CASCADE OF MULTI-LEVEL MULTI-INSTANCE CLASSIFIERS FOR IMAGE ANNOTATION

Cam-Tu Nguyen[1], Ha Vu Le [2] and Takeshi Tokuyama[1]
[1]*Graduate School of Information Sciences, Tohoku University, Sendai, Japan*
[2]*VNU University of Engineering and Technology, Hanoi, Vietnam*

Keywords: Image annotation, Cascade algorithm, Multi-level feature extraction.

Abstract: This paper introduces a new scheme for automatic image annotation based on cascading multi-level multi-instance classifiers (CMLMI). The proposed scheme employs a hierarchy for visual feature extraction, in which the feature set includes features extracted from the whole image at the coarsest level and from the overlapping sub-regions at finer levels. Multi-instance learning (MIL) is used to learn the "weak classifiers" for these levels in a cascade manner. The underlying idea is that the coarse levels are suitable for background labels such as "forest" and "city", while finer levels bring useful information about foreground objects like "tiger" and "car". The cascade manner allows this scheme to incorporate "important" negative samples during the learning process, hence reducing the "weakly labeling" problem by excluding ambiguous background labels associated with the negative samples. Experiments show that the CMLMI achieve significant improvements over baseline methods as well as existing MIL-based methods.

1 INTRODUCTION

Only after a couple of years, online photo-sharing websites (Flickr, Picassa web, Photobucket, etc.), which host hundreds of millions of pictures, have quickly become an integral part of the Internet. As a result, the need for tagging images and multimedia data with semantic labels becomes increasingly important in order to make the Web more well-organized and accessible. On the other hand, the enormous amount of photos taken everyday makes the task of manual annotation an extremely time-consuming and expensive task. Automatic image annotation therefore receives significant interest in image retrieval and multimedia mining.

Although *image classification* and *object recognition* also assign meta data to images, the difference of image annotation from classification and recognition defines its typical challenging issues. In general, the number of labels (classes/objects) is usually larger in image annotation compared to classification and recognition. Because of the dominating number of negative examples, both the one-vs-one and one-vs-all schemes in multi-class supervised learning do not scale very well for image annotation. Unlike object recognition, image annotation is "weakly labeling" (Carneiro et al., 2007), that is a label is assigned to one image without indication of the region corresponding to that label. Moreover, scalability requirement prevents researchers to investigate feature extraction for every label in image annotation. This, however, can be performed with a limited number of objects in object recognition . On the other hand, the variety of visual representations of objects suggests that we should not depend on one feature extraction method to work well with a large number of labels (Akbas and Vural, 2007; Makadia et al., 2010).

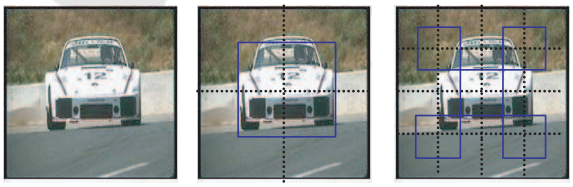

Figure 1: Level 1: the whole image; Level 2: 2x2 grid + 1 subregion in the center; Level 3: 4x4 grid + 5 overlapping subregions (blue border rectangles).

Motivated by the aforementioned issues, we propose a new learning method - a cascade of multilevel multi-instance classifiers (CMLMI) for image annotation. The idea behind our approach is that coarser levels provide better description for background and common concepts such as "forest, building, moun-

Nguyen C., Le H. and Tokuyama T..
CASCADE OF MULTI-LEVEL MULTI-INSTANCE CLASSIFIERS FOR IMAGE ANNOTATION.
DOI: 10.5220/0003634400140023
In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval* (KDIR-2011), pages 14-23
ISBN: 978-989-8425-79-9

tain", while finer levels bring useful information to specific objects such as "tiger, cars, bear". Given an object, the cascade method ensures that we first detect the object's related scene, then focus on the "likely" scene to further recognize the object in that context. Formally, cascading means that learning classifiers at finer levels (e.g. level 3) is dependent of classifiers at coarser levels (e.g. level 1,2) (learning from coarse-to-fine). By so doing, when learning classifiers for specific objects at finer levels, we can ignore (negative) samples of non-related scenes, thus reduce training time. Since negative examples are those of the same scene without the considered object, there is more chance for us to separate the object from the background. For instance, since a "tiger" usually appears in a forest, the negative examples of forest background, which does not contain "tiger", helps recognize "negative" regions (forest regions) in the positive examples of "tiger". As a result, it improves the selection of regions corresponding to "tiger", and reduces the ambiguity of "weakly labeling".

Specifically, our propose contains two main parts: 1) multi-level feature extraction; and 2) cascade of multi-instance classifiers over multiple levels. Multi-level means we divide images into different levels of granularity from the coarsest one (the whole image) to increasingly fine subregions (Figure 1). Several feature extraction algorithms are performed at each level, each algorithm produces a set of feature vectors corresponding to subregions of the image. Given a label, a cascade of multi-level multi-instance classifiers is then built across levels, from cheapest (coarsest) features to the most expensive (finest) features. Here, features extracted from the whole image (level 1) are called global features.

