Unsupervised Modeling and Recognition of Object Categories with Combination of Visual Contents and Geometric Similarity Links

Gunhee Kim
School of Computer Science
Carnegie Mellon University
gunhee@cs.cmu.edu

Christos Faloutsos
School of Computer Science
Carnegie Mellon University
christos@cs.cmu.edu

Martial Hebert
School of Computer Science
Carnegie Mellon University
hebert@cs.cmu.edu

ABSTRACT

This paper proposes a probabilistic approach for unsupervised modeling and recognition of object categories which combines two types of complementary visual evidence, visual contents and inter-connected links between the images. By doing so, our approach not only increases modeling and recognition performance but also provides possible solutions to several problems including modeling of geometric information, computational complexity, and the inherent ambiguity of visual words. Our approach can be incorporated in any generative models, but here we consider two popular models, pLSA and LDA. Experimental results show that the topic models updated by adding link analysis terms significantly improve the standard pLSA and LDA models. Furthermore, we presented competitive performances on unsupervised modeling, ranking of training images, classification of unseen images, and localization tasks with MSRC and PASCAL2005 datasets.

Categories and Subject Descriptors

I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding—*perceptual reasoning, visions*

General Terms

Algorithms, experimentation

Keywords

Object recognition, image retrieval

1. INTRODUCTION

Generative topic models based on the *bag-of-words* representation have been successful modeling and recognition tools in computer vision [7, 25, 24]. These models originated from statistical text analysis to automatically discover latent topics (*i.e.* object categories in most cases) in the images based on the distribution of visual words.

MIR'08, October 30–31, 2008, Vancouver, British Columbia, Canada.

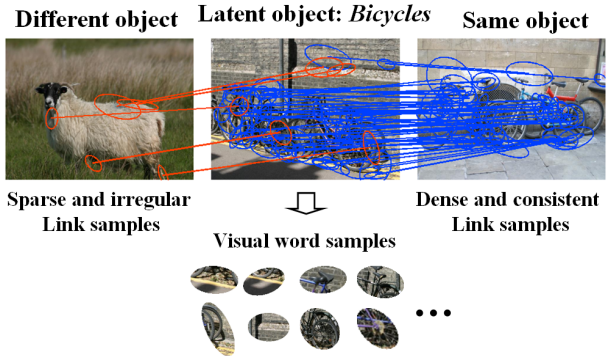


Figure 1: Intuition of the proposed approach. This figure only shows a small part of networks of the image set. The latent topics are involved in the generations of not only visual words but also links between images.

This paper introduces the use of one general idea, *link analysis*, which has been used extensively in other research areas but which has been largely ignored before in computer vision. By combining link analysis with conventional topic models such as pLSA [11] and LDA [2], we not only propose reasonable solutions to several key drawbacks of current topic models but we also report significant improvement of modeling and recognition performances. The underlying observation is that all documents in a corpus have referential relationships with one another, which are as valuable clues as the contents of documents. By analogy, images that contain similar object categories tend to generate a larger number of correspondences when compared by a matching algorithm. We will call such images as *strongly linked* with each other. As a result, analyzing which images is strongly linked with which other image (the equivalent of documents referencing each other) provides useful information in addition to the statistics of visual words (the equivalent of statistics on the content of each document). In this paper, we propose an approach to implementing this analogy for category discovery, ranking, classification, and detection applications.

Fig. 1 illustrates the intuition of our approach. The topic models are based on the observation that samples of visual words are generated from a mixture model of latent topics. Assuming that we have an image matcher which is reasonable in most cases, the distributions of the links generated

by the matcher are highly likely to be governed by the same latent topics as well. In other words, if two images share similar latent topics, then a matcher should generate a large number of consistent correspondences between features in the two images. Otherwise, the correspondences between the images may be sparse and irregular.

By combining the visual contents with link analysis, the proposed approach has three advantages over conventional topic models as follows.

(1) *Easy plug-in of geometric information into topic models*: One persistent problem of the topic models based on bag-of-words is that it is not straightforward to incorporate geometric information into the model, primarily because of the computational complexity of the model (See Section 2.) We indirectly formulate the geometric information in the form of geometrically consistent matches between a pair of images, which requires only a very simple modification of any generative models. By doing so, the complexity of our model is independent of the number of parts to be considered. We take advantage of the recent successes of a lot of off-the-shelf geometrically consistent matching approaches such as spectral matching of [15, 16], deformable matching of [1], or Pyramid matching [9]. Although there is no limitation on the choice of matching algorithms, our approach is based on the spectral matching of [15, 16].

(2) *Relaxation of the ambiguous definition of visual words*: In the visual domain, the definition of visual words is not intuitive. There are no natural boundaries, orders, and clear semantic meaning of visual words unlike in the text analysis domain from which the bag-of-words representation originated. Because of this difficulty, under/over-clustering is unavoidable during the dictionary formation. For instance, two words A and B may be quantized into different clusters even though they are semantically similar. In our approach, this effect can be relaxed by similarity links across the visual words in different images. Without quantization, the matching is based on the appearance affinity (*i.e.* L_2-norm of differences between feature vectors) in the context of geometric consistency.

