A Multiple Instance Learning Approach to Image Annotation with Saliency Map

Tran Phuong Nhung[1], Cam-Tu Nguyen[2], Jinhee Chun[1], Ha Vu Le[3] and Takeshi Tokuyama[1]
[1] *Graduate School of Information Sciences, Tohoku University, Sendai, Japan*
[2] *National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China*
[3] *VNU University of Engineering and Technology, Hanoi, Vietnam*

Keywords: Visual Saliency, Image Annotation, Multiple Instance Learning.

Abstract: This paper presents a novel approach to image annotation based on multi-instance learning (MIL) and saliency map. Image Annotation is an automatic process of assigning labels to images so as to obtain semantic retrieval of images. This problem is often ambiguous as a label is given to the whole image while it may only corresponds to a small region in the image. As a result, MIL methods are suitable solutions to resolve the ambiguities during learning. On the other hand, saliency detection aims at detecting foreground/background regions in images. Once we obtain this information, labels and image regions can be aligned better, i.e., foreground labels (background labels) are more sensitive to foreground areas (background areas). Our proposed method, which is based on an ensemble of MIL classifiers from two views (background/foreground), improves annotation performance in comparison to baseline methods that do not exploit saliency information.

1 INTRODUCTION

Image annotation is an automatic process of finding appropriate semantic labels for images from a predefined vocabulary. In other words, an image is assigned with a few relevant text keywords that reflect the image's visual content. Image annotation has become a prominent research topic in the domain of medical image interpretation, computer vision and semantic scene classification. In particular, it is useful towards image retrieval as annotated keywords greatly narrow the semantic gap between low level (visual) features and high level semantics.

Automatic image annotation is a challenging task due to various imaging conditions, complex and hard-to-describe objects, as well as a highly textured background (Qi and Han, 2007). Unlike object recognition, image annotation is a "weak labeling" problem. That means a label is assigned to the whole image without any alignment of regions and the label (Carneiro et al., 2007). As a result, the multi-instance approach (Zhou and Zhang, 2007) is a natural solution to resolve the ambiguities in training data. Unlike traditional supervised learning, an example (e.g. an image) in MIL is not described by a feature vector (a single instance) but a set of feature vectors (a bag of instances). By considering an image as an example and its instances as feature vectors extracted from subregions of the image, the problem of image annotation naturally fits the multi-instance setting. There were several studies (Carneiro et al., 2007; Nguyen et al., 2010; Yang et al., 2006) that have successfully applied MIL to the problem of image annotation. One disadvantage of those methods is that they treated instances (sub-regions in images) equally for every label. Our observation is that we can weight instances differently with respect to different labels. An example is given in Figure 1, the instance (feature vector) that falls into the salient region is more relevant the foreground object (polar bear), while the other instances (background regions) are more important towards the background label (snow). Even though we do not really have the correspondence between the instances (regions) and the labels (polar bear, snow) due to the "weak labeling" problem, the observation can be useful towards reducing noises in learning an image annotation system.

In order to apply the above idea, there are several questions that need to be addressed. The first question is how to obtain visual saliency for image annotation. Fortunately, this problem has been well studied in computer vision (Hou et al., 2012) where the main assumption is that "the image energy focuses on the locations of a spatially sparse foreground, relative to a

Phuong Nhung T., Nguyen C., Chun J., Le H. and Tokuyama T..
A Multiple Instance Learning Approach to Image Annotation with Saliency Map.
DOI: 10.5220/0004543901520159
In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval and the International Conference on Knowledge Management and Information Sharing* (KDIR-2013), pages 152-159
ISBN: 978-989-8565-75-4

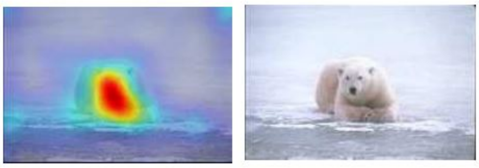

Figure 1: An example image from Corel5K data set with its detected saliency map.

spectrally sparse background". The second question is that given a set of labels, how can we know a label is a foreground label or background label without additional information from humans. Our approach is that we let the training data decide. More specifically, we firstly obtain two kinds of learners from two different views (background instances/foreground instances). Next, we base on the performances on the training data to decide the weights, and then combine two kind of learners to obtain the final learner for annotation. To the best of our knowledge, this is the first attempt that exploits saliency map for image annotation. Experimental results show the effectiveness of our proposed method.

