Structured Max Margin Learning on Image Annotation and Multimodal Image Retrieval

Zhen Guo† Zhongfei (Mark) Zhang† Eric P. Xing‡ Christos Faloutsos‡
†Computer Science Department, SUNY at Binghamton, Binghamton, NY 13905
‡ School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, 15213
†{zguo,zhongfei}@cs.binghamton.edu ‡{epxing,christos}@cs.cmu.edu

1 Introduction

Image retrieval plays an important role in information retrieval due to the overwhelming multimedia data brought by modern technologies, especially the Internet. One of the notorious bottlenecks in the image retrieval is the semantic gap (Smeulders et al., 2000). Recently, it is reported that this bottleneck may be reduced by the multimodal approach (Barnard et al., 2003; Feng et al., 2004) which takes advantage of the fact that in many applications image data typically co-exist with other modalities of information such as text. The synergy between different modalities may be exploited to capture the high level concepts.

In this chapter, we follow this line of research by further considering a max margin learning framework. We assume that we have multiple modalities of information in co-existence. Specifically, we focus on imagery and text modalities whereas the framework may be easily extended to incorporate other modalities of information. Accordingly, we assume that we have a database consisting of imagery data where each image has textual caption/annotation. The framework is not just for image retrieval, but for more flexible across-modality retrieval (e.g., image-to-image, image-to-text, and text-to-image retrieval). Our framework is built upon the max margin framework and is related to the model proposed by Taskar et al. (Taskar et al., 2005). Specifically, we formulate the image annotation and image retrieval problem as a structured prediction problem where the input $\mathbf{x}$ and the desired output $\mathbf{y}$ are structures. Furthermore, following the max margin approach the image retrieval problem is formulated as a quadratic programming (QP) problem. Given the multimodal information in the image database, the dependency information between different modalities is learned by solving for this QP problem. Across-modality retrieval (image annotation and word querying) and image retrieval can be done based on the dependency information. By properly selecting the joint feature representation between different modalities, our approach captures the dependency information between different modalities which is

independent of specific words or specific images. This makes our approach scalable in the sense that it avoids retraining the model starting from scratch every time when the image database undergoes dynamic updates which include image and word space updates.

While this framework is a general approach which can be applied to multimodal information retrieval in any domains, we apply this approach to the Berkeley Drosophila embryo image database[1] for the evaluation purpose. Experimental results show significant performance improvements over a state-of-the-art method.

2 Related Work

Multimodal approach has recently received substantial attention since Barnard and Duygulu et al. started their pioneering work on image annotation (Barnard et al., 2003; Duygulu et al., 2002). Recently there have been many studies (Blei & Jordan, 2003; Pan et al., 2004; Feng et al., 2004; Chang et al., 2003; Datta et al., 2006; Wu et al., 2005) on multimodal approaches.

The structure model covers many natural learning tasks. There have been many studies on the structure model which include conditional random fields (Lafferty et al., 2001), maximum entropy model (McCallum et al., 2000), graph model (Chu et al., 2004), semi-supervised learning (Brefeld & Scheffer, 2006) and max margin approach (III & Marcu, 2005; Tsochantaridis et al., 2004; Taskar et al., 2003; Altun et al., 2003). The max margin principle has received substantial attention since it was used in the support vector machine (SVM) (Vapnik, 1995). In addition, the perceptron algorithm is also used to explore the max margin classification (Freund & Schapire, 1999).

Our main contribution is to develop an effective solution to the image annotation and multimodal image retrieval problem using the max margin approach under a structure model. More importantly, our framework has a great advantage in scalability over many existing image retrieval systems.

3 Supervised Learning

We begin with the brief review of the supervised learning in the max margin framework. Suppose that there is a probability distribution P on $\mathcal{X} \times \mathcal{Y}, \mathcal{X} \subset \mathbb{R}^n$ according to which data are generated. We assume that the given data consist of l labeled data points $(\mathbf{x}_i, y_i), 1 \leq i \leq l$ which are generated according to P. For the purpose of simplicity, we assume the binary classification problem where the labels $y_i, 1 \leq i \leq l$, are binary, i.e., $y_i = \pm 1$.

