Knowledge-Based Topic Model for Unsupervised Object Discovery and Localization

Zhenxing Niu, *Member, IEEE*, Gang Hua, *Senior Member, IEEE*, Le Wang, *Member, IEEE*, and Xinbo Gao, *Senior Member, IEEE*

***Abstract*— Unsupervised object discovery and localization is to discover some dominant object classes and localize all of object instances from a given image collection without any supervision. Previous work has attempted to tackle this problem with vanilla topic models, such as latent Dirichlet allocation (LDA). However, in those methods no prior knowledge for the given image collection is exploited to facilitate object discovery. On the other hand, the topic models used in those methods suffer from the topic coherence issue—some inferred topics do not have clear meaning, which limits the final performance of object discovery. In this paper, prior knowledge in terms of the so-called must-links are exploited from Web images on the Internet. Furthermore, a novel knowledge-based topic model, called LDA with mixture of Dirichlet trees, is proposed to incorporate the must-links into topic modeling for object discovery. In particular, to better deal with the polysemy phenomenon of visual words, the must-link is re-defined as that one must-link only constrains *one* or *some* topic(s) instead of *all* topics, which leads to significantly improved topic coherence. Moreover, the must-links are built and grouped with respect to specific object classes, thus the must-links in our approach are *semantic-specific*, which allows to more efficiently exploit discriminative prior knowledge from Web images. Extensive experiments validated the efficiency of our proposed approach on several data sets. It is shown that our method significantly improves topic coherence and outperforms the unsupervised methods for object discovery and localization. In addition, compared with discriminative methods, the naturally existing object classes in the given image collection can be subtly discovered, which makes our approach well suited for realistic applications of unsupervised object discovery.**

***Index Terms*— Object discovery, object localization, topic model, latent Dirichlet allocation.**

Manuscript received October 9, 2015; revised February 13, 2016 and May 7, 2016; accepted July 13, 2016. Date of publication June 22, 2017; date of current version October 17, 2017. This work was supported in part by NSFC under Grant 61432014, Grant U1605252, Grant 61402348, and Grant 61672402, in part by the Key Industrial Innovation Chain in Industrial Domain under Grant 2016KTZDGY-02, and in part by the National High-Level Talents Special Support Program of China under Grant CS31117200001. The work of G. Hua was supported by NSFC under Grant 61629301. The work of L. Wang was supported by NSFC under Grant 61503296. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Jing-Ming Guo. *(Corresponding author: Gang Hua.)*

Z. Niu and X. Gao are with the School of Electronic Engineering, Xidian University, Xi'an 710071, China (e-mail: zhenxingniu@gmail.com; xinbogao@mail.xidian.edu.cn).

G. Hua is with Microsoft Research, Redmond, WA 98052 USA (e-mail: ganghua@gmail.com).

L. Wang is with the Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: lewang@mail.xjtu.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2017.2718667

I. Introduction

OBJECT discovery and localization is to discover dominant object classes frequently present in a given image collection, and localize corresponding object instances in each image. For example, given an image collection of objects, we need to discover what object classes exist (*e.g.*, 'airplane', 'motorbike', 'horse', *etc.*) and to localize the corresponding object instances in each image with pixel-wise segmentations or bounding boxes, as shown in Fig.9. Moreover, the object classes to be discovered can be generalized as any object of interest such as scene elements (*e.g.*, 'tree', 'building', 'mountain', *etc.*), as shown in Fig.1.

This is a challenging problem since object discovery and object localization are coupled together. If the image classes are given, it becomes solvable to localize object instances in images (*i.e.*, an object detection/localization problem). Conversely, if each object instance in an image is clearly annotated with a bounding box but not its object class, it is also solvable to discover its object class (*i.e.*, an unsupervised image clustering/recognition problem). However, it is challenging to solve them simultaneously. Moreover, we tackle this problem in a fully unsupervised setting, *i.e.*, without even image-level class annotations or any assumption on the object classes to be discovered. In other words, this problem is much more challenging than the weakly-supervised localization problem [1] (*i.e.*, image-level class annotations are known) and the co-localization problem [2] (*i.e.*, with an assumption of only one single dominant class present in the given dataset).

Probabilistic models, especially topic models, seem to be well suited for tackling the unsupervised object discovery task and have been studied by Sivic *et al.* [3] and Russell *et al.* [4]. For example, Sivic *et al*. [3] have proposed a method that builds on Probabilistic Latent Semantic Analysis (PLSA) [5] to separate images into four different object classes (faces, airplanes, rear cars, and motorbikes). Furthermore, they replace PLSA with the Latent Dirichlet Allocation (LDA) [6] and use multiple image segments as the equivalent of documents, so as to better localize the objects in images [4].

However, the topic models used in those approaches (*e.g.*, LDA) suffer from a common issue—the *topic coherence* issue [7], [8]: some inferred topics do not have clear meaning, *i.e.*, the representative words for such a topic (the top-K words in a topic-specific distribution) are incoherent with each other, and hence it is hard to interpret what the topic means [7]. As a result, when directly applying such models to object

Fig. 1. Visualization of five topics inferred from LabelMe dataset. They are generated according to our visualization method (refer to Section III-D in detail). 3 topics (*i.e.*, top-3 rows) are coherent and associated to 'building', 'sky', and 'mountain' respectively; and the other 2 topics (*i.e.*, the 4-th and 5-th rows) are incoherent, where the 4-th topic associated to 'tree' is interfered with some image segments about 'mountain'; the 5-th topic is even not associated to any object class since it is represented by many unrelated image segments.

discovery, the incoherent topics do not correspond to any meaningful object class, which limits the final performance of object discovery. As shown in Fig.1, it illustrates three coherent topics inferred by LDA-MDT (*i.e.*, the 1st, 2nd and 3rd topics) and two incoherent topics inferred by LDA (*i.e.*, the 4th and 5th topics) from the LabelMe dataset, where each topic is visualized as a ranked list of image segments. Obviously, the coherent topics clearly correspond to the object classes 'building', 'sky', and 'mountain' respectively. In contrast, the 4-th topic basically corresponds to the object class 'tree' but is still interfered with some image segments about 'mountain'; Even more, the 5-th topic is too incoherent (it is represented by many unrelated image segments) to be interpreted as any specific object class.

Recently, in the text analysis domain, it is found that topic coherence can be improved by leveraging prior knowledge of words [8]–[11]. Particularly, the prior knowledge is expressed as some *Must-Link* constraints between words [9]–[11]. If two words are related to each other (*i.e.*, they have similar semantic meaning), a Must-Link between them can be built. By encouraging two words of a Must-Link having similar probability for topics we can significantly improve topic coherence [9]–[11]. A typical model is the LDA-DF which incorporates Must-Links into LDA by using the Dirichlet Forest [10]. Experimental results have demonstrated that it can improve topic coherence and produce more meaningful topics.

Inspired by those methods, we introduce those *knowledge-based topic models* into computer vision domain for object discovery and localization. Particularly, the visual knowledge, expressed by Must-Links between visual words, are exploited and incorporated into topic modeling, which can improve topic coherence and lead to better object discovery performance.

However, our preliminary experimental results reveal that the performance of object discovery cannot be significantly improved by directly employing those models such as LDA-DF for object discovery (refer to SectionIV-C). In our belief, this is because the Must-Link defined in those methods cannot properly deal with the *polysemy phenomenon* of words (*i.e.*, a word has several different semantic meaning). For example, the word 'apple' has two meaning: an electronic product and a kind of fruit. As an electronic product, a Must-Link between it and words such as 'iphone' can be established. As a kind of fruit, a Must-Link between it and words such as 'banana' can be established. Recall that the Must-Link used in all previous methods is defined as follow: *if two words have a Must-Link, it indicates that the probability of the two words for* ***all*** *topics must be similar (high or low) to each other* [10]. Due to this definition, the Must-Links must satisfy the *transitive property*: if there are two Must-Links (A, B) and (A, C), there must be another Must-Link (B, C). To satisfy the transitive property, some Must-Links will be automatically built by those methods if they do not exist.

As a result, for the previous example, another Must-Link between 'iphone' and 'banana' will be automatically built, but obviously it is undesirable. Thus, we need to pay more attention to the polysemy phenomenon of word when building Must-Link constraints.

Moreover, when applying topic models to computer vision tasks, it becomes a more serious concern since the polysemy phenomenon of visual words exists much more widely than that of words in text analysis [12], [13]. Specifically, words of a specific language exist naturally, *e.g.*, there is a pre-defined and fixed vocabulary for English. In contrast, topic models are based on the bag of visual words (BoW) image representation, where visual words are generated by clustering image local features. Due to the descriptive limitation of image local features, it is difficult to assign a clear semantic meaning to each visual word. In other words, a visual word often has multiple and ambiguous semantic meaning [13].

Although incorporating Must-Links is beneficial to improve topic coherence, it is necessary to carefully consider the polysemy phenomenon of visual words when building Must-Links. If some undesirable Must-Links are generated, the undesirable supervision information will deteriorate the improvement of topic coherence.

In fact, the weakness of previous methods stems from the definition of Must-Link. Therefore, in this paper we re-define Must-Link in a better way—*if two words have a Must-Link, the probability of the two words for* ***one*** *or* ***some*** *topic(s) must be similar (high or low) to each other*. In other words, one Must-Link only constrains ***one*** or ***some*** topic(s) instead of ***all*** topics. In practice, (1) we build and group Must-Links with respect to distinct object classes; (2) we only make Must-Links transitive within groups rather than across groups; and (3) each topic is assigned with a Must-Link group. As a result, the Must-Links within one group are related to the same object class, and hence they need to be transitive and get together to only constrain the assigned topic(s) instead of all topics.

As shown in Fig 2, the Must-Links ('apple', 'banana') and ('apple', 'iphone') are separated into two groups, which constrain the two topics 'fruits' and 'electronics' respectively. Since Must-Links need not to be transitive across multiple groups, the undesirable Must-Link ('iphone, banana') will not be automatically built anymore.