The rest of the paper is organized as follows. Section 2 reviews related works. Section 3 and 4 describe the fundamental introduction to two building theories (MIL and visual saliency detection) in our approach. Section 5 introduces the proposed image annotation approach based on MIL and visual saliency. Section 6 reports and discusses the experimental results. Section 7 summarizes the main idea of the paper and concludes some remarks.

2 RELATED WORKS

Image annotation is a typical Multi-instance Multi-label problem (MIML) (Zhou and Zhang, 2007) where an image is represented by a set of regions (instances) and assigned with multiple labels. The most successful method is based on propagating labels from the nearest neighbors in the training data (Lavrenko et al., 2003; Guillaumin et al., 2009). The methods, however, degenerate the task of image annotation from a MIML problem into a set of single-instance multi-label (SIML) problems; thus, they do not explicitly cope with noises in training process. Although the propagation approach is simple, the annotation time increases linearly with the size of the training data set.

Another approach is to degenerate the image annotation problem into a set of multi-instance single-label problems. Considering one specific label, we build a classifier that determines whether to assign a given label to an image. This classification problem is a multi-instance problem as a label (e.g. tiger) is assigned to an image but only relevant to some subregions of the image (the other subregions are relevant to other labels such as "forest", "tree"). The MIL approach explicitly takes into account these noises in training. In the following, we will follow the terminologies in MIL and respectively refer to "instance" and "bag" as a feature vector corresponding to a subregion of an image and the image itself. MISVM and miSVM algorithms (Andrews et al., 2002) were adapted from single-instance learning version of SVM in order to cope with multiple instance data version. On the other hand, Yang et al. introduced Asymmetric SVM (ASVM) to pose different asymmetrical loss function to two types of errors (false positive and false negative) in order to improve the accuracy of annotation process (Yang et al., 2006). In addition, Supervised Multi-class Labeling (SML) (Carneiro et al., 2007) is based on MIL and density estimation in order to measure the conditional distribution of feature given a specific word. SML uses a bag of image examples annotated by a particular word and estimate the distribution of image features extracted from the bag of images. The distribution is fitted by mixture Gaussian distribution in a hierarchical manner. SML does not consider the negative examples in the learning binary examples. Since SML only uses positive examples for each concept in training, the training complexity is reduced considerably. On another attempt, CMLMI (Nguyen et al., 2011) proposed a cascade of multi-level multi-instance classifiers to reduce class imbalance in MIL, and exploited multiple modalities to improve image annotation. In this paper, we propose an approach to image annotation that uses saliency information (visual saliency map). Our main idea is that the salient regions are more important towards assigning foreground labels to images.

Visual saliency detection has attracted a lot of interest in computer vision as it provides fast solutions to several complex processing. There are many studies (Ueli et al., 2004) (Navalpakkam and Itti, 2006) that show the effectiveness of visual saliency map for object recognition, tracking and detection. To the our best knowledge, however, such kind of information has not been used for image annotation before. This paper shows that incorporating visual saliency information to image annotation can reduce the noises associated with "weak labeling" problem, thus improve the performance annotation process. We demonstrate our method with miSVM, the multi-instance version of SVM, with the support of visual saliency. Thus, in the following, we will give a brief introduction to these methods.

3 MULTIPLE INSTANCE LEARNING WITH miSVM

MIL is a variation of supervised learning problems with incomplete knowledge about labels of training examples (Zhou and Zhang, 2007). The majority of the work in MIL is concerned with binary classification problems, where each example has a classification label that assigns it into one of two categories "positive" or "negative". The goal is to learn a model based on the training examples that are effective in predicting the classification labels of future examples.

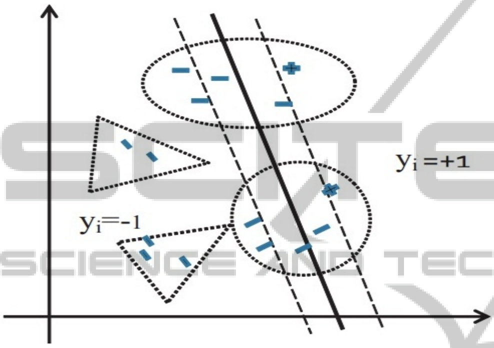

Figure 2: Multiple Instance Learning: positive and negative bags are denoted by circles and triangles respectively.