In the literature, cascade of classifiers were successfully used to design fast object detectors (Viola and Jones, 2001) while multi-level features were applied to image classification (Lazebnix et al., 2009) and object recognition (Torralba et al., 2010). To the best of our knowledge, however, this is one of the first attempts that adopts the hierarchy of multi-level feature extraction to group features according to acquisition cost so as to develop a cascade learning algorithm for image annotation. In comparison with previous cascading algorithms, we take into account the "weakly labeling" problem by using MIL and make the cascading algorithm suitable to image annotation. In addition, our approach is more robust than previous MIL methods because we consider multi-level feature extraction which allows us to cope with the variety in visual representation among labels. The advantages, thus, lie in threefold: 1) reducing training time by a cascade learning algorithm; 2) relaxing the ambiguity of "weakly labeling" problem of image annotation; and 3) obtaining strong classifiers, which are robust to multiple resolution.

The rest of this paper is organized in 6 sections. Section 2 summarizes typical approaches to image annotation and related tasks. Multi-level feature extraction and multi-instance learning are presented in Section 3 and Section 4. Our proposed method for image annotation is given in details in Section 5. Experimental results will be given in Section 6. Finally, Section 7 concludes the important remarks of our work.

2 RELATED WORKS

Image annotation and related tasks (object recognition, image classification and retrieval) have been the active topics for more than a decade and led to several noticeable methods. In the following, we present an overview of typical approaches, which are categorized into 1) classification-based methods; and 2) joint-distribution based methods.

2.1 Classification-based Approach

The early effort in the area is to formalize image annotation as the task of binary classification. Some examples are to classify images into "indoor" or "outdoor" (Szummer and Picard, 1998). In object recognition, Viola and Jones (Viola and Jones, 2001) proposed a method for face detection (face/non face classification) using Adaboost, which is very fast in dropping non face windows in images, thus results in fast face detectors.

For image retrieval, the two-class formalization is not enough to meet searching requirements. Lyndon et al. (Kennedy and Chang, 2007) used a reranking method to combine binary classifiers. Akbas et al. (Akbas and Vural, 2007) fused binary classifiers by learning a new meta classifier from category-membered vectors, which are generated from the binary classifiers. Nguyen et al. (Nguyen et al., 2010) proposed a feature-word-topic model in which one individual classifier is learned for each label based on visual features. By modeling topics of words, the authors then refine the results from binary classifiers to obtain topic-oriented annotation for later image retrieval.

In order to apply classification approach to image annotation, we need to take the "weakly labeling" problem into account. Typically, this can be done by adopting multi-instance learning (MIL) instead of single-instance learning. Ansdrew et al. (Andrews et al., 2002) adapted single-instance learning version

of Support Vector Machine (SVM) to multi-instance learning versions namely MI-SVM and mi-SVM and applied to image annotation with 3 classes (tiger, fox, elephant). On the other attempt, Yang et al. (Yang et al., 2006) introduced Asymmetric SVM (ASVM) to pose different loss functions to 2 types of error (false positive and false negative) for annotation. ASVM has been applied to 70 common labels of Corel5K, which is the common benchmark for image annotation, and shown comparative results. Also following the idea of MIL but supervised multiclass labeling (SML) [5] does not consider negative examples in learning binary classifiers. Given a label, SML is based on hierarchical Gaussian mixture to train a binary classifier using only positive examples. Since only global features are used in SML, it is not clear whether SML works well for specific objects or not although on average it showed state-of-the-art performance on Corel5K. All in all, current MIL-based image annotation systems do not exploit the benefit of combining global and region-based features.

2.2 Joint Distribution-based Approach

Statistical generative models introduce a set of latent variables to define a joint distribution between visual features and labels for image annotation. Jeon et al. (Jeon et al., 2003) proposed Cross-Media Relevance Model (CMRM) for image annotation. This work relies on normalized cuts to segment images into regions then clusters visual descriptors of segments to build blobs. CMRM uses training images as latent variables to estimate the joint distribution between blobs and words. Continuous Relevance Model (CRM) (Lavrenko et al., 2003) is another relevance model but different from CMRM by the fact that it models directly the joint distribution between words and continuous visual features using non-parametric kernel density estimate. As a result, it is not as sensitive to quantization errors as CMRM. These methods (CMRM, CRM) are also referred as keyword propagation methods since they transfer keywords of the nearest neighbors (in the training dataset) to the given new image. The drawback of those methods is that the annotation time depends linearly on the number of training set, thus have the scalable limitation (Carneiro et al., 2007).

Following this approach, topic model-based methods (Blei and Jordan, 2003; Monay and Gatica-Perez, 2007) do not use training images but hidden topics (concepts/aspects) as latent variables. These methods also rely on either quantized features (Monay and Gatica-Perez, 2007) or continuous variables (Blei and Jordan, 2003). The main advantages of the topic model-based approach lies in two points: 1) the better scalability in comparison with propagation methods; and 2) the ability to encode scene settings (via topics) into image annotation.