(3) *Integration between two different recognition approaches - topic model based and prototype example based recognitions*: Our unsupervised modeling learns the parameters of word-topic and link-topic distributions. At the same time, our ranking method provides fixed number of prototype example images per object class. For each unseen image, we estimate the distribution of visual words and carry out image matching with exemplars to obtain link distributions. These two contributors are separately used in most previous work. The former is the topic model based recognition (*e.g.* [7, 25, 3]) and the latter is the prototype example based recognition (*e.g.* [14, 8]). In sum, we merge two recognition approaches into a single framework. Obviously, the statistics of the visual content of each image and the relationships between other images are equally valuable to describe the image.

Among diverse existing generative models, this paper only considers two standard models in computer vision such as pLSA [11] and LDA [2] although, in principle, there is no limitation to integration with any generative models. The link analysis techniques are very popular in other research areas such as text analysis, web applications, and bioinformatics. Therefore, the pLSA and LDA based models which combine topic contents with link analysis are already used in other research communities [4, 5] and the statistical models used here are inspired from this earlier work.

2. RELATED WORK

A lot of generative topic models have been proposed for modeling and recognition of object categories [7, 25, 24, 3, 6]. Due to the vast amount of previous work in this area, here we limit ourselves to topic models based on bag-of-words with *spatial information*. A general overview of this line of research can be found in [6].

The modeling of spatial information can be roughly classified into two classes of approaches. The first approach is to impose a spatial coherency constraint without suffering from computational complexity such as Spatial-LTM[3], Spatial LDA[28], and [17]. The basic idea of the first two papers is that neighboring visual words are labeled by the same latent topics if they have similar appearance. In [17], the *reward map* is constructed by feature correspondences with geometric consistency and used as extra evidence to weigh more on the features with more matches.

The other approach is to explicitly represent the spatial relations between parts in the model such as [22] and [26]. Although this approach may be computationally expensive, it can achieve more discriminative modeling power. In order to reduce the complexity issue, [22] proposes a hierarchical model which consists of a part layer and a feature layer. In [26], the theoretical inference cost linearly increases with the number of parts, but they still used a small number of parts (*i.e.* less than ten) per an object and the training images were manually aligned for learning.

Our method for incorporating geometric information is different from those two classes of approaches. Rather than explicitly plugging geometric information into the model, we indirectly model the spatial evidence in the form of links with weights across the visual features by using the output of geometric consistent matching algorithms, which are *independent* of the model. By doing so, our approach can benefit from the performance of the matching algorithms with no increase of complexity with respect to the number of parts. Also, we do not need to rely on a definition of neighborhood between the parts, and, at the same time, our approach does not sacrifice the discriminative power of geometric information of whole parts. Our approach is similar to [17] in the sense that we also take advantage of feature correspondences to design the topic models. However, the formulation of the correspondence is different. In [17], the statistics of feature matching is summed up in the form of the reward map, but here the correspondences are modeled as links between features.

This work is inspired by some notable success in combining topic contents with link analysis in web applications such as intelligent surfers [23], web crawlers [18], and the analysis of blog topic influences [21]. Recently, Kim et al [13] showed that link analysis techniques can be used for unsupervised inference of object category models. They represented visual information in the form of a large-scale network and formulated the unsupervised classification and localization in the modeling as the problem of finding *hubs* and *communities*. The *hubs* behave like important class-specific visual information and the *communities* map to a set of object categories. The basic differences of this work with [13] are 1) 1) [13] deals with only the problem of extracting categories

through unsupervised learning, but this paper addresses the ranking of images to each object and the use of the learned models for recognition in novel images. A key contribution is to show how the models learned in an unsupervised manner can be used effectively for recognition; 2) [13] is based solely on link analysis, but ours exploits not only that but also on the topic content of visual words.

3. VISUAL WORDS AND LINKS

3.1 Visual words

Following the standard approach to obtain visual words, we apply the Harris-Affine interest point detector [20] and the SIFT descriptor [19] to each image. In turn, the codebook of size W is created by K-means clustering to all feature descriptors.

3.2 Links

For the representation of linking information, we adopt Kim et al [13]'s *Visual Similarity Network* (VSN), which explicitly expresses pairwise similarities across all n features in a set of training images $\mathbb{I}$. The vertices of the VSN are features extracted from Images, and the edges are correspondences between the features in different images which are discovered by an image matcher. The weights of the edges measure the degree of appearance and geometric similarities between features. Mathematically, the VSN $\mathbf{V}$ is represented by a $n \times n$ sparse matrix in which $\mathbf{V}_{ij}$ is a non-negative similarity value. If $\mathbf{V}_{ij} = 0$, there is no similarity links between node i and j. Please note that n is a quite large number which means the total number of features in the image set but the matrix $\mathbf{V}$ is very sparse. In general, in each image hundreds of features (*i.e.* nodes) are extracted but only tens of correspondences between a pair of image are discovered by image matching. Here, the VSN is built by first matching all of the training images against each other by using the spectral matching of [15, 16] and by recording all the pairs of features that are matched by this procedure.

3.3 Co-occurrence matrices

Once the visual words are computed for all M images, the $W \times M$ term-image co-occurrence matrix ($\mathbf{N}$) is generated. Instead of using simple counts of words, we *weighted* each word before adding it to a term. Intuitively, if a feature has more inlinks (*i.e.* more matches), it receives more weight in $\mathbf{N}$. Therefore, the weight of a word w in image j, $\mathbf{N}_{wj}$, is not simply the number of occurrences of w in image j, but instead: $\sum_j \mathbf{V}_{jw}$ (*i.e.* the sum of weights of inlinks to node w). And then, each column of $\mathbf{N}$ (*i.e.* a word histogram of each image) is normalized such that the sum of original word counts of an image is preserved.