miSVM (Andrews et al., 2002; Nguyen et al., 2011) extends the notion of the margin from an individual instance to a set of instances (Figure 2). Let $D_y = \{(X_i,Y_i)|i=1,\ldots,N, X_i=\{\mathbf{x}_j\}; Y_i=\{+1,-1\}\}$ be a set of images (bags), where a bag X_i of instances ($\mathbf{x}_j$) is positive ($Y_i = 1$) if at least one instance $\mathbf{x}_j \in X_i$ has its label Y_i positive (the subregion in the image corresponds to positive label). As shown in Figure 2, positive bags are denoted by circles and negative bags are marked as triangles. The relationship between instance labels and bag labels can be compressed as $Y_i = \max(y_j), j = 1,\ldots,|X_i|$. The functional margin of a bag with respect to a hyperplane is defined in (Andrews et al., 2002) as follows:

$$Y_i \max_{\mathbf{x}_j \in X_i} (\mathbf{a}\mathbf{x}_j + b)$$

The prediction then has the form $Y_i = sgn \max_{\mathbf{x}_j \in X_i} (\mathbf{a}\mathbf{x}_j + b)$. The margin of a positive bag is the margin of the most positive instance, while the margin of a negative bag is defined as the "least negative" instance. Keeping the definition of bag margin in mind, the Multiple Instance SVM (miSVM) is defined as following:

$$\text{minimize: } \frac{1}{2}||\mathbf{a}|| + C\sum_{i=1}^{N} \xi_i$$

$$\text{subject to: } Y_i \max_{\mathbf{x_j} \in X_i} (\mathbf{a}\mathbf{x}_j + b) \geq 1 - \xi_i, i = 1,\ldots,N, \xi_i \geq 0$$

By introducing selector variables s_i which denotes the instance selected as the positive "witness" of a positive bag X_i, Andrews et al. has derived an optimization heuristics. The general scheme of optimization heuristics alternates two steps: 1) for given selector variables, train SVMs based on selected positive instances and all negative ones; 2) based on current trained SVMs, updates selector variables. The process finishes when no change in selector variables.

4 VISUAL SALIENCY

Research of visual psychology has shown that when observing an image, people do not have the same interest in all of it. The Human Visual System (HVS) has a remarkable ability to automatically focus on only salient regions. Saliency at a given location is determined by the degree of difference between that location and its surrounds in a small neighborhood. Thus, saliency map is obtained from summing up differences of image pixels with their respective surrounding pixels.

In this paper, we propose an approach to image annotation that uses salient interesting points. We use a vector space representation of the local descriptors of the salient regions to describe the image in an invariant manner, and a classifier which is achieved based on learning processes generates the correct annotation for the images. In particular, for our algorithm, we select salient regions using the method which is called "image signature" (Hou et al., 2012). It is defined as the sign function of the Discrete Cosine Transform (DCT) of an image. An overview of the complete algorithm is presented in the Figure 3. Firstly, the image is decomposed into three channels (RGB). Then, a saliency map is computed for each color channel independently. Finally, saliency map is simply summed across three color channels.

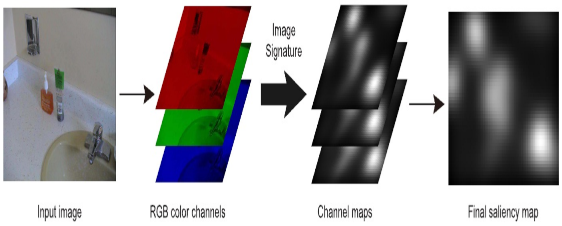

Figure 3: Overview of the process of finding salient regions (Hou et al., 2012).

Initially, they consider gray-scale images which are presented as follows:

$$x = f + b \qquad (x, f, b \in \mathbb{R}^N) \qquad (1)$$

where, f represents the foreground and is assumed to be sparsely supported in the standard spatial basis; b represents the background and is assumed to be sparsely supported in the basis of DCT. That means both f and $\hat{b}$ which is DCT(b) have only a small number of nonzero components in the standard spatial basis and the DCT domain respectively.

An illustration of image is standard spatial basis and DCT is shown in Figure 4. The first row: f, b, x in the spatial domain. The second row: the same signals are represented in the DCT domain: $\hat{f}$, $\hat{b}$, $\hat{x}$.

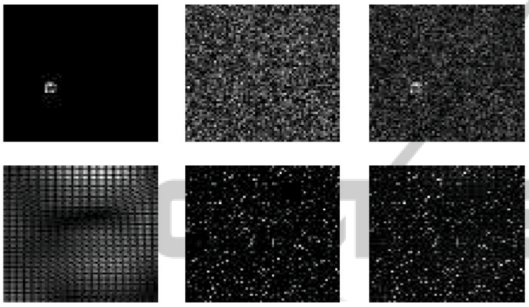

Figure 4: An illustration of the randomly generated images in the spatial domain and Discrete Cosine Transform domain (Hou et al., 2012).