In the supervised learning scenario, the goal is to learn a function f to minimize

[1] http://www.fruitfly.org

the expected loss called risk functional

$$R(f) = \int L(\mathbf{x}, y, f(\mathbf{x}))dP(\mathbf{x}, y) \tag{1}$$

where L is a loss function. A variety of loss functions have been considered in the literature. The simplest loss function is 0/1 loss

$$L(\mathbf{x}_i, y_i, f(\mathbf{x}_i)) = \begin{cases} 0 & \text{if } y_i = f(\mathbf{x}_i) \\ 1 & \text{if } y_i \neq f(\mathbf{x}_i) \end{cases} \tag{2}$$

In Regularized Least Square (RLS), the loss function is given by

$$L(\mathbf{x}_i, y_i, f(\mathbf{x}_i)) = (y_i - f(\mathbf{x}_i))^2$$

In SVM, the loss function is given by

$$L(\mathbf{x}_i, y_i, f(\mathbf{x}_i)) = \max(0, 1 - y_i f(\mathbf{x}_i))$$

For the loss function Eq. (2), Eq. (1) determines the probability of a classification error for any decision function f. In most applications the probability distribution P is unknown. The problem, therefore, is to minimize the risk functional when the probability distribution function $P(\mathbf{x}, y)$ is unknown but the labeled data $(\mathbf{x}_i, y_i), 1 \leq i \leq l$ are given. Thus, we need to consider the empirical estimate of the risk functional (Vapnik, 1998)

$$R_{emp}(f) = C \sum_{i=1}^{l} L(\mathbf{x}_i, y_i, f(\mathbf{x}_i)) \tag{3}$$

where $C > 0$ is a constant. We often use $C = \frac{1}{l}$. Minimizing the empirical risk Eq. (3) may lead to numerical instabilities and bad generalization performance (Schölkopf & Smola, 2002). A possible way to avoid this problem is to add a stabilization (regularization) term $\Theta(f)$ to the empirical risk functional. This leads to a better conditioning of the problem. Thus, we consider the following regularized risk functional

$$R_{reg}(f) = R_{emp}(f) + \gamma\Theta(f)$$

where $\gamma > 0$ is the regularization parameter which specifies the tradeoff between minimization of $R_{emp}(f)$ and the smoothness or simplicity enforced by small $\Theta(f)$. A choice of $\Theta(f)$ is the norm of the RKHS representation of the feature space

$$\Theta(f) = \|f\|_K^2$$

where $\|.\|_K$ is the norm in the RKHS $\mathcal{H}_K$ associated with the kernel K. Therefore, the goal is to learn the function f which minimizes the regularized risk functional

$$f^* = \arg\min_{f \in \mathcal{H}_K} C \sum_{i=1}^{l} L(\mathbf{x}_i, y_i, f(\mathbf{x}_i)) + \gamma \|f\|_K^2 \tag{4}$$

The solution to Eq. (4) is determined by the loss function L and the kernel K. A variety of kernels have been considered in the literature. Three most commonly-used kernel functions are listed in the Table 1 where $\sigma > 0, \kappa > 0, \vartheta < 0$. The following classic Representer Theorem (Schölkopf & Smola, 2002) states that the solution to the minimization problem Eq. (4) exists in $\mathcal{H}_K$ and gives the explicit form of a minimizer.

Theorem 1 *Denote by* $\Omega : [0, \infty) \to \mathbb{R}$ *a strictly monotonic increasing function, by* $\mathcal{X}$ *a set, and by* $\Lambda : (\mathcal{X} \times \mathbb{R}^2)^l \to \mathbb{R} \cup \{\infty\}$ *an arbitrary loss function. Then each minimizer* $f \in \mathcal{H}_K$ *of the regularized risk*

$$\Lambda((\mathbf{x}_1, y_1, f(\mathbf{x}_1)), \cdots, (\mathbf{x}_l, y_l, f(\mathbf{x}_l))) + \Omega(\|f\|_K)$$

admits a representation of the form

$$f(\mathbf{x}) = \sum_{i=1}^{l} \alpha_i K(\mathbf{x}_i, \mathbf{x}) \tag{5}$$

with $\alpha_i \in \mathbb{R}$.

According to Theorem 1, we can use any regularizer in addition to $\gamma\|f\|_K^2$ which is a strictly monotonic increasing function of $\|f\|_K$. This allows us in principle to design different algorithms. The simplest approach is to use the regularizer $\Omega(\|f\|_K) = \gamma\|f\|_K^2$. Given the loss function L and the kernel K, we substitute Eq. (5) into Eq. (4) to obtain a minimization problem of the variables $\alpha_i, 1 \leq i \leq l$. The decision function f^* is immediately obtained from the solution to this minimization problem.

Different loss functions lead to different supervised learning algorithms. In the literature, two of the most popular loss functions are the squared loss function for RLS and the hinge loss function for SVM.