It is worth noticing that the Must-Links in our approach are different from those in previous methods. The Must-Links

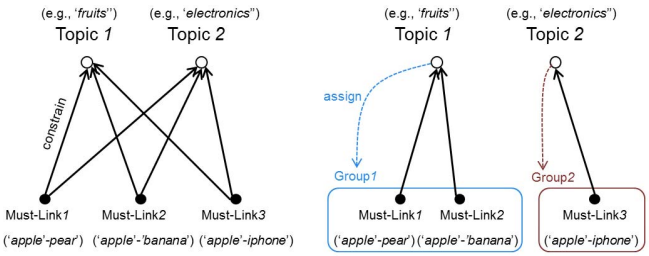


Fig. 2. The comparison between two definitions of Must-Links. The original one (left) indicates each Must-Link constrains all topics. For the proposed definition (right), Must-Links are grouped, and hence each topic is only constrained by a *group* of Must-Links instead of *all* Must-Links. In other words, each Must-Link constrains *one* or *some* topics instead of *all* topics.

in our model are *semantic-specific*, since they are exploited and grouped with respect to distinct object classes (semantics). In contrast, the Must-Links in previous methods are exploited independently of specific object classes (semantics), *i.e.*, it is a kind of *general* or *universal* prior knowledge for all object classes. Obviously, the semantic-specific Must-Links in our approach are more discriminative, which has potential to elevate the performance of object discovery and localization.

On the other hand, in order to build Must-Links with respect to the target object classes, it is necessary to have some images related to the target object classes. However, neither training images nor side information (supervision) for the target object classes are available due to the fully unsupervised setting. Nevertheless, in this paper the Must-Links with respect to the target object classes are built by exploiting rich Web images on the Internet. In particular, some images in the collection are treated as queries and submitted to an image search engine. Then, the retrieved images are usually related to the target object classes, and hence can be utilized to build Must-Links.

For example, an image about *horse* is treated as a query and submitted to Google image search engine, then we can gather lots of Web images which are related to *horse*. By measuring the correlation between visual words from those retrieved images, we can find some frequently co-occurrent visual words, where Must-Links for them can be built to benefit the discovery of the object class *horse*. However, there are many noisy images in those retrieved images, which are not related to the query image. Therefore, this paper proposes a method to efficiently select query images from given image collection and to effectively extract Must-Links from those noisy retrieved images.

Briefly, our approach contains two components: the first is to exploit Must-Link constraints from the Web images which is related to the object classes to be discovered; the second is to incorporate those Must-Link constraints into the process of topic modeling, so that topic coherence could be improved, which is beneficial to object discovery and localization. Our contributions are three-folds:

1) Visual knowledge in terms of Must-Links is mined by exploiting Web images on the Internet. Since the Must-Links are built and grouped with respect to distinct object classes, the visual knowledge is *semantic-specific* in our approach;
2) A new definition of Must-Link (*i.e.*, a Must-Link only constrains *some* topics instead of *all* topics) is proposed so that the polysemy issue of visual words could be effectively alleviated;
3) A novel knowledge-based topic model, called Latent Dirichlet Allocation with Mixture of Dirichlet Trees (LDA-MDT), is proposed to incorporate *multiple* Must-Link groups into topic modeling, which allows that a Must-Link only constrains *some* topics instead of *all* topics. It can improve topic coherence and lead to better object discovery and localization performance. To the best of our knowledge, this is the first work to introduce a knowledge-based topic model to computer vision tasks.

II. Related Work

In this section, we review related works for object discovery and localization, weakly supervised localization, and knowledge-based topic model.

A. Object Discovery and Localization

Unsupervised object discovery methods have been surveyed by Tuytelaars *et al.* [14] in detail. Here, we briefly discuss the related work from a different point of view. Probabilistic models have been extensively exploited for object discovery in the last decade. Sivic *et al.* [3] have proposed a method that builds on PLSA to separate images into four distinct object categories. They later extended their work by replacing PLSA with LDA and using multiple image segmentations, so as to better localize object instances in images [4]. Recently, the method of Clustering-by-Composition is proposed for unsupervised discovery of image categories [15]. In particular, if images can be easily composed using pieces from each other, they are clustered together and represent an image category. Wang *et al.* proposed a method of regularized k-means clustering to leverage context information for object discovery [16]. However, these methods focus on exploiting the internal information of the given image collection, but our approach exploits visual knowledge from rich Web images.

Another group of methods exploit the pairwise affinities between images according to their partial-matching feature correspondences. For example, Grauman and Darrell [17] and Lee and Grauman [18] measure the image similarity by matching their SIFT features, and spectral clustering is adopted for final classification. Similarly, link analysis techniques such as PageRank are employed to model the images similarity network in [13] and [19]. However, the features can only be effectively matched when the two object have similar poses. For example, it is hard to match a bicycle in 'frontal' pose with a bicycle in 'side' pose. However, our approach is based on the BoW representation, so it can properly handle the deformation of objects.

B. Knowledge-Based Topic Model

To improve topic coherence, many methods have been proposed to incorporate prior knowledge into the process of

topic modeling [9]. The prior knowledge for a topic can be expressed with *seed* words, *i.e.*, to assign some seed words to a topic to encourage the topic containing those seed words with high probability [8], [11], [20]. On the other hand, prior knowledge for topics is usually expressed in terms of Must-Links or Cannot-Links. Existing works such as [10], [11], and [21], only considered the Must-Link type of knowledge while [10] also used the Cannot-Link type of knowledge. Generally, Cannot-Links can also be converted and expressed as Must-Links [10]. Thus, this paper only expresses all prior knowledge in terms of Must-Links. Moreover, to the best of our knowledge, none of knowledge-based topic models has been used in computer vision tasks in the past.

C. Weakly-Supervised Localization

Recently, some works focus on the task of object localization by assuming either the object classes are known (i.e., weakly supervised localization problem) or there is only one unknown object class (i.e., co-localization problem).

For weakly supervised localization problem, each image is annotated with image-level labels that indicate whether a target object class appears in the image or not [1], [22]–[24]. Given those image-level annotations, the task is to figure out where the object instance is in each image. These labels enable to learn more discriminative localization models, *e.g.*, by mining negative images [22]. Recent work on discriminative patch discovery [23], [24] learns mid-level visual representations in a weakly-supervised manner, and use them for discovery. Wang *et al.* proposed a latent category learning to exploit the latent information contained in the image backgrounds to facilitate weakly-supervised object localization [25]. Recently, a more general and challenging problem is addressed by Cho *et al.* [26], where the image collection may contain multiple object classes and the image-level annotations are also unavailable.

For co-localization problem, it is often assumed that the image collection only contains a single dominant object class while its label is unknown [2]. The task is to localize the common objects in each image with a bounding box or a pixel-wise segmentation. Furthermore, this problem is further extended that some noisy images which do not contain the dominant object class are allowed in the image collection [27]–[29]. So that it can be used to handle noisy Web images. Tang *et al.* [28] use the discriminative clustering framework of [2] to localize common objects in a set of noisy images, and Joulin *et al.* [29] extend it to the co-localization of video frames.

However, those methods only focus on object localization, *i.e.*, for each image it can only output an object bounding box but the object class is unknown. In other words, it cannot simultaneously conduct object discovery and localization.

III. Our Approach

To discover and localize objects simultaneously, our approach follows the framework of [4]: multiple segmentations of each image are first obtained by using Normalized Cuts algorithm with varied parameters. And then, each segment is represented as a bag-of-words. Then, a topic model is used to discover the latent topics. At last, the inferred topics are associated to the target object classes, meanwhile the extent of each segment indicates the object location, as shown in Fig 3.

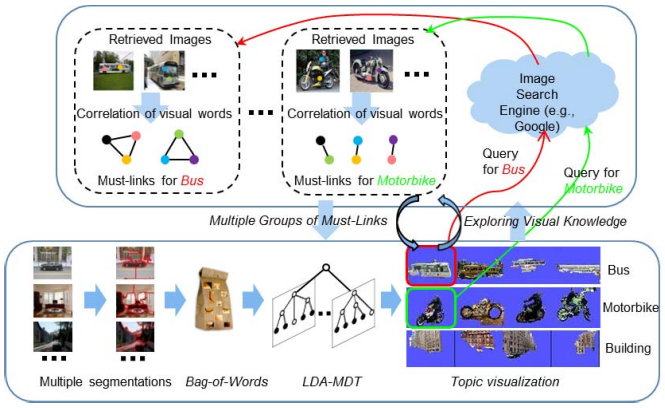


Fig. 3. The Framework of object discovery and localization. First, the visual knowledge in terms of Must-Links is mined by exploiting Web images on the Internet. Secondly, for a given image collection, multiple segmentations of each image are obtained by using Normalized Cuts algorithm with varied parameters. And then, they are represented by using the BoW representations. Thirdly, the proposed LDA-MDT is used to infer latent topics. At last, the inferred topics are visualized and they correspond to the target object classes (*i.e.*, object discovery). Meanwhile, the extent of each segment indicates the object location (*i.e.*, object localization).

However, different from [4], the visual knowledge in terms of Must-Links is extracted from Web images in our approach; moreover, the LDA is replaced with our novel LDA-MDT to incorporate multiple groups of Must-Links into topic modeling. As a result, topic coherence is significantly improved and hence we can better discover and localize objects.

A. Exploiting Visual Knowledge

Although no supervision information is available for the object classes to be discovered, we can still mine visual knowledge from rich Web images on the Internet. Intuitively, if we have an exemplar image associated to a target object class, we can easily gather more images from an image search engine by taking the exemplar image as a query. After that, if two visual words frequently co-occur across those retrieved images, they are likely related to each other and associated to the target object class. Thus, a Must-Link between them could be established, which is beneficial to discover the target object class.

