Unsupervised Object Class Discovery via Saliency-Guided Multiple Class Learning

Jun-Yan Zhu[1,2], Jiajun Wu[3,1], Yichen Wei[1], Eric Chang[1] and Zhuowen Tu[1,4]
[1]Microsoft Research Asia
[2]Dept. of Computer Science and Technology, Tsinghua University
[3]Institute for Interdisciplinary Information Sciences, Tsinghua University
[4]Lab of Neuro Imaging and Dept. of Computer Science, UCLA
{junyanzhu89, jiajunwu.cs}@gmail.com, {yichenw, echang, zhuowent}@microsoft.com

Abstract

Discovering object classes from images in a fully unsupervised way is an intrinsically ambiguous task; saliency detection approaches however ease the burden on unsupervised learning. We develop an algorithm for simultaneously localizing objects and discovering object classes via bottom-up (saliency-guided) multiple class learning (bMCL), and make the following contributions: (1) saliency detection is adopted to convert unsupervised learning into multiple instance learning, formulated as bottom-up multiple class learning (bMCL); (2) we utilize the Discriminative EM (DiscEM) to solve our bMCL problem and show DiscEM's connection to the MIL-Boost method[34]; (3) localizing objects, discovering object classes, and training object detectors are performed simultaneously in an integrated framework; (4) significant improvements over the existing methods for multi-class object discovery are observed. In addition, we show single class localization as a special case in our bMCL framework and we also demonstrate the advantage of bMCL over purely data-driven saliency methods.

1. Introduction

The computer vision field has witnessed milestone achievements in learning object models with full supervision [33, 11, 30]. However, a large amount of labeled data is required for these methods to train practically working systems. Recently, many unsupervised approaches have also been proposed for the object localization and categorization tasks [26, 15, 22, 18, 20, 38, 19, 32, 21]. While important progresses have been made with the encouraging performance reported on datasets such as Caltech-101 [10], ETHZ [13], and MSRC2 [31], most of the existing approaches work under certain conditions (many have strict constraints about the problem settings). Such conditions include, *e.g.*, large occupation of the foreground objects; no "irrelevant" other object types; and clean background. However, in practice objects are often small and not centered in the image. The background could also be cluttered and present non-uniformly, as suggested in the unsupervised scene discovery research [23]. How to apply previous approaches in a general unsupervised setting is not very clear.

A closely related direction to unsupervised learning is multiple instance learning (MIL) [34, 1, 8] where only image (bag) level labels are given without the detailed annotation of where the objects are. MIL significantly reduces the efforts in manual labeling for object detection. Furthermore, when multiple object classes are present, it is desirable to automatically discover them simultaneously in the current MIL scheme.

In the machine learning literature, several multiple instance clustering (MIC) algorithms [37, 36] are designed to perform localized content-based image clustering, which has relatively less constraints. These methods introduce multiple instance concept into standard clustering methods such as K-means and MMC [9, 35]. However, most existing MIC solutions reported discouraging purity result (37.1%) [37, 36] in a benchmark dataset SIVAL [25] and they do not perform simultaneous localization. Therefore, despite the novel multi-instance concept to unsupervised object discovery problem, the low performance questions the practical aspects of the current MIC methods. In comparison, state-of-the-art unsupervised object discovery methods [18, 20] could achieve about 98% purity result in Caltech-101; however, when applied to SIVAL, they only obtained 28.3%.

In this paper, we argue that unsupervised object discovery in a general setting might be an ill-posed problem. This is due to the intrinsic ambiguity of the complex object appearances and the background clutter. However, it is still desirable to build such an unsupervised object class discovery with relatively loose constraints. A recent method

[7] used a classifier trained on several classes of objects as "meta information" to learn other object types. From a different angle, saliency detection has been an active research area [16, 4, 12] where objects of interest are assumed to be "salient" in an image. It appears (arguably) that being salient is a more general concept than classifiers trained on a specific number of object classes.

Here, we propose a system adopting saliency detection in a multiple instance learning framework, which has the following new aspects: (1) unlike the direct top-down discovery of object classes in [37, 36] or using specifically trained classifiers on a number of supervised object classes [7], we utilize saliency detection (bottom-up method) to guide the unsupervised object discovery; (2) object localization, object class discovery, and object detector training are performed simultaneously in an integrated framework, bMCL; (3) we develop a general scheme, Discriminative EM (DiscEM), to perform optimization for bMCL, and we show its connection to MIL-Boost[34]; (4) significant improvements on challenging benchmark datasets over the exiting systems are obtained by integrating saliency detection with bMCL.

Although being an active area, saliency detection has yet to justify its usage for high-level vision tasks and we show that the saliency-guided notion can indeed be of great help in the unsupervised object discovery task.

2. Related Work

Tuytelaars et al. [32] surveyed recent unsupervised object learning method, but with the focus on the probabilistic latent models. Here, we briefly discuss the related work from several different angles.

As stated before, several unsupervised approaches have recently been proposed for the object localization and categorization task [26, 15, 22, 18, 20, 38, 19, 32, 21]. Zhu et al. [38] learned a probabilistic grammar for object classes (mostly weakly-supervised) but with results reported on a subset of the Caltech dataset [10], in which the foreground objects are mostly centered and they occupy a significant portion of each image. Lee and Grauman [20] grouped edge/contour fragments into objects without supervision, but under the requirement of having well-defined (strong) shape cues for objects.

In the machine learning literature, multiple instance learning (MIL) [34] and multiple instance clustering (MIC) [37, 36] are used to learn single and multiple object classes respectively. Since no "negative" images are present and no specific prior information about foreground objects is used, MIC [37, 36] reported poor results on challenging datasets like SIVAL [25].

