# Towards Developing a Unified Multimodal Image Retrieval Framework

*Zhongfei (Mark) Zhang  Zhen Guo*

SUNY Binghamton, Binghamton, NY, USA

*Ruofei Zhang*

Yahoo! Labs, Santa Clara, CA, USA

**ABSTRACT**

Built upon the previous work on automatic image annotation and multimodal image retrieval, in this paper we present a unified multimodal image retrieval framework, called UPMIR. The contributions of this paper include: (1) the development of the UPMIR framework; and (2) extensive evaluations of UPMIR including evaluations against a state-of-the-art image retrieval system in a large scale, visually and semantically diverse database crawled from the Web to demonstrate that UPMIR not only has more effective and efficient retrieval performance, but also facilitates more enhanced retrieval modalities.

***Index Terms***— UPMIR, Multimodal Image Retrieval

## 1. INTRODUCTION

Image retrieval, due to its foreseen great potential in many applications including the Web search, has received extensive attention in the literature for years. One of the current research foci in this area is to attempt to reduce the notorious "semantic gap" [9]. One of the effective approaches to solving for this problem is to use multimodal information [12]. As a specific example of this approach in a previous effort, we proposed a probabilistic semantic model for automatic image annotation and multimodal image retrieval [11]. Built upon this model, we have developed a unified framework for multimodal image retrieval. Since this framework is based on the posterior probability, we call this unified framework as UPMIR. The contributions of this paper include: (1) the development of the UPMIR framework; and (2) extensive evaluations of UPMIR including evaluations against a state-of-the-art image retrieval system in a large scale, visually and semantically diverse database crawled from the Web to demonstrate that UPMIR not only has more effective and efficient retrieval performance, but also facilitates more enhanced retrieval modalities.

Related work in the literature includes the Bayes point machine by Chang et at [4], the latent semantic indexing model by Zhao and Grosky [13], and the iterative similarity propagation approach by Wang et al [10]. Another related topic to multimodal image retrieval is the automatic image annotation, which refers to the retrieval of the relevant words to a given query image. Examples on this topic include Barnard et al [1], Blei and Jordan [2], Duygulu et al [6], Feng et al [7], and Li and Wang [8].

## 2. UPMIR Framework

We first introduce the notations and the assumptions we use in this paper. As a multimodal image retrieval framework, UPMIR may be applied to any scenarios where there is any modality of collateral information available to image data for image retrieval. However, in this paper, we assume that the modality of the collateral information to image data is text only. We further assume that the collateral text is processed using the standard text processing techniques to become a collection of annotation words. The database consists of two components: the training database and the testing database. Both components may have different image collections while they share the same annotation word vocabulary. The image collections of the two components are assumed to follow the same distributions of the feature vectors. We assume that the whole database consists of *N* images and *M* annotation words. Each image in the database is represented as a visual feature vector $f_i, i \in [1, N]$ where UPMIR is independent of any specific visual features used. In the rest of the paper, we use the terms image and visual feature vector interchangeably. Each distinct annotation word in the annotation vocabulary is represented as $w^j, j \in [1, M]$.

In addition, we assume that the visual feature vectors of images in the database, $f_i, i \in [1, N]$ are known i.i.d. samples from an unknown distribution. We also assume that the specific pairs of visual feature vectors and annotation words $(f_i, w^j), i \in [1, N], j \in [1, M]$ are known i.i.d. samples from an unknown distribution. Furthermore we assume that these samples are associated with an unobserved *semantic concept* class variable $z \in Z = \{z_1, ...z_K\}$ where *K* is the number of the semantic concept class variables for the whole database. Each observation of one visual feature vector $f \in F = \{f_i, f_2, ..., f_N\}$ belongs to one or more concept classes $z_k$ and each observation of one word

$w \in V = \{w^1, w^2, ..., w^M\}$ in one image $f_i$ belongs to one concept class.