In this paper, we build Must-Links between visual words according to their correlations across retrieved images, *i.e.*, the Pearson's product-moment coefficient [30] is used to measure the correlations between visual words. In particular, those associated images are firstly represented with the BoW representations. Thus we have a count matrix M, where each entry $M(v, i)$ indicates the number of the visual words v present in the image i. Then, the correlation between every pair of visual words is measured according to Eq.1:

$$corr(v_1, v_2) = \frac{C(v_1, v_2)}{\sqrt{C(v_1.v_1)C(v_2, v_2)}} \quad (1)$$

Input: the image collection $\mathcal{I} = \{I_n\}_{n=1}^{N}$

- Initialization stage:
  - For each I_n, multiple segmentations are obtained by using Normalized Cuts algorithm with varied parameters. Thus, we get a collection of image segments $\mathcal{S} = \{S_d\}_{d=1}^{D}$
  - Each segment S_d is treated as a document by using the BoW representation.
  - The original **LDA** model is used to model $\mathcal{S}$. Thus, we get T inferred topics $\{\phi_t^0\}_{t=1}^{T}$
  - For each topic ϕ_t, it is visualized as a sorted list of image segments. The top-1 image for each topic is treated as the query q_t. Thus, we have a query set $\mathcal{Q}^0 = \{q_t\}_{t=1}^{T}$.
- Loop for $l = 1, ..., L$:
  - For each query in the $\mathcal{Q}^{(l-1)}$, some retrieved images are gathered from the Google image search engine, and hence a group of Must-Links is built.
  - With the multiple groups of Must-Links, the proposed **LDA-MDT** model is used to model $\mathcal{S}$ according to Eq.5 and Eq.6. Thus, we get T updated topics $\{\phi_t^l\}_{t=1}^{T}$
  - For each topic ϕ_t^l, it is visualized as a sorted list of image segments. The 'top-ranked and never used' image in the list is used to update the query set $\mathcal{Q}^l$.

Output: the image segments of a same topic are regarded as belonging to the corresponded object class (*i.e.*, object discovery). Meanwhile the extent of each segment indicates the object location (*i.e.*, object localization).

Fig. 4. The iterative process of object discovery and localization.

where $C(v_1, v_2)$ indicates the covariance between visual words v_1 and v_2 for the M. At last, all pairs of visual words are ranked in a descending order, and the top ranked pairs of visual words form a set of Must-Links.

However, since the target object classes themselves are unknown and need to be discovered, it is hard to directly find exemplar images (*i.e.*, query images) associated to them. In this paper, we address this *chicken-and-egg* problem in an iterative manner, as shown in Fig.3 and 4. At the initialization stage, original LDA is employed for topic inference. Although the inferred topics are somewhat incoherent (*i.e.*, some images in the list may be incoherent with the object classes), for each topic we can take the Top-1 image in its list as the initial exemplar since it is more related to the object class than others. After that we can gather some Web images and then build some Must-Links.

In the following iterations, on the one hand the LDA model is replaced with the proposed LDA-MDT to incorporate the built Must-Links into topic inference, on the other hand the 'top ranked but never used' images are treated as new queries to gather more Web images. As a result, the set of Web images as well as the set of Must-Links are updated, which leads to improved topic coherence.

In addition, from the procedure of building Must-Links, we can see that Must-Links are built and grouped with respect to different object classes. As shown in Fig. 3, for object classes 'bus' and 'motorbike', we can build two independent Must-Link groups. Moreover, the proposed LDA-MDT can make Must-Links only be transitive within one group instead of across all groups. As a result, our approach can properly solve the polysemy issue of visual words, since Must-Links within a group are associated to one object class and they should be transitive, whereas Must-Links from different groups are associated to different object classes and need not to be transitive.

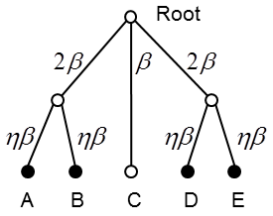


Fig. 5. The Dirichlet tree for words $\{A, B, C, D, E\}$ and corresponding Must-Links $\{(A, B), (D, E)\}$.

B. Latent Dirichlet Allocation With Mixture of Dirichlet Trees (LDA-MDT)

1) Dirichlet Tree: Similar to LDA-DF, if all of the Must-Links satisfy the transitive property, a Dirichlet tree can easily encode these Must-Links as follows: the transitive closures of Must-Links are computed, and each transitive closure is a subtree, with one internal node and the words in the closure as its leaves. A transitive closure is a subset of nodes where there is a Must-link for every pair of nodes in the subset. The weights from the internal node to its leaves are $\eta\beta$. Then, the root connects to those internal nodes s with weight $|L(s)|\beta$, where $|L(s)|$ represents the number of leaves in the subtree under s. In addition, the root directly connects to other words not in any closure, with weight β. The parameter η is the 'strength' parameter which indicates how strong the Must-Links should be satisfied. For example, for Must-Links $\{(A, B), (D, E)\}$ over words $\{A, B, C, D, E\}$, the corresponding Dirichlet Tree is built as shown in Fig. 5.

2) Mixture of Dirichlet Trees: Obviously, the previous Dirichlet tree can only be used to encode the transitive Must-Links. In other words, it can only be used to encode a Must-Link group. So, it is necessary to propose a new topic model to model multiple Must-Link groups, where the Must-Links are transitive within groups rather than across multiple groups.

To this end, we proposed to use mixture of Dirichlet trees to encode multiple Must-Link groups. Briefly, each Must-Link group is encoded with a Dirichlet tree, then those Dirichlet trees are mixed to model all Must-Link groups. Specifically, if there are K groups of Must-Links $ML^k, k = 1, 2, \ldots, K$, the codebook of all words V is separated into two subsets: one set V_{ML} contains all words associated to any Must-Links; another $V_{nML} = V - V_{ML}$ contains all words not associated to any Must-Links.

The mixture of Dirichlet trees is constructed as follows: firstly each Dirichlet tree $DT^k, k = 1, 2, \ldots, K$ over V_{ML} is constructed according to each Must-Link group ML^k. Secondly, the K Dirichlet trees are mixed according to a multinomial distribution $p(q_t) \sim \text{Multi}(\boldsymbol{\gamma})$, and connected to the root node. At last, another flat Dirichlet tree over V_{nML} is directly connected to the root node.

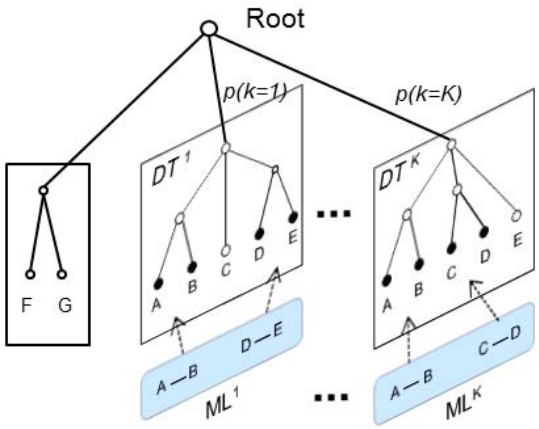


Fig. 6. An illustration to construct the mixture of Dirichlet trees. For a codebook $V = \{A, B, C, D, E, F, G\}$, there are K Must-Link groups where $ML^1 = \{(A, B), (D, E)\}$ and $ML^K = \{(A, B), (C, D)\}$. Thus, we have $V_{ML} = \{A, B, C, D, E\}$ and $V_{nML} = \{F, G\}$. So, K Dirichlet trees $DT^k, k = 1, \ldots, K$ are constructed and mixed according to probability distribution $\{p(k), k = 1, \ldots, K\}$.

Fig. 6 is an example for the construction of a mixture of Dirichlet trees. Over codebook $V = \{A, B, C, D, E, F, G\}$, if there are K Must-Link groups, *e.g.*, $ML^1 = \{(A, B), (D, E)\}$ and $ML^K = \{(A, B), (C, D)\}$. We have $V_{ML} = \{A, B, C, D, E\}$ and $V_{nML} = \{F, G\}$. As described above, K Dirichlet trees $DT^k, k = 1, \ldots, K$ are constructed and mixed according to probability distribution $\{p(k), k = 1, \ldots, K\}$.

3) The LDA-MDT Model: In this paper, a new topic model, called Latent Dirichlet Allocation with Mixture of Dirichlet Trees (LDA-MDT) is proposed to incorporate multiple Must-Link groups into topic modeling. Briefly, the Dirichlet prior over visual words in LDA is replaced with the Mixture of Dirichlet trees in our LDA-MDT. Specifically, K Dirichlet trees $DT^k, k = 1, 2, \ldots, K$ are firstly built for K groups of Must-Links, where there is one-to-one mapping between DT^k and ML^k. Secondly, for each topic t, a tree q_t is selected from K trees and used as the prior to generate the topic-specific distribution ϕ_t. Thirdly, the visual words in each image is generated as similar as the LDA model.

Given K groups of Must-Links $ML^k, k = 1, 2, \ldots, K$, the proposed LDA-MDT assumes that D images with visual words $\{w_d\}_{d \in D}$ are generated by the following process:

1) For each Must-Link group $ML^k, k = 1, 2, \ldots, K$:
   a) Build a Dirichlet tree DT^k, whose corresponding distribution is $DirichletTree(\beta, \eta(k), \kappa(k))$
2) For each topic $t \in T$:
   a) Draw its corresponding Dirichlet tree q_t over K trees DT^k according to $p(q_t = k) \sim \text{Multi}(\boldsymbol{\gamma})$.
   b) According to q_t, draw a topic distribution over visual words $\phi_t \sim DirichletTree(\beta, \eta, \kappa(q_t))$.
3) For each image $d \in D$:
   a) Draw topic proportions $\theta_d \sim \text{Dir}(\alpha)$.
   b) For each visual word $w_{d,n}$:
      i) Select a topic $z_{d,n} \sim \text{Multi}(\theta_d)$.
      ii) Draw a visual word $w_{d,n} \sim \text{Multi}(\phi_{z_{d,n}})$.

From the generative process, it is obvious that each topic is only drawn from one Dirichlet tree, and each Dirichlet tree is constructed for one Must-Link group. So, each topic is only constrained by a Must-Link group rather than all Must-Links. Because of this, our LDA-MDT can more properly handle the polysemy issue of visual words.