Another recent active research field is saliency detection. Impressive results have been reported using mostly bottom-up (data-driven) processes [16, 4, 12]. In addition to estimate pixels' saliency [16, 4], window saliency is proposed and computed in [12]. Multiple images are utilized to perform co-saliency in [3] but they are mostly focused on the single-class unsupervised co-segmentation task rather than localizing and learning multiple object models. Despite the substantial interests in computer vision, saliency detection has received relatively less attention to the object discovery community. In [27], the most "salient" regions are selected to update the models based on a fixed matching threshold.

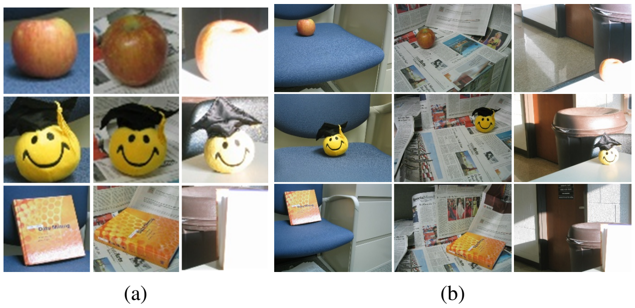

Figure 1: (a) The localization results obtained by our algorithm, bMCL. (b) The original images from SIVAL[25].

We adopt the bottom-up saliency cues into an intergraded learning framework to localize objects, discover object classes, and train discriminative object models at the same time, which differs our method from all the previous approaches. Other work using bottom-up cues focuses on multiple segmentations [26] or emphasizes on self-paced discovery that progressively accumulates models[21].

3. Saliency-Guided Notion

We argue that the problem of unsupervised object discovery is an ambiguous task in general conditions. Thus, we utilize bottom-up saliency detection to guide the learning process and turn unsupervised learning into weakly supervised learning.

Ambiguity of unsupervised object discovery

In an investigation experiment on the SIVAL dataset, we asked 10 participants to divide two groups of images into three categories. While all the participants divided first group (Figure 1a) into three object classes without a second thought, they felt confused and spent much more time on the second group (Figure 1b). Finally, 7 people divided the second group into apples, toys, and books while the other 3 people categorized images as different scenes. The difficulty that human vision encounters here reveals the strong ambiguity in unsupervised object discovery, especially for complex objects and backgrounds. Without a prior knowledge, clustering based algorithms [9, 35, 37, 36] may fail to tell the objects from the background clutter.

Window-based saliency detection

Saliency detection (mostly bottom-up) can be used to guide the object discovery task for two reasons: (1) conceptually, it is a strong prior that many observed objects are

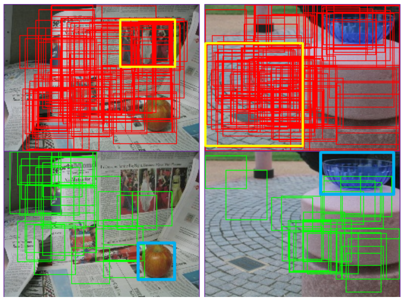

Figure 2: Example of bags and instances: Blue rectangle: the desired object window. Yellow rectangle: the most salient window obtained by [12]; Red rectangles: top salient windows as instances in positive bag; Green rectangles: randomly sampled and least salient windows as instances in negative bag.

salient in the training images; 2) practically, simple, efficient and effective saliency detection methods [4, 16, 12] deliver encouraging results on challenging images.

In particular, window-based saliency [12] is proposed to measure how likely an image window contains a salient object. This method computes the saliency scores for all windows and outputs the locally optimal ones as object candidates. Although the complex background may create many false detections, we observe that the objects are mostly covered in the top ranked windows, as shown in Figure 2.

From unsupervised object discovery to weakly supervised learning

The observation is validated in SIVAL dataset and we find that 98% objects are covered in the top 70 salient windows. This nice property naturally allows us to define positive and negative bags, which are then used in multiple instance learning. Specifically, for each image a positive (object) bag consists of detected top salient windows and a negative (background) bag consists of randomly sampled and least salient windows, as illustrated in Figure 2. In this way, we convert unsupervised object discovery into a weakly supervised learning problem.

4. Our formulation

In the following sections, we introduce a new learning method, bottom-up multiple class learning (bMCL) and show how to perform its optimization.

4.1. Review of Multiple Instance Learning

Multiple instance learning (MIL) is a popular approach in weakly supervised learning. Here we give a brief overview and focus on the boosting based MIL approaches [34, 1, 8]. In MIL, each bag $x_i \in \mathcal{X}^m$ consists of a set of instances $\{x_{i1}, \ldots, x_{im}\}(x_{ij} \in \mathcal{X})$. While each bag x_i has a class label $y_i \in \mathcal{Y} = \{-1, 1\}$ as training input, instance labels $y_{ij} \in \mathcal{Y}$ are unknown and treated as hidden variables. A bag is positive if at least one instance is positive and a bag is negative if its all instances are negative, *i.e.* $y_i = \max_j (y_{ij})$. For notation simplicity, we assume each bag has the same number of instances, *i.e.* $n_i = m$ $(i = 1, \ldots, n)$.