Despite of topic-based methods or propagation methods, the disadvantage of joint distribution-based approach is its lack of direct modeling between visual features and labels, which makes it difficult to optimize annotation (Carneiro et al., 2007). In order to study the impact of feature extraction on different types of labels,it is more appropriate to follow the multiple instance learning methods as mentioned in the section of classification-based approach.

3 MULTI-LEVEL FEATURE EXTRACTION

As stated previously, our method consists of 2 main parts: 1) multi-level feature extraction; and 2) cascade of multi-instance classifiers over levels. This section reviews noticeable methods to extract visual descriptors for image annotation, classification and retrieval as a fundamental for our multi-level feature extraction described later. We distinguish 3 types of visual descriptors, which are global features, region-based features, and hybrid.

Global Feature Extraction: an image is not divided into subregions. As a result, we obtain only one feature vector (one instance) for each image. Many low-level features can be extracted and concatenated from the whole image such as color histogram, texture, or edge histogram (Deselaers et al., 2008; Makadia et al., 2010; Douze et al., 2009; Jégou et al., 2010; Akbas and Vural, 2007). Bag-of-feature (Hofmann, 1999; Deselaers et al., 2008) obtained by quantizing features at interest points can also be classified to this category because one image is not divided into smaller regions, and an image has only one histogram feature vector. Recent baseline in image annotation (Makadia et al., 2010) also relied on global feature extraction. However, they did not concatenate feature vectors but combined similarities from different feature types to measure similarity between images for K-nearest-neighbor based image annotation.

Local Feature Extraction: an image is divided into smaller regions using image segmentation (Barnard et al., 2003; Duygulu et al., 2002; Jeon et al., 2003) or grid-based division. A feature vector is then extracted from each subregion (Feng et al., 2004). As a result, an image has several feature vectors, one corresponds to one subregion. Since image segmentation is still a difficult task, many of current works avoid this

task and divide images into grids instead (Feng et al., 2004; Jeon et al., 2004). Previous study (Feng et al., 2004) have shown that grid-based division can obtain better results than segmentation on Corel5K benchmark.

Hybrid Method: Spatial pyramid method (Lazebnix et al., 2009) can be considered as a hybrid of local and global representations. Informally, an image is divided to increasingly coarser grids. We are then able to concatenate weighted histograms of all cells (of the grids) into one vector. This method has been applied to scene classification and image classification with little ambiguity, which does not have "weakly labeling" as in image annotation. Even our approach also divides images into different coarse grids (coarse levels) and extract features from levels, the difference is that we do not concatenate the feature vectors from different levels but exploit the hierarchy to group feature sets according to acquisition cost. As a result, we are able to develop a cascade algorithm for image annotation.

4 MULTI-INSTANCE LEARNING WITH SUPPORT VECTOR MACHINES

Multi-instance learning is essential in our propose. This section begins with standard supervised learning with Support Vector Machine (SVM), which is single instance learning, then presents one extension to turn SVM into multi-instance SVM.

In standard supervised learning, it is often the case that we are given a training set of labeled instances (samples) $D = \{(\mathbf{x}_i, y_i) | i = 1, \ldots, N; \mathbf{x}_i \in R^d; y_i \in \mathcal{Y} = \{+1, -1\}\}$ and the objective is to learn a classifier, i.e., a function from instances to labels: $h : R^d \rightarrow \mathcal{Y}$. This class of supervised learning belongs to single-instance learning, where Support Vector Machine (SVM) (Schölkopf et al., 1999) is one of the most successful methods.

Multiple Instance Learning (MIL) generalizes the single instance learning to cope with the ambiguity in training dataset. Instead of receiving a set of labeled instances, we are given a set of negative/ positive bags, each contains many instances. A negative bag contains all negative instances, while a positive bag has at least one positive instance but we do not know which one it is. The formalization of MIL naturally fits the "weakly labeling" in image annotation where a positive bag (w.r.t a label) corresponds to an image annotated with that label. There were several methods for MIL. For simplicity, we will discuss one simple formalization to apply SVM for MIL namely MI-SVM (Andrews et al., 2002).

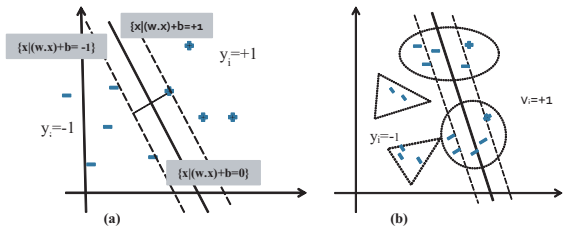

Figure 2: Support Vector Machines: (a) Single Instance Learning; (b) Multiple Instance Learning: positive and negative bags are denoted by circles and triangles respectively.