Similar to the term-image matrix $\mathbf{N}$, we define a $M \times M$ link-image co-occurrence matrix $\mathbf{A}$ as the other input to the algorithm. Since the VSN describes pairwise relationships at the feature level, we need to summarize them at the image level. Given the VSN $\mathbf{V}$, $\mathbf{A}_{ab}$ (*i.e.* the weights of the similarity inlinks from the image a to image b) is obtained as follows; $\mathbf{A}_{ab} = \sum_{i \in a} \sum_{j \in b} \mathbf{V}_{ij}$. It is a simple sum of all weights of links associated with the image a and b. In short, the link-image matrix $\mathbf{A}$ can be thought of an affinity matrix between all pairs of images.

In general, the term-image matrix $\mathbf{N}$ is quite sparse since only a small number of words in the dictionary occur in an image. In order to make the link-image matrix $\mathbf{A}$ be similarly sparse, we limit the maximum number of nonzero elements in each column of $\mathbf{A}$ to $10\log(M)$. In other words, we only consider k-neighbor neighbors for each image. (*i.e.* inlinks with top-k largest weights). k is set to $10\log(M)$ by following the recommendation of [27, 13]. In practice, this heuristic dramatically decreases the computation time and it is known to be less sensitive parameter setting (as shown in [27]). Finally, each column of $\mathbf{A}$ is normalized to be $\sum_a \mathbf{A}_{ab} = \sum_a \mathbf{N}_{ab}$. This is intended for the two factors of content and links to have the same influences on the model.

Intuitively, our representation of links and their weights are *exchangeable* in the sense that their orders are not important, which is a necessary condition for using the LDA model [2].

4. THE PROPOSED GENERATIVE MODELS

In this section, we describe the updated models which combine visual contents with links based on pLSA and LDA. The underlying assumption is that the image-specific topic distribution not only generates visual words in the image but also governs geometrically consistent similarity matching between images as shown in Fig. 1. Intuitively, the images that share similar topics tend to share similar visual appearances and matching behaviors.

4.1 pLSA-based Model

Our pLSA-based model is based on the joint probabilistic model of [4], which combines term-based pLSA with link-based pHITS. As shown in Eq.1, pLSA and pHITS have similar mathematical forms. The only difference is that the pLSA models the distribution of terms w_n in an image d_j as $P(w_n|d_j)$, whereas the pHITS models the probability of in-links c_l (*i.e.* the citation to the image l) by an image d_j as $P(c_l|d_j)$. In our application, the citation to the image l by the image j means how well the image l is matched by the image j. These two equations share the same topic-image term $P(z_i|d_j)$, which is assumed to generate terms in an image $P(w_n|z_i)$ and links with other images $P(c_l|z_i)$, respectively.

$$P(w_n|d_j) = \sum_i P(w_n|z_i)P(z_i|d_j), \quad (1)$$

$$P(c_l|d_j) = \sum_i P(c_l|z_i)P(z_i|d_j). \quad (2)$$

The parameters we are interested in are $P(z_i|d_j)$, $P(w_n|z_i)$, $P(c_l|z_i)$. They can be obtained by EM iterations to maximize the log-likelihood function (Eq.3), which is a simple extension of that of pLSA [11] by introducing the relative weight α between the two contributions. In the following experiments, we use $\alpha = 0.5$, which means the two contributions are equally weighted.

$$\mathcal{L} = \sum_j \Bigg[\alpha \sum_n \frac{\mathbf{N}_{nj}}{\sum_{n'} \mathbf{N}_{n'j}} \log \sum_i P(w_n|z_i)P(z_i|d_j) + (1-\alpha) \sum_l \frac{\mathbf{A}_{lj}}{\sum_{l'} \mathbf{A}_{l'j}} \log \sum_i P(c_l|z_i)P(z_i|d_j) \Bigg]. \quad (3)$$

where $\mathbf{N}_{nj}$ denotes how often a term w_n occurs in image d_j and $\mathbf{A}_{lj}$ indicates the weights of links c_l in image d_j (See Section 3.3).

4.2 LDA-based Model

Here, we only introduce the key equations that define our LDA-based model. They can be derived directly by following the procedures proposed by the original LDA paper [2]. We consistently follow the notations of [2] for readability. For the inference and parameter estimation, we use the variational approximation method [2]. Conceptually, our model resembles the mixed-membership models of [5] which are used for the field and subtopic classifications of papers in PNAS using *words* in abstract and *references* in bibliographies.

The joint distribution of $\{\theta, \mathbf{z}, \mathbf{w}, \mathbf{c}\}$ given the parameters $\{\lambda, \alpha, \beta\}$ and its variational distribution are given in Eq.4. These are direct extensions to the standard LDA model by introducing the link-topic distributions which are almost identical to the word-topic distributions. As shown in Eq.4, the first term is the standard expression used in the topic model based on distributions of words, the second term is the similar term obtained by using link distributions instead of word distributions. This parallel between the two terms can be carried over in the rest of the model, including, in particular in the update equations below.