In practice, if there are K target object classes in the given image collection, we will build K Must-Link groups and K corresponding Dirichlet trees, where there is a one-to-one mapping between the target object classes and the Dirichlet trees. On the other hand, we often generate $T > K$ topics from the K Dirichlet trees. The relation between a topic t and Dirichlet trees is described with a variable q_t. For example, if we have $q_1 = 1$, $q_2 = 1$, $q_3 = 2$, $\cdots$, $q_T = K$, it indicates that the 1st and 2nd topics are both generated from the 1st Dirichlet tree, the 3rd topic is generated from the 2nd Dirichlet tree, and the T-th topic is generated from the K-th Dirichlet tree.

Thus, the correspondences between discovered topics and object classes can be summarized as follows: (1) some discovered topics may correspond to the same one target object class. This can be regarded as an advantage of our method since we can better find fine-grained object classes. As shown in SectionIV-E, although only one object class 'airplane' is labeled for the dataset, our model could find two fine-grained object classes 'airplane in the sky' and 'airplane at the airport'. (2) some discovered topics may correspond to uninterested object classes. Taking the PASCAL07-6 $\times$ 2 as an example, although there are only 6 target object classes manually labeled (*e.g.*, 'aeroplane', 'bicycle', *etc*), we can find some uninterested object classes such as 'sky', 'tree', *etc*. (3) some discovered topics are too incoherent to correspond to any object class. It is because our method cannot totally eliminate incoherent topics, although it can improve topic coherence.

In this paper, the sampling of topics from Dirichlet trees is modeled with a Multinomial distribution $p(q_t) \sim \text{Multi}(\boldsymbol{\gamma})$. Here, the $\boldsymbol{\gamma}$ is a K-dimensional vector $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_K)$ indicating the importance of those Must-Link groups, *i.e.*, if we wants more topics to satisfy the Must-Links ML^k, the element γ_k will be given a relatively larger value than other elements in $\boldsymbol{\gamma}$. In addition, we have a K-dimensional vector $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_K)$, where each entry η_k indicates the Must-Link strength for the corresponding Must-Link group ML^k.

Each Must-Link group is modeled by a Dirichlet tree $DirichletTree(\beta, \eta, \kappa)$. Each Dirichlet tree DT implicitly defines its edge weight using $\{\beta, \eta\}$ and its tree structure κ. In particular, let $\xi^{(w)}$ be the Dirichlet tree edge weight *leading to word* w. Thus, given $\{\beta, \eta\}$, it is easy to compute the edge weight $\xi(s)$. On the other hand, the tree structure κ can be described with $\{C(.), I, L(.)\}$, where $C(s)$ is the immediate children of node s in the tree, I is the internals nodes, and $L(s)$ is the leaves in the subtree under s. According to the definition of the Dirichlet tree [31], [32], the probability $p(\boldsymbol{w}|\boldsymbol{z}, \boldsymbol{q}_{1:T}, \beta, \eta)$ of a word $w \in L$ is described as similar

to [10],

$$p(\boldsymbol{w}|\boldsymbol{z}, \boldsymbol{q}_{1:T}, \beta, \eta) = \prod_{t=1}^{T}\prod_{s}^{I_t}\left(\frac{\Gamma(\sum_{w}^{C_t(s)}\xi_t^{(w)})}{\Gamma(\sum_{w}^{C_t(s)}(\xi_t^{(w)}+n_t^{(w)}))}\prod_{w}^{C_t(s)}\frac{\Gamma(\xi_t^{(w)}+n_t^{(w)})}{\Gamma(\xi_t^{(w)})}\right) \quad (2)$$

Finally, the complete generative model is

$$p(\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{q}_{1:T}|\alpha, \beta, \boldsymbol{\eta}, \boldsymbol{\gamma}) = p(\boldsymbol{w}|\boldsymbol{z}, \boldsymbol{q}_{1:T}, \beta, \boldsymbol{\eta})p(\boldsymbol{z}|\alpha)p(\boldsymbol{q}_{1:T}|\boldsymbol{\gamma}) \quad (3)$$

It is noticed that the proposed LDA-MDT is the generalized version of LDA-DF. If the mixture of multiple Dirichlet trees is replaced with a Dirichlet tree, the LDA-MDT model degrades to the original LDA-DF model. So, LDA-MDT can be used in more realistic applications or more general scenarios than the LDA-DF.

C. Inference for LDA-MDT

The mixture of Dirichlet trees is conjugated to multinomial distribution, we can efficiently perform inference by Markov Chain Monte Carlo (MCMC) method. In particular, the collapsed Gibbs sampling method [33] is employed for the inference. Different from LDA, the latent topics $\boldsymbol{z}$ as well as the tree indices $\boldsymbol{q}_{1:T}$ need to be inferred in the LDA-MDT. We alternatively sample $\boldsymbol{z}$ and $\boldsymbol{q}_{1:T}$ as follows.

1) Sampling q_t: According to the generative process of the LDA-MDT, one Dirichlet tree is selected for each topic. So, we select T Dirichlet trees $q_t, t = 1, 2, \ldots, T$ corresponding to T topics. Since they are selected independently, we have $p(\boldsymbol{q}_{1:T}|\boldsymbol{\gamma}) = \prod_{t=1}^{T} p(q_t|\boldsymbol{\gamma})$. Moreover, since $p(q_t|\boldsymbol{\gamma}) \sim$ Multi($\boldsymbol{\gamma}$), thus

$$p(\boldsymbol{q}_{1:T}|\boldsymbol{\gamma}) = \prod_{t=1}^{T} p(q_t|\boldsymbol{\gamma}) = \prod_{k=1}^{K}\gamma_k^{n(k)}, \quad (4)$$

where $n(k)$ indicates the number of topics assigned to Dirichlet tree DT^k.

Combining the Eq.3 and Eq.4, the conditional probability for sampling q_t is

$$p(q_t = k|\boldsymbol{z}, \boldsymbol{q}_{1:T}^{-t}, \boldsymbol{w}, \alpha, \beta, \eta, \boldsymbol{\gamma}) = \gamma_k \times \prod_{s}^{I_t}\left(\frac{\Gamma(\sum_{w}^{C_t(s)}\xi_t^{(w)})}{\Gamma(\sum_{w}^{C_t(s)}(\xi_t^{(w)}+n_t^{(w)}))}\prod_{w}^{C_t(s)}\frac{\Gamma(\xi_t^{(w)}+n_t^{(w)})}{\Gamma(\xi_t^{(w)})}\right) \quad (5)$$

2) Sampling z_i: Let $n_{-i,t}^{(d)}$ be the number of word tokens in document d assigned to topic t, excluding the word w_i. Let $n_{-i,t}^{(w)}$ be the number of word tokens that are under node w in the Dirichlet tree of topic t, excluding the word w_i. the conditional probability for sampling z_i is as follow

$$p(z_i = t|\boldsymbol{z}^{-i}, \boldsymbol{q}_{1:T}, \boldsymbol{w}, \alpha, \beta, \eta, \boldsymbol{\gamma}) = \prod_{s}^{I_t(\uparrow i)}\left(\frac{\xi_t^{C_t(s\downarrow i)}+n_{-i,t}^{C_t(s\downarrow i)}}{\sum_{w}^{C_t(s)}(\xi_t^{(w)}+n_{-i,t}^{(w)})}\right), \quad (6)$$

where $I_t(\uparrow i)$ indicates the subset of internal nodes in the Dirichlet tree of topic t that are ancestors of leaf w_i, and $C_t(s \downarrow i)$ indicates the unique node that is s's immediate child and an ancestor of w_i.

According to Eq.5 and Eq.6, the sampling of q_t as well as z_i is alternatively conducted for sufficient iterations. In practice, the procedure of the alternative sampling converges after 200 iterations in our experiments, which is evaluated according to the log-likelihood of Eq.3.

In addition, we can estimate the ϕ with the last sample $\boldsymbol{z}$ as following [34]:

$$\phi_t = \frac{n_{w,t}+\beta}{\sum_t n_{w,t}+W\beta} \quad (7)$$

where $n_{w,t}$ indicates the number of times that topic t is assigned to w.

D. Topic Visualization for Object Discovery

Since each inferred topic is assumed to correspond to an object class, the object discovery can be achieved by assigning all the image segments to the inferred topics, *i.e.*, the image segments assigned to a same topic can be regarded as belonging to the corresponded object class. Meanwhile the extent of each segment indicates the object location (*i.e.*, object localization). From another perspective, this assignment procedure can be regarded as *topic visualization* because after that each topic can be visualized as some assigned image segments.

In our approach, each image segment is described by a distribution over visual words (*i.e.*, normalized histogram $\boldsymbol{w}_i$). Recall that each topic is also described by a distribution over visual words (*i.e.*, ϕ_t of Eq.7). Thus, the similarity between an image segment and a topic can be measured in terms of their Kullback-Leibler (KL) distance as follows:

$$sim(\boldsymbol{w}_i, \phi_t) = 1 - KL(\boldsymbol{w}_i, \phi_t). \quad (8)$$

According to the similarity between image segments and inferred topics, each image segment is assigned to its closest topic. As a result, each topic can be visualized as a sorted list of the assigned image segments. Fig.10 illustrates the top-10 image segments for some inferred topics of the LDA-MDT. From Fig.10, we can clearly see what object class one topic corresponds to. For example, the first topic (the first row) corresponds to the object class 'building', and the last topic (the last row) corresponds to the 'sea'.

It is worth noticing that topic coherence as well as object discovery performance can also be vividly evaluated from the results of topic visualization. If an inferred topic is more coherent, the assigned image segments are not only more closely related to each other but also clearly correspond to a specific object class, which leads to better object discovery performance. Otherwise, the image segments for an incoherent topic are usually unrelated to each other and it is difficult to explain what object class it corresponds to, which results in bad object discovery performance.

IV. Experiments

In this section, three image collections (*i.e.*, Caltech4, LabelMe, and PSACAL07-6 $\times$ 2) are used to evaluate the proposed method. Firstly, we qualitatively compare the inferred topics of original LDA, LDA-DF, and the proposed

LDA-MDT, which illustrates that topic coherence can be significantly improved with our model. Secondly, our approach is quantitatively compared to other object discovery methods including topic model based methods [4] and matching based methods [13], [17], [18]. Thirdly, the performance of object localization for our approached are evaluated and compared to some state-of-the-art object localization methods.