4.1 Support Vector Machines

In Support Vector Machines (Schölkopf et al., 1999), a class of hyperplanes that separate negative and positive patterns (Figure 2) is considered. For separable case, the hyperplane represented by a pair $(\mathbf{a}, b)$ ($\mathbf{a} \in R^N$ and $b \in R$) satisfies:

$$\begin{cases} \mathbf{ax} + b \geq +1 & \text{if } y_i = +1 \\ \mathbf{ax} + b \leq -1 & \text{if } y_i = -1 \end{cases}$$

The corresponding decision function becomes $f(\mathbf{x}) = sgn(\mathbf{ax} + b)$. Among the hyperplanes that are able to separate positive and negative patterns, the optimal hyperplane is the one with maximum margin and most likely to have minimum test error (Schölkopf et al., 1999). It has been proved that the margin of a hyperplane is reversely proportional to $||\mathbf{a}||$. In practice, a separating hyperplane may not exist, i.e. data is non-separable, slack (positive) variables ξ are introduced to allow misclassified examples. The optimization turns into:

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

$$\text{subject to: } y_i(\mathbf{ax} + b) \geq 1 - \xi_i, i = 1, \ldots, N$$

where C is the constant determining the trade-off. SVMs also can carry out the nonlinear classification by using kernel functions that embed data into higher space where the nonlinear pattern now appears linear.

4.2 Multiple Instance Support Vector Machines

Let $D_w = \{(X_i, Y_i) | i = 1, \ldots, N, X_i = \{\mathbf{x}_j\}; Y_i = \{+1, -1\}\}$ be a set of images (bags) with/without word w, 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_j positive (the subregion in the image corresponds to word w). As shown in Figure 2b, 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 $Y_i = \max(y_j), j = 1,\ldots,|X_i|$.

MI-SVM (Andrews et al., 2002) extends the notion of the margin from an individual instance to a set of instances (Figure 2b). 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 (MI-SVM) 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.

5 CASCADE OF MULTI-LEVEL MULTI-INSTANCE CLASSIFIERS

5.1 Notation and Learning Algorithm

Let $\mathcal{D} = \{(I_1, \mathbf{w_1}), \ldots, (I_N, \mathbf{w_N})\}$ be a training dataset, in which $\mathbf{w_n}$ is a set of words associated with image I_n and sampled from a vocabulary $V = \{w_1, w_2, \ldots, w_{|V|}\}$. The objective is to learn a mapping function from visual space to word space so that we can index and rank new images for text-based retrieval. The two main components of our propose are described as follows:

- **Extracting Multi-level Features:** we divide each image in T different levels then perform M feature extraction algorithms $\mathcal{F}_m$ as in Figure 3. Here, we can choose any suitable feature extraction such as color, texture, shape description, gist, etc. for $\mathcal{F}_m$. Let $\mathcal{M}(l)$ $(l = 1,\ldots,T)$ be indexes of feature extractions at level l, e.g. $\mathcal{M}(1) = 1,2; \mathcal{M}(2) = 3,4,5$ (Figure 3). From this notation, we have $\sum_{l=1}^{T} |\mathcal{M}(l)| = M$. Also, we can infer that all the feature extraction algorithms at previous levels of level l are indexed from 1 to $\min\{\mathcal{M}(l)\} - 1$.

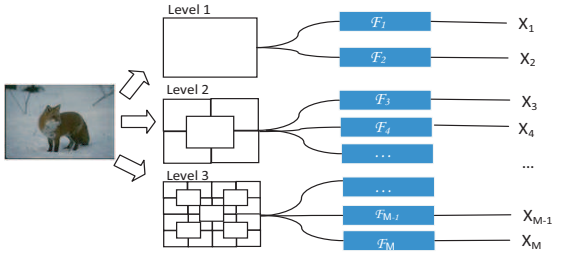

Figure 3: An image is divided into different levels of granularity. For a level, we perform one or more feature extraction methods. We then obtain M feature extraction methods.

- **Cascade of Multi-instance Classifiers Over Levels:** given a label w, $D_w = \{B^+, B^-\}$ denotes a training dataset where B^+ (B^-) is the set of images with (without) w. Let Y be a vector of corresponding classes of images in D_w, i.e. $Y_n = 1$ if $I_n \in B^+$ and $Y_n = -1$ otherwise. Let *score* be the output (confidence) vector generated by machines (classifiers), where $score_n > 0$ (or absolute value of $score_n$ (< 0)) is the confidence of assigning (not assigning) w to $I_n \in D_w$. We denote h_m the weak classifier, which maps from feature space X_m of feature extraction algorithm $\mathcal{F}_m$ to $\{-1, 1\}$. The confidence score posed by h_m on the image I is denoted by $h_m(\mathcal{F}_m(I))$, that is we apply h_m on feature vectors obtained by $\mathcal{F}_m$ on I. Based on these notations, CMLMI is presented in Algorithm 1. Note that multi-instance learning turns into single-instance learning at the coarsest level when global feature vector is in use.

For global feature extractions at level $l = 1$, an image has one instance (one feature vector), the problem turns into normal supervised learning. We applied SVM for this case. At finer level ($l > 1$), one image has a set of instances, one corresponds to one subregion. Due to weakly labeling, we do not know which instance best represents the given label. The multiple-instance version of SVM (MI-SVM) (see Section 4) is used to address this ambiguity.

We update scores of images in D_w at level l using the following recursion:

$$score = H_l = \gamma_l * H_{l-1} + \sum_{m \in \mathcal{M}(l)} \alpha_m * h_m + c_l$$

Algorithm 1: A Cascade of Multi-Level Multi-Instance Classifiers.