In Eq.5 the Dirichlet parameter γ, the multinomial parameters $(\phi_1, ...\phi_N)$ and $(\varphi_1, ...\varphi_L)$ are the variational parameters.

$$P(\theta, \mathbf{z}, \mathbf{w}, \mathbf{c}|\lambda, \alpha, \beta) = P(\theta|\lambda)\Big(\prod_{n=1}^{N} P(z_n|\theta)P(w_n|z_n, \alpha)\Big) \Big(\prod_{l=1}^{L} P(z_l|\theta)P(c_l|z_l, \beta)\Big), \quad (4)$$

$$q(\theta, \mathbf{z}|\gamma, \phi, \varphi) = q(\theta|\gamma)\prod_{n=1}^{N} q(z_n|\phi_n)\prod_{l=1}^{L} q(z_l|\varphi_l). \quad (5)$$

Eq.6-8 show the parameters to be estimated by variational EM iteration. These are iteratively updated until convergence. The detailed procedures are described in [2].

$$\phi_{ni} \propto \alpha_{iw_n} \exp\{\Psi(\gamma_i)\}, \quad \varphi_{li} \propto \beta_{ic_l} \exp\{\Psi(\gamma_i)\}, \quad (6)$$

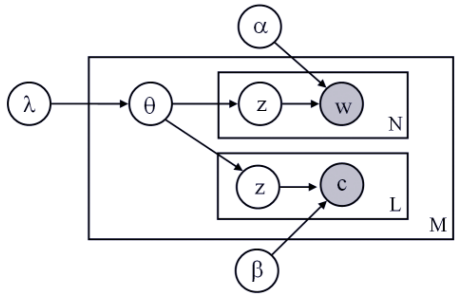
$$\gamma_i = \lambda_i + \sum_{n=1}^{N} \phi_{ni} + \sum_{l=1}^{L} \varphi_{li}, \quad (7)$$

$$\alpha_{ij} \propto \sum_{d=1}^{M}\sum_{n=1}^{N_d} \phi_{dni} w_{dn}^{j}, \quad \beta_{ij} \propto \sum_{d=1}^{M}\sum_{l=1}^{L_d} \varphi_{dli} c_{dl}^{j}. \quad (8)$$

5. MODELING, RANKING, CLASSIFICATION, AND LOCALIZATION

The main tasks we consider here are to 1) automatically generate object models in an unsupervised way, 2) find out prototype example images per object class by ranking 3) classify the unseen images, and 4) localize the probable regions of the object in a novel image. Thereafter, we refer to each of them as the *unsupervised modeling*, *ranking*, *classification*, and *localization* tasks, respectively. All related equations of pLSA and LDA-based models for theses tasks are summarized in Table 1.

Figure 2: The modified LDA model [5]. This model is a simple extension of the standard LDA by adding the link generation process which shares the same topic distributions with the word generation.



1. Sample $\theta \sim Dir(\theta|\lambda)$.
2. For each word $w_n, n \in \{1, ...N\}$
   (a) Sample $z_n \sim Mult(\theta)$
   (b) Sample $w_n \sim Mult(w|z_n, \alpha)$.
3. For each link $c_l, l \in \{1, ...L\}$
   (a) Sample $z_l \sim Mult(\theta)$
   (b) Sample $c_l \sim Mult(c|z_l, \beta)$.

In the *unsupervised modeling*, a set of M unlabeled images is classified into object classes with a single given piece of information (*i.e.* the number of object categories K). In other words, it assigns the most probable class membership to each unlabeled training image. In the pLSA model, the modeling is intuitive since we can easily obtain the distribution of latent topics, $P(z_i|d_j)$, for all images. For each image j, we select the topic i which maximizes $P(z_i|d_j)$. In the LDA model, the modeling is obtained from one of variational parameters, γ_i, which is proportional to the posterior probability that each image contains topic i.

In our framework, the *ranking* task is important for following two reasons. First, our algorithm requires pairwise image matching to generate $\mathbf{A}$ as an input. Since it is inefficient to match each test image to all training images, we select fixed number of example images (*e.g.* 30 in our experiments) per class in the training set. In other words, the ranking leads significant recognition time saving from $O(M)$ to $O(K)$. Second, since our modeling is carried out in an unsupervised way, there might be some misclassification in low-ranked training images. Therefore, our ranking algorithm can remove out those potentially inaccurate training information.

As a measure of ranking of images with respect to each topic, we use $P(c_l|z_i)$ in the pLSA model and β_{ij} in the LDA model. They indicate how likely the image is to be cited (*i.e.* matched) from within the community of topic i. In other words, if a image has high value of $P(c_l|z_i)$, then it can be interpreted as a highly influential (*i.e.* authoritative) image with respect to its object category i.

The *classification* task involves discovering the object classes of unseen images. As inputs, we need to measure the samples of visual words in the image and link samples by an image matching with $30 \times K$ example images (which correspond to $\mathbf{N}$ and $\mathbf{A}$ in Section 3.3) Generally, it is done by running

the same process using the trained word-topic distributions: $P(w|z)$ and $P(c|z)$ in the pLSA model and λ, α, β. In the fold-in heuristics [25, 3], these values are fixed during the EM inference. However, we just use the learnt parameters for the initialization and allow the updates. Experimentally, the classification performances did not change whether they are updated or not, but for localization in the new image, the update may be helpful because it allows more opportunity to fit to new image data. Since the learnt parameters should be almost same to the parameters for the test image set, the EM iterations were quickly converged.