3.1 Regularized Least Square Approach

We first outline the RLS approach which applies to the binary classification and the regression problem. The classic RLS algorithm is a supervised method where we solve:

$$f^* = \arg\min_{f \in \mathcal{H}_K} C \sum_{i=1}^{l} (y_i - f(\mathbf{x}_i))^2 + \gamma\|f\|_K^2$$

Table 1: Three most commonly-used kernel functions

kernel name	kernel function
polynomial kernel	$K(\mathbf{x}, \mathbf{x}_i) = (\langle \mathbf{x}, \mathbf{x}_i \rangle + c)^d$
Gaussian radial basis function kernel	$K(\mathbf{x}, \mathbf{x}_i) = \exp(-\frac{\|\mathbf{x}-\mathbf{x}_i\|^2}{2\sigma^2})$
sigmoid kernel	$K(\mathbf{x}, \mathbf{x}_i) = \tanh(\kappa\langle \mathbf{x}, \mathbf{x}_i \rangle + \vartheta)$

where C and γ are the constants.

According to Theorem 1, the solution is of the following form

$$f^*(\mathbf{x}) = \sum_{i=1}^{l} \alpha_i^* K(\mathbf{x}_i, \mathbf{x})$$

Substituting this solution in the problem above, we arrive at the following differentiable objective function of the l-dimensional variable $\boldsymbol{\alpha} = [\alpha_1 \cdots \alpha_l]^\top$:

$$\boldsymbol{\alpha}^* = \arg\min C(Y - K\boldsymbol{\alpha})^\top (Y - K\boldsymbol{\alpha}) + \gamma \boldsymbol{\alpha}^\top K \boldsymbol{\alpha}$$

where K is the $l \times l$ kernel matrix $K_{ij} = K(\mathbf{x}_i, \mathbf{x}_j)$ and $\mathbf{Y}$ is the label vector $\mathbf{Y} = [y_1 \cdots y_l]^\top$.

The derivative of the objective function over $\boldsymbol{\alpha}$ vanishes at the minimizer

$$C(KK\boldsymbol{\alpha}^* - K\mathbf{Y}) + \gamma K \boldsymbol{\alpha}^* = 0$$

which leads to the following solution.

$$\boldsymbol{\alpha}^* = (CK + \gamma \mathbf{I})^{-1} C\mathbf{Y}$$

3.2 Max Margin Approach

In the max margin approach, one attempts to maximize the distance between the data and classification hyperplane. In the binary classification problem, the classic SVM attempts to solve the following optimization problem on the labeled data.

$$\begin{aligned} \min \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{l} \xi_i \\ s.t. \quad & y_i\{\langle \mathbf{w}, \Phi(\mathbf{x}_i)\rangle + b\} \geq 1 - \xi_i \\ & \xi_i \geq 0 \quad i = 1, \cdots, l \end{aligned} \tag{6}$$

where Φ is a nonlinear mapping function determined by the kernel and b is a regularized term.

Again, the solution is given by

$$f^*(\mathbf{x}) = \langle \mathbf{w}^*, \Phi(\mathbf{x})\rangle + b^* = \sum_{i=1}^{l} \alpha_i^* K(\mathbf{x}_i, \mathbf{x}) + b^*$$

To solve Eq. (6) we introduce one Lagrange multiplier for each constraint in Eq. (6) using the Lagrange multipliers technique and obtain a quadratic dual problem of the

Lagrange multipliers.

$$\begin{aligned} \min \quad & \frac{1}{2}\sum_{i,j=1}^{l} y_i y_j \mu_i \mu_j K(\mathbf{x}_i, \mathbf{x}_j) - \sum_{i=1}^{l} \mu_i \\ s.t. \quad & \sum_{i=1}^{l} \mu_i y_i = 0 \\ & 0 \le \mu_i \le C \quad i = 1, \cdots, l \end{aligned} \tag{7}$$

where μ_i is the Lagrange multiplier associated with the i-th constraint in Eq. (6).

We have $\mathbf{w}^* = \sum_{i=1}^{l} \mu_i y_i \Phi(\mathbf{x}_i)$ from the solution to Eq. (7). Note that the following conditions must be satisfied according to the Kuhn-Tucker theorem (Vapnik, 1998):

$$\mu_i(y_i(\langle \mathbf{w}, \Phi(\mathbf{x}_i)\rangle + b) + \xi_i - 1) = 0 \quad i = 1, \cdots, l \tag{8}$$

The optimal solution of b is determined by the above conditions.

Therefore, the solution is given by

$$f^*(\mathbf{x}) = \sum_{i=1}^{l} \alpha_i^* K(\mathbf{x}_i, \mathbf{x}) + b^*$$

where $\alpha_i^* = \mu_i y_i$.