Standard boosting [14, 24] assumes an additive model on instance-level decisions: $h_{ij} = h(x_{ij})$ where $h(x_{ij}) = \sum_t \lambda_t h_t(x_{ij})$ is a weighted vote of weak classifiers $h_t : \mathcal{X} \to \mathcal{Y}$. Assuming that $y_{ij} \in \mathcal{Y}$ is the hidden instance label, its probability as positive is given by:

$$p_{ij} = \Pr(y_{ij} = 1 | x_{ij}; h) = \frac{1}{1 + \exp(-h_{ij})}. \qquad (1)$$

The bag-level probability is computed via a Noisy-OR (NOR) model:

$$p_i = \Pr(y_i = 1 | x_i; h) = 1 - \prod_{j=1}^{m} (1 - p_{ij}). \qquad (2)$$

Since the bag label is given in the training set, we can optimize the negative log-likelihood function: $\mathcal{L}_{MIL} = -\sum_{i=1}^{n} (\mathbf{1}(y_i = 1) \log p_i + \mathbf{1}(y_i = -1) \log(1 - p_i))$, by greedy search for h^t over a weak classifier candidate pool, followed by a line search for λ_t. $\mathbf{1}(\cdot)$ is an indicator function. According to the AnyBoost[24] framework, the weight w_{ij} on each instance x_{ij} is updated as:

$$w_{ij} = -\frac{\partial \mathcal{L}_{MIL}}{\partial h_{ij}} = \begin{cases} -\dfrac{1}{1 - p_{ij}} \dfrac{\partial p_{ij}}{\partial h_{ij}} & \text{if } y_i = -1 \\ \dfrac{1 - p_i}{p_i(1 - p_{ij})} \dfrac{\partial p_{ij}}{\partial h_{ij}} & \text{if } y_i = 1 \end{cases} \qquad (3)$$

4.2. Bottom-up Multiple Class Learning

Previous MIL solutions cannot be directly applied in unsupervised object discovery since they assume the single class among positive bags. While multiple instance clustering (MIC) approaches [37, 36] are designed to explore hidden patterns in multiple classes, their performance is poor because they treat all the images as positive bags and there are no negative bags.

In bMCL, we propose a maximum margin clustering concept [35] into the MIL scheme. The overall formulation tries to (1) discriminate the positive (object) instances from negative (background) instances; (2) maximize the difference between different object classes in the positive bags.

Given K object classes and N unlabeled images, we obtain $n = 2N$ bags (N positive bags and N negative bags based on bottom-up saliency detection). There are two kinds of hidden variables in bMCL: 1) the instance-level label y_{ij} for each instance x_{ij} in bag x_i and 2) the class latent label $k_{ij} \in \mathcal{K} = \{0, 1, \ldots, K\}$ for the instance x_{ij} that belongs to the k^{th} class (we denote $k_{ij} = 0$ and $k_i = 0$ for the negative instance and bag respectively). Here, we assume the existence of only one foreground object class in each positive bag; that is, we allow one class of objects to appear in each image. Thus, the class label k_i for each positive bag is defined based on the class labels of its instances as

$$k_i = k \iff \forall j,\ k_{ij} \in \{0, k\} \text{ and } \max_j (k_{ij}) = k \tag{4}$$

where $k \in \{1, ..., K\}$. Throughout the paper, we denote $H = (H_K, H_I)$ as hidden variables where $H_K = \{k_i, i = 1, .., n\}$ and $H_I = \{y_{ij}, i = 1, .., n, j = 1, .., m\}$ (Please notice that $k_{ij} = y_{ij} \cdot k_i$).

For bags $X = \{x_1, \ldots, x_n\}$ with their corresponding labels $Y = \{y_1, \ldots, y_n\}$, we define the overall negative log-likelihood function $\mathcal{L}(\theta; Y, X)$ as

$$\begin{aligned}\mathcal{L}(\theta; Y, X) &= -\log \Pr(Y|X;\theta) = -\log \sum_{H_K} \Pr(Y, H_K|X;\theta) \\ &= -\log \sum_{H_K} \sum_{H_I} \Pr(Y, H|X;\theta),\end{aligned} \tag{5}$$

where the model parameter $\theta = \{h^1, .., h^k, .., h^K\}$ and h^k is the appearance model for the k^{th} object class. The evaluation score for x_{ij} to the k^{th} class is computed as $q^k_{ij} = q^k(x_{ij}) = \frac{1}{1+\exp(-h^k_{ij})}$ where $h^k_{ij} = h^k(x_{ij})$. Thus, we compute the instance-level probability as

$$p^k_{ij} = \Pr(k_{ij} = k|x_{ij};\theta) \propto \prod_{t=1}^{K} (q^t_{ij})^{\mathbf{1}(t=k)} (1 - q^t_{ij})^{\mathbf{1}(t \neq k)}. \tag{6}$$

Next, we derive the probability $\Pr(Y, H_K|X;\theta)$; we assume all the bags being conditionally independent:

$$\Pr(Y, H_K|X;\theta) = \prod_{i=1}^{n} \Pr(y_i, k_i|x_i;\theta) = \prod_{i=1}^{n} [\Pr(k_i|x_i;\theta) \cdot s_i], \tag{7}$$

where $s_i = \mathbf{1}((y_i = -1 \wedge k_i = 0) \vee (y_i = 1 \wedge k_i \neq 0))$.

For each positive or negative bag, because the full derivation is combinatorial, we approximate its probability as

$$Pr(k_i = k|x_i;\theta) \approx \prod_{t=1}^{K} \left[(q^t_i)^{\mathbf{1}(t=k)} (1 - q^t_i)^{\mathbf{1}(t \neq k)}\right] \tag{8}$$

where $q^t_i = \Pr(\exists j, k_{ij} = t|x_i;\theta) = 1 - \prod_{j=1}^{m} (1 - p^t_{ij})$ denotes the measure for at least one instance x_{ij} in bag x_i belonging to the t^{th} class. Then $\Pr(Y, H_K|X;\theta)$ can be denoted in a class-wise manner:

$$\Pr(Y, H_K|X;\theta) \propto \prod_{t=1}^{K} \prod_{i=1}^{n} \left[(q^t_i)^{\mathbf{1}(t=k_i)} (1 - q^t_i)^{\mathbf{1}(t \neq k_i)} \cdot s_i\right]. \tag{9}$$

We could further explicitly use the instance-level hidden variables H_I and denote $\Pr(Y, H|X;\theta)$. Similar to the overall loss function $\mathcal{L}(\theta; Y, X)$, we also define the bag-level loss function $\mathcal{L}(\theta; Y, X, H_K) = -\log \Pr(Y, H_K|X;\theta)$ and the instance-level loss function $\mathcal{L}(\theta; Y, X, H) = -\log \Pr(Y, H|X;\theta)$, which will be later used in our Discriminative EM (DiscEM) algorithm (See the next section).