Input : A set $D_w = \{B^+, B^-\}$ of positive and negative examples for word w.
Output: A strong classifier H_w for w

1 Initialize $score_n = 0$, $\theta_i = 1/|B^-|$, $c = 0$, and $\alpha_m = 0$ for $n = 1,\ldots,|B|$, $i = 1,\ldots,|B^-|$;, and $m = 1,\ldots,M$.

2 **//Learning weak classifiers over T levels**
3 **for** $l \leftarrow 1$ **to** T **do**
4 **if** $l == 1$ **then**
5 Learn classifiers h_m using SVM from D_w for all $m \in \mathcal{M}(l)$
6 **if** $l > 1$ **then**
7 Sample a smaller set SB^- from B^- according to θ
8 Learn classifiers h_m using MI-SVM from $SD_w = \{B^+, SB^-\}$ for all $m \in \mathcal{M}(l)$
9 **end**
10 **//Update *score* for all images in D_w**
11 Set $score_n = \gamma_l * score_n + \sum_{m \in \mathcal{M}(l)} \alpha_m * h_m(\mathcal{F}_m(I_n)) + c_l$ for $n = 1,\ldots,|D_w|$
12 Find coefficients $\gamma_l > 0$, α_m and c_l to minimize $||score - Y||_2$
13 **//Update coefficients of classifiers in previous levels**
14 **for** $m' = 1$ **to** $\min\{\mathcal{M}(l)\} - 1$ **do**
15 $\alpha_{m'} = \alpha_{m'} * \gamma_l$
16 **end**
17 Update the overall threshold $c = c * \gamma_l + c_l$
18 Sort *score* in descending order, and let r_j be the ranking position of $I_j \in B^-$ in sorted *score*
19 Update $\theta_j \leftarrow \theta_j * 1/r_j$ for all $j = 1,\ldots,|B^-|$ and normalize θ so that $\sum_j \theta = 1$
20 **end**
21 Final robust classifier:

$$H_w = \frac{\sum_{m=1}^{M} \alpha_m * h_m + c}{\sum_{m=1}^{M} \alpha_m + c}$$

Since we have the constraint that $\gamma_l > 0$, the ranking of images is based on previous ranking (H_{l-1}) but modified by the additional classifiers of current level (the second term). The constant term c_l is used as the constant threshold for level l. We then find coefficients for classifiers of level l using linear regression that is minimizing square error $||H - Y||_2$ (lines from 10 to 11 in Algorithm 1). Here, scores for images in D_w are accumulated from level 1 to level $l-1$ and stored in *score*.

Unlike previous boosting methods, the sampling distribution θ on B^- is updated based on the ranking positions of negative samples on the sorted *score* instead of the *score* itself (line 18,19). As a result, a negative example at higher rank will be weighted more than negative examples at lower ranks. From the experiments, we see that this ranking-based scheme is better than score-based for unbalanced training set.

5.2 Detailed Analysis

This section presents theoretical analysis for our algorithm, which focuses on the benefit of CMLMI in training time and shows that our algorithm is suitable to image annotation.

Based on cascading scheme, it is obvious that our method requires less training time than learning all individual classifiers independently. The training time of MI-SVM depends on $|B^+| + NR * |B^-|$, where NR is the number of subregions per image. That NR is larger on finer levels makes the domination of negative instances over positive ones even more serious. Training MI-SVM in cascade with SB_w (Line 7 in Algorithm 1) is more efficient than training an independent one with D_w.

Not only having advantage in training time, but also our method is suitable to image annotation and able to reduce the ambiguity of weakly labeling. When the coarse levels are in charge of detecting related context of the given level, the finer levels are able to focus on sample images of similar scene to separate the object from the background, and reduce ambiguity caused by weakly labeling. Figure 4 demonstrates our idea. Here, circles still denote positive bags, in which we know positive instances are available but do not know which ones, and triangles denote negative bags, of which we have guarantee that all instances are negative. The negative bag selected here is the one with instances close to some other instances of one positive bag (the red circle). The com-

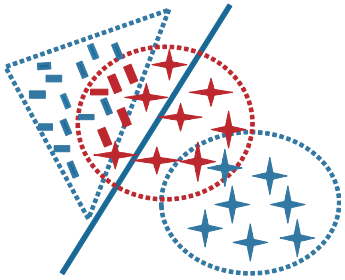

Figure 4: Negative bags that share common negative instances with positive bags reduce ambiguity. Here the stars denote unknown classes (either positive (+) or negative (-).

mon/similar instances correspond to subregions of the shared/similar background of the two bags. Since we have the knowledge that all instances of the negative bag are negative, we can conclude that the instances of the red circle, which are close to or even included in the negative bag, are negative. Along with the similarity among positive bags, which contain the same object, this information helps to obtain better hyperplane to separate negative and positive instances. To our best knowledge, this is one of the first attempts that makes use of the similarity between negative bags and positive bags to reduce ambiguity in MIL. Most of previous approaches in MIL only made use of similarity among positive bags to deal with the ambiguity. For example, (Carneiro et al., 2007) only uses positive bags to generalize a dominating distribution over positive bags. (Maron and Lozano-Pérez, 1998) finds regions in the instance space with instances from many different positive bags and far away from instances from negative bags. In (Yang et al., 2006; Andrews et al., 2002), negative bags are sampled randomly only to cope with the domination of negative examples over positive examples without giving notice to negative bags that share backgrounds with positive bags. Recently, (Deselaers and Ferrari, 2010) also follows the idea that the significant portion of positive instances will result in a reasonable classifier performing better than by change. However, we observe that some negative instances also amount to significant portion, which are the instances corresponding to common backgrounds. This problem becomes more serious when more and more labels are taken into consideration like those in image annotation.