4 Structured Max Margin Learning

In an image database where each image is annotated by several words, the word space is a structured space in the sense that the words are interdependent on each other. As shown later, the feature space of images is also a structured space. Therefore, it is not trivial to apply the max margin approach to image databases and several challenges exist. In this chapter, we focus on the max margin approach in the structured space and apply it to the learning problem in image databases.

Assume that the training set consists of a set of training instances $S = \{(I^{(i)}, W^{(i)})\}_{i=1}^{L}$, where each instance consists of an image object $I^{(i)}$ and the corresponding annotation word set $W^{(i)}$. We define a block as a subimage of an image such that the image is partitioned into a set of blocks and all the blocks of this image share the same resolution. For each block, we compute the feature representation in the feature space. These blocks are interdependent on each other in the sense that adjacent blocks are similar to each other and nonadjacent blocks are dissimilar to each other. Therefore, the feature space of images is actually a structured space.

Since the image database may be large, we apply k-means algorithm to all the feature vectors in the training set. We define VRep (visual representative) as a representative of a set of all the blocks for all the images in the database that appear

visually similar to each other. A VRep is used to represent each cluster and thus is represented as a feature vector in the feature space. Consequently, the training set becomes VRep-annotation pairs $S = \{(\mathbf{x}^{(i)}, \mathbf{y}^{(i)})\}_{i=1}^{N}$, where N is the number of the clusters, $\mathbf{x}^{(i)}$ is the VRep object and $\mathbf{y}^{(i)}$ is the word annotation set related to this VRep object. We use $\mathcal{Y}$ to represent the whole set of words and $\mathbf{w}_j$ to denote the j-th word in the whole word set. $\mathbf{y}^{(i)}$ is the M-dimensional binary vector ($M = \|\mathcal{Y}\|$) in which the j-th component $\mathbf{y}_j^{(i)}$ is set to 1 if word $\mathbf{w}_j$ appears in $\mathbf{x}^{(i)}$, and 0 otherwise. We use $\mathbf{y}$ to represent an arbitrary M-dimensional binary vector.

We use score function $\mathbf{s}(\mathbf{x}^{(i)}, \mathbf{w}_j)$ to represent the degree of dependency between the specific VRep $\mathbf{x}^{(i)}$ and the specific word $\mathbf{w}_j$. In order to capture the dependency between VReps and words it is helpful to represent it in a joint feature representation $\mathbf{f} : \mathcal{X} \times \mathcal{Y} \rightarrow \Re^d$. The feature vector between $\mathbf{x}^{(i)}$ and $\mathbf{w}_j$ can be expressed as $\mathbf{f}(\mathbf{x}^{(i)}, \mathbf{w}_j)$ and the feature vector between $\mathbf{x}^{(i)}$ and word set $\mathbf{y}$ is the sum for all the words: $\mathbf{f}_i(\mathbf{y}) = \mathbf{f}(\mathbf{x}^{(i)}, \mathbf{y}) = \sum_{j=1}^{M} \mathbf{y}_j \mathbf{f}(\mathbf{x}^{(i)}, \mathbf{w}_j)$. In this feature vector, each component may have a different weight in determining the score function. Thus, the score function can be expressed as a weighted combination of a set of features $\alpha^\top \mathbf{f}(\mathbf{x}^{(i)}, \mathbf{w}_j)$, where α is the set of parameters.

The learning task then is to find the optimal weight vector α such that:

$$arg \max_{\mathbf{y} \in \mathcal{Y}^{(i)}} \alpha^\top \mathbf{f}(\mathbf{x}^{(i)}, \mathbf{y}) \approx \mathbf{y}^{(i)} \quad \forall i$$

where $\mathcal{Y}^{(i)} = \{\mathbf{y} | \sum \mathbf{y}_j = \sum \mathbf{y}_j^{(i)}\}$. We define the loss function $l(\mathbf{y}, \mathbf{y}^{(i)})$ as the number of different words between these two sets. In order to make the true structure $\mathbf{y}^{(i)}$ as the optimal solution, the constraint is reduced to:

$$\alpha^\top \mathbf{f}_i(\mathbf{y}^{(i)}) \geq \alpha^\top \mathbf{f}_i(\mathbf{y}) + l(\mathbf{y}, \mathbf{y}^{(i)}) \; \forall i, \forall \mathbf{y} \in \mathcal{Y}^{(i)}$$

We interpret $\frac{1}{\|\alpha\|} \alpha^\top [\mathbf{f}_i(\mathbf{y}^{(i)}) - \mathbf{f}_i(\mathbf{y})]$ as the margin of $\mathbf{y}^{(i)}$ over another $\mathbf{y} \in \mathcal{Y}^{(i)}$. We then rewrite the above constraint as $\frac{1}{\|\alpha\|} \alpha^\top [\mathbf{f}_i(\mathbf{y}^{(i)}) - \mathbf{f}_i(\mathbf{y})] \geq \frac{1}{\|\alpha\|} l(\mathbf{y}, \mathbf{y}^{(i)})$. Thus, minimizing $\|\alpha\|$ maximizes such margin.

