Unsupervised Object Category Discovery via Information Bottleneck Method

Zhengzheng Lou[†], Yangdong Ye[†], Dong Liu[‡]
[†] School of Information Engineering, Zhengzhou University, Zhengzhou, 450052, P. R. China
[‡] School of Computer Sci. & Tec., Harbin Institute of Technology, Harbin, 150001, P. R. China
iezzlou@gmail.com, yeyd@zzu.edu.cn, dongliu.hit@gmail.com

ABSTRACT

We present a novel approach to automatically discover object categories from a collection of unlabeled images. This is achieved by the *Information Bottleneck* method, which finds the optimal partitioning of the image collection by maximally preserving the relevant information with respect to the latent semantic residing in the image contents. In this method, the images are modeled by the *Bag-of-Words* representation, which naturally transforms each image into a visual document composed of visual words. Then the sIB algorithm is adopted to learn the object patterns by maximizing the semantic correlations between the images and their constructive visual words. Extensive experimental results on 15 benchmark image datasets show that the Information Bottleneck method is a promising technique for discovering the hidden semantic of images, and is superior to the state-of-the-art unsupervised object category discovery methods.

Categories and Subject Descriptors: H.3.1 [Information Storage and Retrieval]: Content Analysis and Indexing

General Terms: Algorithms, Performance, Experimentation

Keywords: Unsupervised Object Category Discovery, Information Bottleneck, Bag-of-Words

1. INTRODUCTION

The existing approaches for object category discovery can be mainly classified into two paradigms. The first one is supervised learning [1, 2], which needs a large amount of labeled images to train the classifiers. However, it is a labor-intensive and time-consuming task to manually label the images, and the labeling process would often invite subject biases or mistakes by human labelers. Therefore, it is technically more feasible to use the second paradigm, i.e., the unsupervised learning [3, 4], which relies solely on the unlabeled images, to perform the recognition. In this paradigm, we aim to find coherent clusters that are highly correlated with the true object categories of the images. Such a task has been fruitfully explored in image content understanding discipline, which is also the main focus of our work in this paper.

Two key issues need to be well addressed before we can automatically discover object categories from the unlabeled image collections. The first is the image processing techniques. Note that the images in the same object category may take on diversified visual appearances while the images from different object categories may also have ambiguous visual pattern. Our goal is to find the common visual pattern for each object category, even though no prior knowledge is provided. Intuitively, the global image features that describe the holistic image content are not appropriate for this given task due to their sensitivity to the scale or orientation of the object appearances. Instead, we look for low level features that are invariant to the types of degradation. In this work, we employ the local features to describe the image content. Specifically, we extract the SIFT features [5] from the images and then adopt the *Bag-of-Words* (BoW) [1, 2, 3] model as the low-level feature representation.

The second issue associated with unsupervised object category discovery is that even after obtaining robust image representations, we still need a reliable mechanism to learn the object categories from the visual contents. The existing unsupervised learning algorithms such as K-means and Affinity Propagation [6] typically try to solve this problem in a two-step manner: (1) building an affinity matrix to reflect the image relations based on image features; (2) partitioning the images into different groups based on the affinity values. Here a basic assumption in the learning procedure is that two images with high affinity value should be in the same object category. However, due to the "semantic gap" between the low-level features and the semantic concepts, such an assumption does not hold in real scenarios, which limits the performance of the existing unsupervised learning algorithms. To relieve this difficulty, an information-theoretic approach is utilized to discover the visual patterns of similar semantics in the image collection. Instead of building mapping between the representative features and semantic concepts, we conduct the object discovery process based on the *Information Bottleneck* (IB) method [7], which uses the information-theoretic optimization to automatically learn the latent semantic correlations between the images and their constructive visual words. As the visual words carry the semantic clues about the underlying concepts (e.g., images in the same objective category can be identified via a set of visual words that are informative to the given cate-

MM'10, October 25–29, 2010, Firenze, Italy.
Copyright 2010 ACM 978-1-60558-933-6/10/10 ...$10.00.

gory), the obtained latent semantic relations provide a more reliable clue for the object category discovery. In the experiments, we will show that this algorithm outperforms the state-of-the-art unsupervised object category discovery algorithms.

