# Multi-Modal Multiple-Instance Learning with the Application to the Cannabis Webpage Recognition


Yinjuan Wang
College of aviation automation
Civil Aviation University of China
Tianjin, China
yjwang1_yjs09@cauc.edu.cn

Nianhua Xie, Weiming Hu
National laboratory of pattern Recognition
Chinese Academy of Sciences
Beijing, China
(nhxie, wmhu)@nlpr.ia.ac.cn

Jinfeng Yang
College of aviation automation
Civil Aviation University of China
Tianjin, China
jfyang@cauc.edu.cn



*Abstract*—**With the development of the World Wide Web, there exists more and more illicit drug Webpages. Thus, how to screen cannabis Webpages on the internet is a quite important issue. Conventional methods that only use the keyword-based or image-based approaches are not sufficient. We propose a Multi-Modal Multiple-Instance Learning (MMMIL) approach combining both text and image information for cannabis webpage recognition. The main technical contributions of our work are two-fold. First, the text information associated with images is used to build a pre-classifier, which can pre-select pseudo positive training bags from new Webpages to update multi-modal classifier. This can be seen as a pseudo active learning process. Second, we design an efficient instance selection technique by utilizing text information to speed up the training process without compromising the performance. The experiments on a dataset containing over 40,000 images for more than 4,000 Webpages demonstrate the effectiveness and efficiency of the proposed approach.**



***Keywords- Cannabis Webpage Recognition;Multi-Modal; MIL***


## I. INTRODUCTION

The amount of Webpages has been increasing dramatically with the rapid development of the digital world. The Internet provides great convenience for users to obtain all sorts of information that they need. However there also exists illicit drug Webpages on the Internet. This kind of harmful information has a bad influence on users, especially teenagers. Just like the filtering of online pornographic content, either text-based or image-based filtering is insufficient. It is therefore important to take full advantage of image and text information for classification. In this paper, we focus on simultaneously utilizing the image information and the text information around images. The texts associated with an image in a Webpage are usually related to the semantic description of the image. In [10, 11, 12], the experiments have demonstrated that the text associated with the image is valid for image classification.

We propose a Multi-Modal Multiple-Instance Learning (MMMIL) approach for cannabis webpage recognition. One modality is the text associated the image, the other is image. Since the text information is accurate, reliable and related to the semantic description of image, through the text information, we can get a better understanding of the corresponding images. Initial instance selection technique is designed by the text information to construct the instance space. It dramatically decreases the dimension of the instance space in contrast to MILES [8]. Also, pseudo active learning method is proposed. It utilizes the text information to build a pre-classifier which selects the pseudo positive training bags from new bags. Figure 1 shows the process of selecting the pseudo positive training bags. Each webpage is treated as a bag, and each bag has different number of instances (images). Based on the existence of the pseudo positive instance in the bag, we decide whether a new bag is a pseudo positive training bag or not. A bag is pseudo positive bag if at least one of its instances is pseudo positive instance. Here the pseudo positive instance is defined as an image associated with certain keywords. Because of the existence of these certain keywords which are generally indicative of true positive instance, we tentatively treat it as a positive instance although it may in fact be a negative instance. Then the pseudo positive training bags are used to update the SVM Model. Note that although we do not have the labels for pseudo positive training bags, we assume these bags are positive because certain keywords appear in these bags. As a result, we pre-classify some new bags as pseudo positive training bags and use them in the online learning phase.