The goal now is to solve the optimization problem:

$$\begin{aligned} \min \quad & \|\alpha\|^2 \\ s.t. \quad & \alpha^\top \mathbf{f}_i(\mathbf{y}^{(i)}) \geq \alpha^\top \mathbf{f}_i(\mathbf{y}) + l(\mathbf{y}, \mathbf{y}^{(i)}) \; \forall i, \forall \mathbf{y} \in \mathcal{Y}^{(i)} \end{aligned}$$

4.1 Min-max formulation

The above optimization problem is equivalent to the following optimization problem:

$$\begin{aligned} \min \quad & \|\alpha\|^2 \\ s.t. \quad & \alpha^\top \mathbf{f}_i(\mathbf{y}^{(i)}) \geq \max_{\mathbf{y} \in \mathcal{Y}^{(i)}} (\alpha^\top \mathbf{f}_i(\mathbf{y}) + l(\mathbf{y}, \mathbf{y}^{(i)})) \quad \forall i \end{aligned} \tag{9}$$

We take the approach proposed by Taskar et al. (Taskar et al., 2005) to solve it. We consider the maximization sub-problem contained in the above optimization problem.

We have

$$\begin{aligned}\alpha^\top \mathbf{f}_i(\mathbf{y}) + l(\mathbf{y}, \mathbf{y}^{(i)}) &= \alpha^\top \sum_j \mathbf{y}_j \mathbf{f}(\mathbf{x}^{(i)}, \mathbf{w}_j) + \sum_j \mathbf{y}_j^{(i)}(1 - \mathbf{y}_j) \\ &= \mathbf{d}_i + (\mathbf{F}_i\alpha + \mathbf{c}_i)^\top \mathbf{y}\end{aligned}$$

where $\mathbf{d}_i = \sum_j \mathbf{y}_j^{(i)}$ and $\mathbf{F}_i$ is a matrix in which the j-th row is $\mathbf{f}(\mathbf{x}^{(i)}, \mathbf{w}_j)$; $\mathbf{c}_i$ is the vector in which the j-th component is $-\mathbf{y}_j^{(i)}$.

This maximization sub-problem then becomes:

$$\begin{aligned}\max \quad & \mathbf{d}_i + (\mathbf{F}_i\alpha + \mathbf{c}_i)^\top \mathbf{y} \\ s.t. \quad & \sum_j \mathbf{y}_j = \sum_j \mathbf{y}_j^{(i)}\end{aligned}$$

We map this problem to the following linear programming(LP) problem:

$$\begin{aligned}\max \quad & \mathbf{d}_i + (\mathbf{F}_i\alpha + \mathbf{c}_i)^\top \mathbf{z}_i \\ s.t. \quad & \mathbf{A}_i\mathbf{z}_i \le \mathbf{b}_i \quad \mathbf{z}_i \ge 0\end{aligned}$$

for appropriately defined $\mathbf{A}_i, \mathbf{b}_i$, which depend only on $\mathbf{y}, \mathbf{y}^{(i)}$; $\mathbf{z}_i$ is the relaxation for $\mathbf{y}$. It is guaranteed that this LP program has an integral (0/1) solution.

We consider the dual program of this LP program:

$$\begin{aligned}\min \quad & \mathbf{d}_i + \mathbf{b}_i^\top \lambda_i \\ s.t. \quad & \mathbf{A}_i^\top \lambda_i \ge \mathbf{F}_i\alpha + \mathbf{c}_i \quad \lambda_i \ge 0\end{aligned} \tag{10}$$

Now we can combine (9) and (10) together:

$$\begin{aligned}\min \quad & \|\alpha\|^2 \\ s.t. \quad & \alpha^\top \mathbf{f}_i(\mathbf{y}^{(i)}) \ge \mathbf{d}_i + \mathbf{b}_i^\top \lambda_i \quad \forall i \\ & \mathbf{A}_i^\top \lambda_i \ge \mathbf{F}_i\alpha + \mathbf{c}_i \quad \forall i\end{aligned} \tag{11}$$