The main contributions of this paper can be summarized as follows:

- We propose an effective method to learn the latent semantic correlations between the images and their low-level features. In particular, we demonstrate that the learnt correlations are essentially semantic related, which alleviates the semantic gap in the existing unsupervised learning techniques.
- We introduce an novel, simple and effective algorithm to perform unsupervised object category discovery by exploiting the latent semantic correlations between the images and the visual words.

2. RELATED WORK

Some research efforts have been dedicated into the task of unsupervised object category discovery. Several works [3, 8] have applied topic models, such as PLSA, LDA, to discover object categories from a set of unlabeled images. However, these algorithms typically learn latent semantic topics based on visual words correlation while ignore the correlations between the images and the visual words. Different from these algorithms, the IB-based method can effectively exploit the semantic correlations between the images and the visual words, which results in more promising results in the task of unsupervised object category discovery. Some other methods try to solve the problem based on affinity-based algorithm. For example, Grauman *et al.* [9] employed the spectral clustering and Dueck *et al.* [6] employed Affinity Propagation to address this problem. Tuytelaars *et al.* [4] have experimented with various methods to discover object categories, including baseline methods, two latent variable models, as well as two spectral clustering methods. However, as aforementioned, these algorithms rely on typically affinity matrix which is difficult to estimate due to the "semantic gap" issues. On the contrary, the IB-based method can relieve this difficulty in the task.

Winston *et al.* [10] have utilized the IB method for video reranking and achieved good results, which proves IB is a promising method for semantic learning, but the task focuses solely on video search reranking. Goldberger *et al.* [11] have applied the aIB algorithm [12] to unsupervised image clustering. They assumed that the image colors and their spatial distribution in the image content are generated by a mixture of Gaussians, and represented the images by Gaussian mixture model. However, this method can only utilize color information to perform clustering, which is not applicable for object images with diverse visual contents. On the contrary, our method uses the BoW model as the low-level representation, which is more appropriate for the object category discovery task.

3. OUR APPROACH

In this section, we introduce the IB-based object category discovery method. We first discuss the low-level representation for the images, and then present how IB method is utilized to learn the optimized category assignment.

3.1 Image Representation

We take the BoW model as our low-level feature representation. Generally, the BoW model represents each image as a feature vector, which contains the occurrence number of the individual visual words in the image. The construction of BoW model can be implemented through the following steps.

- Extracting local patches from each image and representing them by SIFT descriptors [5].
- Building a visual vocabulary by vector quantization via the K-means algorithm in which each cluster centroid is applied as a visual word.
- Mapping the SIFT descriptors into the vocabulary so that the descriptors can be described by the visual word index.
- Counting the occurrence number of the individual visual words in each image and using a histogram to represent each image.

It is worth noting that such representation is analogous to document representation in the text analysis domain, where images can be treated as documents while the visual words represent the keywords.

3.2 IB for Object Category Discovery

We now introduce the IB-based object category discovery method. Assume we are given an image collection $\mathcal{X} = \{x_1, x_2, \cdots, x_n\}$ and obtained the visual vocabulary $\mathcal{Y} = \{y_1, y_2, \cdots, y_m\}$, where n and m are the total number of images and the size of vocabulary respectively. Based on the BoW model, each image can be represented by a histogram. Then we can define the conditional distribution of the visual words as $p(y|x) = \frac{n(y|x)}{\sum_{y' \in Y} n(y'|x)}$, where $n(y|x)$ denotes the number of occurrences of the visual word y in the image x. The prior distribution of $p(x)$ for each image is set as an uniform distribution, i.e., $p(x) = \frac{1}{n}$. Using the above definitions, the joint distribution between image variable X and visual word variable Y can be obtained by $p(x, y) = p(y|x)p(x)$.