Since we use interest region detectors as our unit visual information, the *localization* consists essentially in the selection of features which are most probable on the object in the image. Following [25], we select the feature by using $P(z_i|w_n, d_j)$ in the pLSA. In the LDA, the corresponding measure is ϕ_{ni}, which is the posterior probability that the word w_n in an image is generated from topic i. For each image, we select the features whose $P(z_i|w_n, d_j)$ are close enough to the maximum of the image. In other words, we choose the features with $P(z_i|w_n, d_j) \geq \rho \times \max_j P(z_i|w_n, d_j)$. In the experiments, $\rho = 0.8$ is used to be consistent with [13].

Table 1: Equations of pLSA and LDA-based models for unsupervised modeling(M), ranking(R), classification(C), and localization(L).

	pLSA-based model	LDA-based model
M	$i^* = \arg\max_i P(z_i\|d_j)$	$i^* = \arg\max_i \gamma_i$
R	$P(c_l\|z_i)$	β_{ij}
C	$i^* = \arg\max_i P(z_i\|d_{test})$	$i^* = \arg\max_i \gamma_{i,test}$
L	$P(z_i\|w_n, d_{test})$	ϕ_{ni}

6. EXPERIMENTS

We designed two different experiments to evaluate the proposed methods. First, in order to justify the usefulness of link analysis, we performed comparison tests between the standard pLSA and LDA models and their linked versions for the unsupervised modeling task. Second, we present results of unsupervised modeling, ranking of training images, classification, and localization of unseen images with more challenging datasets such as MSRC [12] and PASCAL05/ETHZ dataset[1].

For better comparison tests, we updated publicly available pLSA and LDA software[2].

6.1 Comparison tests

For comparison tests, we used one of the experimental setups of [13]. Specifically, we randomly selected 100 images per object for the five object classes of Caltech-101 - {*airplane*, *rear cars*, *faces*, *motorbikes*, *watches*}. The task is the unsupervised modeling, in which 500 training images are classified according to the categories with only the number of object classes (K=5) given.

We compared the performances of three different versions of topic models - (1) Standard pLSA and LDA models, (2) pLSA and LDA models with matching weighted co-occurrence matrices (Section 3.3), and (3) Linked pLSA and LDA with weighted co-occurrence matrices. Fig.3 represents variations of learned category accuracies of the three different versions with the 1000 codebook size. They represent how well the unlabeled training images are clustered according to their categories by measuring the agreement of topic discovery with ground truth labels. It clearly shows that the proposed approach (*i.e.* matching weighted counts of words and the combination of visual contents and links) leads to significant performance increase. Our LDA-based unsupervised modeling achieved 97.8% on average, which is very competitive compared to 97.3% in [13] and 86% with four object classes in [10].

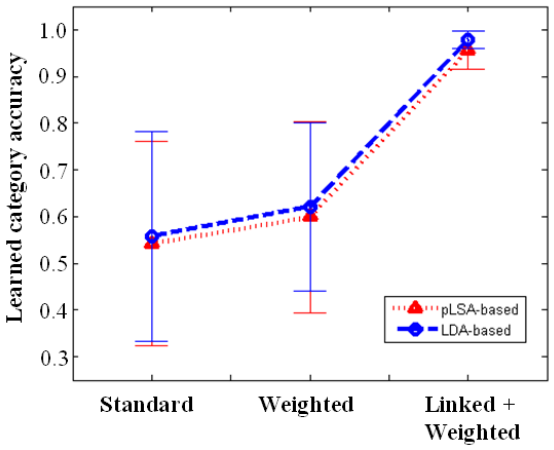


Figure 3: Performance comparison between three different versions of topic models on the five objects of Caltech-101 dataset. The means and standard deviations of 10 runs for each are shown in percentile. The accuracies of (pLSA, LDA) are (1) *Standard*: ($54.2 \pm 21.8, 55.8 \pm 22.5$), (2) *Weighted*: ($59.9 \pm 20.4, 62.1 \pm 18.1$), (3) *Linked+Weighted*: ($95.5 \pm 4.1, 97.8 \pm 1.9$).

6.2 Results of modeling, Ranking, Classification, and Detection

We evaluate the proposed unsupervised modeling and recognition method using two different datasets, which are Five objects of MSRC dataset {272 *Bicycles*, 505 *Cars*, 166 *Doors*, 190 *Sheep*, and 165 *Signs*} and four objects of PASCAL05 / ETHZ *Giraffes* {95 ETHZ *motorbikes*, 100 ETHZ *cars*, 168 TUGraz *person*, 88 ETHZ *Giraffes*}. Since all images in the MSRC dataset are 640×480, they are resized to 320×240 for better computational speed. However, the PASCAL05 consists of diverse sizes of images and they are used without rescaling. For the MSRC and PASCAL05/ETHZ datasets, we randomly selected 75 and 40 images for training and test sets, respectively. The models learned using the training images in an unsupervised way are used for classification and localization of test images. We iterated ten runs of experiments. Unlike ours, most prior work evaluates the performance of unsupervised modeling, ranking, classification, and detection in separate experiments [25, 3]. Also, the MSRC and PASCAL05 datasets are challenging in the sense that they have not been much used for unsupervised modeling.

[1]The PASCAL dataset is available at http://www.pascal-network.org/challenges/VOC/ and ETHZ *Giraffes* at http://www.vision.ee.ethz.ch/datasets.