In DiscEM, if the expectation of $H = \{H_K, H_I\}$ is estimated, we could decompose the minimization of the overall loss function $\frac{d}{d\theta}\mathcal{L}(\theta; Y, X)$ into $\frac{d}{d\theta}\mathcal{L}(\theta; Y, X, H)$ and optimize K standard boosting additive models on instance-level decisions: $h^k_{ij} = h^k(x_{ij})$, where $h^k(x_{ij}) = \sum_t \lambda_t h^k_t(x_{ij})$ is a weighted vote of weak classifiers $h^k_t : \mathcal{X} \rightarrow \mathcal{Y}$. In this way, if we could well estimate the hidden variables H, bMCL can be solved with standard boosting framework[24]. Dealing with the hidden variables H falls into an EM flavor of solution[6]. Next, we discuss the details of our solution to eqn.(5).

5. Discriminative EM

The optimization of eqn.(5) deals with the hidden variables H. To solve the problem, we give a general formulation of Discriminative EM (DiscEM) algorithm, which performs discriminative learning in the presence of hidden variables. We directly apply the DiscEM to explore the hidden variables H in bMCL. We also observe that under the MIL assumption, MIL-Boost[34] is equivalent to this formulation. Based on this observation, the EM step for the instance-level hidden variables H_I is dealt with in a standard MIL-Boost and we only tackle the class labels H_K explicitly. Furthermore, because DiscEM is a general discriminative learning framework in the presence of hidden variables, it can be applied to other situations with hidden space of explicit forms.

5.1. General DiscEM Formulation

Now we will do discriminative learning with the presence of hidden variables. Our step is similar to standard EM[6] while the primary difference is that we are given labels $Y = \{y_1, \ldots, y_n\}$ in addition to observations $X = \{x_1, \ldots, x_n\}$, and we want to estimate the model θ that minimizes the negative log-likelihood function $\mathcal{L}(\theta; Y, X)$ As before, we proceed by integrating H out:

Theorem 1. *The discriminative expectation maximization (DiscEM) algorithm optimizes the training set log likelihood $\mathcal{L}(\theta; Y, X)$ w.r.t. model parameters θ in the presence of hidden variable H, via*

$$\frac{d}{d\theta}\mathcal{L}(\theta; Y, X) = E_{H \sim \Pr(H|Y,X;\theta)} \frac{d}{d\theta}\mathcal{L}(\theta; Y, X, H), \tag{10}$$

where $\mathcal{L}(\theta; Y, X, H) = -\log \Pr(Y, H|X;\theta)$. Notice that $\Pr(H|Y, X;\theta) = \frac{\Pr(Y,H|X;\theta)}{\Pr(Y|X;\theta)}$ and X, Y are given.

The general form of DiscEM is similar to the standard EM. We iteratively improve an initial estimate θ_0 with suc-

cessively better estimates $\theta_1, \theta_2, ...$, and so on until convergence. Each phase r consists of two steps:
E step: Compute $\Pr(H|Y,X;\theta)$ via previous estimate θ_r.
M step: Update θ_{r+1} by minimizing $\mathcal{L}(\theta;Y,X)$.

Note that in the above formulation, parameter θ can be purely discriminative, *i.e*. they are parameters of classifiers. In this way, DiscEM can take the advantages of discriminative learning algorithms. This contracts DiscEM to other conditional-EM frameworks[17, 28], where the task is to learn generative parameters through a discriminative objective. Compared with standard supervised algorithms, DiscEM can better handle hidden variables and embrace the weakly supervised learning setting.

Assuming all the data are conditionally independent *i.e.* $\Pr(Y|X;\theta) = \prod_{i=1}^{n} \Pr(y_i|x_i;\theta)$, we give the main insight connecting MIL-Boost[34] and DiscEM:

Theorem 2. *When the instance-level model (1) and the bag-level model (2) are used, MIL-Boost's update rule (3) is equivalent to DiscEM, which reads:*

$$\frac{d}{d\theta}\log\Pr(y_i|x_i;\theta) = \begin{cases} \displaystyle\sum_{j=1}^{m} \frac{-1}{1-p_{ij}} \frac{d}{d\theta} p_{ij} & \text{if } y_i = -1 \\ \displaystyle\sum_{j=1}^{m} \frac{1-p_i}{p_i(1-p_{ij})} \frac{d}{d\theta} p_{ij} & \text{if } y_i = 1 \end{cases} \tag{11}$$

Proof: due to the space limit, we show the proof of two theorems in the supplementary material.

The above DiscEM formulation of MIL-Boost partly explains its success. However, since MIL-Boost combines weak classifiers, which can not easily attain the optimum in the **M** step, it has to incorporate a gradient descent strategy in the function space [24].

5.2. DiscEM for bMCL

DiscEM could be directly applied to bMCL since bMCL forms an optimization problem for discriminative cost function $\mathcal{L}(\theta;Y,X)$ under the complex hidden variables $H = (H_K, H_I)$ in eqn.(5) . Based on **Theorem 1**, we could alternate between **E** step (applying model θ_r to obtain the probability estimation of instance labels H_I^r and class labels H_K^r, and sampling) and **M** step (train new classifiers based on sampled data). Furthermore, taking advantage of the equivalence between DiscEM and MIL-Boost, we could replace the **EM** step for the instance labels H_I by a standard MIL-Boost[34] and only need to integrate H_K out.