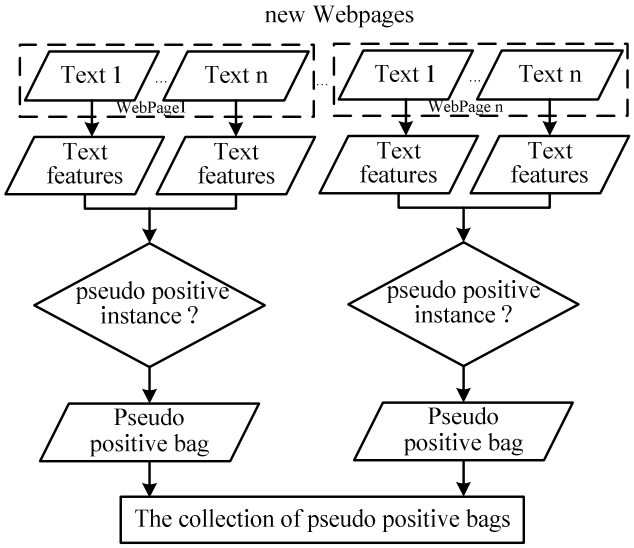


Figure 1. The process of pre-classification

The rest of the paper is organized as follows. Section II discusses the related work. In Section III, we introduce the

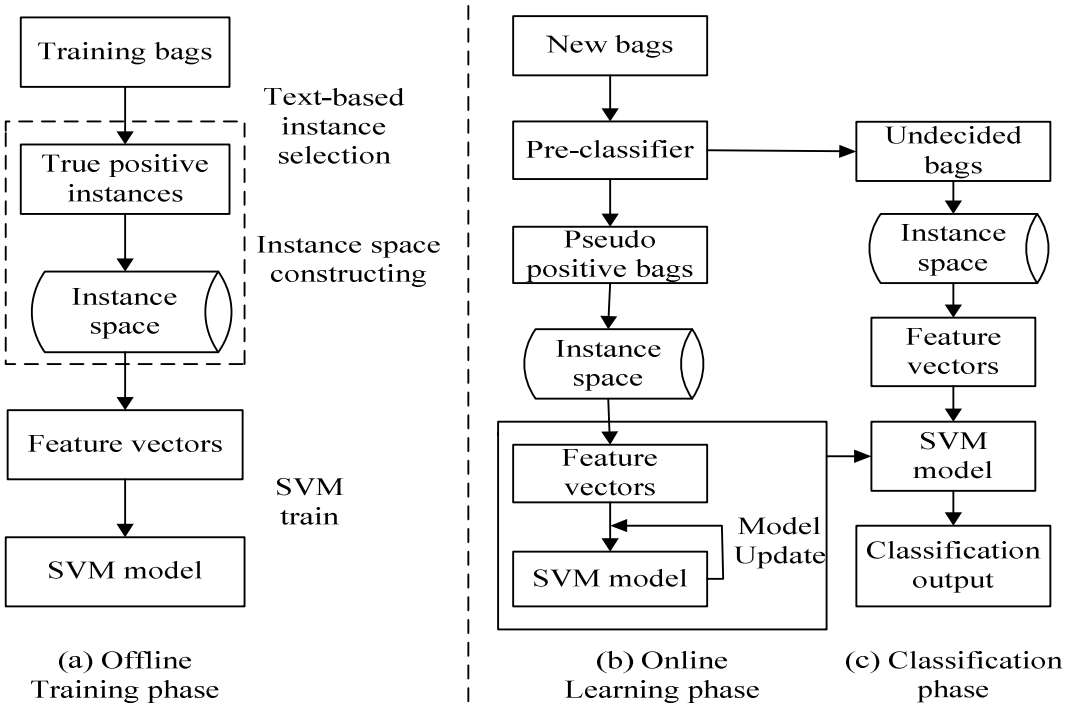


Figure 2. The framework of our method

Multi-Modal Multiple Instance learning algorithm for recognition cannabis webpage recognition. Experimental results are reported in Section IV. We conclude and discuss the future work in Section V.