[2]The pLSA and code is available at http://people.csail.mit.edu/fergus/iccv2005/bagwords.html The LDA code is at http://chasen.org/~daiti-m/dist/lda/

6.2.1 Unsupervised modeling

Evaluating unsupervised modeling is a bit tricky since, by definition of the task, there is really no objective *right answer* of ground truth. In these experiments, however, the number of categories is small enough and the images are simple enough (*e.g.* most of them contain only one object class) that it is reasonable to expect the unsupervised learning algorithm to recover a clustering of the training set into the desired categories. In this case, it is acceptable to evaluate performance by using a confusion matrix since we have a single ground truth category label for each training image. We do recognize that this is only possible for this type of experiment and that indirect measurement methods will be necessary in the future.

Table 2 represents the confusion matrices for the unsupervised modeling accuracies of the pLSA and LDA models. Experimental results showed that our performance is competitive since we achieved 85.44% for the PASCAL dataset and 90.29% for the MSRC dataset. Also, we observed that the LDA based model slightly outperforms the pLSA based model.

6.2.2 Ranking of training images

As introduced in Table 1, we can rank the training images with respect to each topic by using $P(c_l|z_i)$ in the pLSA and β_{ij} in the LDA based model. In practice, this ranking is quite useful since in many cases we need to find some representative images for each object category. For example, for each topic i, we can sort $P(c_l|z_i)$ or β_{ij} of all images and select top-k images with highest values as prototype images.

Fig.4 shows how the top-k example images per object class agree with ground truth category labels by varying the k from 5 to 30. Obviously, as k increases, the ratios drop slightly. One interesting observation is that for $k = 30$ the accuracy is slightly worse than the unsupervised modeling accuracy reported in Table 2. For example, for the case of $k = 30$, the LDA based model in the PASCAL dataset (*i.e.* the rightmost bar in the right picture in the Fig.4), the accuracy is 77.5%. However, the unsupervised modeling ratio for all 45 images per class is 85.44% (See the Table 2). This discrepancy occurs because $P(c_l|z_i)$ and β_{ij} are purely link analysis terms whereas the results of Table 2 are contributed by the combination of visual contents and link analysis. Therefore, this can be interpreted as a strong evidence of superiority of the combination over relying on a single aspect.

6.2.3 Classification of unseen images

For each object class, we selected the top 30 training images as exemplars as described in the previous section. In addition to words-topic/links-topic parameters learned during the training, we can also take advantage of the matches between exemplar images, which makes our method more discriminative. Table 3 shows the results of classifying test images. Even though this task is more challenging than the unsupervised modeling step in the sense that we do not have full comparison between images and can be affected by modeling errors, we only observed a slow degradation in performance. For both datasets, we achieved more than 80% success ratios.

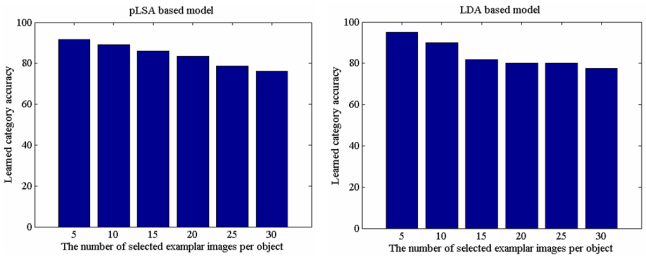


Figure 4: Learnt category accuracies of top-k selected prototype example images per object in PASCAL05 dataset (Left: pLSA based model, right: LDA based model).

6.2.4 Localization of unseen images

Fig.5 shows some localization examples on the MSRC and PASCAL05 datasets. For each image, we selected the features which satisfy $P(z_i|w_n, d_j) \geq 0.8 \times \max_j P(z_i|w_n, d_j)$. (For LDA model, ϕ_{ni} is used instead.) The topic i is assigned to each word by $i^* = \arg\max_i P(z_i|w_n, d_j)$. We draw the features by different colors according to the assigned topic. As shown in the pictures, the majority of topics assigned to high confident features are consistent with the topic of the image.

As discussed in Section 3.3, our method is based on the counts of words *weighted* by an image matcher with geometric consistency. Contrary to the standard bag-of-words representation, we can take advantage of the geometric information inferred by the matcher for the localization.

7. CONCLUSION

We introduced the use of *link analysis*, one general idea to improve the generative topic models for modeling and recognition of object categories. The approach is based on the observation that the relationships between visual information across images are as valuable as the visual contents in the images. Inspired by the statistical framework developed for web applications, our experimental results show that the combination of contents and links is a promising approach for visual inference problems, too. Since the proposed approach covers the four important visual tasks including unsupervised modeling, ranking, classification of unseen images, and localization, it can be used as a building block to the applications such as automatic image annotation, image retrieval, and object detection.

