On the Scalability and Adaptability for Multimodal Retrieval and Annotation

Z. Zhang[1], Z. Guo[1], C. Faloutsos[2], E.P. Xing[2], J.-Y. Pan[3]
[1]Computer Science Department, SUNY Binghamton, USA
[2]School of Computer Science, Carnegie Mellon University, USA
[3]Google Inc., Pittsburgh, PA, USA

Abstract

This paper presents a highly scalable and adaptable co-learning framework on multimodal image retrieval and image annotation. The co-learning framework is based on the multiple instance learning theory. While this framework is a general framework that may be used in any specific domains, to evaluate this framework, we apply it to the Berkeley Drosophila ISH embryo image database for the evaluations of the retrieval and annotation performance. In addition, we also apply this framework to across-stage inferencing for the embryo images for knowledge discovery. We have compared the performance of the framework for retrieval, annotation, and inferencing on the Berkeley Drosophila ISH database with a state-of-the-art multimodal image retrieval and annotation method to demonstrate the effectiveness and the promise of the framework.

1 Introduction

Image retrieval has been studied for over a decade. The notorious bottlenecks of image retrieval include the semantic gap, the scalability, and the adaptability issues [15, 24]. Recently, it is reported that these bottlenecks may be effectively reduced using multimodal approaches [8, 3, 9, 24] by taking the advantage that in many applications imagery data do not exist in isolation but typically co-exist with other modalities of information such as text. It is demonstrated in the literature [9, 24] that given such a presence of the multimodal information, there are effective methods to breakthrough these bottlenecks by exploiting the synergy among the different modalities of the information.

In this paper, we follow this line of research by further proposing a highly scalable and adaptable co-learning framework. We assume that we have multiple modalities of information in co-existence. Specifically in this paper, 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 or annotation (e.g., image-to-image, image-to-text, and text-to-image retrieval). The co-learning framework is based on the multiple instance learning (MIL) theory [6, 13, 2]. This framework has a strong scalability in the sense that the retrieval or annotation can be done in a constant time, independent of the database scale; at the same time, this framework has also a strong adaptability in the sense that it allows incrementally updating the database indexing with a constant operation when the database is periodically updated with new information. These advantages excel many of the existing image retrieval methods in the literature that do not scale or adapt at all, i.e., their retrieval time is dependent upon the database scale and they must redo the indexing (or training) from the scratch even if the database is only updated incrementally.

While this co-learning framework is a general multimodal information retrieval framework that may be used in any specific domains, to demonstrate the effectiveness of this framework, we apply it to the Berkeley Drosophila ISH embryo image database [1] for the evaluations of the retrieval and annotation performance. In addition, we also apply this framework to across-stage inferencing of the embryo images for knowledge discovery (e.g., given an embryo image in stage 5, what is the corresponding image in stage 7 for image-to-image three-stage inferencing? and what is the corresponding annotation for this image in stage 7 for image-to-word three-stage inferencing?). We have compared the performance of the framework for both retrieval and inferencing on the Berkeley Drosophila ISH database with a state-of-the-art multimodal image retrieval and annotation method to demonstrate the effectiveness and the promise of the framework.

2 Related Work

In the machine learning community, MIL has become a focused topic in recent years and has received extensive attention in the literature ever since the classic work of Dietterich et al [6], Auer [2], and Maron and Lozano-Perez [13]. Yang and Lozano-Perez [21] and Zhang et al [23] were among the first to apply MIL to image retrieval, which led to more subsequent work on this topic [22, 25].

Multimodal image retrieval and annotation have recently received substantial attention since Barnard and Duygulu et al started their pioneering work on image annotation [8, 3]. Recent work includes [9, 4, 16, 14, 24, 12, 18, 11, 10, 5, 17].

14th International Conference of Image Analysis and Processing - Workshops (ICIAPW 2007)
0-7695-2921-6/07 $25.00 © 2007 IEEE

Yang et al [19, 20] proposed similar frameworks in the sense that MIL theory was used for image retrieval where text annotations were available in the training set. However, they focused on decomposing an image into regions, and more importantly, they failed to address the scalability and adaptability issues at all. The main contribution of this work is that this co-learning framework can not only be applied to multimodal image retrieval and annotation, but more importantly, this framework is highly scalable and adaptable; these advantages of the co-learning framework excel many existing image retrieval methods in the literature that do not scale or adapt at all.