Based on the joint distribution $p(x, y)$, we employ the IB method [7] to discover the object categories in unlabeled image collections. The IB method is a probabilistic distributional clustering method, which aims to extract a meaningful representation by compressing the image space X into a "bottleneck" variable T, while maximally preserving the revelent information with respect to visual words Y. We use mutual information $I(T; X)$ to denote the compression of images X into clusters T and use mutual information $I(T; Y)$ to denote the preserving information of T with respect to visual words Y. Then there is a tradeoff between these two mutual information values, which can be mathematically expressed as:

$$\mathcal{L}_{min} = I(T; X) - \beta I(T; Y), \tag{1}$$

where β is a trade-off parameter, and the mutual information $I(T; Y)$ can be defined as

$$I(T; Y) = \sum_{t \in T} \sum_{y \in Y} p(t, y) \log \frac{p(t, y)}{p(t)p(y)}. \tag{2}$$

The formal solution to $\mathcal{L}_{min}$ can be characterized by three distributions [7]: $p(t|x)$, the membership probability of image x belonging to cluster t; $p(y|t)$, the distribution t over the visual word variable Y; $p(t)$, the probability of the image cluster t, whose formulation can be shown as follows:

$$\begin{cases} p(t|x) = \frac{p(t)}{Z(x,\beta)} e^{-\beta D_{KL}(p(y|x)||p(y|t))} \\ p(y|t) = \frac{1}{p(t)} \sum_x p(x,y,t) = \frac{1}{p(t)} \sum_x p(x,y)p(t|x) \\ p(t) = \sum_{x,y} p(x,y,t) = \sum_x p(x)p(t|x), \end{cases} \tag{3}$$

where $Z(x,\beta)$ is a normalization function, and $D_{KL}(\cdot||\cdot)$ is the *Kullback-Leibler* divergence.

Constructing the optimal solution to the IB objective function is a NP-hard problem, and several approximate algorithms have been proposed to find the optimal solution. In this work, we adopt the sIB algorithm [13] to solve the problem, which aims to optimize an equivalent objective function as $\mathcal{L}_{max} = I(T;Y) - \beta^{-1} I(T;X)$. It takes a "draw-and-merge" procedure to find optimal solution of the function and starts with some random partition of images X into $|T|$ clusters. At each step, a single image $x \in X$ is drawn from its current cluster $t(x)$ as a new singleton cluster, and then is merged into the cluster t^{new} such that $t^{new} = \arg\min_{t \in T} d(x,t)$. The merge criterion $d(x,t)$ is computed by

$$d(x,t) = (p(x) + p(t)) \cdot JS_{\Pi}(p(y|x), p(y|t)), \tag{4}$$

where $JS_{\Pi}(p(y|x), p(y|t))$ is the *Jensen-Shannon* divergence. After sequential "draw-and-merge" procedure for all images, we can get a stable solution, where no more assignment updates can further improve $\mathcal{L}_{max}$.

4. EXPERIMENTS

4.1 Datasets and Methodologies

We use two groups of benchmark image datasets to evaluate the performance of our proposed method. The first group is constructed by [6], which includes two subsets from Caltech101. These two sets are referred to as DataSet1 and DataSet2 respectively, and their detailed information are shown in Table 1. The other group is constructed in [4] with 13 datasets in total. Among these datasets, the first one is a subset of Caltech256 with 20 categories, which is named as Dataset3 in Table 1. The other 12 datasets are generated by dividing the whole Caltech256 into 12 subsets according to the alphabetical order of the object category names. Specifically, we use the DataSet1-20 to represent the subset of the first 20 categories in Caltech256 and similarly use DataSet21-40, $\cdots$ DataSet220-240 to represent other datasets. In summary, 15 datasets (DataSet1, DataSet2, DataSet3, DataSet1-20, $\cdots$, DataSet220-240) are involved in our experiments. It should be noted that the datasets used in this paper all contain 20 categories except DataSet1, so they are much more challenging than the datasets used in the recent object category discovery works [3, 6].

In the experiments, different methods are compared in the following.

- The IB-based object category discovery method.
- K-means algorithm, where the distance between an image and a category centroid is measured by squared Euclidean distance.
- Normalized Cuts algorithm [14], where the graph edge wight matrix is calculated by negative squared Euclidean distance.
- The PLSA object category discovery method [3].

Table 2: Comparison of different unsupervised learning algorithms, evaluated by condition entropy (lower is better).