We believe that link analysis techniques for computer vision are quite new and thus much remains to be explored. First, in this paper the links are defined as *relations between images*, which is largely due to the fact that the framework used here is developed for the analysis on the documents and their references to others. However, closer investigation on the feature level interactions may be necessary since an image may contain semantically different objects. For example, a scene may consist of buses, humans, sky, tree, and buildings, which are not necessarily related. Second, we would like to explore how scalable our approach is. Since the techniques used in this work have been applied to large-scale systems such as WWW, social networks, and bioinformatics, there may not be any fundamental limitation in scala-

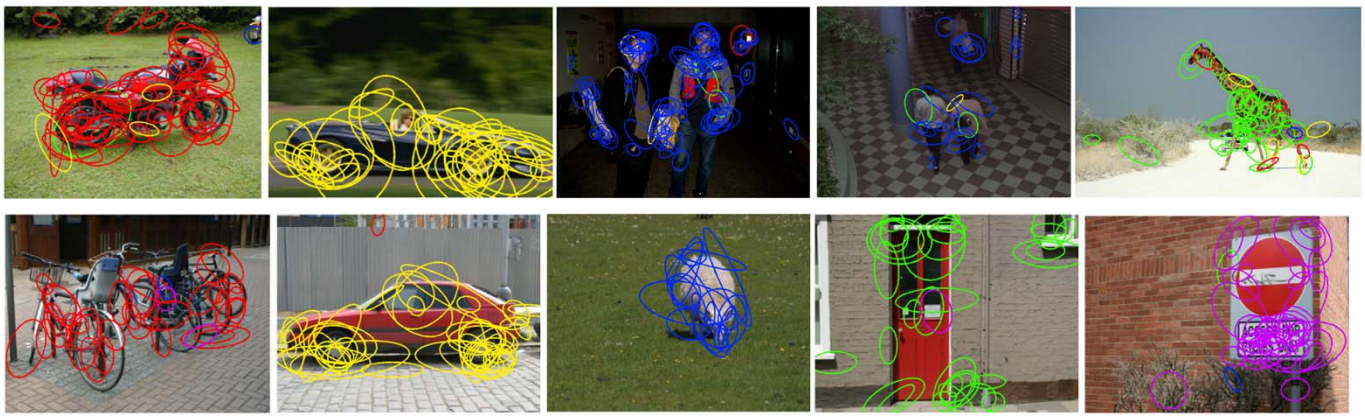


Figure 5: Localization results. The first row show examples of the PASCAL05/ETHZ, and the second row shows the MSRC dataset. The colors of the features are assigned according to the topics. (1) For PASCAL05/ETHZ, red, yellow, blue, and green are assigned to topics of *motorbikes*, *cars*, *people*, and *giraffes*, respectively. (2) For MSRC dataset, red, yellow, green, blue, and purple colors are used for *bicycles*, *cars*, *doors*, *sheep*, and *signs* topics, respectively. (These figures are best viewed in color.)

Table 2: Confusion tables for unsupervised modeling. The left and right columns show the results of pLSA and LDA based models, respectively. The means and standard deviation values of 10 runs for each are shown. The modeling accuracies of (pLSA, LDA) based models are (*83.13%*, *85.44%*) with PASCAL05/ETHZ and (*85.55%*, *90.29%*) with MSRC on average. (PG: *Giraffes*, PM: *Motorbikes*, PC: *Cars*, PP: *Persons* of PASCAL05/ETHZ dataset, MB: *Bicycles*, MC: *Cars*, MD: *Doors*, MS: *Sheep*, MG: *Signs* of MSRC dataset)

	PG	PM	PC	PP
PG	**71.3**±6.4	**13.7**±5.9	**2.5**±1.7	**12.5**±3.3
PM	**1.3**±1.8	**85.5**±5.0	**6.7**±6.8	**6.5**±5.3
PC	**0.0**	**0.0**	**100**	**0.0**
PP	**10.7**±6.5	**5.5**±4.4	**8.0**±5.9	**75.8**±8.2

	PG	PM	PC	PP
PG	**76.5**±5.8	**7.5**±5.5	**0.3**±0.8	**15.7**±6.1
PM	**1.8**±2.1	**86.0**±5.0	**1.8**±2.6	**10.4**±6.1
PC	**0.0**	**0.5**±1.1	**97.3**±2.5	**2.2**±2.2
PP	**13.0**±6.3	**3.2**±3.1	**1.8**±3.1	**82.0**±4.4

	MB	MC	MD	MS	MG
MB	**77.8**±3.7	**1.7**±1.9	**2.5**±1.9	**16.8**±4.7	**1.2**±1.0
MC	**0.0**	**74.4**±13.8	**0.7**±0.7	**10.8**±12.0	**14.1**±13.9
MD	**0.0**	**0.0**	**95.1**±1.7	**1.7**±1.1	**3.2**±1.6
MS	**0.0**	**0.1**±0.4	**1.6**±2.2	**98.3**±2.1	**0.0**
MG	**0.4**±0.9	**1.3**±1.4	**9.2**±2.6	**6.8**±2.4	**82.3**±4.0

	MB	MC	MD	MS	MG
MB	**96.1**±2.3	**0.3**±0.6	**1.2**±1.3	**2.4**±2.0	**0.0**
MC	**0.1**±0.4	**82.0**±18.1	**0.1**±0.4	**11.1**±15.3	**6.7**±12.1
MD	**1.2**±1.2	**0.1**±0.4	**94.5**±1.9	**1.9**±1.6	**2.3**±1.7
MS	**0.4**±0.6	**0.0**	**0.3**±0.6	**99.2**±1.1	**0.1**±0.4
MG	**3.4**±1.6	**2.8**±3.4	**8.5**±3.2	**5.7**±3.3	**79.6**±6.8

Table 3: Confusion tables for classification of test images. The left and right columns show the results of pLSA and LDA based models, respectively. The means and standard deviation values of 10 runs for each are shown. The classification accuracies of (pLSA, LDA) based models are (*83.63%*, *80.5%*) with PASCAL05/ETHZ and (*82.19%*, *82.16%*) with MSRC on average.

