20)					Feedback Iteration (Top-50)				
	1	2	3	4	5	1	2	3	4	5
Color + Jeffrey	*0.583*	0.673	0.683	0.688	0.692	*0.461*	*0.526*	*0.592*	*0.637*	0.679
Texture + L_2	0.407	0.566	0.576	0.606	0.628	0.261	0.374	0.456	0.510	0.563
NCut + PAs	0.482	0.629	0.639	0.682	0.704	0.352	0.464	0.528	0.585	0.629
k-means + IRM	**0.590**	*0.684*	*0.694*	0.723	*0.766*	0.437	0.499	0.586	0.630	*0.687*
salient + EMD	0.458	0.637	0.647	*0.731*	0.744	0.321	0.436	0.518	0.582	0.648
salient + EMDS	0.486	0.646	0.646	0.695	0.722	0.340	0.445	0.521	0.592	0.656
Ours	0.556	**0.801**	**0.811**	**0.852**	**0.890**	**0.473**	**0.606**	**0.731**	**0.794**	**0.851**

VI. CONCLUSION

We have proposed a kernel-based framework to fuse multimodal information sources for retrieval. Unlike the previous techniques that the fusion is often done in the domains of feature representation, or in the space of modal-wise outputs, our scheme conveniently works with kernel matrices, and leads to a unified approach for dealing with information fusion among modalities that contain large intra-varieties.

In view of the innate properties of CBIR with relevance feedback, the quantity-wise unbalance between labeled and unlabeled data (a.k.a. the small sample problem) often forms the bottleneck for performance improvement. A typical solution to this problem relies on learning a classifier based on the cluster assumption. In our work, two important issues on the design of kernel matrices are both taken into account. One is about the alignment between labeled data, and the other concerns the realization of the data intrinsic structure based on manifold assumption. We satisfactorily address these two aspects of consideration by considering a boosting algorithm. Indeed boosting with our design of *monotone base kernel* (MBK) matrices plays the central role of our approach. As we have described, through boosting the proposed algorithm can simultaneously address *intrinsic structure preserving*, *kernel alignment*, and *multimodal fusion*. All in all, our kernel-based multimodal learning method appears to be promising, and is general enough for handling multimedia data other than images.

REFERENCES

[1] M. Belkin and P. Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In Advances in Neural Information Processing Systems, 2001.

[2] A. Blum and T. Mitchell. Combining labeled and unlabeled sata with co-training. In Annual Conf. on Learning Theory, pages 92–100, 1998.

[3] C.-C. Chang and C.-J. Lin. LIBSVM: a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[4] O. Chapelle, J. Weston, and B. Schölkopf. Cluster kernels for semisupervised learning. In Advances in Neural Information Processing Systems, 2002.

[5] F. Chung. Spectral graph theory. In Regional Conf. Series in Mathematics, 1997.

[6] K. Crammer, J. Keshet, and Y. Singer. Kernel design using boosting. In Advances in Neural Information Processing Systems, 2002.

[7] N. Cristianini, J. Shawe-Taylor, A. Elisseeff, and J. Kandola. On kerneltarget alignment. In Advances in Neural Information Processing Systems, 2001.

[8] K. Grauman and T. Darrell. Efficient image matching with distributions of local invariant features. In Int'l Conf. on Computer Vision and Pattern Recognition, pages 627–634, 2005.

[9] J. He, M. Li, H.-J. Zhang, H. Tong, and C. Zhang. Manifold-ranking based image retrieval. In ACM Conf. on Multimedia, pages 9–16, 2004.

[10] X. He. Incremental semi-supervised subspace learning for image retrieval. In ACM Conf. on Multimedia, pages 2–8, 2004.

[11] T. Joachims. Transductive inference for text classification using support vector machines. In Int'l Conf. on Machine Learning, 1999.

[12] Y. Ke, R. Sukthankar, and L. Huston. Efficient near-duplicate detection and sub-image retrieval. In ACM Conf. on Multimedia, 2004.

[13] G. Lanckriet, N. Cristianini, P. Bartlett, L. Ghaoui, and M. Jordan. Learning the kernel matrix with semi-definite programming. In Int'l Conf. on Machine Learning, pages 323–330, 2002.

[14] J. Li, J. Wang, and G. Wiederhold. Irm: Integrated region matching for image retrieval. In ACM Conf. on Multimedia, pages 147–156, 2000.

[15] Y.-Y. Lin, T.-L. Liu, and H.-T. Chen. Semantic manifold learning for image retrieval. In ACM Conf. on Multimedia, pages 249–258, 2005.

[16] Y.-Y. Lin, T.-L. Liu, and C.-S. Fuh. Local ensemble kernel learning for object category recognition. In Int'l Conf. on Computer Vision and Pattern Recognition, 2007.

[17] Y.-Y. Lin, T.-L. Liu, and C.-S. Fuh. Dimensionality reduction for data in multiple feature representations. In Advances in Neural Information Processing Systems, 2008.

[18] D. Lowe. Distinctive image features from scale-invariant keypoints. Int'l Journal of Computer Vision, 60(2):91–110, 2004.

[19] C. Ong, A. Smola, and R. Williamson. Hyperkernels. In Advances in Neural Information Processing Systems, 2002.

[20] G. Pass, R. Zabih, and J. Miller. Comparing images using color coherence vectors. In ACM Conf. on Multimedia, pages 65–73, 1996.

[21] Y. Rubner, C. Tomasi, and L. Guibas. The earth mover's distance as a metric for image retrieval. Int'l Journal of Computer Vision, 2000.

[22] J. Shi and J. Malik. Normalized cuts and image segmentation. IEEE Trans. Pattern Analysis and Machine Intelligence, 22(8):888–905, 2000.

[23] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In Int'l Conf. on Computer Vision, pages 1470–1477, 2003.

[24] A. Smola and R. Kondor. Kernels and regularization on graphs. In Annual Conf. on Learning Theory, pages 144–158, 2003.

[25] H. Tong, J. He, M. Li, C. Zhang, and W.-Y. Ma. Graph based multimodality learning. In ACM Conf. on Multimedia, pages 862–871, 2005.

[26] L. Wang, Y. Gao, K. L. Chan, P. Xue, and W.-Y. Yau. Retrieval with knowledge-driven kernel design: An approach to improving svm-based cbir with relevance feedback. In Int'l Conf. on Computer Vision, 2005.

[27] Y. Wu, E. Chang, K. Chang, and J. Smith. Optimal multimodal fusion for multimedia data analysis. In ACM Conf. on Multimedia, 2004.

[28] O. Yamaguchi, K. Fukui, and K. Maeda. Face recognition using temporal image sequence. In Int'l Conf. on Face & Gesture Recognition, pages 318–323, 1998.

[29] R. Yan, J. Ynag, and A. Hauptmann. Learning query-class dependent weights in automatic video retrieval. In ACM Conf. on Multimedia, pages 548–555, 2004.

[30] X. Zhu, J. Kandola, Z. Ghahramani, and J. Lafferty. Nonparametric transforms of graph kernels for semi-supervised learning. In Advances in Neural Information Processing Systems, 2004