A. Image Datasets

We tested on progressively more difficult datasets. For the Caltech4 set, an image contains a single dominant object (*e.g.*, a face) appearing in flat or cluttered background; For the LabelMe set, an image may contain many dominant object classes, *e.g.*, an image about *city* contains *building*, *road*, *sky*, *etc*; For the PSACAL07-6 $\times$ 2 set, the objects are usually not centered in an image, only cover a small region of an image, and even have complex background clutters.

1) Caltech4 [13]: It contains four categories, *i.e.*, *faces*, *airplane*, *motorbikes*, and *cars rear*. We compare against the methods of [13] and [17] because they share the same goal of discovering object categories in an image collection.

2) LabelMe [35], [36]: This LabelMe dataset has 2688 images and is larger than that in [35]. Some scene elements such as *trees*, *building*, *sea*, *mountain*, *windows*, *road* and *sky* are frequently present in those scene images. Thus, they are treated as dominant object classes to be discovered. Note that this LabelMe dataset is much more challenging than the LabelMe dataset used in [4] since it contains much more images and distinct object classes.

3) PASCAL07-6 $\times$ 2 [1]: It is a subset of the PASCAL VOC 2007 train+val dataset, which has 463 images containing 6 object classes (*aeroplane*, *bicycle*, *boat*, *bus*, *horse*, and *motorbike*) from the left and right viewpoint each. This is a challenging dataset since an object may only cover a small region of an image, and there are lots of background clutters in images.

B. Preprocessing and Representation

Following the approach of [4], we represent images using affine covariant regions, which are described by SIFT [37] descriptors and quantized into visual words. We use the same codebook as [4] with the size of 2224. To produce multiple segmentations, we use the Normalized Cuts algorithm [38], and vary two parameters of the algorithm: the number of segments is set as $K = 3, 5, 7, 9$; and the input image is resized at 2 scales: 100$-$ and 150$-$pixels. After that, each image segment is represented as a document by using the BoW representation [39].

C. Topic Visualization and Coherence

To illustrate that topic coherence can be significantly improved by incorporating prior knowledge into topic modeling, we visualize and compare the topics inferred from the original LDA [4], the LDA-DF [10], and the proposed LDA-MDT models.

Specifically, we take 8 frequently present scene elements in the LabelMe dataset as the target object classes: *building*, *window*, *clear sky*, *cloudy sky*, *trees*, *mountain*, *road*, and *sea*. And then, 8 groups of Must-Links are built by exploiting Web images on the Internet, where each Must-Link group corresponds to one target object class. Next, a LDA-MDT model with 9 mixed Dirichlet trees is constructed as follows: 8 trees are constructed according to the 8 Must-Link groups independently; and the last tree is constructed without any Must-Link among visual words (*i.e.*, a one-layer flat tree), which is used to model the other scene elements (*i.e.*, un-interested scene elements) present in those scene images.

In contrast, for the LDA-DF model all the Must-Links across all groups are mixed together to construct one Dirichlet tree; for the LDA model, all the Must-Links are neglected.

For all the three models, we have similar parameter setting, *i.e.*, the topic number is $T = 20$, the hyper-parameters are $\alpha = 0.5$ and $\beta = 0.5$. For the LDA-MDT, the parameters $\gamma_k = 10(k = 0, 1, 2 \ldots, 8)$ for 9 Dirichlet trees are set as the same to each other, which indicates all the 8 target object classes and the rest one have the same probability to present in the image collection. In addition, the strength parameters $\eta_k = 100(k = 0, 1, 2 \ldots, 8)$ are also set as the same to each other, which indicates all the 8 groups of Must-Links should be strongly satisfied.

As shown in Fig.7, for the object class 'mountain' the 3 corresponding topics inferred from the LDA, LDA-DF, and LDA-MDT are visualized respectively. Each topic is visualized as a sorted list of image segments, where only top-30 image segments are shown due to the limited space.

Topic coherence can be easily checked from the results of topic visualization, *i.e.*, if there are less unrelated image segments in the list, the topic is more coherent. Moreover, if the unrelated image segments are lower ranked in the list, the topic is more coherent. From Fig.7, topic coherence of LDA-MDT is much better than that of LDA and LDA-DF. Particularly, 2 out of the 30 image segments are unrelated to 'mountain' for LDA-MDT (denoted with red boxes in the Fig.7). In contrast, there are 8 image segments unrelated to 'mountain' for LDA.

Compared to the LDA, topic coherence can only be slightly improved by using LDA-DF (there are 5 image segments unrelated to 'mountain' for LDA-DF). This means that it is critical for exploiting and modeling multiple semantic-specific Must-Link groups. We have the similar conclusions for other topics/objects. This experimental results reveal that the performance of object discovery cannot be significantly improved by directly employing those models such as LDA-DF for object discovery. The qualitative results will be shown in SectionIV-D. More visualizations of topics for LDA-MDT are shown in Fig.10 and Fig. 8.

D. Object Class Discovery

Next, we quantitatively evaluate the performance of our approach for object discovery on three datasets respectively (*i.e.*, Caltech4, LabelMe, and PSACAL07-6 $\times$ 2), and compare to other object discovery methods including matching based methods [13], [17], [18].

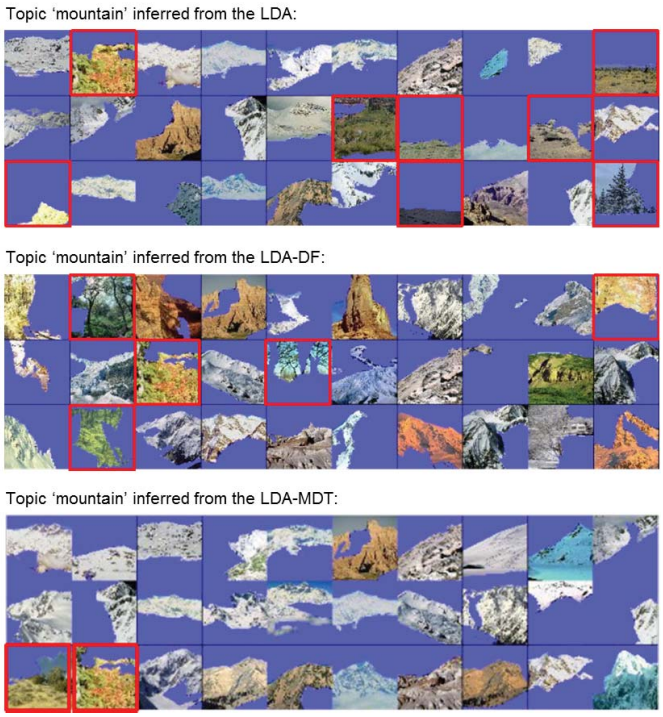


Fig. 7. The visualization of 3 topics about 'mountain' on LabelMe dataset, which are inferred from LDA, LDA-DF, and LDA-MDT respectively (from top to bottom, three rows for each model). Each topic is visualized as a sorted list of image segments (only top-30 image segments are shown here), where the unrelated image segments are denoted with red boxes. In particular, for LDA-MDT, there are 2 out of 30 image segments unrelated to 'mountain' (*i.e.*, the 21th and 22th segments). In contrast, there are 8 image segments unrelated to 'mountain' for LDA (*i.e.*, the 2nd, 10th, 16th, 17th, 19th, 21th, 27th, and 30th segments). Note that both the number and the rank position of the unrelated segments indicate the quality of topic coherence.

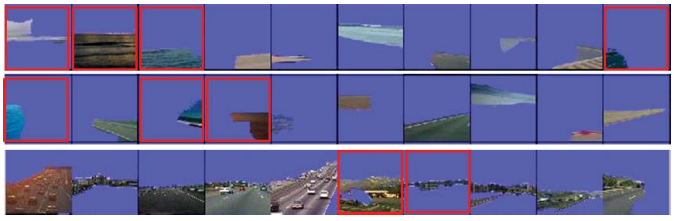


Fig. 8. The visualization of topics 'road' on LabelMe dataset, which are inferred from LDA, LDA-DF, and LDA-MDT respectively (from top to bottom).

1) Caltech4 Dataset: For the Caltech4 dataset, since each image only contains a single dominant object, we use the *Purity* to measure the performance of object classes discovery [17]. Purity measures the extent to which a cluster contains images of a single dominant class [17].

In particular, the topic number is set as 10 for all of LDA, LDA-DF, and LDA-MDT. By visualizing those topics, 4 out of 10 topics are manually selected, which corresponds to the target object classes (*i.e.*, *faces*, *airplane*, *motorbikes*, and *cars rear*). And then, for each image only one image segment is selected and assigned to one of the four topics according to the KL distance. In other words, each image will only belong to one of the four topics, which indicates its object class.

Fig. 9. The visualization of topics 'airplane', 'faces', 'motorbikes', and 'cars' for LDA-MDT (from top to bottom).

TABLE I

THE OBJECT DISCOVERY PERFORMANCE MEASURED BY THE MEAN PURITY ON THE CALTECH4 DATASET

Methods	Performance
Russell *et al.* [4] (LDA)	90.37
Andrzejewski *et al.* [10] (LDA-DF)	92.28
Grauman *et al.* [17]	86.00
Lee *et al.* [18]	88.82
Kim *et al.* [13]	98.55
Ours (LDA-MDT)	98.12

The purity for the three models and some other methods are summarized in Table I. We can see that the performance of object discovery cannot be significantly improved by directly replacing LDA with LDA-DF. However, the proposed LDA-MDT could better leverage the visual knowledge to guide the topic model to discover latent topics.

From Table I, we find that the proposed method outperforms the feature-matching based methods [17], [18] by a large margin, and is comparable with the link-analysis based method of [13]. However, the link-analysis based method suffers from a scalability issue, since it needs to consider all the pairwise image relations for the given image collection. In contrast, our method is scalable to large scale image collections.