3 Co-Learning Framework

We first consider the scenario that the whole database is initially used as the training set to build up the database indexing before the database is allowed to evolve under the assumption that the word vocabulary stays the same. We will relax this assumption in the scalability analysis in Sec. 4.

In the rest of the paper, we use calligraphic letters to denote the set variables or functions, and to use regular letters to denote regular variables or functions. A database $\mathcal{D} = \{\mathcal{I}, \mathcal{W}\}$ consists of two parts, an image collection $\mathcal{I}$ and a vocabulary collection $\mathcal{W}$. A collection of images $\mathcal{I} = \{I_i, i = 1, ..., N\}$ is the whole database images used as the training set; $N = |\mathcal{I}|$; for each image I_i, there are a set of words annotating this image $\mathcal{W}_i = \{w_{ij}, j = 1, ..., N_i\}$; the whole vocabulary set of the database is $\mathcal{W}$; $M = |\mathcal{W}| = |\bigcup_{i=1}^{N} \mathcal{W}_i|$. 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. We define a 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 of an image may be represented as a feature vector in a feature space.

Before we present the framework, we first make a few assumptions.

- A semantic concept corresponds to one or multiple words, and a word corresponds to one semantic concept for each image. Consequently, semantic concepts may be represented in words.
- A semantic concept corresponds to one or multiple VReps, and a VRep corresponds to one or multiple semantic concepts.
- A word corresponds to one or multiple VReps, and a VRep corresponds to one or multiple words.
- An image may have one or more words for annotation.

3.1 Establish the mapping between the word space and the image-VRep space

For each image I_i, we partition it into a set of exclusive blocks B_{ij}, i.e.,

$$I_i = \bigcup_j B_{ij}, j = 1, ..., n_i \qquad B_{ij} \cap B_{ih} = \emptyset, j \neq h \qquad (1)$$

where n_i is a function of the resolution of I_i such that the resolution of B_{ij} is no less than a threshold. If all the images in the database are in the same resolution, all the n_i's are the same as a constant. Since each block may be represented as a feature vector in a feature space, for all the blocks of all the images in the database, a nearest neighbor clustering in the feature space leads to a partition of the whole block feature vectors in the feature space into a finite number of clusters such that each cluster is represented by its centroid; let L be the number of such clusters. This centroid is a VRep corresponding to this cluster for all the images in the database. Consequently, the whole VRep set in the database is

$$\mathcal{V} = \{v_i | i = 1, ..., L\} \qquad (2)$$

Thus, each image I_i may be represented by a subset of $\mathcal{V}$. Each VRep is represented as a feature vector in the feature space and corresponds to a subset of all the images in the database such that this VRep appears in those images in the subset, i.e., for each VRep v_i, there is a subset $\mathcal{I}_{v_i}$ of the images in the database such that

$$\mathcal{I}_{v_i} = \{I_h | h = 1, ..., n_{v_i}\} \qquad (3)$$

where $n_{v_i} = |\mathcal{I}_{v_i}|$.

Once we have obtained all the VReps for the images in the database, we sort all the textual vocabulary words in W (say alphabetically), and for each word w_k, there is a corresponding set of images, $\mathcal{S}_k$, such that this word w_k appears in the annotation of each of the images in the set. Since each image is represented as a set of decomposed blocks, $\mathcal{S}_k$ may be represented as

$$\mathcal{S}_k = \{I_{k_i} | I_{k_i} = \bigcup_j B_{k_{ij}}, j = 1, ..., n_{k_i}\} \qquad (4)$$