## II. RELATED WORK

The Multiple Instance Learning (MIL) was first proposed by Dietterich et al. [1] in the context of drug activity prediction. In MIL, an individual example is called an instance and each bag is a collection of instances. The key assumption in MIL is that: a bag is positive if at least one of its instances is a positive instance, whereas a negative bag only contains negative instances. In order to solve the MIL problem, many methods have been proposed. An abundance of methods try to convert MIL into a standard single-instance problem and then solve it with conventional methods by representing the bags as feature vectors. Chen and Wang [2] proposed DD-SVM, which learns multiple target distributions for positive instances given by the local extrema of the EM-DD optimization to all the training instances. Fu [7] proposed the Multiple Instance Learning with Instance Selection (MILIS). Initial instance selection is done by a simple yet effective kernel density estimator on the negative instances to speed up the training process. Chen et al. [8] carried out Multiple-Instance Learning via Embedded Instance Selection (MILES), based on mapping each bag to a feature space defined by all the instances in the training bags and performing joint feature selection and classification by using the 1-norm SVM [3].

Although, this MIL approaches has been proved effective, they only use one modality and the learning process is complicated and the computation cost is expensive. Our work takes full advantage of two modalities to improve the computation and the performance of the classifier. Text feature is used to construct the instance space in the offline training phase. This can solve the computationally expensive in MILSE. In the online learning phase, we propose a pseudo active learning process to pre-select some pseudo positive training bags and then use them to update a SVM model. This can improve the performance of the SVM model.

## III. MULTI-MODAL MULTIPLE INSTANCE LEARNING

Our classification framework is illustrated in Figure 2. Figure 2(a) shows the offline training process, where text-based instance selection is first performed to construct instance space. Then all the training bags can be embedded into the instance space. This means that the MIL problem is converted into a supervised problem via similarity based feature mapping using the selected instances. Then training bags are used to train the initial SVM classifier. Figure 2(b) shows the online learning process. For some new bags, a pre-classifier which only uses the text vector is designed to select pseudo positive training bags. Then these pseudo positive training bags are embedded into the instance space which is constructed in the offline learning process. Finally, the SVM model is updated by these pseudo positive training bags. For other undecided bags, updated SVM is used to classify them.

### A. Notation

To describe the MMMIL algorithm, we need to introduce some notations. In the following, each Webpage is treated as a bag, the training set can be denoted as $B=\{B_1^+,B_2^+,...,B_{l^+}^+,B_1^-,B_2^-,B_{l^-}^-\}$ where $B_i^{+(-)}$ denotes the $i$th bag from the positive (negative) bag

and $l^{+(-)}$ denotes the number of positive (negative) bags. For the sake of convenience, we omit the sign+/- when there is no need for distinction. The label of the bag $B_i$ is $l(B_i) \in \{+1,-1\}$. The bag $B_i = \{x_{i1}, x_{i2}, ..., x_{in_i}\}$, the number of instances $n_i$ varies, for different bags contain different number of instances. Each image in a webpage is an instance $x_{ij} = \{v_j, w_j\}$ $j = (1,2,...,n_i)$ which consists of the image vector $v_i = \{v_{j1}, v_{j2}, ..., v_{j\gamma}\}$ and the text vector $w_j = \{w_{j1}, w_{j2}, ..., w_{ju}\}$. The parameter $\gamma$ is the number of visual words in the Bag-of-Visual-Word. The parameter $u$ is the number of the keywords. The entire instances in the training set are represented as $x^k$, $k = 1,2,...,n$ where

$$n = \sum_{i=1}^{l^+} n_i^+ + \sum_{i=1}^{l^-} n_i^- \tag{1}$$

## B. Review of MILES

Our work extends ideas from MILES, so we first review it before introducing our method. MILES is mainly divided into two steps: one is instance-based feature mapping, for building instance space; the other is 1-Norm support vector machines learning, for constructing classifiers and selecting important features simultaneously. The whole instance $C = \{x^k : k = 1,2...n\}$ in the training bag is used to build the instance space, and then all of the training bags are embedded into the instance space via a similarity based feature mapping. The measure of similarity between the instance $x^k$ and the bag $B_i$ is

$$s(x^k, B_i) = \max_j \exp\left(\frac{\| x_{ij} - x^k \|}{\sigma^2}\right) \tag{2}$$