We use **Theorem 1** to rewrite $\frac{d}{d\theta}\mathcal{L}(\theta;Y,X)$ as

$$\frac{d}{d\theta}\mathcal{L}(\theta;Y,X) = E_{H_K\sim\Pr(H_K|Y,X,\theta)}[\frac{d}{d\theta}\mathcal{L}(\theta;Y,X,H_K)]. \tag{12}$$

The loss function could be decomposed in a class-wise manner, *i.e.*$\mathcal{L}(\theta;Y,X,H_K) = \sum_{k=1}^{K}\mathcal{L}^k(h^k;Y,X,H_K)$. Using eqn.(9), $\mathcal{L}^k(h^k;Y,X,H_K)$ can be computed as:

$$\begin{aligned}\mathcal{L}^k(h^k;Y,X,H_K) = -\sum_{i=1}^{n}[&\mathbf{1}(k=k_i)\log q_i^k \\ &+ \mathbf{1}(k\neq k_i)\log(1-q_i^k)],\end{aligned} \tag{13}$$

which is valid when all the (y_i, k_i) in (Y, H_k) satisfy the condition $s_i = \mathbf{1}((y_i = -1 \wedge k_i = 0) \vee (y_i = 1 \wedge k_i \neq 0))$, as shown in eqn.(9). Note that there is a normalization term in eqn.(9); we ignore it here for computational simplicity since it is close to 1; ignoring it does not affect the general formulation of DiscEM in eqn. (12).

Eqn.(13) essentially builds K classifiers with each classifier h^k takes bags labeled class k as positive bags and all the rest as negative bags, and minimizes $\mathcal{L}^k(h^k;Y,X,H_K)$ separately. This formulation can one way be understood as maximizing margins among positive bags of different classes and also the negative bags, since both SVM and Boosting maximize the margin explicitly and implicitly respectively.

For each $\mathcal{L}^k(h^k;Y,X,H_K)$, hidden instance variables H_I could be further integrated out:

$$\begin{aligned}&\frac{d}{d\theta}\mathcal{L}^k(h^k;Y,X,H_K) = \\ &E_{H_I\sim\Pr(H_I|Y,H_K,X;\theta)}[\frac{d}{d\theta}\mathcal{L}^k(h^k;Y,X,H)].\end{aligned} \tag{14}$$

However, since $\mathcal{L}^k(h^k;Y,X,H_K)$ is the same cost function discussed in **Theorem 2**, rather than integrating H_I out in eqn.(14), we use a standard boosting based MIL approach[34] to minimize the cost function.

Algorithm 1 Bottom-up Multiple Class Learning

```
**Input**: Bags $\{x_1,\dots,x_n\},\{y_1,\dots,y_n\},T,K,H_K^0$.
**Output**: $K$ discriminative classifiers: $h^1,\dots,h^K$.
$r = 0$ .
**repeat**
  $r = r + 1$.
  **for** $k = 1 \to K$ **do** {**M** Step}
    Given class variables $H_K^{r-1}$, group terms $\mathcal{L}^k(h_r^k;Y,X,H_K^{r-1})$ by class indices.
    Train a strong MIL classifier $h_r^k$ to minimize $\mathcal{L}^k(h_r^k;Y,X,H_K^{r-1})$ via MIL-Boost. $T$ is the number of weak classifiers in MIL-Boost.
  **end for**
  **for** $i = 1 \to n$ **do** {**E** Step}
    Compute $\Pr(y_i = 1, k_i = k|x_i;\theta_r)$ using estimated model $\theta_r = \{h_r^1,\dots,h_r^K\}$. Sample $k_i$ via $\Pr(k_i = k|y_i = 1, x_i;\theta_r) \sim \Pr(y_i = 1, k_i = k|x_i;\theta_r)$.
  **end for**
**until** $H_K^r = H_K^{r-1}$
```

Details of bMCL are illustrated in Algorithm 1. We iterate between **M** step and **E** step until no class labels H_K^r are changed. To obtain a good initialization, we collect all the top salient S_c windows in all the images, obtain K initial clusters using K-means[9], and sample the H_K^0 based on $\Pr(y_i = 1, k_i = k|x_i; \theta_0) \propto 1/(1 + \exp(-\sigma||x_i - c_k||^2))$ where c_k is the centroid of the k^{th} cluster.

6. Experiments

Dataset Our goal is to perform unsupervised object class discovery under general conditions. As most frequently used databases[10, 13, 31] in this problem have specific constraints as discussed earlier, we turn to more challenging vision benchmarks, briefly described below.

SIVAL dataset [25] is frequently used in MIL, semi-supervised learning, and image retrieval. It is a difficult set because the scenes are highly diverse and often complex and the objects may occur anywhere spatially in the image and also may be photographed with different orientations. Similar as in [36, 37], we randomly partition the 25 object classes into 5 groups, named from *SIVAL1* to *SIVAL5*.

CMU-Cornell iCoseg dataset [2] is designed for co-segmentation and contains 38 object classes. We construct a challenging subset, named *CC*, containing five classes with certain similarities in object appearances and backgrounds: two kinds of red planes, helicopter, kite, and hot balloon.

3D object category dataset [29] contains 10 object classes, where each class contains 10 different object instances imaged under different viewpoints and distances. We randomly select one object instance from each class and partition the 10 sub-classes into two datasets, named *3D1* and *3D2*. To increase the challenge, only images with the smallest object scale are included.

Parameters and features: In bMCL, the positive bag is fixed to the top 70 salient windows returned by [12], and the negative bag is the 40 least salient windows from a large set of random windows. The other parameters are fixed as $K = 5, T = 100, S_c = 3, \sigma = 0.1$. We use Color Moment, Edge Histogram, and GIST as window representation. Decision stump is used as the weak classifier throughout the experiment.