where $B_{k_{ij}}$ is the jth block in image I_{k_i}. For each block $B_{k_{ij}}$ in image I_{k_i}, a feature vector $f_{k_{ij}}$ in the feature space is used to represent the block. In order to establish the relationship between the word space and the image-VRep space, we map the problem to an MIL problem [6]. A general MIL problem is to learn a function $y = F(x)$, where we are given by multiple samples of x represented as bags, and each bag has ambiguities represented by the multiple instances of x. Here the problem is that each bag is an image, and all the instances of this bag are the blocks represented by the corresponding feature vectors; the y here is a word vector instead of a value in the range of [0, 1] in the classic version of MIL, consisting of all the words given in the training set that correspond to a specific VRep; the function to be learned each time is the function of a VRep mapping to the words. Specifically, for each word $w_k \in \mathcal{W}$, we use MIL to apply to the whole image database to obtain the optimal block feature vector t_k. Given the distribution of all the $f_{k_{ij}}$ corresponding to the image set $\mathcal{S}_k$ in the feature space, using the Diverse Density algorithm of MIL [13], we are able to immediately obtain the optimal block feature vector as the t_k

$$t_k = \arg\max_t \prod_t P(t \in I | I \in \mathcal{S}_k) \prod_t P(t \in I | I \notin \mathcal{S}_k) \qquad (5)$$

14th International Conference of Image Analysis and Processing - Workshops (ICIAPW 2007)
0-7695-2921-6/07 $25.00 © 2007 IEEE

where $P(.|.)$ is a posterior probability. Now we have established the one-to-one mapping between the word w_k and the block feature vector t_k. Then we use the nearest neighbor clustering to identify all the closest VReps v_{k_l} such that

$$\|t_k - v_{k_l}\| < T_k \tag{6}$$

where T_k is a threshold. Denote the set of those VReps that satisfy this constraint as $\mathcal{V}_k$,

$$\mathcal{V}_k = \{v_{k_l} | l = 1, ..., n_{w_k}\} \tag{7}$$

where n_{w_k} is the number of such VReps satisfying this constraint. Thus, for each word w_k, there is a corresponding set of VReps $\mathcal{V}_k$ that are close to t_k subject to the threshold T_k. In addition, according to Eq. 3, each such VRep v_{k_l} has an associated image set $\mathcal{I}_{k_l}$ such that all the images in the set have this VRep. For each such image $I_{k_{l_i}} \in \mathcal{I}_{k_l}$, using the mixture of Gaussian model [7], we compute the posterior probability $P(I_{k_{l_i}}|w_k)$. Then, we rank all the images in the set $\mathcal{I}_{k_l}$ by the posterior probability $P(I_{k_{l_i}}|w_k)$. We denote such a ranked list of images in the database as $\mathcal{L}_k$. Hence, for each word w_k, there is a corresponding ranked list of images in the database

$$\mathcal{L}_k = \{I_{k_h} | h = 1, ..., |\mathcal{L}_k|\} \tag{8}$$

i.e.,

$$w_k \leftrightarrow \mathcal{L}_k \tag{9}$$

Similarly, we use MIL to learn the function $y = F'(x)$ where x's are the ambiguous instances of the annotation words for an image and y is the set of the corresponding VReps to a word; here again the bag is an image. Specifically, for each VRep v_i, according to Eq. 3, there is a corresponding image set $\mathcal{I}_{v_i}$; and for each image $I_{v_{ij}} \in \mathcal{I}_{v_i}$, there is a corresponding annotation word set $\mathcal{W}_{v_{ij}}$

$$\mathcal{W}_{v_{ij}} = \{w^h_{v_{ij}} | h = 1, ..., |\mathcal{W}_{v_{ij}}|\} \tag{10}$$

Thus, using the Diverse Density algorithm of MIL [13] again, we are able to obtain the optimal annotation word w_k corresponding to the image set $\mathcal{I}_{v_i}$

$$\begin{aligned} w_k &= \arg\max_w \prod_w P(w \in \mathcal{W}_{v_{ij}} | I \in \mathcal{I}_{v_i}) \\ &\quad \times \prod_w P(w \in \mathcal{W}_{v_{ij}} | I \notin \mathcal{I}_{v_i}) \end{aligned} \tag{11}$$

Similarly, we may use the same algorithm to compute the ith best annotation word corresponding to VRep v_i. Consequently, for every VRep v_i, there is a corresponding ranked list of annotation words $\mathcal{L}_{v_i}$, i.e.,

$$v_i \leftrightarrow \mathcal{L}_{v_i} \tag{12}$$

Finally, For every VRep $v_i \in \mathcal{V}$, we compute the prior probability $P(v_i)$ by determining the relative occurrence frequency of v_i in the whole image database. Similarly, for every word $w_k \in \mathcal{W}$, we compute the prior probability $P(w_k)$ by determining the relative occurrence frequency of w_k for all the images in the database.