Based on the probabilistic semantic model we proposed in [11], we have the following relationship:

$$P(f_i, w^j) = \sum_{k=1}^{K} P_z(z_k) P_F(f_i \mid z_k) P_v(w^j \mid z_k) \quad (1)$$

where $P_F(\bullet)$ and $P_V(\bullet)$ are the probabilistic distributions of the visual feature vectors and the annotation words, respectively. We also assume that the probabilistic distribution of the unknown semantic class variables is $P_Z(\bullet)$. All the probabilistic distributions may be estimated using the method described in [11].

The main contribution of the previous effort [11] is to explicitly exploit and represent the synergy between the imagery and text modalities as the semantic space and then map the problem of image annotation and retrieval in the original space to that in this space which has a much lower dimension than that of the original image space or the text vocabulary space, i.e., *K << min(N, M)*. Specifically, based on the relationship (1), given an image query $f_i$

$$P(w^j \mid f_i) \approx \frac{\sum_{k=1}^{K} P_v(w^j \mid z_k) P_F(f_i \mid z_k)}{\sum_{h=1}^{K} P_F(f_i \mid z_h)} \quad (2)$$

Thus, the words with the top highest $P(w^j \mid f_i)$ are returned as the generated annotation words for the query image $f_i$. Similarly, we retrieve images for a word query $w^j$ by determining the conditional probability $P(f_i \mid w^j)$

$$P(f_i \mid w^j) \approx \frac{\sum_{k=1}^{K} P_v(w^j \mid z_k) p_F(f_i \mid z_k)}{\sum_{h=1}^{K} P_v(w^j \mid z_h)} \quad (3)$$

The images in the database with the top highest $P(f_i \mid w^j)$ are returned as the retrieval result for the query word $w^j$.

Therefore, we have the UPMIR framework as follows. If the query is given as a textual word $q_w$, assuming that $q_w$ is in the database vocabulary, based on (3), we can immediately retrieve the relevant images through ranking the posterior probabilities $P(f_i \mid q_w)$.

If the query is given as an image $q_f$, assuming that the feature vector $q_f$ follows the same probabilistic distribution of the feature vectors of the images in the database, based on (2), we can immediately obtain the top *m* most relevant annotation words $w^j$ weighted by $P(w^j \mid q_f)$ for the query image $q_f$; then for each returned annotation word $w^j$, using (3) we obtain the retrieved images through ranking the posterior probabilities $P(w^j \mid q_f)\ P(f_i \mid w^j)$; finally, we merge the *m* ranked lists based on the posterior probabilities $P(w^j \mid q_f)\ P(f_i \mid w^j)$ for the *m* different $w^j$.

In the general case where the query is multimodal with both a textual component and an image component, we assume that the query consists of *s* query words $q_w^1, ..., q_w^s$ with the user specified weights $\alpha_1, ..., \alpha_s$ and *t* query images $q_f^1, ..., q_f^t$ with the user specified weights $\beta_1, ..., \beta_t$. For each query word $q_w^j$, $j = 1, ..., s$, by calling the query by word, we have a retrieved image list $A_j$; similarly, for each query image $q_f^i$, $i = 1, ..., t$, by calling the query by image, we have a retrieved image list $B_i$; the final, overall retrieval is then the ranked list

$$C = (\mathop{\oplus}_{j=1,...,s} \alpha_j A_j) \quad \oplus \quad (\mathop{\oplus}_{i=1,...,t} \beta_i B_i) \quad (4)$$

where $\oplus$ means the merge based on the posterior probabilities $P(f_i \mid w^j)$ or $P(w^j \mid q_w^i)\ P(f_i \mid w^j)$.

## 3. EMPIRICAL EVALUATIONS

We have implemented UPMIR framework as a prototype system, and in the rest of the paper for the purpose of reference, we also call the prototype as UPMIR.