6 EXPERIMENTS

6.1 Corel5K Dataset

The Corel5k benchmark is obtained from Corel image database and commonly used for image annotation (Duygulu et al., 2002). It contains 5,000 images and were pre-divided into a training set of 4,000 images, a validation set of 500 images, and a test set of 500 images. Each image is labeled with from 1 to 5 captions from a vocabulary of 374 distinct words.

6.2 Evaluation

Given a testing dataset, we can measure the effectiveness of the algorithm. Regarding a label w, the typical measures for retrieval are precision P_w, recall R_w:

$$P_w = \frac{\text{Number of images correctly annotated with } w}{\text{Number of images annotated with } w}$$

$$R_w = \frac{\text{Number of images correctly annotated with } w}{\text{Number of images manually annotated with } w}$$

We calculate P and R, which are means of P_w and R_w over all labels. To balance the trace-off between P and R, $F_1 = 2 * P * R/(P+R)$ is usually used as another measure for evaluation. In order to measure retrieval performance, we also calculate the average precision (AP) for one label w as follows:

$$AP_w = \frac{\sum_{r=1}^{N} P(r) \times rel(r)}{\text{Number of images annotated manually with } w}$$

where r is a rank, N is the number of retrieved images, $rel(r)$ is a binary function to check the word at r is in the manual list of words or not, and $P(r)$ is the precision at r. Note that, the denominator of AP is independent with N. Finally, mAP is obtained by averaging *APs* over all labels of the testing dataset.

Table 1: Feature extractions & classifiers.

Level 1	$\mathcal{F}_1$: "gist" of scene	SVM-GIST
-	$\mathcal{F}_2$: color histogram	SVM-color
Level 2	$\mathcal{F}_3$: color histogram	MISVM-color
-	$\mathcal{F}_4$: Gabor texture	MISVM-texture

6.3 Experimental Settings

For the experiments, we performed a cascade of 4 classifiers with 2 levels. Here, we worked with only 2 levels because the images of Corel5K are all in small size. Moreover, we would like to focus on the basic case to analyze the impact of global features on reducing the weakly labeling problem. At the first level, global features were extracted from the whole image. We exploited Gist (Oliva and Torralba, 2001), and color histogram in RGB color space with 16 channels. For each region in the second level, we also performed color histogram extraction but with 8 channels

Table 2: CMLMI vs. various MIL methods.

(a) In comparison with other standalone MIL methods. Results of ASVM-MIL and mi-SVM are reported in (Yang et al., 2006)

Method	P	R	F1	mAP
ASVM-MIL	**0.31**	0.39	0.35	-
mi-SVM	0.28	0.35	0.31	-
MISVM-Color	0.13	**0.55**	0.21	0.19
MISVM-Texture	0.07	0.36	0.13	0.86
CMLMI	0.30	0.52	**0.38**	**0.35**

(b) In comparison with standalone SVM with global features

Method	P	R	F1	mAP
SVM-Color	0.20	0.39	0.27	0.19
SVM-Gist	0.27	0.47	0.34	0.28
CMLMI	0.30	0.52	0.38	0.35

and texture extraction using Gabor filter as in (Makadia et al., 2010). Summary of feature extraction methods and their relationship with levels are is given in Table 1. The numbers of dimension in corresponding feature spaces of algorithms $\mathcal{F}_1, \mathcal{F}_2, \mathcal{F}_3$, and $\mathcal{F}_4$ are 960; 4096; 192; and 512 respectively.

We name classifiers trained on feature spaces of $\mathcal{F}_1, \mathcal{F}_2, \mathcal{F}_3$, and $\mathcal{F}_4$ as SVM-Gist, SVM-color, MISVM-color, and MISVM-texture. Conventionally, CMLMI is used to indicate the strong classifier H_w learned according to Algorithm 1, in which classifiers of level 2 (MISVM-color, and MISVM-texture) are dependent of classifiers of level 1 (SVM-Gist, and SVM-color). In the following, we refer to, for example, MISVM-color (or standalone MISVM-color) to indicate an independent classifier trained on D_w, and MISVM-color of CMLMI to imply the MISVM-color learned in the cascade according to Algorithm 1. In the other words, MISVM-color of CMLMI is the classifier trained on SD_w sampled based on the results of level 1 (SVM-Gist and SVM-color of CMLMI).

6.4 Experimental Results on 70 Most Common Labels

Like (Yang et al., 2006), we selected 70 most common labels from Corel5K dataset for experiments. The reason is that labels with a small number of the positive samples (for example: 5 10 positive samples) are not efficient to train a classifier.