2) LabelMe Dataset: The LabelMe dataset is a collection of scene images. 7 scene elements (*i.e.*, *trees*, *building*, *sea*, *mountain*, *windows*, *road*, *sky*) are frequently present in those scene images. Thus, they are treated as dominant object classes to be discovered in this experiment.

Different from Caltech4, each scene image in LabelMe dataset usually contains several scene elements, *e.g.*, a scene image about *city* contains *building*, *road*, *sky*, *etc*. Therefore, each image is associated to multiple object classes, and we cannot use *purity* to measure the performance of object classes discovery. Similar to [4], average precision (AP) is used to evaluate the performance of object discovery for LabelMe. Particularly, each topic is visualized as a ranked list of image segments, where each image is assigned to all topics but with different positions in lists. Obviously, such sorted lists are well suited to computing the AP.

In particular, the topic number is set as 20 for all the LDA, LDA-DF, and LDA-MDT. By visualizing the 20 topics, 7 topics are manually selected which are mostly related to the 7 scene elements. For each selected topic, the sorted list of image segments is reduced to the sorted list of images by only keeping the top sorted segment for each image, *i.e.*, only

TABLE II

THE OBJECT DISCOVERY PERFORMANCE ON LabelMe DATASET INCLUDING OBJECTS 'MOUNTAIN', 'CLOUDY SKY', 'CLEAR SKY', 'WINDOWS/PANE', 'BUILDING', 'TREES', 'SEA', AND 'ROAD'. THE PERFORMANCE IS EVALUATED IN TERMS OF AP

Methods	Mountain	Cloudy Sky	Clear Sky	Window/Pane	Building	Trees	Sea	Road
Russell *et al.* [4] (LDA)	39.78	88.60	93.55	67.30	88.12	80.30	41.68	36.18
LDA + Web	38.12	86.30	92.40	66.50	89.20	78.50	38.80	37.50
Andrzejewski *et al.* [10] (LDA-DF)	41.17	89.79	94.17	69.27	90.12	80.37	42.05	40.39
Ours (LDA-MDT)	48.37	90.73	94.67	73.77	91.01	82.31	57.32	76.45

Fig. 10. The visualization of topics 'building', 'window/pane', 'clear sky', 'cloudy sky', 'trees','dark trees', 'grass', 'mountain', 'road', and 'sea water' for LDA-MDT on LabelMe dataset (from top to bottom).

one image segment is kept for each image. As a result, the AP of the selected topic (*i.e.*, for a scene element) is computed with the sorted list of images and their ground truth.

As shown in Table II and Fig.10, the proposed LDA-MDT can better improve the performance of object discovery against both LDA and LDA-DF. Once again, it is obvious that topic coherence can only be slightly improved by directly replacing LDA with LDA-DF.

More importantly, the proposed LDA-MDT could more effectively discover the objects which are hard to be discovered by LDA such as 'Window', 'Road', *etc*. For example, the AP for 'Window' is improved from 67% to 73% from Table II. Generally, those objects hard to be discovered usually cover a small region of an image. In fact, topic models such as LDA usually fail when documents are too short [7]. Since small objects usually contain little local features, the LDA cannot effectively model the small objects. However, by leveraging the visual knowledge for the target object classes, the knowledge-based topic models could alleviate this weakness of standard topic models.

3) PASCAL07-6 × 2 Dataset: The PASCAL07-6 × 2 dataset is a subset of the PASCAL VOC 2007 train+val dataset, which has 463 images containing 6 object classes (*aeroplane*, *bicycle*, *boat*, *bus*, *horse*, and *motorbike*). This dataset has been extensively used to evaluate the methods of weakly supervised localization [1]. In this paper, we use it to evaluate our approach for fully unsupervised object discovery and localization. This is a challenging dataset since an object may only cover a small region of an image, and there are lots of background clutters in images.

In particular, the topic number is set as 30 for all of LDA, LDA-DF, and LDA-MDT. After topic inference and visualization, we find that there are only 21 topics related to the 6 target object classes, as shown in Fig.12. For the unrelated 9 topics, they are all related to either un-interested object classes (*e.g.*, 'sky', 'grass', *etc*) or too incoherent to be explained.

Due to the large intra-class variation of the target object classes, there are one-to-many mappings between the 6 target object classes and the 21 inferred topics, as shown in Fig.12. As shown in Fig.11, there are two topics related to object class 'horse': one topic is more likely about 'horse in the hurdle racing', and another is about 'horse on the grass'. Taking the object class 'airplane' as another example, there are two related topics: one is about 'airplane at the airport', and another is about 'airplane in the sky'. Thus, we manually associate the 21 inferred topics to the 6 target object classes.

To compute the AP for each object class, the inferred topics associated to a same object class are merged. And the results are shown in Table III. We also compare with some state-of-the-art methods for object discovery [13], [15]. From the Table III, our model outperforms LDA and LDA-DF. Compared to matching based methods, our approach outperforms the PageRank based method of [13] on 5 out of 6 object classes. Even compared with the state-of-the-art method of Clustering-by-Composition [15], we still achieve better performance on 4 out of 6 object classes.

E. Relationship of the Inferred Topics

From Fig.11, we find that there are one-to-many mappings between target object classes and the inferred topics. This is due to that the 6 target object classes are *manually* defined, however, the inferred topics are learned from the image collection in a unsupervised manner. Thus, the inferred topics can subtly discover the *naturally existing* object classes in the image collection. For example, there is subtle difference between the semantics of 'airplane in the sky' and 'airplane at the airport' although both of them are manually annotated as 'airplane'. And our approach can subtly discover those naturally existing and fine-grained object classes. In realistic

TABLE III

THE OBJECT DISCOVERY PERFORMANCE ON PASCAL07 6 × 2 DATASET INCLUDING OBJECTS 'AEROPLANE', 'BICYCLE', 'BOAT', 'BUS', 'HORSE', AND 'MOTORBIKE'. THE PERFORMANCE IS EVALUATED IN TERMS OF AP

Methods	Aeroplane	Bicycle	Boat	Bus	Horse	Motorbike
Russell *et al.* [4] (LDA)	34.60	61.27	39.31	35.74	43.78	47.33
LDA + Web	35.50	57.30	38.44	31.34	45.92	48.76
Andrzejewski *et al.* [10] (LDA-DF)	34.75	62.12	41.07	36.29	44.25	47.73
Kim, *et al.* [13]	45.74	40.33	31.01	35.44	29.10	30.23
Faktor, *et al.* [15]	**47.51**	44.62	45.73	34.51	**50.31**	41.40
Ours (LDA-MDT)	36.71	**62.93**	**46.84**	**45.39**	46.61	**53.59**

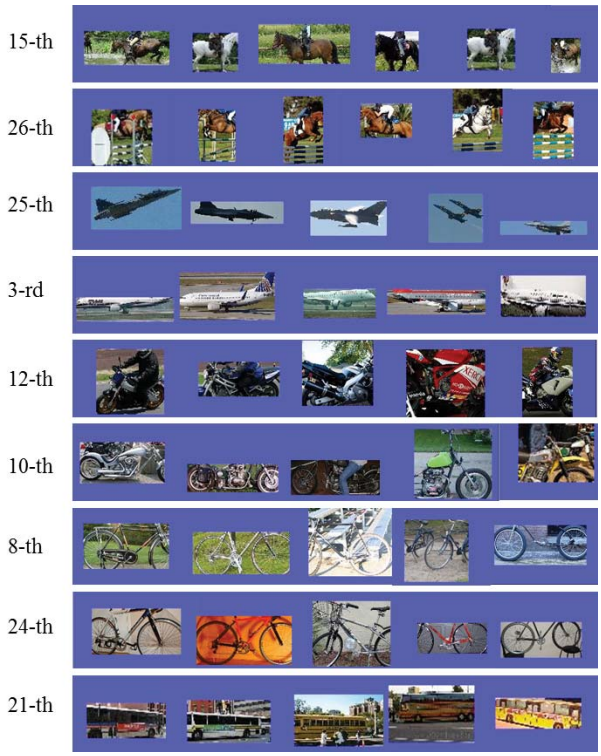


Fig. 11. The visualization of 9 topics for the LDA-MDT including 'horse(on the grass)', 'horse(in the hurdle racing)', 'airplane(in the sky)', 'airplane(at the airport)', 'motorbike', '(light) motorbike', 'bicycle', and 'bus'.

applications, it is more important to subtly discover the fine-grained image classes for the understanding of an unknown image collection. For example, for an on-line shopping recommendation system, it is more helpful to discover 'boots' from a user's image collection rather than just discovering 'shoes' from it.

The advantage of discovering naturally existing image classes for our approach is stemmed from the semantical clustering ability of topic models. And this is the substantial advantage for applying topic models to the task of unsupervised object discovery, especially when compared to those discriminative methods.

Furthermore, the inferred topics are usually not independent to each other, and their relationship can be discovered in our approach. Since each topic is visualized as a list of image segments, we can discover the relationship of two topics by checking the embedding of their image segments. Specifically, according to our LDA-MDT model, image segments are embedded into a T-dimensional space, where each image

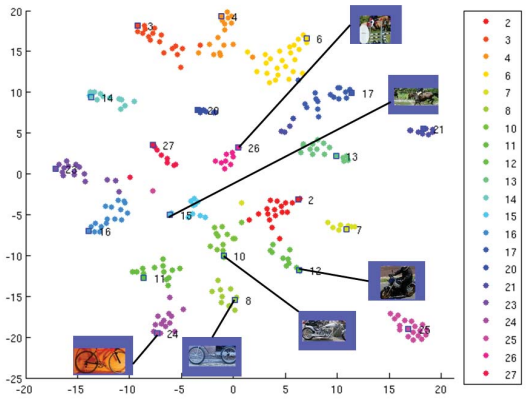


Fig. 12. The t-SNE 2D embedding of image segments for our LDA-MDT. There are 21 inferred topics related to the 6 target object classes, which are shown in different colors and indexes. Best viewed in color.

segment is represented as a T-dimensional vector θ_d. And then the $\theta_d, d = 1, \ldots, D$ are visualized in a 2D space by using t-SNE [40]. As a result, each embedding of a image segment is visualized as a point in the 2D space, where the distance between two point indicates the similarity between the two image segments. After that, we can discover the relationship between the inferred topics by checking the embedding of image segments.