To effectively evaluate UPMIR, we elect to use the same data set we used in the previous effort [11]. This data set consists of the imagery as well as the surrounding text data automatically crawled from the Web. The images and the surrounding text describing the image contents in the Web pages are extracted from the blocks containing the images by using the VIPS algorithm [3]. The surrounding text is processed using the standard text processing techniques to obtain the annotation words. Apart from images and annotation words, the weight of each annotation word for the images is computed by using a scheme incorporating TF, IDF, and the tag information in VIPS, and is normalized to the range (0,10). The reason why we use this data set instead of using a commonly used data set such as Corel data set is due to the fact that Corel images are professionally made and therefore have high quality and limited semantics conveyed with relatively small variations of visual contents; consequently, Corel images are easier for image annotation and retrieval than the unconstrained image data in many real world applications such as the Web data. Compared with images in the Corel database, the images in our data set are more diverse both on semantics and on

visual appearance, which serve as an example of minimally constrained real world data and reflect the true nature of image search in many real world applications.

The whole data set consists of 17,000 images and 7,736 stemmed annotation words, of which 12,000 images with the whole annotation word vocabulary are used for the training to obtain all the posterior probabilities. With these probabilities, UPMIR is evaluated for the whole database of the 17,000 images and the 7,736 words as the testing database. To effectively demonstrate the strength and promise of the multimodal querying of UPMIR, we have evaluated UPMIR with three different weightings between the text and image components of a multimodal query, along with the pure text and pure image scenarios. To simplify the evaluations, we have set $s=1$ and $t=1$ in (4). Consequently, we have randomly selected 600 images from the whole image database and for each of the images we have randomly selected one of its annotation words; we take these 600 (word, image) pairs as the query set. For each image query component, we take the top 20 returned annotation words as the "relevant" words, and then call query by text for each of the 20 words to obtain the retrieved image list; the final retrieval is the merge of the 20 lists and the list from the word query component in terms of the posterior probabilities. Figs. 1 and 2 report the averaged retrieval precision and recall over the 600 queries, respectively, with all the five different (word, image) weighting scenarios, where $(\alpha_1, \beta_1)$ represents the text component weight and the image component weight defined in (4). From the Figures, we observe that UPMIR image retrieval is biased towards text component querying; this is clear in the Figures that the pure text querying has the best performance while the pure image querying has the worst performance in the whole text vs. image querying weighting spectrum. This is due to the following reason: the annotation words to the images in the testing database are actually not manually ground-truthed due to the large scale of the database. This means two things. First, given a multimodal query in this experiment, the text component may not actually be relevant to the content of the image component. Second, for the image component of a query, the generated 20 annotation words may not actually be relevant to the image content. Due to this reason, since an image query requires two retrievals in UPMIR (an image-to-text retrieval and a text-to-image retrieval) while a text query only requires one retrieval, errors are propagated and accumulated more substantially in image querying than in text querying.

Nevertheless, we demonstrate that UPMIR outperforms a state-of-the-art pure image querying system UFM [5]. Figs. 3 and 4 report the averaged precision and recall as the performance comparison with UFM using the pure image querying mode for the 600 query images. It is clear that with the pure image querying mode, UPMIR performs at least the same as UFM and in most cases better than UFM (e.g., with top 2 images retrieved, UPMIR has 10% better retrieval precision than UFM). To further demonstrate that UPMIR is also more efficient than UFM in image retrieval, we note that UPMIR and UFM are both implemented and evaluated in the environment of Pentium IV 2.26 GHz CPU with 1G memory. Given the scale of 17,000 images and 7,736 word vocabulary, the average response time for each query for UPMIR is 0.936 second while that for UFM is 9.14 seconds. Clearly UPMIR beats UFM substantially. This is due to the fact that UPMIR has much lower complexity than UFM as UFM is region-based and for each comparison between two images UFM requires a combinatorial complexity while for UPMIR it is only a constant complexity.