For all the training set of $l^+$ positive bags and $l^-$ negative bags, the mapping yields the following matrix representation of all training bags in the instance space:

$$[m_1^+, ..., m_{l^+}^+, m_1^-, ..., m_{l^-}^+] = \begin{bmatrix} s(x^1, B_1^+) & \cdots & s(x^1, B_{l^-}^-) \\ s(x^2, B_1^+) & \cdots & s(x^2, B_{l^-}^-) \\ \cdots & \cdots & \\ s(x^n, B_1^+) & \cdots & s(x^n, B_{l^-}^-) \end{bmatrix} \tag{3}$$

After feature mapping, the MIL problem is converted to the supervised problem, then the classification problem is to find a linear classifier $y = \text{sign}(w^T m + b)$, where w and b are model parameters and m corresponds to a bag. They rewrote the parameter $\|w\|_1 = \sum_k w_k$, $|w_k| = u_k - v_k$ where $u_k$, $v_k \geqslant 0$. If either $u_k$ or $v_k$ has to equal to 0, we have $|w_k| = u_k + v_k$. The total Loss function is defined as

$$\mu \sum_{i=1}^{l^+} \xi_i + (1-\mu) \sum_{j=1}^{l^-} \eta_j \tag{4}$$

Where $\mu$ and $1-\mu$ penalize differently on false negatives and false positives, and $0 < \mu < 1$. Then the SVM approach constructs classifiers based on hyperplanes by minimizing a regularized training error,

$$\begin{aligned} &\min_{u,v,b,\xi,\eta} \lambda \sum_{k=1}^{n} (u_k + v_k) + \mu \sum_{i=1}^{l^+} \xi_i + (1-\mu) \sum_{j=1}^{l^-} \eta_j \\ &[(u - v)]^T m_i^+ + b] + \xi_i \geq 1, i = 1, ..., l^+, \\ &-[(u - v)]^T m_i^- + b] + \eta_j \geq 1, i = 1, ..., l^-, \\ &u_k, v_k \geq 0, k = 1, ..., n, \\ &\xi_i, \eta_j \geq 0, i = 1, ..., l^+, j = 1, ...l^-. \end{aligned} \tag{5}$$

where $\xi$, $\eta$ are hinge losses, $\lambda$ is called the regularization parameter. Finally, the classification of testing bag $B_i$ is computed as

$$y = sign\left(\sum_{k \in I} w_k^* s(x^k, B_i) + b^*\right) \tag{6}$$

Let $w^* = u^* - v^*$ and $b^*$ be the optimal solution of (5).The set of selected features is given as $\{s(x^k, \cdot) : k \in I\}$ where $I = \{k : |w_k^*| > 0\}$ is the index set for nonzero entries in $w^*$.

## C. Offline Learning

From the review of the MILES, we can see that it uses all the instances in the training set directly to construct the bag-level feature vector. This gives rise to a very high-dimensional feature vector even if the dataset is not large. As the negative instances add to the bags, the label of the bags will not be changed, only positive instances make a contribution to the classification. Therefore instead of using all the training instances, in the offline learning process, an efficient instance selection approach is proposed to select instances for construction of the instance space.

The text associated with an image is usually very accurate and reliable. The negative bag usually contains a few keywords, but the positive bag may contain a lot. So the text information is useful to determine the true positive instances. Let $key = \{key_1, key_2, ..., key_u\}$ be a set of keywords. The parameter $u$ represents the number of keywords. Each instance of the text feature $w_j = \{w_{j1}, w_{j2}, ...w_{ju}\}$ is represented by term frequencies of keywords. The element $w_{jk}$ represents term frequency of the keyword $key_k$.