	PG	PM	PC	PP
PG	**72.8**±3.2	**11.5**±4.7	**4.2**±2.1	**11.5**±4.9
PM	**1.7**±2.1	**86.8**±5.8	**5.5**±3.1	**6.0**±4.3
PC	**0.0**	**2.0**±3.1	**95.8**±3.1	**2.2**±2.5
PP	**11.0**±3.4	**3.5**±2.1	**6.2**±6.5	**79.3**±5.9

	PG	PM	PC	PP
PG	**66.5**±6.3	**12.2**±7.7	**10.3**±5.1	**11.0**±6.7
PM	**0.5**±1.1	**80.0**±9.1	**12.7**±6.3	**6.8**±4.4
PC	**0.0**	**0.5**±1.1	**97.8**±2.8	**1.7**±2.6
PP	**9.2**±3.7	**5.3**±3.4	**7.7**±8.4	**77.8**±8.1

	MB	MC	MD	MS	MG
MB	**76.5**±6.1	**3.5**±3.0	**5.5**±3.5	**12.8**±5.0	**1.7**±0.9
MC	**0.0**	**69.3**±11.4	**1.1**±1.2	**22.7**±14.4	**6.9**±6.6
MD	**0.0**	**0.5**±0.7	**88.2**±3.7	**3.7**±1.1	**7.6**±2.7
MS	**0.5**±0.7	**0.1**±0.4	**1.2**±1.0	**97.9**±2.0	**0.3**±0.6
MG	**0.7**±0.7	**2.1**±0.9	**7.5**±1.3	**10.6**±4.1	**79.1**±5.1

	MB	MC	MD	MS	MG
MB	**79.1**±8.2	**4.4**±3.0	**5.5**±3.0	**10.1**±5.4	**0.9**±0.9
MC	**0.0**	**69.1**±10.0	**1.5**±1.3	**25.7**±12.3	**3.7**±4.1
MD	**0.0**	**0.4**±0.6	**91.2**±3.3	**3.5**±1.4	**4.9**±2.4
MS	**0.7**±0.9	**0.7**±1.1	**1.3**±2.4	**96.9**±3.3	**0.4**±0.6
MG	**0.7**±0.9	**2.9**±2.5	**11.5**±3.2	**10.4**±3.4	**74.5**±5.1

bility. However, all pairwise image matching are required during the modeling step, and accordingly a cleverer image matching scheme may be necessary to relax the quadratic computation.

Acknowledgement This research was performed in part for the Intelligent Robotics Development Program, one of the 21st Century Frontier R&D Program funded by the Ministry of Commerce, Industry and Energy of Korea.

8. REFERENCES

[1] A. C. Berg, T. L. Berg, and J. Malik. Shape matching and object recognition using low distortion correspondence. In *CVPR*, 2005.

[2] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *JMLR*, 3:993–1022, 2003.

[3] L. Cao and L. Fei-Fei. Spatial coherent latent topic model for concurrent object segmentation and classification. In *ICCV*, 2007.

[4] D. Cohn and T. Hofmann. The missing link - a probabilistic model of document content and hypertext connectivity. In *NIPS*, 2001.

[5] E. Erosheva, S. Fienberg, and J. Lafferty. Mixed-membership models of scientific publications. *PNAS*, 101(1):220–5227, 2004.

[6] L. Fei-Fei. Bag of words models: Recognizing and learning object categories. In *CVPR Short Courses*, 2007.

[7] L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *CVPR*, 2005.

[8] A. Frome, Y. Singer, F. Sha, and J. Malik. Learning globally-consistent local distance functions for shape-based image retrieval and classification. In *ICCV*, 2007.

[9] K. Grauman and T. Darrell. The pyramid match kernel: Discriminative classification with sets of image features. In *ICCV*, 2005.

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

[11] T. Hofmann. Probabilistic latent semantic analysis. In *NIPS*, 1999.

[12] A. C. John Winn and T. Minka. Object categorization by learned universal visual dictionary. In *ICCV*, 2005.

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

[14] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.

[15] M. Leordeanu and M. Hebert. A spectral technique for correspondence problems using pairwise constraints. In *ICCV*, 2005.

[16] M. Leordeanu and M. Hebert. Efficient map approximation for dense energy functions. In *ICML*, 2006.

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

[18] H. Liu, E. Milios, and J. Janssen. Probabilistic models for focused web crawling. In *WIDM*, 2004.

[19] D. Lowe. Distinctive image features from scale invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[20] K. Mikolajczyk and C. Schmid. Scale and affine invariant interest point detectors. *IJCV*, 60(1):63–86, 2004.

[21] R. Nallapati and W. Cohen. Link-plsa-lda: A new unsupervised model for topics and influence in blogs. In *ICWSM*, 2008.

[22] J. C. Niebles and L. Fei-Fei. A hierarchical model of shape and appearance for human action classification. In *CVPR*, 2007.

[23] M. Richardson and P. Domingos. The intelligent surfer: Probabilistic combination of link and content information in pagerank. In *NIPS*, 2002.

[24] 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.

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

[26] E. B. Sudderth, A. Torralba, W. T. Freeman, and A. S. Willsky. Learning hierarchical models of scenes, objects, and parts. In *ICCV*, 2005.

[27] U. von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.

[28] X. Wang and E. Grimson. Spatial latent dirichlet allocation. In *NIPS*, 2007.