6.1. Simultaneous categorization and localization

We show the superior performance of bMCL over two recent multiple instance clustering (MIC) approaches M^{3}IC [36] and BAMIC [37] (we compare with the best distance metric of those used in BAMIC), and one state-of-the-art unsupervised object discovery approach [18] (UnSL), which achieves top performance (about 98% purity) on a Caltech-101 subset [10]. We use their implementations and the same parameters as those in the original work. The same feature space for bMCL is provided to BAMIC and M^{3}IC.

There has been little work on exploiting saliency for the task, except [27]. We implement a saliency detection based baseline (SD) by selecting the most "salient" window obtained by [12] in each image and clustering such windows by K-means [9]. This baseline is more reasonable than the straightforward and greedy method in [27].

	bMCL	SD	M^{3}IC	BAMIC	UnSL
SIVAL1	**95.3**	80.4	39.3	38.0	27.0
SIVAL2	**84.0**	71.7	40.0	33.3	35.3
SIVAL3	**74.7**	62.7	37.3	38.7	26.7
SIVAL4	**94.0**	86.0	33.0	37.7	27.3
SIVAL5	**75.3**	70.3	35.3	37.7	25.0
CC	**80.0**	73.9	46.1	47.8	60.0
3D1	**81.1**	64.0	46.9	43.2	37.3
3D2	**85.6**	82.9	52.3	51.4	37.5

Table 1: Object categorization performance measured by the mean purity. We compare bMCL with recent MIC approaches M^{3}IC[36], BAMIC[37], one state-of-the-art unsupervised discovery method, UnSL[18] and SD (saliency detection baseline), more reasonable than [27].

In bMCL, we use learned object detectors to evaluate the densely sampled (multi-scale, multi-size) image windows and output the class label k_i and the instance x_{ij} (window) with the highest probability p_{ij}^k for each bag x_i (image).

Similar as in previous work [20, 19], purity is used as the evaluation metric for the categorization problem, which measures the extent to which a cluster contains images of a single dominant class. Table 1 reports the mean purity results of 10 runs for all the methods, and shows that bMCL outperforms all the other methods by a large margin.

The performance gap can be well explained by the illustrative results shown in Figure 3. MIC approaches (BAMIC [37] and M^{3}IC [36]) do not explicitly differentiate objects from backgrounds (no negative bags) and can be easily confused by similar backgrounds. Besides, MIC methods cannot perform object localization. The keypoint based UnSL [18] approach lacks a spatial constraint on keypoints and the found object keypoints are scatted over the entire image. By contrast, bMCL finds an object class only when multiple salient windows from different images agrees with each other, which is a better constraint under more general conditions (different object sizes and complex background). Note that the saliency detection alone is far from perfection, as shown in the first row in Figure 3.

6.2. Detecting novel objects using learned detector

Previous unsupervised object discovery methods cannot obtain discriminative object models in an integrated manner. They are either restricted to only categorization (no object localization) [22, 18, 38], or have to resort to a separate detector training process using their localization results [15, 19, 21], or only obtain specialized detectors such as chamfer distance based shape templates [20]. By contrast, bMCL integrates the detector training into the framework for generic object classes.

To validate the generalization ability of such learned detectors, we randomly withdraw 5 images from each object

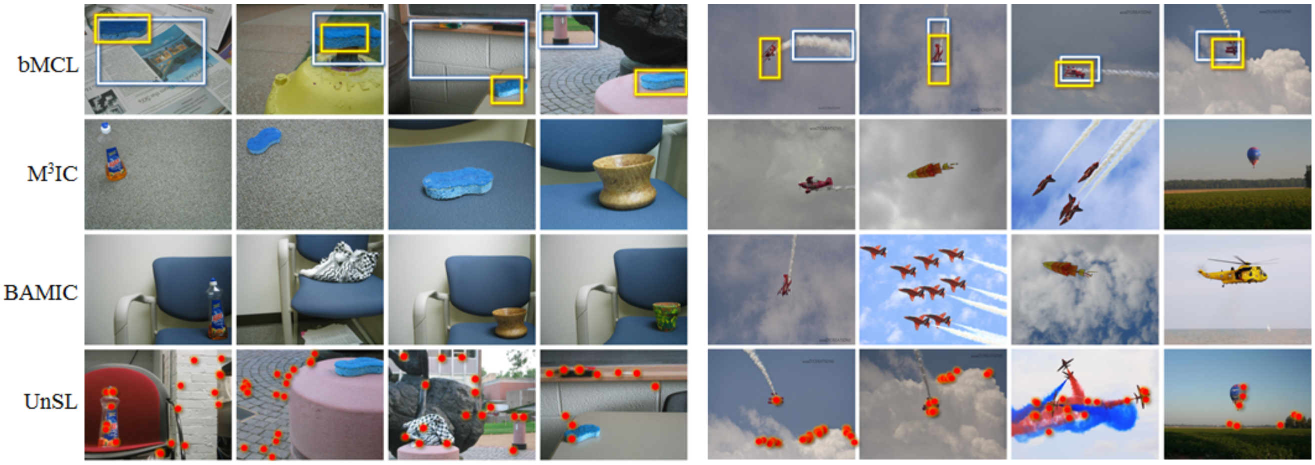

Figure 3: Illustrative categorization results of four methods in two object classes, left from SIVAL1 [25] and right from CC [2]. From top to down: bMCL, M^{3}IC [36], BAMIC [37] and UnSL [18]. In bMCL, the yellow rectangle is the localized object and the white rectangle is the most salient window computed by [12]. In UnSL, the learned object keypoints are overlayed (red points). See Section 6.1 for detailed discussions.