In Figure 4, f is randomly generated in the standard spatial basis, then f is presented in DCT domain by computing DCT(f). Since b is assumed sparsely supported in DCT domain, firstly generate DCT(b), and then b is achieved by inversely transformed back into spatial domain b=IDCT(DCT(b)). This idea is the same to represent for x in spatial domain and DCT domain. Based on the property that f and b are sparsely supported in different domains, they can isolate the set of fixes for f which is nonzero.

Given an image x, the main idea of image signature algorithm is described as follows: firstly, they separate the support of f by taking the sign of the mixture signal x in the DCT domain. The purpose of this step is to sharpen the image by keeping only the high frequency components.

$$ImageSignature(x) = sign(DCT(x)) = sign(\hat{x}) \quad (2)$$

Then, inversely transform it back into the spatial domain by computing the reconstructed image.

$$\bar{x} = IDCT[sign(\hat{x})] \quad (3)$$

Finally, the saliency map m of an image is generated by smoothing the squared reconstructed image.

$$m = g * (\bar{x} \circ \bar{x}) \quad (4)$$

where, g is a Gaussian kernel. In this case, a Gaussian smoothing is necessary since a salient object is not only spatially sparse but also localized in a contiguous region.

5 IMAGE ANNOTATION BASED ON MIL & VISUAL SALIENCY

In this section, we formulate image annotation as a supervised learning problem under multiple instance learning frameworks. In order to improve the accuracy and efficient in image annotation, we introduce the approach to it based on visual saliency.

5.1 Overview of proposed Model

Our model contains two processes and it is showed in the Figure 5.

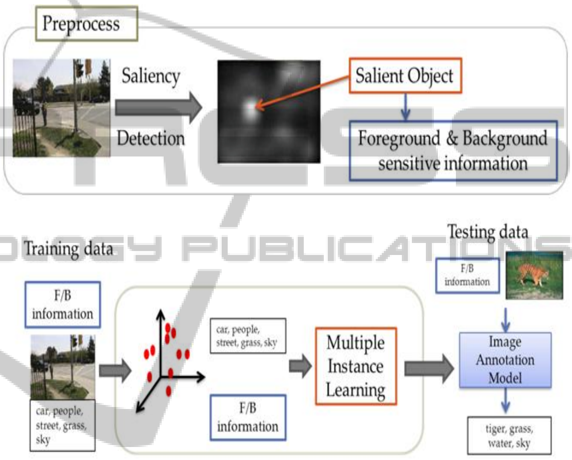

Figure 5: Overview of proposed model.

In the pre-process, we apply image signature algorithm (Hou et al., 2012) to detect the salient object of images. Based on the saliency map, we achieve the foreground and background sensitive information of the image. In the next step, we construct an automatic image annotation system based on saliency map. As mentioned earlier, we decompose the problem of image annotation into a set of multi-instance single-label problems. In other words, a MIL binary classifier is learned in order to link a image region to a specific keyword (e.g. tiger) , i.e. an image is assigned with "tiger" if it is classified "positive" according to the "tiger" MIL classifier . Then the learned classifiers are used to produce the annotation for new input images. During learning MIL classifiers, we further reduce the ambiguity in image annotation by using the saliency information as a supporting factor for MIL. The idea is that the region is more salient is sampled more for foreground labels. A simple solution is based on sampling/weighting method. A salient region is weighted more, thus it have higher ability to be annotated for foreground label.

5.2 Image Annotation with Foreground and Background Decomposition

We assume that every image is divided into regions, and each region is described by a feature vector (instance). Thus, an image X_i is presented by a set of feature vectors $X_i = \{x_{i1}, x_{i2}, ..., x_{in}\}$, where n is the number of regions in image X_i.

Based on the saliency map which is achieved by image signature algorithm (Hou et al., 2012), we assign the saliency values to regions of one image. Note again that we have a one-to-one correspondence between regions and instances. Supposed that m is the saliency map, where $m(l,k)$ represents the saliency value of pixel (l,k) in the image, δ_{ij} that a probability of $region_j$ (or instance j) of image X_i is foreground. Therefore, δ_{ij} is defined as follows:

$$\delta_{ij} = \frac{\sum_{(l,k)\ in\ region\ r_j} m(l,k)}{\sum_{(l',k')\ in\ image\ X_i} m(l',k')} \tag{5}$$

Then, $(1-\delta_{ij})$ indicates the probability for $region_j$ of image X_i to be a background region.