Given this learned correspondence relationship between the word space and the imagery space, we have completed the co-learning for indexing the database as part of the framework. Now we are ready for across-modality retrieval.

3.2 Word to image retrieval

If a query is given by several words for retrieving images from the database, we assume that the query consists of words $w_{q_k}, k = 1, ..., p$. We also assume that all the query words are within the textual vocabulary of the training data. Since each w_{q_k} has a corresponding ranked list of images $\mathcal{L}_k$, we just need to merge these p ranked lists $\mathcal{L}_k$, $k = 1, ..., p$ by $P(I_{k_i}|w_{q_k})$ for all the different images I_{k_i}.

3.3 Image to image retrieval

If a query is given by several images for retrieving images from the database, we assume that the query consists of images $I_{q_k}, k = 1, ..., p$. These images may or may not be necessarily from the database; however, we assume that these images follow the same feature distributions of those in the database. For each query image I_{q_k}, we partition it into p_k blocks following the definition in Eq. 1, and extract the feature vector $f_{q_{kl}}$ for each block $B_{q_{kl}}$. For each $f_{q_{kl}}$, we compute the similarity distances to all the VReps v_i in the feature space. Based on the similarity distances and the assumption that the features in the query images follow the same distributions of those of the images in the database, each $B_{q_{kl}}$ is replaced with the corresponding closest VRep v_i in the feature space. From Eq. 3, each v_i has a corresponding image set $\mathcal{I}_{v_i}$; we assume that there are in total r_k such VReps v_i found in the query image I_{q_k} and $r_k \leq p_k$. Let $\mathcal{S}_{q_k}$ be the largest common set of the r_k image sets $\mathcal{I}_{v_i}$.

On the other hand, for each VRep v_i of I_{q_k}, we immediately have a ranked word list $\mathcal{U}_{v_i}$ based on $P(w_k|v_i)$. We merge the r_k ranked lists based on $P(w_k|v_i)P(v_i|I_{q_k})$ to form a new ranked list $\mathcal{U}_{q_k}$, where $P(v_i|I_{q_k})$ is the occurrence frequency of the VRep v_i appearing in the image I_{q_k}. For all the words in the list $\mathcal{U}_{q_k}$ (in the implementation we may truncate to the top few words for the list), we use the word-to-image retrieval scheme in Sec. 3.2 to generate a ranked image list $\mathcal{L}_{q_k}$. $\mathcal{L}_{q_k}$ is then further trimmed such that only those images that are in $\mathcal{S}_{q_k}$ survive with the same relative ranked order in $\mathcal{L}_{q_k}$. Finally, we merge the p ranked lists $\mathcal{L}_{q_k}, k = 1, ..., p$.

3.4 Image to word retrieval

If the query is given by several images for word retrieval, i.e., for automatic annotation, we assume that the query consists of p images, $I_{q_k}, k = 1, ..., p$. Similar to the image-to-image retrieval in Sec. 3.3, each query image I_{q_k} is decomposed into several VReps, and assume that the p query images have in total s_k VReps $v_i, i = 1, ..., s_k$. Let $P(v_i|I_q)$ be the relative frequency of the VRep v_i in all the query images $I_{q_k}, k = 1, ..., p$. Since each VRep v_i has a corresponding ranked list of words $\mathcal{U}_{v_i}$ based on $P(w_k|v_i)$, the final retrieval is the merged ranked list of words based on $P(w_k|v_i)P(v_i|I_q)$ from the s_k ranked lists $\mathcal{U}_{v_i}$.

3.5 Multimodal querying for image retrieval

If the query is given by a combination of a series of words and a series of images for information retrieval, without loss of generality, we perform image retrieval as follows. We use the word-to-image retrieval in Sec. 3.2 and the image-to-image retrieval in Sec. 3.3, respectively, and finally merge the retrievals together based on their corresponding posterior probabilities.