Our initial instance selection is described as follows: for each instance $x_{ij}$ in positive training bags $B_i$, the total term frequency $\sum_{k=1}^{u} w_{jk}$ is first computed. If the maximum total term frequency $\max_j \sum_{k=1}^{u} w_{jk} > 0$, we take the corresponding instance as true positive instance. As we know, a positive bag may contain different number of true positive instances, thus if the other nonzero total term frequencies are close to the maximum total term frequency, they may also be taken as the true positive instances. The degree of the closeness is determined by the parameter $\alpha$.If they meet the condition：

$$\begin{cases} \max_j \sum_{k=1}^{u} w_{jk} - \sum_{k=1}^{u} w_{jk} < \alpha \\ \sum_{k=1}^{u} w_{jk} \neq 0 \\ \max_j \sum_{k=1}^{u} w_{jk} \neq 0 \end{cases} \tag{7}$$

these instances can be taken as true positive instances as well. If none of the total term frequency $\sum_{k=1}^{u} w_{jk}$ is nonzero in a bag, then all of the instances in this bag will not be chosen. Though none of instances is selected in this bag, we can select more instances in the bags which have more than one nonzero total term frequency.

We use the $IP = \{IP_1, IP_2, ..., IP_\omega\}$ as the collection of true positive instances, where $\omega$ is the number of the true positive

TABLE I. AVERAGE EQUAL ERROR RATE AND TIME COST

| *Method* | *EER* | *Time cost* |
|---|---|---|
| ***MILES*** | 22.22% | 85.39s |
| ***Only instance selection*** | 21.82% | 7.65s |
| ***Instance selection and pseudo active learning*** | 19.04% | 8.52s |

TABLE II. AVERAGE EQUAL ERROR RATE AND TIME COST

| *Method* | *EER* | *Time cost* |
|---|---|---|
| ***MILES*** | Out of memory | Out of memory |
| ***Only instance selection*** | 20.53% | 99s |
| ***Instance selection and pseudo active learning*** | 18.24% | 168s |

instances. Then each true positive instance in the collection of $IP$ corresponds to one dimension of the instance space.

After initial instance selection, each bag can be embedded to the instance space with the coordinates $m(B_i)$ :

$$m(B_i) = [s(IP_1, B_i), s(IP_2, B_i), ...s(IP_\omega, B_i)]^T \quad (8)$$

And the measure of similarity between the concept $IP_\varphi$ and $B_i$ is defined as

$$s(IP_\varphi, B_i) = \max_j \exp(\frac{\| x_{ij} - IP_\varphi \|}{\sigma^2}) \quad (9)$$

Then each bag in the instance space is represented as a vector. This means that the MIL problem converts to the supervised problem when all the bags have been embedded to the instance space. Finally, we train the initial SVM model as the MILES.

The major advantage of our offline learning lies in the initial instance selection. There is no complicated computation in the process of selection. Compared with the dimension of the MILES in the bag-level feature, ours is dramatically decreased.

### *D. Online learning*

The phase of online learning is divided into two parts. One is pseudo active learning and the other is SVM update. In the process of offline learning, we have trained the initial SVM model. When new bags come, we can pre-select some pseudo positive training bags by the pseudo active learning, and then update the SVM model which has been formed in offline learning by these bags to boost the performance of the classification.

In the process of pseudo active learning, a pre-classifier is first designed to pre-select some pseudo positive training bags by its text feature. However, how to judge the label of the instance in the new bag? The method is similar to the initial instance selection. For an instance $x_{ij}$, if the text feature $w_{jk}$ meets the condition as (7), the instance $x_{ij}$ is a pseudo positive instance. As long as a bag contain at least one pseudo positive instance, we consider it as the pseudo positive bag. The active learning is a form of supervised machine learning in which the learning algorithm is able to interactively query the user (or some other information source) to obtain the desired outputs at new data points [9]. However, in our method, instead of interactively querying the user, a pre-classifier is designed to select pseudo training bags, and the labels of these bags are fixed to positive. Thus, we called this process as a pseudo active learning process. When there are some pseudo positive training bags, we update SVM model. For other undecided bags, we use the updated SVM model for classification.