Table 2(a) shows that CMLMI outperforms other MIL methods. As observable from the table, we obtain improvements of 17.35% in F_1 measure and 16.14% in mAP compared to MISVM-color. In contrast to MISVM-texture, CMLMI significantly increases F_1 measure by 25.64% and mAP by 26.71%. Comparing to previous works, CMLMI obtains better results than mi-SVM both in precision and recall, which leads to a raise of 7.54% in F_1 measure. Also,

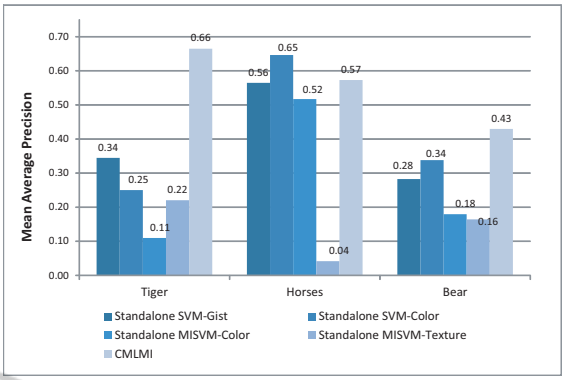

Figure 5: mAP of CMLMI in comparision with different standalone methods.

CMLMI outperforms ASVM-MIL in recall while obtaining comparable precision (P of 0.30 with CMLMI, and P of 0.31 with ASVM-MIL). This results in an improvement 3.54% of our method over ASVM-MIL in F_1 measure.

Table 2(b) compares CMLMI to SVM with global features. We can see that CMLMI also obtain better results in F_1 and mAP (F_1 of 0.38 and mAP of 0.35) compared with SVM-color (F_1 of 0.21 and mAP of 0.20), and SVM-Gist (F_1 of 0.34, mAP of 0.27). Among the standalone classifiers (SVM-color, SVM-gist, MISVM-color, and MISVM-texture), SVM with global features outperform MISVM with region-based feature extractions. Interestingly, SVM-Gist is even comparable to ASVM-MIL although image segmentation, which is more expensive than global feature, has been used in ASVM-MIL. However, combining the classifiers in our cascading algorithm yields the best results.

6.5 Experimental Results on Sample Foreground Labels

We conducted carefully analysis for “tiger”, “horse” and “bear” in Corel5K since the concepts correspond to foreground objects which might benefit from finer levels. Figure 5 shows mAP of standalone classifiers and CMLMI for three labels. It can be seen that individual feature types have different influences on different labels. Except for Gist ($\mathcal{F}_1$) that shows its importance for all three labels, global color histogram ($\mathcal{F}_2$) has more impact on annotating images with “horses” and “bear” than with “tiger”. Texture feature at level 2 (of MISVM-texture) performs better than the other feature extraction methods only with “tiger”. CMLMI significantly outperforms other standalone classifiers on “tiger” and “bear” while falls a little on “horses” compared with SVM-color. Interestingly, standalone MISVM-color is comparable to

(a) Subregions selected by standalone MISVM-color

(b) Subregions selected by MISVM-color of CMLMI

Figure 6: The subregions selected by standalone MISVM-color for label “tiger”, and the subregions selected by MISVM-color of CMLMI from the corresponding images. Here, the numbers under each subregion indicate image IDs.

(a) Subregions selected by standalone MISVM-texture

(b) Subregions selected by MISVM-texture of CMLMI

Figure 7: The subregions selected by standalone MISVM-texture for label “horses” and the subregions (of corresponding images) selected by MISVM-texture at the 2-nd level of CMLMI.

CMLMI for “horses”. In order to uncover the question in the “horse” case, we conducted detailed analysis, and found that MISVM-color and SVM-color captured grass fields in the background instead of horses. Indeed, no subregion with the color of a horse was considered in MISVM-color. Thus, the good performance of standalone MISVM-color and SVM-color owes to special feature of the Corel5K dataset in which horses are on grass fields in most of pictures.

As previously mentioned, the negative examples for finer levels are drawn based on the ambiguity of coarser levels, which are able to detect the background better. By considering the negative examples of similar background, we are able to add “negative instances”, which usually appear with the real positive instances of positive examples. As a result, there is more chance for us to separate the ”positive instance” from “negative instance” in positive examples. Figure 6 and Figure 7 show the examples of selecting positive instances from corresponding positive bags with standalone MISVM and MISVM of CMLMI. We can see from the figures that MISVM of CMLMI is able to select more relevant subregions. For the case of “tiger”, MISVM-color of CMLMI is given more information about background (grass, forest, stone, water), it has successfully avoided selecting background-related instances as positive ones.

7 CONCLUDING REMARKS

In this paper, we have presented an overview of image annotation: its typical problems, feature extraction methods and typical methodologies. By analyzing the main problems of image annotation, we proposed a method based on cascading multi-level multi-instance classifiers, which has main advantages as follows:

- Our cascade of MLMI classifiers is able to reduce training time since we can remove some negative examples, which are “easily” detected as negative based on the scene, in finer levels.
- Multi-level feature extractions allow us to annotate images with multiple resolutions. One example is that a photo of tiger might be a close-up photo or the photo of a tiger in its context. Multi-level feature extractions bring more chance to capture all of this variety.