4 Scalability and Adaptability Analysis

From the analysis between Secs. 3.2 and 3.5, it is clear that it only takes a constant time to process any type of query for retrieval or annotation. Thus, this co-learning framework is highly scalable. Due to the space limitation, we omit the proof and analysis for the adaptability issue, and directly state the conclusion that this co-learning framework is also highly adaptable, with only a constant time updating the whole database indexing while the database has a local change.

5 Empirical Evaluations

While this co-learning framework is a general multimodal information retrieval framework that can be applied to any specific domains, in order to demonstrate and evaluate its effectiveness, we use the Berkeley Drosophila ISH embryo image database [1] as the testbed for the evaluations of the retrieval performance of the framework. In addition, we also apply the framework to this testbed for across-stage inferencing for knowledge discovery. To demonstrate the effectiveness and the promise of the framework in the retrieval and inferencing capabilities, we compare the performance of this framework with that of a state-of-the-art multimodal image retrieval and annotation method MBRM [9].

In the testbed, there are in total 16 stages of the embryo images archived in six different folders with each folder contains two to four stages of the images; there are in total 36,628 images and 227 words in all the six folders; not all the images have annotation words. For the retrieval evaluations, we use folder 5884, which corresponds to stages 11 and 12. There are about 5,500 images with annotation words and 64 annotation words in this folder. We split the whole folder's images into two parts (one third and two thirds), with the two thirds used in training to build up the indexing and the one third used as evaluation testing. For across-stage inferencing evaluations, we use folders 5884 and 5885 for the two-stage inferencing evlauations, and use folders 5883, 5884, and 5885 for the three-stage inferencing evaluations. Since images in each folder are mixed together across the stages, the two-stage inferencing may be up to four stages' inferencing, and the three-stage inferencing may be up to six stages' inferencing. In each of the inferencing scenarios, we use the same split for each folder for training and evaluations.

In order to facilitate the across-stage inferencing capabilities, we handcraft the ontology of the words involved in the evaluations. For example, *cardiac mesoderm primordium* in folder 5884 is considered as the same as *circulatory system* in folder 5885. With this ontology and the simple word matching, images in different stages are well-connected and thus across-stage inferencing becomes possible. Note that if we use a pure image similarity across different stages for inferencing instead of using this framework with the ontology, there would be much more ambiguities introduced. Therefore, for the two stage inferencing, regardless of what query modality is, a retrieval in text is first conducted within the first folder; then through the word matching with the built up ontology to the second folder, a final retrieval to the expected modality is conducted in the second folder. Similar inferencing is performed for the three stage inferencing. In the following figures, the dashed lines are for precisions and the solid lines are for recalls.

Figs. 1 and 2 report both the retrieval and the two-stage inferencing performances for word to image and image to word scenarios in comparison with the MBRM method. Clearly, for word-to-image scenario, our framework outperforms MBRM substantially in retrieval performance, and performs slightly better than MBRM in most cases for the two-stage inferencing; for image-to-word scenario, our framework has almost the same performance as that of MBRM in retrieval, but MBRM performs slightly better than the framework in the two-stage inferencing. Since MBRM does not have the capability for image-to-image retrieval, we adapt it to the across-stage image-to-image inferencing using the same ontology we have manually built. Fig. 3 reports the performance comparison between our framework and MBRM for the three-stage inferencing, where our framework preforms slightly better than MBRM overall. Fig. 4 showcases such an inferencing example across at least three stages based on our framework. Fig. 4(a) is an embryo image in folder 5883 (stages 9 to 10) and Fig. 4(b) is the corresponding embryo image found by the framework in folder 5885 (stages 13 to 16). Note the substantial visual difference between the two images which would fail any purely visual feature based retrieval method for this type of inferencing.

To demonstrate the scalability and adaptability advantages, Figs. 5 and 6 document the empirical evaluations for incrementally adding new images and words for the folder 5885 (from the 30% of the original data up to the 100% of all the data). Since MBRM does not allow incrementally updating the indexing, its performance is not reported in the figures for a fair comparison. On the other hand, the co-learning framework has almost about the same performance when the database scales up from 30% to 100%, which indirectly verifies the strong adaptability. The strong scalability is supported by noting that the co-learning framework has the same average response times (0.0051 second for image-to-word and 0.026 second for word-to-image) for all the three scales while MBRM has about linear average response times (0.013, 0.018, 0.021 seconds for image-to-word and 0.41, 0.53, 0.60 seconds for word-to-image) under the same Linux environment.