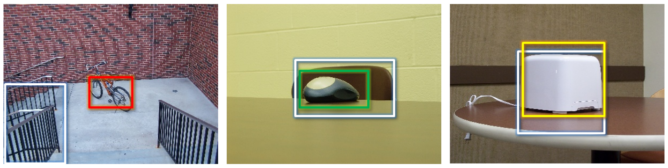

Figure 4: Novel object detection results in 3D object dataset[2]. Color rectangles: the bMCL's localization and classification result. White rectangles: saliency detection results by [12].

	apple	book	candle	note	scrunge
bMCL	0.65	**0.75**	**0.66**	**0.65**	**0.68**
[12]	**0.69**	0.74	0.62	0.54	0.61
[4]	0.49	0.71	0.43	0.62	0.52

Table 2: Comparison of co-saliency using bMCL with state-of-the-art saliency detection methods [12, 4] in F-measure of five SIVAL classes: apple, dataminingbook (book), candlewithholder (candle), stripednotebook (note), and bluescrunge (scrunge).

class, train bMCL models using the remaining images, and detect the object in the withdrawn images as described in Section 6.1. The detection accuracy[1] over SIVAL (averaged over 5 SIVAL datasets), CC and 3D (averaged over 3D1 and 3D2) is 74.4%, 72.0% and 76.0%, respectively. Note that such average categorization purity results on the three datasets in Table 1 are 84.7%, 80.0% and 83.4%, respectively. We consider the detection accuracy as satisfactory enough since the detectors are trained on a smaller training set. Figure 4 shows exemplar detection results.

6.3. Extra properties about bMCL

bMCL performs multi-class object discovery in an integrated framework. Here we briefly illustrate that some specific tasks are indeed special cases of our bMCL formulation. These include methods like co-saliency [3] and weakly supervised object localization and learning [26, 5, 7].

	bMCL	[7]	[5]	[26]
PASCAL 06	45	**49**	34	27
PASCAL 07	**31**	28	19	14

Table 3: Comparison with previous weakly supervised learning methods, measured in CorLoc [7]. We follow the same experimental setting as [7] and cite their reported results for [26, 5, 7].

Co-saliency is a relatively new concept proposed in [3] and used to perform co-segmentation of the same object instance in multiple images. However, the advantage of using co-saliency prior for co-segmentation task is not very clear since the improvement over single-image saliency prior is marginal (about 1%).

Our framework can be used to find the same salient object in multiple images, when $K = 1$. Its efficacy is validated using the two state-of-the-art saliency methods [12, 4][2]. The most salient window in bMCL is obtained as described in Section 6.1. For [4], the smallest rectangle containing 95% of total saliency pixels is regarded as the most salient window in each image. In [12], it is the window with the best saliency score. Comparison of F-measure [4] of three methods in all 25 SIVAL [25] classes clearly shows that bMCL outperforms methods in [12, 4] (better than [4] in all classes and better than [12] in 22 classes). Table 2 reports the results on five randomly selected classes. Superior performance of bMCL shows that utilizing the inter-image knowledge in a top-down manner can improve the bottom-up saliency detection.

Weakly supervised learning with a single object class Previous work [26, 5, 7] addresses the problem of localizing objects of a single class and learning a corresponding

[1] A detected object is correct if its category is correct and its overlap with ground truth object is larger than 50%

[2] We use the implementations from the authors.

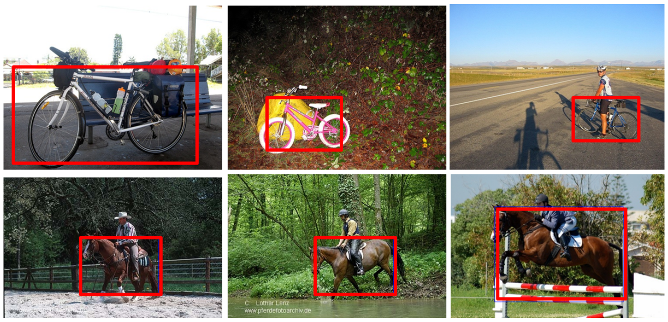

Figure 5: Red rectangles: object localization results of bMCL with a single object class (bicycle or horse) on the challenging PASCAL 07.

detector. Similarly, bMCL can actualy fit this task by setting $K = 1$. Table 3 shows that bMCL outperforms [26, 5] and is comparable with [7] on the challenging PASCAL datasets. Note that the method in [7] trains varying meta-information classifiers for different datasets whereas bMCL adopts bottom-up saliency detection to discover multi-class objects, which is more general, efficient and convenient in practice. Figure 5 illustrates exemplar object localization results on PASCAL VOC 07.

7. Conclusion

In this paper, we have introduced a new learning algorithm, bottom-up multiple class learning, which performs object localization, object class discovery, and object detector training in an integrated framework. We show the great advantage of the proposed method on a variety of benchmark datasets. We also demonstrate that our method achieves comparative results on a range of extra tasks including co-saliency and weakly supervised learning with the single class. Moreover, our saliency-guided notion may arouse more attention on utilizing the saliency measure for high-level vision applications in the future.

Acknowledgements: This work was supported by Microsoft Research Asia and also supported by ONR N000140910099 and NSF CAREER award IIS-0844566.