The Fig.12 is the t-SNE embedding of the image segments in the PASCAL07-6 × 2, where the image segments assigned to a same topic are visualized as points in the same color. By checking the embedding of image segments, we have following conclusions: (1) the proposed topic visualization make sense with respect to the embedding of image segments, *i.e.*, the image segments assigned to a same topic (*i.e.*, the points in the same color) are clustered in the embedding space; and more importantly, (2) the distance between two clusters indicates the *relationship of the two corresponding topics*. For example, the cluster of topic 'light motorbike' is relatively close to the cluster of topic 'bicycle', which indicates they are somewhat like to each other. In contrast, the cluster of topic 'horse in the hurdle racing' is far from the cluster of topic 'horse on the grass', which indicates they describe two distinct actions.

It is noticed that the manual selection of coherent topics is necessary for our method. Our method can only reduce the number of incoherent topics or make incoherent topics more coherent, but we cannot totally eliminate incoherent topics.

TABLE IV
THE COMPARISON OF OBJECT LOCALIZATION ON PASCAL07 6×2 DATASET INCLUDING OBJECTS 'AEROPLANE', 'BICYCLE', 'BOAT', 'BUS', 'HORSE', AND 'MOTORBIKE'. THE PERFORMANCE IS MEASURED IN TERMS OF CorLoc

Methods	Aeroplane		Bicycle		Boat		Bus		Horse		Motorbike	
	Left	Right	Left	Right	Left	Right	Left	Right	Left	Right	Left	Right
Russell *et al.* [4] (LDA)	40.47	58.97	16.66	23.91	11.90	17.50	14.28	34.78	27.65	31.11	40.54	51.51
LDA + Web	41.67	58.15	15.87	26.08	11.36	18.34	15.93	37.21	30.31	32.28	41.15	52.33
Andrzejewski *et al.* [10] (LDA-DF)	47.61	71.79	25.00	26.08	14.28	22.50	33.33	43.47	38.29	40.00	54.05	63.63
Tang, *et al.* [28]	41.86	51.28	25.00	24.00	11.36	11.63	38.10	56.52	43.75	52.17	51.28	64.71
Cho, *et al.* [26]	62.79	66.67	**54.17**	**56.00**	**18.18**	18.60	42.86	**69.57**	**70.83**	**71.74**	**69.23**	44.12
Ours (LDA-MDT)	**64.28**	**87.17**	35.41	39.13	16.66	**30.00**	**42.86**	60.86	46.80	53.33	59.45	**75.75**

TABLE V
THE OBJECT LOCALIZATION PERFORMANCE ON LabelMe DATASET INCLUDING 'MOUNTAIN', 'SKY', 'BUILDING', AND 'ROAD'. THE LOCALIZATION SCORE IS AS THE SAME AS [4]

Methods	Mountain	Sky	Building	Road
Russell *et al.* [4] (LDA)	0.31	0.73	0.62	0.21
LDA + Web	0.28	0.72	0.63	0.26
Andrzejewski *et al.* [10]	0.32	0.73	0.64	0.24
Ours (LDA-MDT)	0.55	0.75	0.76	0.35

F. Object Instance Localization

1) LabelMe Dataset: On LabelMe dataset, the accuracy of object localization is evaluated in a similar way as [4]. Specifically, one topic is selected for each object class, and the top twenty returned image segments for the topic are used for the evaluation of localization. Let R as well as GT be the returned image segment and the ground truth region for an object of interest. The ratio of the intersection of R and GT to the union of R and GT (*i.e.*, $\rho = \frac{GT \cap R}{GT \cup R}$) is regarded as the performance score of localization for the returned image segment. The average score of the top twenty returned segments is regarded as the final score for the object of interest.

As shown in TableV, the proposed LDA-MDT can better improve the performance of object localization against both LDA and LDA-DF.

2) PASCAL07-6×2 *Dataset:* On PASCAL07-6×2 dataset, bounding boxes are more preferred to measure the performance of object localization. So, we find the tightest rectangle around the inferred image segment as its bounding box. And the correct localization (CorLoc) metric is used for evaluation on this dataset, which is defined as the percentage of images correctly localized according to the PASCAL criterion: $\rho = \frac{GT \cap R}{GT \cup R} > 0.5$, where R is the predicted box and GT is the ground-truth box.

We run our method on a collection of all class/view images in PASCAL07-6×2, and evaluate its CorLoc performance in Table IV. We compare our method with topic model based methods (*i.e.*, LDA and LDA-DF model), and find out that the localization performance can be significantly improved (about 12% improvement against LDA and 4% improvement against LDA-DF).

Moreover, we also compare with both methods for weakly-supervised localization [1] and methods for co-localization [26], [28]. We find that our method outperforms the method of [1] even though image-level annotations are unknown for our approach. In addition, our method also outperforms the method of [28] about 3%. The state-of-the-art method on PASCAL07-6×2 is proposed by Cho *et al.* [26]. We also can achieve comparable results against this method. It is noticed that those methods only focus on object localization, *i.e.*, for each image it can only output an object bounding box but the object class is unknown. For example, the method of [26] can only discover the underlying 'topology' (nearest-neighbor structure) of image collections instead of discovering object classes. However, our approach can conduct object class discovery and object instance localization simultaneously.

G. Implementation Details of Our Approach

To build Must-Links, we exploit Web images on the Internet. Specifically, for each topic (*i.e.*, an object class), we collect top-$3,000$ retrieved images from the Google image search engine. After computing and sorting the correlation of visual words based on those retrieved images, the top-500 pairs of visual words are selected to form the Must-Link group for this topic. In practice, the performance of our approach is not very sensitive to those parameters. For the number of retrieved images, we find that the top-$3,000$ retrieved images are sufficient for evaluating the correlation of visual words. We try to change the number of retrieved images from $1,000$ to $6,000$ with step size of 500, the results reveal that collecting more retrieved images will not consistently improve the performance, which is likely due to that the low-ranked images are usually unrelated to the target object class.

In addition, we also evaluate the performance of object discovery with respect to the number of selected Must-Links N. On both the LabelMe and PASCAL datasets by increasing the N from 50 to $3,000$ with step size of 50, we find that the AP increases radially at the beginning ($N < 500$), becomes stable from 500 to $2,000$, and decreases slightly after $N > 2000$. To make a trade-off between the computing efficiency and performance, we select $N = 500$ in this paper.

As aforementioned, we exploit Web images on the Internet in an iterative manner as shown in Fig.4. We find that our approach converges rapidly after $3 \sim 5$ iterations. So, we select $L = 5$ in our algorithm (Fig.4).

H. Discussion of the LDA-MDT Model

The proposed LDA-MDT is a generalized version of LDA-DF from two perspectives. The first is as discussed in this paper, *i.e.*, we group Must-Links with respect to their

semantics, and use the LDA-MDT to model those semantic-specific Must-Links.

The second is using the proposed LDA-MDT in another way, which can alleviate another weakness of LDA-DF. For the LDA-DF [10], there is only one (*i.e.*, global) strength parameter η which is used to control the 'strength' of the Must-Links, *i.e.*, it indicates how strong the Must-Links should be satisfied. However, all the Must-Links are consistently controlled by the only one parameter in LDA-DF, and we cannot assign different Must-Links with different strength.

For some applications or scenarios, if we want some Must-Links to be strongly satisfied while some other Must-Links to be weakly satisfied, for example, different Must-Links are built with different confidence, we can group Must-Links with respect to their confidence. And the groups of Must-Links with high confidence are assigned with larger η_k, whereas the groups of Must-Links with low confidence are assigned with smaller η_k. In this way, the proposed LDA-MDT can properly model Must-Links with different strength. This is second way to generalize the LDA-DF as LDA-MDT.

V. Conclusion

In this paper, semantic-specific visual knowledge is mined by exploiting Web images; the Must-Link is re-defined so that the polysemy issue of visual words can be better handled; and a novel LDA-MDT is proposed to incorporate multiple groups of Must-Links into topic modeling. As a result, the visual knowledge is efficiently exploited and incorporated into topic modeling, and hence topic coherence is significantly improved to facilitate object discovery and localization. Our extensive experiments on the Caltech, LabelMe and PASCAL datasets demonstrated the advantages of the proposed model for improving topic coherence. It is shown that our method significantly outperforms the unsupervised methods for object discovery and localization. Even compared to some weakly supervised localization methods, our approach achieves comparable results without using any supervision. Moreover, we expect that the findings we had in this paper in LDA-MDT are general and may also be applied to other tasks beyond computer vision.

References

[1] T. Deselaers, B. Alexe, and V. Ferrari, "Weakly supervised localization and learning with generic knowledge," *Int. J. Comput. Vis.*, vol. 100, no. 3, pp. 275–293, Dec. 2012.

[2] A. Joulin, F. Bach, and J. Ponce, "Discriminative clustering for image co-segmentation," in *Proc. CVPR*, Jun. 2010, pp. 1943–1950.

[3] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman, "Discovering objects and their location in images," in *Proc. ICCV*, Oct. 2005, pp. 370–377.

[4] 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 *Proc. CVPR*, Jun. 2006, pp. 1605–1614.

[5] T. Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," *Mach. Learn.*, vol. 42, no. 1, pp. 177–196, Jan. 2001.

[6] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.

[7] J. Chang, J. L. Boyd-Graber, W. Chong, S. Gerrish, and D. M. Blei, "Reading tea leaves: How humans interpret topic models," in *Proc. NIPS*, 2009, pp. 288–296.

[8] D. Mimno, H. M. Wallach, E. Talley, M. Leenders, and A. McCallum, "Optimizing semantic coherence in topic models," in *Proc. EMNLP*, 2011, pp. 262–272.