Considering the multiple instance Support Vector Machine (miSVM) algorithm (Andrews et al., 2002) that work directly with the bags of instances, the question is that how we modify the instance weights when we have already gained foreground-sensitive and background-sensitive information. The proposed solution is that add more "foreground-sensitive" or "background-sensitive" information to positive bags. Note that we only deal with the instances in the positive bags during training since negative bags have no ambiguous (all the instances in negative bags are negative).

For foreground enhanced bags (images with more "foreground-sensitive" information"), we will enforce foreground instances by adding more foreground-sensitive "pseudo" instances. On the other hand, for background enhanced bags, we will enforce background instances by adding more background-sensitive "pseudo" instances. Note that foreground enhanced bags (background enhanced bags) have more useful information towards foreground (background) labels.

Given a maximum the number of added pseudo instances K, and X_i is the bag of instances/regions; the pseudo code for adding more pseudo instance is presented as in Algorithms 1-2.

Algorithm 1: Generating Foreground-sensitive Bag.

```
Input: $\hat{X}_i = \{X_i, \delta_i\}$, $K$ is the number of added pseudo-instance.
Output: Foreground-enhanced bag $X_i^f$
Initialize $X_i^f = X_i$
Consider the set of C instances in $X_i$
for i=1:K do
    Sample c from Multinominal($\delta_i$)
    Add instance c to bag $X_i^f$
end for
```

Algorithm 2: Generating Background-sensitive Bag.

```
Input: $\hat{X}_i = \{X_i, \delta_i\}$, $K$ is the number of added pseudo-instance.
Output: Background-enhanced bag $X_i^b$
Initialize $X_i^b = X_i$
Consider the set of C instances in $X_i$
for i=1:K do
    Sample c from Multinominal(1-$\delta_i$)
    Add instance c to bag $X_i^b$
end for
```

Foreground-enhanced bag ${X^f}_i$ is enriched with foreground-sensitive information. The probability of instance/region c added to bag ${X^f}_i$ is estimated based on the multinomial distribution of δ; where $\delta = \{\delta_{i1}, \delta_{i2}, , \delta_{in}\}$ is a vector of n elements (n is number of regions), and δ_{ij} $(j = 1:n)$ is the probability of $region_j$ is foreground so that $\sum_{j=1}^{n} \delta_{ij} = 1$. δ_{ij} value is calculated based on the saliency value.

Besides, background-enhanced bag ${X^b}_i$ is enriched with background-sensitive information. The probability of instance/region c added to bag ${X^b}_i$ is estimated based on the multinomial distribution of $(1-\delta)$; where $(1-\delta) = \{\frac{1-\delta_{i1}}{\sum_{j=1}^{n}(1-\delta_{ij})}, \frac{1-\delta_{i2}}{\sum_{j=1}^{n}(1-\delta_{ij})}, ..., \frac{1-\delta_{in}}{\sum_{j=1}^{n}(1-\delta_{ij})}\}$: is a vector of n elements, and $(1-\delta_{ij})$ (with $j = 1, ..., n$) is the probability of $region_j$ is background.

Furthermore, in order to combine foreground-sensitive and background-sensitive information at classifier level, we will build the classifier based on AdaBoost two-step boosting model. Assume that, based on the foreground-sensitive and background-sensitive information we can build two classifiers H^f, H^b respectively. The final classifier of bag X_i is the combination of H^f and H^b:

$$H(X_i) = \lambda \times H^f + (1-\lambda) \times H^b \tag{6}$$

In order to automatically estimate the parameter λ, we build classifier ensemble based on AdaBoost two-step boosting with base learner is miSVM. The pseudo code of AdaBoost two-step boosting is presented as follows:

The method of combining the class predictions from multiple classifiers is known as ensemble learning (Rokach, 2010). AdaBoost is one of the most prominent ensemble learning algorithms. The main idea is that AdaBoost generates multiple training sets from the original training sets and then trains compo-