DataSets	K-means	NCuts	PLSA	IB
DataSet1	1.34	0.90	0.74	**0.68**
DataSet2	2.04	2.02	1.81	**1.51**
DataSet3	2.03	2.13	2.03	**1.81**
DataSet1-20	3.48	3.51	3.41	**3.25**
DataSet21-40	3.53	3.51	3.47	**3.27**
DataSet41-60	3.54	3.65	3.47	**3.33**
DataSet61-80	3.43	3.52	3.42	**3.22**
DataSet81-100	3.22	3.32	3.23	**3.11**
DataSet101-120	3.49	3.49	3.42	**3.27**
DataSet121-140	3.44	3.38	3.34	**3.14**
DataSet141-160	2.79	2.86	2.76	**2.52**
DataSet161-180	3.41	3.36	3.30	**3.14**
DataSet181-200	3.43	3.50	3.44	**3.27**
DataSet201-220	3.53	3.51	3.40	**3.32**
DataSet221-240	3.17	3.21	3.09	**2.93**

In the experiments, 1000 local features are randomly extracted from each image [2], and the visual vocabulary is constructed by K-means algorithm with 1000 visual words.

Following [4], we use conditional entropy to evaluate the performance of different methods, which is defined as

$$H(C|T) = \sum_{t \in T} p(t) \sum_{c \in C} p(c|t) \log \frac{1}{p(c|t)}, \tag{5}$$

where C is the ground truth category labels and T is the obtained cluster labels. The smaller the conditional entropy, the better the performance.

To alleviate the influence caused by random initialization, we ran sIB, K-means, Normalized Cuts and PLSA algorithms 10 times, each with a new random initialization. For the sIB algorithm, we select the one which maximizes the mutual information $I(T;Y)$ as the final result, and for the other algorithms, the result with the smallest conditional entropy is selected as the final result. The number of learned clusters K is taken to be identical with the number of real categories on each dataset.

4.2 Experimental Results and Analysis

The evaluation results on the 15 datasets are illustrated in Table 2, from which we have the following observations. (1) The IB-based object category discovery method consistently outperform all the other algorithms in all 15 datasets, even though the results for these algorithms are the best results selected from 10 times running. (2) Our method is clearly superior to K-means and Normalized Cuts algorithms which discover object categories based on the affinity matrices that are difficult to estimate due to the "semantic gap" issues. On the contrary, the IB-based method can relieve this difficulty by exploiting the semantic correlations between the images and the visual words. (3) The IB-based method can get more promising results than the state-of-the-art object category discovery method PLSA [3] because our method exploit the semantic correlations between the images and the visual words while the PLSA ignores these correlations.

We now demonstrate the performance of the IB-based object category discovery method. Due to the space limitation, only the confusion matrix of DataSet1 and DataSet2 are shown in Table 3 and Figure 1 respectively. Table 3 shows

Table 1: The details of the 15 benchmark datasets. The numbers in parentheses indicate the total images in each category.

DataSet1	Faces(100), Motorbikes(100), dollar_bill(52), garfield(34), snoopy(35), stop_sign(64), windsor_chair(56)
DataSet2	Faces(100), Leopards(100), Motorbikes(100), binocular(33), brain(98), camera(50), car_side(100), dollar_bill(52), ferry(67), garfield(34), hedgehog(54), pagoda(47), rhino(59), snoopy(35), stapler(45), stop_sign(64), water_lilly(37), windsor_chair(56), wrench(39), yin_yang(60)
DataSet3	American flag(97), diamond ring(118), dice(98), fern(110), fire extinguisher(84), fireworks(100), French horn(92), ketch 101(111), killer whale(91), leopards 101(190), mandolin(93), motorbikes 101(798), pci card(105), rotary phone(84), roulette wheel(83), tombstone(91), tower pisa(90), zebra(96), airplanes 101(800), faces easy 101(453)
Other datasets	DataSet1-20, DataSet21-40, $\cdots$, DataSet220-240

Table 3: Confusion matrix of the IB-based object category discovery result on DataSet1.