## IV. EXPERIMENTAL RESULT

### *A. Data collection and processing*

Two datasets are collected for experiments, one is a small dataset, which is collected from more than 200 cannabis Webpages contains over 2,000 images and 500 normal Webpages with about 4,000 images. The other is a large dataset, over 600 cannabis Webpages with about 3,300 images, and more than 3000 normal Webpages with about 26,000 images are collected. The normal Webpages includes a wide range of categories including plants, animal, hemp, cloth, news, education, etc. Different types of Webpages make the classification task more difficult. We discarded the images whose height or width smaller than 100 pixels, as these images are usually advertisements or thumbnails of no significance.

In the processing step, the standard SIFT features [4] are used for local patch description, and the bag of word model [5] is used to construct the histogram for each image. For the positive training bags, all the term frequency of the keywords is ranked. Top 36 keywords are selected such as cannibals, marijuana, etc. The simple and effective term frequency is applied to construct the text vector by the keywords. In the process of instance selection and pre-classification, the text feature is not normalized. If the text and image vector are combined as an instance, they are normalize respectively.

### *B. Experiments on the small dataset*

We make a comparison experiment between the following methods: MILES, offline learning which only uses the initial instance selection and online learning which uses the initial instance selection and pseudo active learning.

There are four parameters $\alpha$, $\sigma^2$, $\lambda$ and $u$ (in the 1-norm SVM) in the experiments, where $u$ and 1- $u$ represent penalization false negatives and false positives respectively. We fixed the $u$ =0.7, because the cost for classifying the true positive bag to the negative bag is more expensive than the cost for classifying the true negative bag to the positive bag. For the parameter $\sigma$, we used the average value of the training set. The parameters $\lambda$ and $\alpha$ were selected via a twofold cross-validation on the training set. We chose $\lambda$ from $2^{-8}$ to $2^8$ with the step size $2^1$, and $\alpha$ from 0 to 9 with the step size 1. Through experiment in our method, the parameter $\alpha$ was chosen as 5 which had the best performance. The parameter $\lambda = 2^{-6}$ gives the best performance in the method of

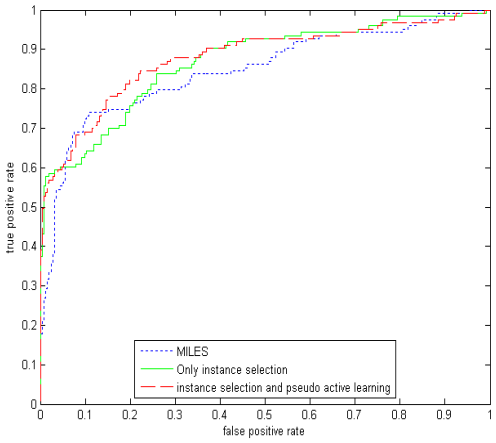


Figure 3. ROC curves for three method

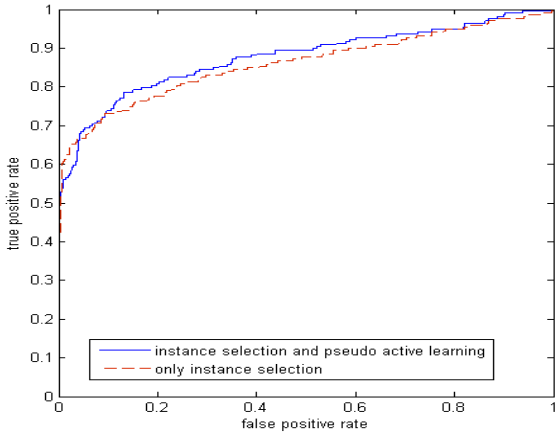


Figure 4. ROC curves of our method

the initial instance selection and pre-classification, for only the instance selection method $\lambda = 2^{-7}$, while for the MILES $\lambda = 2^{-8}$.