Algorithm 3: Two-Step Boosting with MIL.

```
Input:
    Given a training dataset $D = \{(\hat{X}_1, y_{k1}), \ldots, (\hat{X}_d, y_{kd})\}$ w.r.t the label $y_k$
    where $y_{ki} \in \{-1, +1\}$ and $\hat{X}_i = \{X_i, \delta_i\}$;
    A testing image $\hat{X}$;
    The number of iterations $T$ for AdaBoost.
Output:
    $H$ - an ensemble classifier of $2 \times T$ weak classifiers;
    The predictive value for $X$.
TRAINING:
Initialize $D^f = \{\}, D^b = \{\}$
for each positive bag $X_i$ in $D$ do
    Obtain $X_i^f$, $X_i^b$ from $\hat{X}_i$ using Algorithm 1-2.
    $\mathcal{D}^f = D^f \cup X_i^f$ and $\mathcal{D}^b = D^b \cup X_i^b$
end for
Boosting 1st step: foreground-sensitive
$[H^f, \lambda^f] \leftarrow miSVM.Adaboost(D^f, T)$
Boosting 2st step: background-sensitive
$[H^b, \lambda^b] \leftarrow miSVM.Adaboost(D^b, T)$
TESTING:
Obtain $X^f$ and $X_b$ from $\hat{X} = \{X, \delta\}$.
Predictive value for testing image $\hat{X}$
    $H(\hat{X}) = \frac{\sum_{i=1}^{T} \lambda_i^f h_i^f(X^f) + \sum_{i=1}^{T} \lambda_i^b h_i^b(X^b)}{\sum_{i=1}^{T} \lambda_i^f + \sum_{i=1}^{T} \lambda_i^b}$
```

nent learners. It focus on the misclassified from previous round. From each generated training set, AdaBoost induces the ensemble of classifiers by adaptively changing the distribution of the training set based on the accuracy of the previously created classifiers and then use a measure of classifier performance to weight the selection of training examples and the voting. Initially, all weights are set equally, but on each round, the weights of incorrectly classified examples are increased so that the weak learner is forced to focus on the hard examples in the training set. The predictions of the component learners are combined via majority voting, where the class label receiving the most number of votes is regarded as the final prediction.

6 RESULTS AND DISCUSSIONS

6.1 Dataset and Experimental Settings

We use the Corel5k benchmark for the experiment (Duygulu et al., 2002). It contains 5,000 images and each image is segmented into 1-10 regions. The data set is divide into two parts: training set contains 4,500 images and the rest of 500 images for testing. In addition, each image is manually annotated with 1 to 5 keywords from the vocabulary list of 371 distinct words. Prior to modeling, every image in the data set is pre-segmented into sub-regions using normalized cuts algorithms (Shi and Malik, 2000). The feature set consists of 36 features were extracted for each region: 18 color features, 12 texture features and 6 shape features (Duygulu et al., 2002).

We implement classifier ensemble based-on Adaboost with base learner is miSVM (Rokach, 2010). We use the AdaBoost algorithm which is implemented in the WEKA (Hall et al., 2009) with the number of iterations is set to 10. We compare 5 models with base learner is miSVM:

- Baseline_miSVM: baseline model;
- AdaBoost_miSVM: classifier ensemble based on Adaboost;
- F_AdaBoost_miSVM: classifier ensemble based on Adaboost and boosting one step with foreground sensitive;
- B_AdaBoost_miSVM: classifier ensemble based on Adaboost and boosting one step with background sensitive;
- FB_AdaBoost_miSVM: classifier ensemble based on Adaboost and boosting two step with foreground and background sensitive.

The quality of automatic image annotation is also measured through the process of retrieving the test images with a single keyword. For each image, top words are indexes using probabilities of those words generated by image annotation. Given a single-word query, the system returns all images annotated with that word, ordered by probabilities. We limited the number of indexed words per image; in particular, we only obtained top 5 words per image for indexing. Regarding a word w, the number of correctly annotated images is denoted as N_c, the number of retrieved images is denoted as N_r, and the number of truly related images in the testing set is denoted as N_t. The precision (P), recall (R) and F_1 are defined as follows:

$$P(w) = \frac{N_c}{N_r}, \quad R(w) = \frac{N_c}{N_t}, \quad F_1 = \frac{2 \times P \times R}{P + R} \quad (7)$$

We select 70 mostly used keywords in vocabulary list and perform our experiments. The precision and recall averaged over the set of words occurring in the test images are using for evaluating image annotation process.

6.2 Results and Discussions

Our proposed models using foreground and background-sensitive information in image annotation are compared to baseline work in order to demonstrate that image annotation with foreground and background decomposition outperforms other traditional methods.

The average precision, recall and F_1 for baseline work and our four proposed models on 70 mostly used keywords are reported in Table 1.