On the other hand, since UPMIR performance is biased towards the text component querying, we intend to experimentally justify and demonstrate that UPMIR still offers a better image retrieval than a pure text indexing scheme. For this purpose, we manually evaluate UPMIR using the pure text querying mode against Google and Yahoo!. We randomly select 20 words out of the 7,736 vocabulary, and use each of them as a pure text query to pose to UPMIR, Google image search, and Yahoo! image search, respectively. We manually ground truth the precisions. Fig. 5 clearly demonstrates that UPMIR beats Google and Yahoo! for different numbers of the top images retrieved. Since we do not have access to Google or Yahoo! image database, though the comparing databases are different in size and content, this is the best we can do to compare their performances. The purpose is to show that a multimodal image retrieval system such as UPMIR still has clear advantages for image retrieval over a pure text based indexing system.

## 4. CONCLUSIONS

Built upon the previous work on automatic image annotation and multimodal image retrieval, we have presented a unified multimodal image retrieval framework based on the posterior probabilities obtained in the previous work, and we call this framework UPMIR. The contributions of this paper include: (1) the development of the UPMIR framework; and (2) extensive evaluations of UPMIR including evaluations against a state-of-the-art image retrieval system in a large scale, visually and semantically diverse database crawled from the Web to demonstrate that UPMIR not only has more effective and efficient retrieval performance, but also facilitates more enhanced retrieval modalities.

## 5. REFERENCES

[1] K. Barnard, P. Duygulu, N. d.Freitas, D. Blei, and M. I. Jordan. Matching words and pictures. *Journal of Machine Learning Research*, 3:1107–1135, 2003.

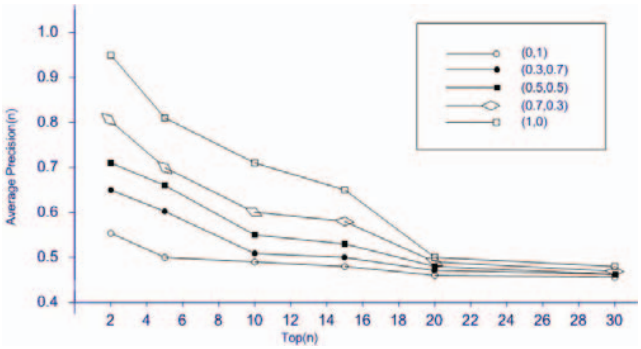


Fig. 1: UPMIR precision with different multimodal weighting combinations.

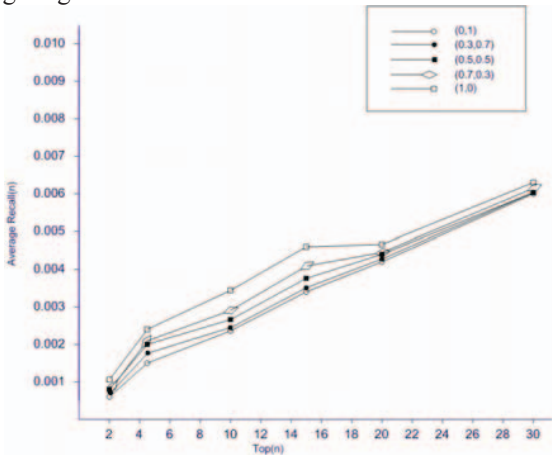


Fig. 2: UPMIR recall with different multimodal weighting combinations.

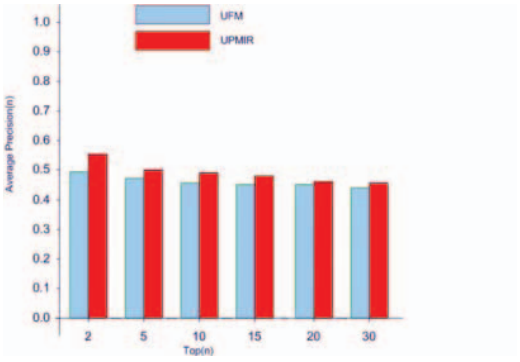


Fig. 3: Precision comparison between UPMIR and UFM.