This formulation is justified as follows. If (10) is not at the minimum, the constraint is tighter than necessary, leading to a sub-optimal solution α. Nevertheless, the training data are typically hardly separable. In such cases, we need to introduce slack variables ξ_i to allow some constraints violated. The complete optimization problem now becomes a QP problem:

$$\begin{aligned}\min \quad & \|\alpha\|^2 + C\sum_i \xi_i \\ s.t. \quad & \alpha^\top \mathbf{f}_i(\mathbf{y}^{(i)}) \ge \mathbf{d}_i + \mathbf{b}_i^\top \lambda_i - \xi_i \quad \forall i \\ & \mathbf{A}_i^\top \lambda_i \ge \mathbf{F}_i\alpha + \mathbf{c}_i \quad \forall i \\ & \alpha \ge 0 \quad inf > \lambda_i \ge 0 \quad inf > \xi_i \ge 0 \quad \forall i\end{aligned} \tag{12}$$

After this QP program is solved, we have the optimal parameters α. Then we have the dependency information between words and VReps by the score function. For each VRep, we have a ranking-list of words in terms of the score function. Similarly we have a ranking-list of VReps for each word.

4.2 Feature representation

For a specific VRep $\mathbf{x}^{(i)}$ and a specific word $\mathbf{w}_j$, we consider the following feature representation $\mathbf{f}$ between them: $(\frac{\delta_{ij}}{n_j}, \frac{n_j}{N}, \frac{\delta_{ij}}{m_i}, \frac{m_i}{M})$. Here we assume that there are N VReps and M words. n_j denotes the number of VReps in which $\mathbf{w}_j$ appears. m_i denotes the number of words which appear in VRep $\mathbf{x}^{(i)}$. δ_{ij} is an indicator function (1 if $\mathbf{w}_j$ appears in $\mathbf{x}^{(i)}$, and 0 otherwise). Other possible features may depend on the specific word or VRep because some words may be more important than others. We only use the features independent of specific words and specific VReps and we will discuss the advantage later.

4.3 Image Annotation

Given a test image, we partition it into blocks and compute the feature vectors. Then we compute the similarity between feature vectors and VReps in terms of the distance. We return the top n most-relevant VReps. Since for each VRep, we have the ranking-list of words in terms of the score function, we merge these n ranking-lists and sort them to obtain the ranking-list of the whole word set. Finally, we return the top m words as the annotation result.

4.4 Word Query

For a specific word, we have the ranking-list of VReps. we return the top n VReps. For each VRep, we compute the similarity between this VRep and each test image in terms of the distance. For each VRep, we have the ranking-list of test images. Finally, we merge these n ranking-lists and return the top m images as the query results.

4.5 Image Retrieval

Given a query image, we annotate it using the procedure in Sec. 4.3. For each annotation word j, there is a subset of images S_j in which this annotation word appears. Then we have the union set $S = \bigcup S_j$ for all the annotation words.

On the other hand, for each annotation word j, the procedure in Sec. 4.4 is used to obtain the related image subset T_j. Then we have the union set $T = \bigcup T_j$. The final retrieval result is $R = S \bigcap T$.

4.6 Database Updates

Now we consider the case where new images are added to the database. Assume that these new images have annotation words along with them. If they do not, we can annotate them using the procedure in Sec. 4.3. For each newly added image, we partition it into blocks and for each block we compute the nearest VRep in terms of the distance and the VRep-word pairs are updated in the database. This also applies to the case where the newly added images may include new word.

Under the assumption that the newly added images follow the same feature distribution as those in the database, it is reasonable to assume that the optimal parameter α also captures the dependency information between the VReps and the newly added words because the feature representation described in Sec. 4.2 is independent of specific words and specific VReps. Consequently, we do not need to re-train the model from scratch. In fact, the complexity of the update is $O(1)$. As the database scales up, so does the performance due to the incrementally updated data. This is a great advantage over many existing image retrieval systems which are unable to handle new vocabulary at all. The experimental result supports and verifies this analysis.

5 Experimental Result

While this approach is a general approach which can be applied to multimodal information retrieval in any domains, we apply this approach to the Berkeley Drosophila embryo image database for the evaluation purpose. We compare the performance of this framework with the state-of-the-art multimodal image annotation and retrieval method MBRM (Feng et al., 2004).