Learned categories→	T1	T2	T3	T4	T5	T6	T7
Faces	**98**	0	0	0	2	0	0
Motorbikes	0	**60**	0	0	**40**	0	0
dollar_bill	0	0	**48**	0	4	0	0
garfield	7	0	0	27	0	0	0
snoopy	1	0	0	27	5	0	2
stop_sign	5	0	0	0	18	**40**	1
windsor_chair	1	0	0	0	7	0	**48**

the confusion matrix between the true categories and the learned categories (denoted by T1, T2, $\cdots$, T7). From this table, we observe that the learned categories T1, T2, T3, T5, T6 and T7 are rather pure, and each can be highly associated with one true object category (Faces, Motorbikes, dollar_bill, Motorbikes, stop_sign and windsor_chair). Almost all images of garfield and snoopy are assigned to cluster T4, and there is no image of other categories in T4. The confusion matrix in Figure 1 demonstrates the encouraging result for DataSet2, where 20 categories are learnt by unsupervised learning method, and most of them can be associated with the true categories with high probabilities. The results of these two datasets show that our proposed IB-based method is able to automatically reveal the meaningful patterns from the image collection without any supervision, and is an effective object category discovery method.

5. CONCLUSIONS

We have introduced an IB-based method for unsupervised object discovery. By representing the images with the BoW model, we treat the images as a set of visual documents. The sIB algorithm is then adopted to learn optimized object categories by maximizing the semantic correlations between the images and the visual words. Extensive experiments on 15 benchmark datasets have confirmed the effectiveness of the proposed method. In the future, we will employ the IB-based object category discovery in more difficult tasks such as Web image recognition and retrieval. We believe this is a promising direction.

6. ACKNOWLEDGEMENT

This work is supported by the National Natural Science Foundation of China under grant No.60773048 with the title "Research on IB Algorithms and its Applications".

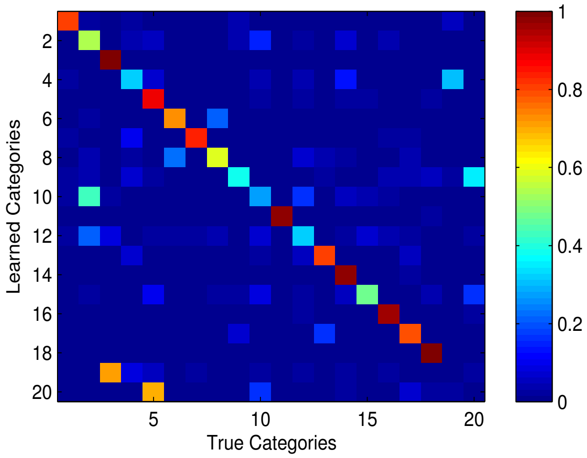

Figure 1: Confusion matrix for DataSet2. Brightness indicates the purity of the learned categories. The ideal is bright along the diagonal.

7. REFERENCES

[1] G. Csurka, C. Dance, L. Fan, J. Willamowski and C. Bray. Visual categorization with bags of keypoints. In *ECCV Workshop on Statistical Learning for Computer Vision*, 2004.

[2] E. Nowak, F. Jurie and B. Triggs. Sampling strategies for bag-of-features image classification. In *ECCV*, 2006.

[3] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman and W. T. Freeman. Discovering objects and their location in images. In *ICCV*, 2005.

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

[5] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2004.

[6] D. Dueck and B. J. Frey. Non-metric affinity propagation for unsupervised image categorization. In *ICCV*, 2007.

[7] N. Tishby, F. C. Pereira and W. Bialek. The information bottleneck method. In *ACCC*, 1999.

[8] J. Sivic, B. C. Russel, A. Zisserman, W. T. Freeman and A. A. Efros. Unsupervised discovery of visual object class hierarchies. In *CVPR*, 2008.

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

[10] W. H. Hsu, L. S. Kennedy and S.-F. Chang. Video search reranking via information bottleneck principle. In *MM*, 2006.

[11] J. Goldberger, S. Gordon and H. Greenspan. Unsupervise image-set clustering using an information theoretic framework. *TIP*, 2006.

[12] N. Slonim and N. Tishby. Agglomerative information bottleneck. In *NIPS*, 1999.

[13] N. Slonim, N. Friedman and N. Tishby. Unsupervised document classification usingsequential information maximization. In *SIGIR*, 2002.

[14] J. Shi and J. Malik. Normalized cuts and image segmentation. *TPAMI*, 2000.