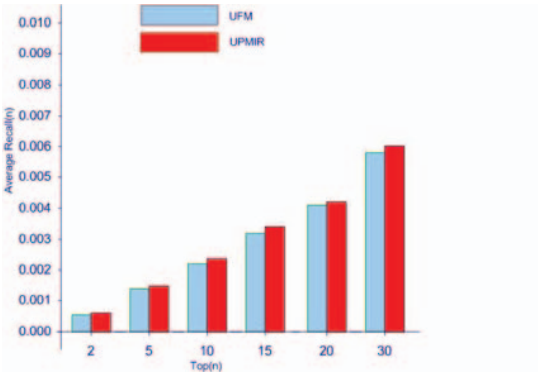


Fig. 4: Recall comparison between UPMIR and UFM.

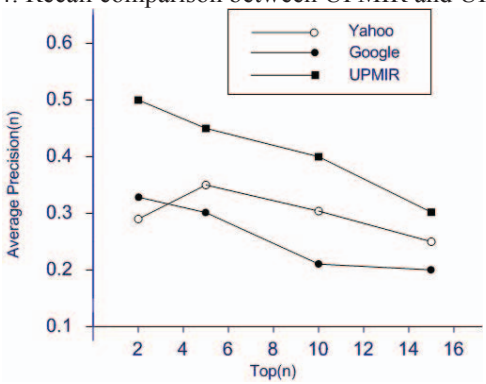


Fig. 5: Average precision comparison among UPMIR, Google image search, and Yahoo! image search.

[2] D. Blei and M. Jordan. Modeling annotated data. In *the 26th International Conference on Research and Development in Information Retrieval (SIGIR)*, 2003.

[3] D. Cai, S. Yu, J.-R. Wen, and W.-Y. Ma. Vips: a vision-based page segmentation algorithm. Microsoft Technical Report (MSRTR-2003-79), 2003.

[4] E. Chang, K. Goh, G. Sychay, and G. Wu. Cbsa: Content-based soft annotation for multimodal image retrieval using bayes point machines. *IEEE Trans. on Circuits and Systems for Video Technology*, 13(1), January 2003.

[5]Y. Chen, J.Z. Wang, A region-based fuzzy feature matching approach to content-based image retrieval, *IEEE T-PAMI*, 24(9), 2002

[6] P. Duygulu, K. Barnard, J. F. G. d. Freitas, and D. A. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *The 7th European Conference on Computer Vision*, volume IV, pages 97–112, Copenhagon, Denmark, 2002.

[7] S. L. Feng, R. Manmatha, and V. Lavrenko. Multiple bernoulli relevance models for image and video annotation. In *The International Conference on Computer Vision and Pattern Recognition*, Washington, DC, June, 2004.

[8] J. Li and J. Z. Wang. Automatic linguistic indexing of pictures by a statistical modeling approach. *IEEE Trans. on PAMI*, 25(9), September 2003.

[9]A.W.M. Smeulders et al, Content-based image retrieval at the end of the early years, *IEEE T-PAMI,* 22, 2000.

[10] X.-J.Wang,W.-Y. Ma, G.-R. Xue, and X. Li. Multi-model similarity propagation and its application for web image retrieval. In *the 12th annual ACM international conference on Multimedia*, pages 944–951, New York City, NY, 2004.

[11] R. Zhang, Z. Zhang, M. Li, W.-Y. Ma, and H.-J. Zhang, A Probabilistic Semantic Model for Image Annotation and Multi-Modal Image Retrieval, to appear in *ACM Multimedia Systems Journal*, 2006.

[12] Z. Zhang, R. Zhang, and J. Ohya, Exploiting the Cognitive Synergy between Different Media Modalities in Multimodal Information Retrieval, *Proc. ICME*, 2004.

[13] R. Zhao and W. I. Grosky. Narrowing the semantic gap—improved text-based web document retrieval using visual features. *IEEE Trans. on Multimedia*, 4(2), 2002.