References

[1] B. Babenko, P. Dollár, Z. Tu, and S. Belongie. Simultaneous learning and alignment: Multi-instance and multi-pose learning. In *ECCV workshop on Faces in Real-Life Images*, 2008. 1, 3

[2] D. Batra, A. Kowdle, D. Parikh, J. Luo, and T. Chen. icoseg: Interactive co-segmentation with intelligent scribble guidance. In *CVPR*, 2010. 6, 7

[3] K. Chang, T. Liu, and S. Lai. From co-saliency to cosegmentation: An efficient and fully unsupervised energy minimization model. In *CVPR*, 2011. 2, 7

[4] M.-M. Cheng, G.-X. Zhang, N. J. Mitra, X. Huang, and S.-M. Hu. Global contrast based salient region detection. In *CVPR*, 2011. 2, 3, 7

[5] O. Chum and A. Zisserman. An exemplar model for learning object classes. In *CVPR*, 2007. 7, 8

[6] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *J. Royal Statist. Soc. Series B*, 39(1):1–38, 1977. 4

[7] T. Deselaers, B. Alexe, and V. Ferrari. Localizing objects while learning their appearance. *Technical report, ETH Zurich*, 2010. 2, 7, 8

[8] P. Dollár, B. Babenko, S. Belongie, P. Perona, and Z. Tu. Multiple component learning for object detection. In *ECCV*, 2008. 1, 3

[9] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification (2nd Edition)*. Wiley-Interscience, 2 edition, Nov. 2001. 1, 2, 6

[10] L. Fei-Fei, R. Fergus, S. Member, and P. Perona. One-shot learning of object categories. *IEEE Trans. PAMI*, 28(4), 2006. 1, 2, 6

[11] P. F. Felzenszwalb, R. B. Girshick, and D. R. David A. McAllester. Object detection with discriminatively trained part-based models. *IEEE Trans. PAMI*, 32(9), 2010. 1

[12] J. Feng, Y. Wei, L. Tao, C. Zhang, and J. Sun. Salient object detection by composition. In *ICCV*, 2011. 2, 3, 6, 7

[13] V. Ferrari, T. Tuytelaars, and L. V. Gool. Object detection by contour segment networks. In *ECCV*, 2006. 1, 6

[14] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *J. of Comp. and Sys. Sci.*, 55(1):119–139, 1997. 3

[15] K. Grauman and T. Darrell. Unsupervised learning of categories from sets of partially matching image features. In *CVPR*, 2006. 1, 2, 6

[16] X. Hou and L. Zhang. Saliency detection: A spectral residual approach. In *CVPR*, 2007. 2, 3

[17] T. Jebara and A. Pentland. Maximum conditional likelihood via bound maximization and the cem algorithm. In *NIPS*, 1998. 5

[18] G. Kim, C. Faloutsos, and M. Hebert. Unsupervised modeling of object categories using link analysis techniques. In *CVPR*, 2008. 1, 2, 6, 7

[19] Y. J. Lee and K. Grauman. Foreground focus: Unsupervised learning from partially matching images. *IJCV*, 85:143–166, 2009. 1, 2, 6

[20] Y. J. Lee and K. Grauman. Shape discovery from unlabeled image collections. In *CVPR*, 2009. 1, 2, 6

[21] Y. J. Lee and K. Grauman. Learning the easy things first: Self-paced visual category discovery. In *CVPR*, 2011. 1, 2, 6

[22] D. Liu and T. Chen. Unsupervised image categorization and object localization using topic models and correspondences between images. In *ICCV*, 2007. 1, 2, 6

[23] N. Loeff and A. Farhadi. Scene discovery by matrix factorization. In *ECCV*, 2008. 1

[24] L. Mason, J. Baxter, P. Bartlett, and M. Frean. Boosting algorithms as gradient descent. In *NIPS*, 2000. 3, 4, 5

[25] R. Rahmani, S. A. Goldman, H. Zhang, J. Krettek, and J. E. Fritts. Localized content based image retrieval. *IEEE Trans. PAMI*, 30(11), 2008. 1, 2, 6, 7

[26] B. C. Russell, A. A. Efros, J. Sivic, W. T. Freeman, and A. Zisserman. Using multiple segmentations to discover objects and their extent in image collections. In *CVPR*, 2006. 1, 2, 7, 8

[27] U. Rutishauser, D. Walther, C. Koch, and P. Perona. Is bottom-up attention useful for object recognition. In *CVPR*, 2004. 2, 6

[28] J. Salojarvi, K. Puolamaki, and S. Kaski. Expectation maximization algorithms for conditional likelihoods. In *NIPS*, 2005. 5

[29] S. Savarese and L. Fei-Fei. 3d generic object categorization, localization and pose estimation. In *ICCV*, 2007. 6

[30] J. Shotton, A. W. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from single depth images. In *CVPR*, 2011. 1

[31] J. Shotton, M. Johnson, and R. Cipolla. Semantic texton forests for image categorization and segmentation. In *CVPR*, 2008. 1, 6

[32] T. Tuytelaars, C. H. Lampert, M. B. Blaschko, and W. Buntine. Unsupervised object discovery: A comparison. *IJCV*, 2009. 1, 2

[33] P. A. Viola and M. J. Jones. Robust real-time face detection. *IJCV*, 57(2), 2004. 1

[34] P. A. Viola, J. C. Platt, and C. Zhang. Multiple instance boosting for object detection. In *NIPS*, 2006. 1, 2, 3, 4, 5

[35] L. Xu, J. Neufeld, B. Larson, and D. Schuurmans. Maximum margin clustering. In *NIPS*, 2005. 1, 2, 3

[36] D. Zhang, F. Wang, L. Si, and T. Li. Maximum margin multiple instance clustering. In *IJCAI*, 2009. 1, 2, 3, 6, 7

[37] M.-L. Zhang and Z.-H. Zhou. Multi-instance clustering with applications to multi-instance prediction. *Applied Intelligence*, 31:47–68, August 2009. 1, 2, 3, 6, 7

[38] L. Zhu, Y. Chen, and A. L. Yuille. Unsupervised learning of probabilistic grammar-markov models for object categories. *IEEE Trans. PAMI*, 31(10), 2009. 1, 2, 6