There are totally 16 stages in the whole embryo image database. We use stages 11 and 12 for the evaluation purpose. There are about 6000 images and 75 words in stages 11 and 12. We split all the images into two parts (one third and two thirds), with the two thirds used as the training set and the one third used as the test set. In order to show the advantage discussed in Sec. 4.6, we use a smaller training subset (110 images) to obtain the optimal parameter α. For these 110 images, there are 35 annotation words. Then we use the test set for evaluation. This experiment result is shown as "Our Framework (1)" in the figures. Then we add the remaining training images to the database and use the test set for evaluations again. This experiment result is shown as "Our Framework (2)" in the figures. When the new images are added to the image database, the new annotation words along with them are also added to the image database.

In the figures, the dashed lines are for precisions and the solid lines are for recalls. In the image annotation result shown in Fig. 1, the performance becomes better when the new images are added to the image database. This is consistent with the analysis in Sec. 4.6. When the image database scales up to the size as the same as that used by the MBRM model, our framework works slightly better than MBRM. In the word query result shown in Fig. 2, our framework performs significantly better than MBRM. Similarly in the image retrieval performance shown in Fig. 3, our framework works much better than MBRM.

6 Conclusion

In this chapter, we discuss a multimodal framework on image annotation and retrieval based on the max margin approach. The whole problem is mapped to a quadratic programming problem. Our framework is highly scalable in the sense that it takes a constant time to accommodate the database updating without needing to retrain the

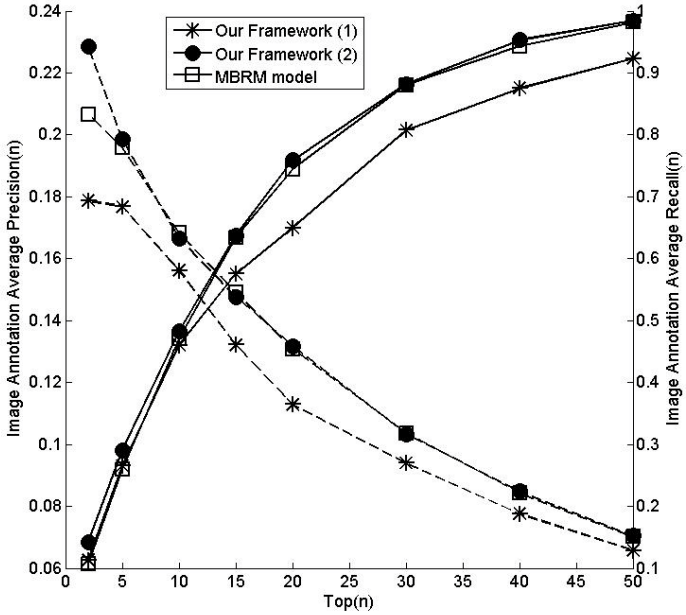

Figure 1: Evaluation of image annotation between our framework and MBRM model.

database from the scratch. The evaluation result shows significant improvements on the performance over a state-of-the-art method.

Acknowledgment

This work is supported in part by the NSF (IIS-0535162, IIS-0812114).

References

Altun, Y., Tsochantaridis, I., & Hofmann, T. (2003). Hidden markov support vector machines. *Proc. ICML*. Washington DC.

Barnard, K., Duygulu, P., Forsyth, D., de Freitas, N., Blei, D. M., & Jordan, M. I. (2003). Matching words and pictures. *Journal of Maching Learning Research*, *3*, 1107–1135.

Blei, D., & Jordan, M. (2003). Modeling annotated data. *Proceedings of the 26th annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 127–134).

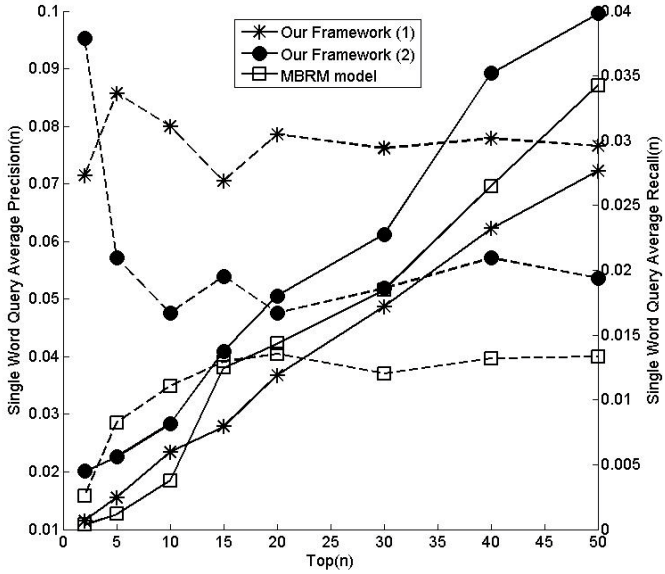

Figure 2: Evaluation of single word query between our framework and MBRM model.