6 Conclusion

We have presented a highly scalable and adaptable co-learning framework based on the MIL theory for multimodal imagery retrieval and annotation. We have conducted an empirical validation on the Berkeley Drosophila ISH embryo image database for the retrieval evaluations. In addition, we have also used this framework for across-stage inferencing evaluations for knowledge discovery. We have compared the performance of this framework with that of

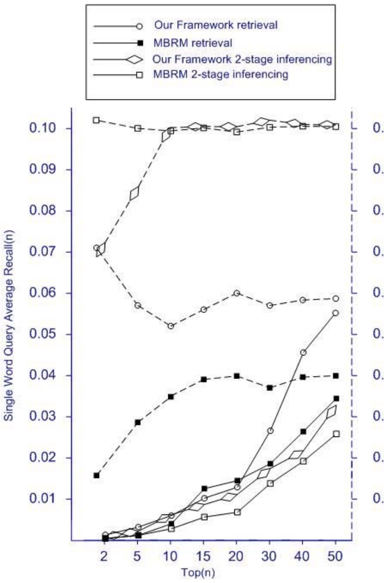

Figure 1. Evaluations of the word to image retrieval and 2-stage inferencing between the co-learning framework and MBRM.

a state-of-the-art multimodal image retrieval and annotation method for retrieval, annnotation, and inferencing to demonstrate the effectiveness and the promise of this framework.

Ackonwledgements

This work is supported in part by NSF (IIS-0535162), AFRL (FA8750-05-2-0284), AFOSR (FA9550-06-1-0327), and a CAREER award to EPX by NSF (DBI-054659).

References

[1] http://www.fruitfly.org/cgi-bin/ex/bquery.pl?qpage=entryqtype=summary.

[2] P. Auer. On learning from multi-instance examples: empirical evaluation of a theoretical approach. In *Proc. ICML*, 1997.

[3] 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.

[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] R. Datta, W. Ge, J. Li, and J. Wang. Toward bridging the annotation-retrieval gap in image search by a generative modeling approach. In *Proc. ACM Multimedia*, 2006.

[6] T. Dietterich, R. Lathrop, and T. Lozano-Perez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89:31 71, 1997.

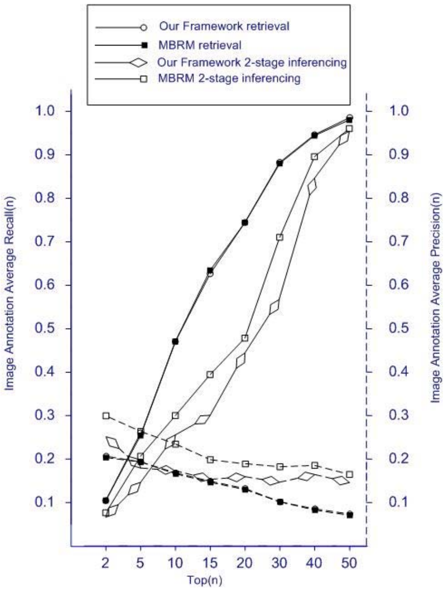

Figure 2. Evaluations of the image to word retrieval and 2-stage inferencing between the co-learning framework and MBRM.

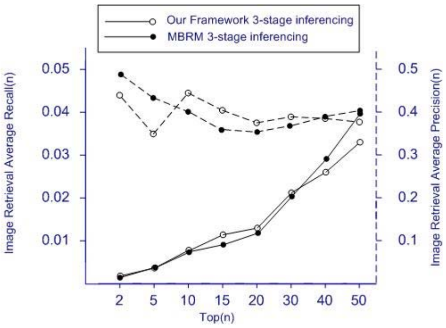

Figure 3. Evaluations of the image to image, 3-stage inferencing between the co-learning framework and MBRM.

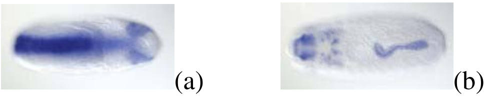

Figure 4. An example of 3-stage image-to-image inferencing.