Table 1: Precision, Recall and F_1 of five models.

	P	R	F_1
Baseline_miSVM	19.5%	37.5%	24.7%
AdaBoost_miSVM	26.7%	37.8%	30.5%
F_AdaBoost_miSVM	27.4%	37.8%	30.9%
B_AdaBoost_miSVM	25.3%	35.8%	28.6%
FB_AdaBoost_miSVM	31.9%	42.2%	36.3%

According to Table 1, it is obviously that all of four proposed models have better results compared to baseline work. Especially, the proposed model FB_AdaBoost_miSVM has the best result in annotation performance, which gains 31.9%, 42.2% and 36.3% on precision, recall and F_1 value respectively. Besides, the model use both foreground and background sensitive information gain better precision, recall and F_1 value than models only use foreground sensitive or background sensitive information. Foreground-sensitive information seems useful than background-sensitive information for image annotation. It is observed that F_AdaBoost_miSVM model has better results in terms of precision, recall and F_1 than B_AdaBoost_miSVM.

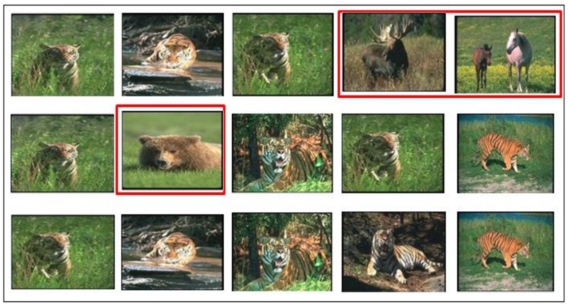

Figure 6: Top five retrieval images for query "tiger". From top to bottom: Baseline_miSVM, F_AdaBoost_miSVM, FB_AdaBoost_miSVM method.

Table 1 demonstrates that our proposed methods have some improvement on baseline work via statistical view. In Figure 6, we illustrate this improvement via image retrieval process. Considering top five retrieval images for query "tiger" by using Baseline_miSVM, F_AdaBoost_miSVM and FB_AdaBoost_miSVM model, the difference is easy to notice. For Baseline_miSVM, two of top five images are not correct. While one image in top five is not exactly corresponded to the query for F_AdaBoost_miSVM, FB_AdaBoost_miSVM successfully retrieved all correct top five images.

In addition, Figure 7 shows the comparison of the ground truth of sample images with their annotation results produced by our four proposed models.

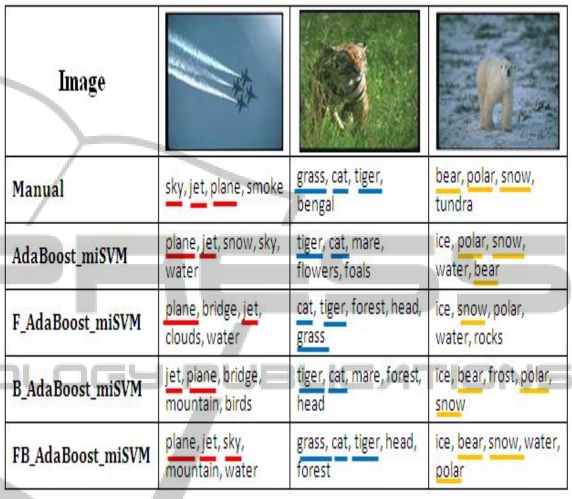

Image			
Manual	sky, jet, plane, smoke	grass, cat, tiger, bengal	bear, polar, snow, tundra
AdaBoost_miSVM	plane, jet, snow, sky, water	tiger, cat, mare, flowers, foals	ice, polar, snow, water, bear
F_AdaBoost_miSVM	plane, bridge, jet, clouds, water	cat, tiger, forest, head, grass	ice, snow, polar, water, rocks
B_AdaBoost_miSVM	jet, plane, bridge, mountain, birds	tiger, cat, mare, forest, head	ice, bear, frost, polar, snow
FB_AdaBoost_miSVM	plane, jet, sky, mountain, water	grass, cat, tiger, head, forest	ice, bear, snow, water, polar

Figure 7: Comparisons of annotations made by our proposed methods and annotations made by human.

In comparison to manual annotation by human, the annotations made by our methods are significantly reliable. In most cases, the proposed models are able to annotate the important keywords to images somehow accurately reflect its semantic meaning. In addition, applying both foreground and background sensitive information give us the best results in annotation compared to model with or without foreground/background sensitive information. As you can see on three examples in Figure 7, while the AdaBoost_miSVM, F_AdaBoost_miSVM and B_AdaBoost_miSVM only products two correct keywords to images, FB_AdaBoost_miSVM model precisely products three out of five keywords.