- We also show experimentally that it is able to reduce the ambiguity of "weakly labeling" in image annotation, and separate the foreground objects from the scene in finer levels of the cascade.

The experiments show promising results of the proposed method in comparison with several baselines on Corel5K. Experiments suggest that as long as the finer levels can bring "new information", they help to obtain better detection of foreground objects. For the future work, we would like to focus more on the role of context in reducing the ambiguity of "weakly labeling".

REFERENCES

Akbas, E. and Vural, F. T. Y. (2007). Automatic image annotation by ensemble of visual descriptors. In *IEEE Conf. on CVPR*, pages 1–8, Los Alamitos, CA, USA.

Andrews, S., Hofmann, T., and Tsochantaridis, I. (2002). Multiple instance learning with generalized support vector machines. In *18th AAAI National Conference on Artificial intelligence*, pages 943–944, Menlo Park, CA, USA.

Barnard, K., Duygulu, P., Forsyth, D., Freitas, N. D., Blei, D. M., K, J., Hofmann, T., Poggio, T., and Shawe-taylor, J. (2003). Matching words and pictures. *Journal of Machine Learning Research*, 3:1107–1135.

Blei, D. M. and Jordan, M. I. (2003). Modeling annotated data. In *Proc. of the 26th ACM SIGIR*, pages 127–134.

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

Deselaers, T. and Ferrari, V. (2010). A conditional random field for multiple-instance learning. In *Proc. of The 27th ICML*, pages 287–294.

Deselaers, T., Keysers, D., and Ney, H. (2008). Features for image retrieval: an experimental comparison. *Inf. Retr.*, 11:77–107.

Douze, M., Jégou, H., Sandhawalia, H., Amsaleg, L., and Schmid, C. (2009). Evaluation of gist descriptors for web-scale image search. In *Proc. of the ACM CIVR*, pages 1–8, New York, NY, USA.

Duygulu, P., Barnard, K., de Freitas, J. F. G., and Forsyth, D. A. (2002). Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Proc. of the 7th ECCV*, pages 97–112, London, UK. Springer-Verlag.

Feng, S. L., Manmatha, R., and Lavrenko, V. (2004). Multiple bernoulli relevance models for image and video annotation. In *Proc. of the 2004 CVPR*.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proc. of the 22nd ACM SIGIR*, pages 50–57, New York, NY, USA.

Jégou, H., Douze, M., and Schmid, C. (2010). Improving bag-of-features for large scale image search. *Int. J. Comput. Vision*, 87(3):316–336.

Jeon, J., Lavrenko, V., and Manmatha, R. (2003). Automatic image annotation and retrieval using cross-media relevance models. In *Proc. of the 26th int. ACM SIGIR*, pages 119–126.

Jeon, J., Lavrenko, V., and Manmatha, R. (2004). Automatic image annotation of news images with large vocabularies and low quality training data. In *Proc. of ACM Multimedia*.

Kennedy, L. S. and Chang, S.-F. (2007). A reranking approach for context-based concept fusion in video indexing and retrieval. In *Proc. of the 6th ACM int. on CIVR*, pages 333–340, New York, NY, USA. ACM.

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

Lazebnix, S., Schmid, C., and Ponce, J. (2009). *Object Categorization: Computer & Human Vision Perspectives*, chapter Spatial Pyramid Matching. Cambridge University Press.

Makadia, A., Pavlovic, V., and Kumar, S. (2010). Baselines for image annotation. *Int. J. Comput. Vision*, 90(1):88–105.

Maron, O. and Lozano-Pérez, T. (1998). A framework for multiple-instance learning. In *Proc. of the Conf. on Advances in Neural Information Processing Systems*, NIPS '97, pages 570–576, Cambridge, MA, USA. MIT Press.

Monay, F. and Gatica-Perez, D. (2007). Modeling semantic aspects for cross-media image indexing. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(10):1802–1817.

Nguyen, C.-T., Kaothanthong, N., Phan, X.-H., and Tokuyama, T. (2010). A feature-word-topic model for image annotation. In *Proc. of the 19th ACM CIKM*, pages 1481–1484.

Oliva, A. and Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int. J. of Comput. Vision*, 42:145–175.

Schölkopf, B., Burges, C. J. C., and Smola, A. J., editors (1999). *Advances in kernel methods: support vector learning*. MIT Press, Cambridge, MA, USA.

Szummer, M. and Picard, R. W. (1998). Indoor-outdoor image classification. In *Proc. of the 1998 Int. Workshop on Content-Based Access of Image and Video Databases*, page 42, Washington, DC, USA.

Torralba, A., Murphy, K. P., and Freeman, W. T. (2010). Using the forest to see the trees: exploiting context for visual object detection and localization. *Commun. ACM*, 53(3):107–114.

Viola, P. and Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *Proc. of IEEE CVPR*, volume 1, pages I–511 – I–518 vol.1.

Yang, C., Dong, M., and Hua, J. (2006). Region-based image annotation using asymmetrical support vector machine-based multiple-instance learning. In *Proc. of the 2006 IEEE CVPR*, pages 2057–2063, Washington, DC, USA.