14th International Conference of Image Analysis and Processing - Workshops (ICIAPW 2007)
0-7695-2921-6/07 $25.00 © 2007 IEEE

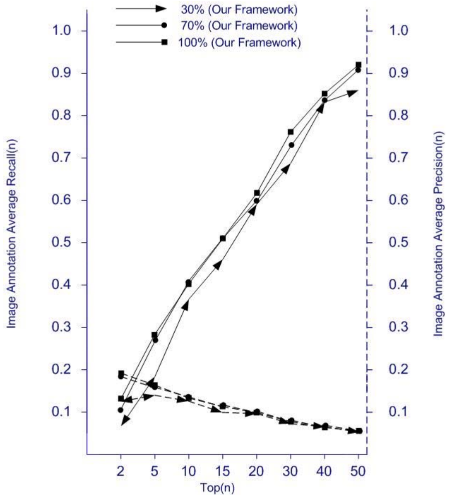

Figure 5. Evaluations of the image to word retrieval scalabilities of the co-learning framework.

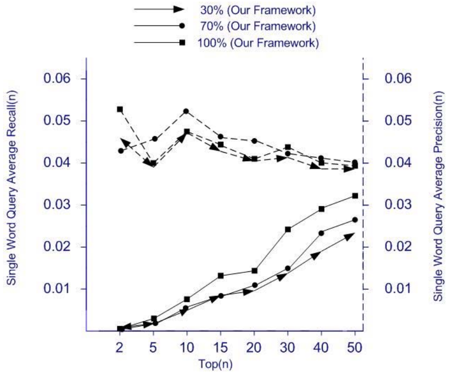

Figure 6. Evaluations of the word to image retrieval scalabilities of the co-learning framework.

[7] W. R. Dillon and M. Goldstein. *Multivariate Analysis, Mehtods and Applications*. John Wiley and Sons, New York, 1984.

[8] 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.

[9] 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.

[10] J. Li and J. Wang. Real-time computerized annotation of pictures. In *Proc. ACM Multimedia*, 2006.

[11] X. Li, L. Chen, L. Zhang, F. Lin, and W.-Y. Ma. Image annotation by large-scale content based image retrieval. In *Proc. ACM Multimedia*, 2006.

[12] W. Liu and X. Tang. Learning an image-word embedding for image auto-annotation on the nonlinear latent space. In *Proc. ACM Multimedia*, 2005.

[13] O. Maron and T. Lozano-Perez. A framework for multiple instance learning. In *Proc. NIPS*, 1998.

[14] J.-Y. Pan, H.-J. Yang, C. Faloutsos, and P. Duygulu. Automatic multimedia cross-modal correlation discovery. In *ACM Proc. KDD*, 2004.

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

[16] 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.

[17] X.-J. Wang, L. Zhang, F. Jing, and W.-Y. Ma. AnnoSearch: Image auto-annotation by search. In *Proc. CVPR*, 2006.

[18] Y. Wu, E. Chang, and B. Tseng. Multimodal metadata fusion using casual strength. In *Proc. ACM Multimedia*, 2005.

[19] C. Yang, M. Dong, and F. Fotouhi. Region based image annotation through multiple instance learning. In *Proc. ACM Multimedia*, 2005.

[20] C. Yang, M. Dong, and J. Hua. Region-based image annotation using asymmetrical support vector machine-based multiple instance learning. In *Proc. CVPR*, 2006.

[21] C. Yang and T. Lozano-Perez. Image database retrieval with multiple-instance learning techniques. In *Proc. ICDE*, 2000.

[22] H. Zhang, R. Rahmani, S. Cholleti, and S. Goldman. Local image representations using pruned salient points with applications to CBIR. In *Proc. ACM Multimedia*, 2006.

[23] Q. Zhang, S. Goldman, W. Yu, and J. Fritts. Content-based image retrieval using multiple instance learning. In *Proc. ICML*, 2002.

[24] 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. In *Proc. ICCV*, 2005.

[25] Q. Zhu, M.-C. Yeh, and K.-T. Cheng. Multimodal fusion using learned text concepts for image categorization. In *Proc. ACM Multimedia*, 2006.