Brefeld, U., & Scheffer, T. (2006). Semi-supervised learning for structured output variables. *Proc. ICML.* Pittsburgh, PA.

Chang, E., Goh, K., Sychay, G., & Wu, G. (2003). Cbsa: content-based soft annotation for multimodal image retrieval using bayes point machines. *IEEE Trans. on Circuits and Systems for Video Technology*, *13*, 26–38.

Chu, W., Ghahramani, Z., & Wild, D. L. (2004). A graphical model for protein secondary structure prediction. *Proc. ICML.* Banff, Canada.

Datta, R., Ge, W., Li, J., & Wang, J. Z. (2006). Toward bridging the annotation-retrieval gap in image search by a generative modeling approach. *Proc. ACM Multimedia.* Santa Barbara, CA.

Duygulu, P., Barnard, K., de Freitas, N., & Forsyth, D. (2002). Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. *Seventh European Conference on Computer Vision* (pp. 97–112).

Feng, S. L., Manmatha, R., & Lavrenko, V. (2004). Multiple bernoulli relevance models for image and video annotation. *International Conference on Computer Vision and Pattern Recognition.* Washington DC.

Freund, Y., & Schapire, R. E. (1999). Large margin classification using the perceptron algorithm. *Maching Learning.*

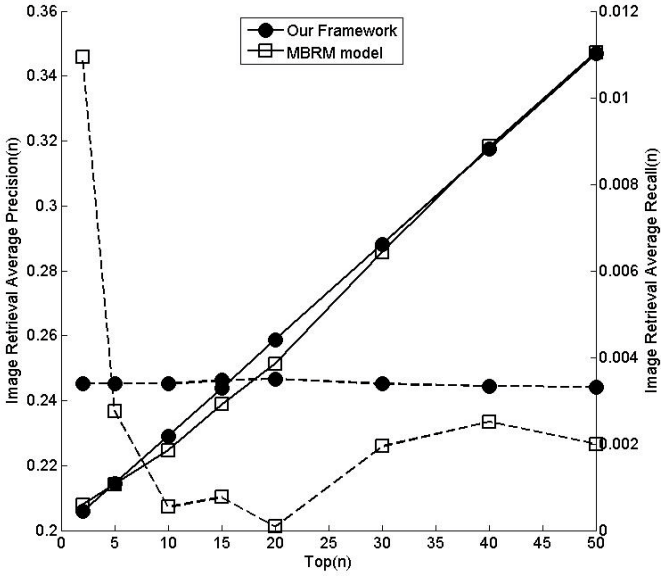

Figure 3: Evaluation of image retrieval between our framework and MBRM model.

III, H. D., & Marcu, D. (2005). Learning as search optimization: Approximate large margin methods for structured prediction. *Proc. ICML*. Bonn, Germany.

Lafferty, J., McCallum, A., & Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Proc. ICML*.

McCallum, A., Freitag, D., & Pereira, F. (2000). Maximum entropy markov models for information extraction and segmentation. *Proc. ICML*.

Pan, J.-Y., Yang, H.-J., Faloutsos, C., & Duygulu, P. (2004). Automatic multimedia cross-modal correlation discovery. *Proceedings of the 10th ACM SIGKDD Conference*. Seattle, WA.

Schölkopf, B., & Smola, A. (2002). *Learning with kernels support vector machines, regularization, optimization and beyond*. MIT Press, Cambridge, MA.

Smeulders, A. W. M., Worring, M., Santini, S., Gupta, A., & Jain, R. (2000). Content-based image retrieval at the end of the early years. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, *22*, 1349–1380.

Taskar, B., Chatalbashev, V., Koller, D., & Guestrin, C. (2005). Learning structured prediction models: A large margin approach. *Proc. ICML*. Bonn, Germany.

Taskar, B., Guestrin, C., & Koller, D. (2003). Max-margin markov networks. *Neural Information Processing Systems Conference*. Vancouver, Canada.

Tsochantaridis, I., Hofmann, T., Joachims, T., & Altun, Y. (2004). Support vector machine learning for interdependent and structured output spaces. *Proc. ICML*. Banff, Canada.

Vapnik, V. N. (1995). *The nature of statistical learning theory*. Springer.

Vapnik, V. N. (1998). *Statistical learning theory*. John Wiley & Sons, Inc.

Wu, Y., Chang, E. Y., & Tseng, B. L. (2005). Multimodal metadata fusion using causal strength. *Proc. ACM Multimedia* (pp. 872–881). Hilton, Singapore.