From Table I and Figure 3, we can see that if we only run the offline learning process without online learning, it increases the result less than 1%, but the computation decreases dramatically. This shows that the initial instance selection can greatly decrease the time complexity. The reason is that the instance space has 3530 dimensions in MILES, but in our method only 375 instances are selected after the initial instance selection. This means that the time of the feature mapping phase decreases about ten times. Table I shows that there are about ten times gap between MILES and our method. This confirms the phase of initial instance selection is reasonable and effective. Meantime our method can preserve or even enhance the performance of the classification. This proves that negative instances make no contribution to classification except increasing time cost. If we run the offline learning and online learning together, the result is increased more than 3%, but the time cost only increases a little than only using the offline learning while decreases dramatically than MILES. This proves that our online learning is effective and efficient for boosting the classification task.

### C. *Experiments on the larger dataset*

When the dataset is increased, MILES cannot work due to its complexity. From Figure 4 and Table II, we can find that our method not only can work, but also make the performance of the classifier increase. This also proves that our initial instance selection and pseudo active learning method is effective and MILES cannot apply in the large dataset.

## V. CONCLUSION AND FUTURE WORK

We have proposed a cannabis webpage classification algorithm using multi-modal multiple instance learning approach. In our approach, text information is used to design a pre-classifier to identify some pseudo positive training bags from the new bags and then use these pseudo positive training bags to update SVM model. To speed up the instance mapping process, text information has been utilized to select the true positive instances for instance space construction. This method shows promising results in comparison with the MILES. In the future, we plan to fusion other useful information (e.g., video information, other text information and so on) to increase the performance, and investigate method for better coordinate of all these information.

## ACKNOWLEDGMENT

This work was supported in part by the NSFC (Grant No. 60825204 and 60935002)

## REFERENCES

[1] T. G. Dietterich, R. H. Lathrop, and T. Lozano-P´erez, “Solving the Multiple Instance Problem with Axis-Parallel Rectangles,” Artificial Intelligence, vol. 89, nos. 1-2, pp. 31-71, 1997.

[2] Y. Chen and J. Z. Wang, “Image categorization by learning and reasoning with regions,” Journal of Machine Learning Research, vol. 5, no. 913-939, 2004.

[3] J. Zhu, S. Rosset, T. Hastie, and R. Tibshirani, “1-norm support vector machines,” in Neural Info. Proc. Systems, 2003.

[4] D. Lowe. “Distinctive image features from scale-invariant keypoints,” IJCV, vol. 60, pp. 91-110, 2004.

[5] G. Csurka, C. Dance, L. Fan, J. Williamowski and C. Bray.”Visual categorization with bags of keypoints,” In ECCV Workshop on SLCV, 2004.

[6] R. Fergus, P. Perona, and A. Zisserman. “Object class recognition by unsupervised scale-invariant learning,” In *Proc. CVPR,* 2:264-271, 2003.

[7] Y Fu, A Robles-Kelly, J Zhou. MILIS: “Multiple Instance Learning with Instance Selection,” IEEE Trans. on PAMI, 2010.

[8] Y. Chen, J. Bi and J.Z. Wang. MILES: “Multiple-instance learning via embedded instance selection,” IEEE Trans. on PAMI, 2006.

[9] Settles, Burr. "Active Learning Literature Survey," in *Computer Sciences Technical Report, 2009*

[10] O. Yakhnenko and V. Honavar. “Multi-modal hierarchical Dirichlet process model for predicting image annotation and image-object label correspondence,” In SIAM SDM, 2009.

[11] M Guillaumin, J Verbeek and C Schmid. ” Multimodal semi-supervised learning for image classification,” in cvpr , 2010.

[12] Q Zhu, M Chen and Y Kwang-Ting Cheng. “Multimodal Fusion using Learned Text Concepts for Image Categorization,” in MULTIMEDIA proceedings of the ACM international conference on Multimedia, 2009.