[9] Z. Chen, A. Mukherjee, B. Liu, M. Hsu, M. Castellanos, and R. Ghosh, "Leveraging multi-domain prior knowledge in topic models," in *Proc. IJCAI*, 2013, pp. 2071–2077.

[10] D. Andrzejewski, X. Zhu, and M. Craven, "Incorporating domain knowledge into topic modeling via Dirichlet forest priors," in *Proc. ICML*, 2009, pp. 25–32.

[11] A. Mukherjee and B. Liu, "Aspect extraction through semi-supervised modeling," in *Proc. ACL*, 2012, pp. 339–348.

[12] J. Yuan, G. Zhao, Y. Fu, Z. Li, A. K. Katsaggelos, and Y. Wu, "Discovering thematic objects in image collections and videos," *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 2207–2219, Apr. 2012.

[13] G. Kim, C. Faloutsos, and M. Hebert, "Unsupervised modeling of object categories using link analysis techniques," in *Proc. CVPR*, Jun. 2008, pp. 1–8.

[14] T. Tuytelaars, C. H. Lampert, M. B. Blaschko, and W. Buntine, "Unsupervised object discovery: A comparison," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 284–302, 2010.

[15] A. Faktor and M. Irani, "'Clustering by composition'—Unsupervised discovery of image categories," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 6, pp. 1092–1106, Jun. 2014.

[16] H. Wang, J. Yuan, and Y. Wu, "Context-aware discovery of visual co-occurrence patterns," *IEEE Trans. Image Process.*, vol. 23, no. 4, pp. 1805–1819, Apr. 2014.

[17] K. Grauman and T. Darrell, "Unsupervised learning of categories from sets of partially matching image features," in *Proc. CVPR*, Jun. 2006, pp. 19–25.

[18] Y. J. Lee and K. Grauman, "Foreground focus: Finding meaningful features in unlabeled images," in *Proc. BMVC*, 2008, pp. 1–10.

[19] G. Kim and A. Torralba, "Unsupervised detection of regions of interest using iterative link analysis," in *Proc. NIPS*, 2009, pp. 961–969.

[20] J. Jagarlamudi, H. Daumé, III, and R. Udupa, "Incorporating lexical priors into topic models," in *Proc. EACL*, 2012, pp. 204–213.

[21] Z. Chen, A. Mukherjee, B. Liu, M. Hsu, M. Castellanos, and R. Ghosh, "Discovering coherent topics using general knowledge," in *Proc. CIKM*, 2013, pp. 209–218.

[22] R. G. Cinbis, J. Verbeek, and C. Schmid, "Multi-fold MIL training for weakly supervised object localization," in *Proc. CVPR*, Jun. 2014, pp. 2409–2416.

[23] C. Doersch, A. Gupta, and A. A. Efros, "Context as supervisory signal: Discovering objects with predictable context," in *Proc. ECCV*, 2014, pp. 362–377.

[24] S. Singh, A. Gupta, and A. A. Efros, "Unsupervised discovery of mid-level discriminative patches," in *Proc. ECCV*, 2012, pp. 73–86.

[25] C. Wang, K. Huang, W. Ren, J. Zhang, and S. Maybank, "Large-scale weakly supervised object localization via latent category learning," *IEEE Trans. Image Process.*, vol. 24, no. 4, pp. 1371–1385, Apr. 2015.

[26] M. Cho, S. Kwak, C. Schmid, and J. Ponce, "Unsupervised object discovery and localization in the wild: Part-based matching with bottom-up region proposals," in *Proc. CVPR*, Jun. 2015, pp. 1201–1210.

[27] M. Rubinstein, J. Kopf, C. Liu, and A. Joulin, "Unsupervised joint object discovery and segmentation in Internet images," in *Proc. CVPR*, Jun. 2013, pp. 1939–1946.

[28] K. Tang, A. Joulin, L. Fei-Fei, and L.-J. Li, "Co-localization in real-world images," in *Proc. CVPR*, Jun. 2014, pp. 1464–1471.

[29] A. Joulin, K. Tang, and L. Fei-Fei, "Efficient image and video co-localization with Frank–Wolfe algorithm," in *Proc. ECCV*, 2014, pp. 253–268.

[30] K. Pearson, "Note on regression and inheritance in the case of two parents," in *Proc. Roy. Soc. London*, vol. 58, pp. 240–242, Jan. 1895.

[31] S. Y. Dennis, III, "On the hyper-Dirichlet type 1 and hyper-Liouville distributions," *Commun. Statist. Theory Methods*, vol. 20, no. 12, pp. 4069–4081, 1991.

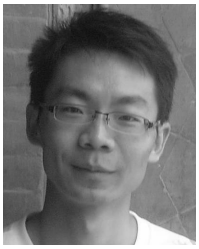
[32] S. Y. Dennis, III, "A Bayesian analysis of tree-structured statistical decision problems," *J. Statist. Planning Inference*, vol. 53, no. 3, pp. 323–344, 1996.

[33] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proc. Nat. Acad. Sci. USA*, vol. 101, no. 1, pp. 5228–5235, 2004.

[34] Z. Niu, G. Hua, X. Gao, and Q. Tian, "Semi-supervised relational topic model for weakly annotated image recognition in social media," in *Proc. CVPR*, Jun. 2014, pp. 4233–4240.

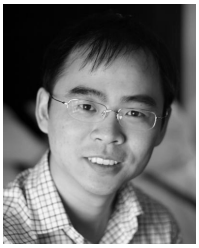
[35] W. Chong, D. Blei, and F.-F. Li, "Simultaneous image classification and annotation," in *Proc. ICCV*, Jun. 2005, pp. 1903–1910.

[36] L. Fei-Fei and P. Perona, "A Bayesian hierarchical model for learning natural scene categories," in *Proc. CVPR*, Jun. 2005, pp. 524–531.
[37] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. ICCV*, Sep. 1999, pp. 1150–1157.
[38] J. Shi and J. Malik, "Normalized cuts and image segmentation," in *Proc. ICCV*, Jun. 1997, pp. 1–7.
[39] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *Proc. ICCV*, Oct. 2003, pp. 1470–1477.
[40] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 2579–2605, 2008.



Zhenxing Niu received the Ph.D. degree in control science and engineering from Xidian University, Xi'an, China, in 2012. From 2013 to 2014, he was a Visiting Scholar with The University of Texas at San Antonio, TX, USA.

He is currently an Associate Professor with the School of Electronic Engineering, Xidian University. His current research interests include computer vision, machine learning, and their application in object discovery and localization. He served as PC member of the IEEE CVPR, the IEEE ICCV, and ACM Multimedia.



Gang Hua was enrolled in the Special Class for the Gifted Young of Xi'an Jiaotong University (XJTU) in 1994 and received the B.S. degree in automatic control engineering and the M.S. degree in control science and engineering from XJTU in 1999 and 2002, respectively, and the Ph.D. degree from the Department of Electrical Engineering and Computer Science, Northwestern University, in 2006. He is currently a Principal Researcher/Research Manager with Microsoft AI & Research Group. He was a Research Staff Member with the IBM Research T. J. Watson Center from 2010 to 2011, a Senior Researcher with the Nokia Research Center, Hollywood, from 2009 to 2010, and a Scientist with the Microsoft Live Labs Research from 2006 to 2009. He also held an academic advisor position with the IBM T. J. Watson Research Center from 2011 to 2014. He was an Associate Professor of Computer Science with the Stevens Institute of Technology. He is currently a Principal Researcher/Research Manager with Microsoft Research.

He has authored over 130 peer reviewed publications in prestigious international journals and conferences. As of March 2017, he holds 19 U.S. patents and has ten U.S. patents pending. He is an IAPR Fellow and a Distinguished Scientist of ACM. He was a recipient of the 2015 IAPR Young Biometrics Investigator Award for his contribution to Unconstrained Face Recognition from Images and Videos, and a recipient of the 2013 Google Research Faculty Award. He is an Associate Editor-in-Chief of the *Computer Vision and Image Understanding*, an Associate Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON CIRCUITS SYSTEM AND VIDEO TECHNOLOGIES, the *IEEE Multimedia*, and the *IAPR Journal on Machine Vision and Applications*. He also served as the Lead Guest Editor on two special issues in IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE and the *International Journal on Computer Vision*, respectively. He is also an Area Chair of the ICCV'2017&2011, CVPR'2017&2015, ICIP'2012&2013&2015, ICASSP'2012&2013, and ACM MM 2011&2012&2015.

Le Wang (M'14) received the B.S. and Ph.D. degrees in control science and engineering from Xi'an Jiaotong University (XJTU), Xi'an, China, in 2008 and 2014, respectively. From 2013 to 2014, he was a visiting Ph.D. student with the Stevens Institute of Technology, Hoboken, USA.

He is currently an Assistant Professor with the Institute of Artificial Intelligence and Robotics, XJTU. His current research interests include computer vision, machine learning, and their application in object discovery and segmentation from web images and videos.

Xinbo Gao (M'02–SM'07) received the B.Eng., M.Sc., and Ph.D. degrees in signal and information processing from Xidian University, Xi'an, China, in 1994, 1997, and 1999, respectively. From 1997 to 1998, he was a Research Fellow with the Department of Computer Science, Shizuoka University, Shizuoka, Japan. From 2000 to 2001, he was a Post-Doctoral Research Fellow with the Department of Information Engineering, The Chinese University of Hong Kong, Hong Kong. Since 2001, he has been with the School of Electronic Engineering, Xidian University. He is currently a Cheung Kong Professor of Ministry of Education, a Professor of Pattern Recognition and Intelligent System, and the Director of the State Key Laboratory of Integrated Services Networks, Xi'an. His current research interests include multimedia analysis, computer vision, pattern recognition, machine learning, and wireless communications.

Prof. Gao is currently a fellow of the Institution of Engineering and Technology. He served as the General Chair/Co-Chair, Program Committee Chair/Co-Chair, or PC Member for around 30 major international conferences. He is on the Editorial Boards of several journals, including *Signal Processing* (Elsevier), and *Neurocomputing* (Elsevier).