From the experimental results, we also observed that labels which are considered sensitive to background-information (e.g. water, grass, sky, coast, etc) have performance of B_Adaboost_miSVM better than F_Adaboost_miSVM. In order to achieve better results, the models like B_AdaBoost_miSVM should be more important for background-sensitive labels. On the contrary, the keywords like tiger, bear, horses and so on are foreground-sensitive labels (F_Adaboost_miSVM is better than B_AdaBoost_miSVM). Models like F_AdaBoost_miSVM should be used in this case. However, if there is no information about foreground/background labels, the FB_Adaboost_miSVM is the most suitable method as it will automatically estimate the weights of foreground/background information in learning the ensemble classifier.

7 CONCLUSIONS

In this paper, we have presented a novel approach for image annotation with the foreground and background decomposition on the view of multiple instance learning. This study is to make use of saliency map to reduce the ambiguity in multiple instance learning for image annotation. Therefore, a simple method based on sampling/weighting method is considered. The main idea is that the salient objects have higher probability to be foreground.

The empirical results in this paper show that applying the foreground and background decomposition to image annotation can yield good performance in most cases. This provides a very simple and efficient solution to weak labeling problem in image annotation.

The results in this paper additionally show that models using both foreground and background information in image annotation outperforms models only use foreground or background information. Besides, it also proves that classifier ensemble based on AdaBoost significantly improves the classification accuracy.

REFERENCES

Andrews, S., Tsochantaridis, I., and Hofmann, T. (2002). Support vector machines for multiple-instance learning. In *NIPS*, pages 561–568.

Carneiro, G., Chan, A. B., Moreno, P. J., and Vasconcelos, N. (2007). Supervised learning of semantic classes for image annotation and retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(3):394–410.

Duygulu, P., Barnard, K., Freitas, J. F. G. d., and Forsyth, D. A. (2002). Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Proceedings of the 7th European Conference on Computer Vision-Part IV*, ECCV '02, pages 97–112, London, UK.

Guillaumin, M., Mensink, T., Verbeek, J., and Schmid, C. (2009). Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. In *Proceedings of the 12th International Conference on Computer Vision (ICCV)*, pages 309–316.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The weka data mining software: An update. In *SIGKDD Explorations*, volume 11.

Hou, X., Harel, J., and Koch, C. (2012). Image signature: Highlighting sparse salient regions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(1):194–201.

Lavrenko, V., Manmatha, R., and Jeon, J. (2003). A model for learning the semantics of pictures. In *Proccedings of the 16th Conference on Neural Information Processing Systems (NIPS'03)*. MIT Press.

Navalpakkam, V. and Itti, L. (2006). An integrated model of top-down and bottom-up attention for optimizing detection speed. In *In IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '06, pages 2049–2056.

Nguyen, C.-T., Kaothanthong, N., Phan, X.-H., and Tokuyama, T. (2010). A feature-word-topic model for image annotation. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, CIKM '10, pages 1481–1484, New York, USA. ACM.

Nguyen, C.-T., Le, H. V., and Tokuyama, T. (2011). Cascade of multi-level multi-instance classifiers for image annotation. In *KDIR '11: Proceedings of the International Conference on Knowledge Discovery and Information Retrieval*, pages 14–23, Paris, France.

Qi, X. and Han, Y. (2007). Incorporating multiple svms for automatic image annotation. *Pattern Recogn.*, 40(2):728–741.

Rokach, L. (2010). Ensemble-based classifiers. *Artif. Intell. Rev.*, 33(1-2):1–39.

Shi, J. and Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8):888–905.

Ueli, R., Dirk, W., Christof, K., and Pietro, P. (2004). Is bottom-up attention useful for object recognition. In *In IEEE Conference on Computer Vision and Pattern Recognition*, CVPR 2004, pages 37–44.

Yang, C., Dong, M., and Hua, J. (2006). Region-based image annotation using asymmetrical support vector machine-based multiple-instance learning. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2*, CVPR '06, pages 2057–2063, Washington, DC, USA. IEEE Computer Society.

Zhou, Z.-H. and Zhang, M.-L. (2007). Multi-instance multi-label learning with application to scene classification. In *Proceedings of the 19th Conference on Neural Information Processing Systems (NIPS)*, pages 1609–1616. Monreal, Canada.