symmetry

MDPI

Article

Weakly Supervised Object Co-Localization via Sharing Parts Based on a Joint Bayesian Model

Lu Wu and Quan Liu *

School of Information Engineering, Key Laboratory of Fiber Optic Sensing Technology and Information Processing, Ministry of Education, Wuhan University of Technology, Wuhan 430070, China; wulv@whut.edu.cn

* Correspondence: quanliu@whut.edu.cn; Tel.: +86-158-2741-1128

Received: 26 February 2018; Accepted: 25 April 2018; Published: 4 May 2018

Abstract: Objects in images are characterized by intra-class variation, inter-class diversity, and noisy images. These characteristics pose a challenge to object localization. To address this issue, we present a novel joint Bayesian model for weakly-supervised object localization. The differences compared to previous discriminative methods are evaluated in three aspects: (1) We co-localize the similar object per class through transferring shared parts, which are pooling by modeling object, parts and features within and between-class; (2) Labels are given at class level to provide strong supervision for features and corresponding parts; (3) Noisy images are considered by leveraging a constraint on the detection of shared parts. In addition, our methods are evaluated by extensive experiments. The results indicated outperformance of the state-of-the-art approaches with almost 7% and 1.5% improvements in comparison to the previous methods on PASCAL VOC 2007 6 × 2 and Object Discovery datasets, respectively.

Keywords: a joint Bayesian model; weakly supervised object co-localization; shared parts; noisy images

1. Introduction

Image recognition has received a great deal of interest in recent years. With images variability in illumination, viewpoint, shape, etc., object localization has been a challenging problem in this domain.

Bounding boxes on objects provide a valuable estimation of the regions of interest in an image. With human-annotated bounding boxes, supervised methods are used to train a large number of images to get promising results. However, image annotations are time-consuming and labor-intensive. Particularly, these manual annotations prevent several important applications (such as object detection [1,2], object tracking [3] and human pose estimation [4], etc.) from going large-scale. For instance, if an excellent object detector asks for 1000 bounding boxes for training, learning detectors for 10 k classes would require 10 million bounding boxes. Unsupervised methods use statistics tools to discover objects automatically, which does not require full annotations [5]. Such results are not promising. The state-of-the-art convolutional neural network (CNN) [6] is an effective method to extract discriminative appearance and transfer knowledge to find the object position, but it requires a large volume of sample training and hardware calculation ability.

To overcome these limitations, our work evaluates weakly-supervised approaches, which have been applied to image classification and localization tasks [7–15]. The main advantage of weakly-supervised learning is that it requires less detailed annotations compared to the fully-supervised method. By allowing extra requirements for annotations instead of applying a set of labels given at the image level, weakly-supervised object localization (WSOL) can be effective for various applications [9,10]. Most approaches of WSOL are discriminative methods, which train noisy and clean images in the same way. Hence, ignoring labels could be noisy at the image level. In addition, these methods attempt to localize each class of objects independently from other classes, which leads to a number of limitations for object localization. On the one hand, ignoring objects and sharing some homogeneity can increase ambiguity for each class. On the other hand, although the appearance of objects can vary in different classes, the backgrounds are still all relevant. Iteratively calculating background information will increase the calculation cost.

To solve the above-mentioned problems, we developed a hierarchical Bayesian model, which consists of an object, parts and features. In order to represent visual objects, a collection of spatially constrained parts were defined as latent variables. This modification captures the dependencies regarding feature location and the appearance so that the objects can reuse the same parts in different proportions. Moreover, some parts transfer the same features in different spatial configurations to share among classes. Those object parts that have a few similar characters are detected, and they are considered as noisy images. For a better illustration of this work, we present the co-localization framework in Figure 1.

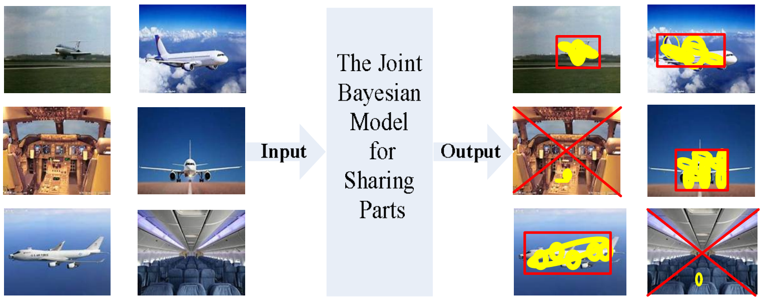


Figure 1. The framework for co-localization task. In this framework, our goal is to localize the airplane within each image. The color yellow represents a salient region. It can be seen that our model can distinguish the right image from noisy images that have a few parts to share.

According to the above discussion, we make three contributions to this work:

(1) We propose a novel framework based on the Bayesian hierarchical topic model for weakly supervised object localization. Without extra requirements of object annotations, latent parts and appearances are given information together at the class level to increase the recognition of an object. In addition, the appearance and correspondence position are modeled jointly to help visualize the object parts.

(2) We show how the joint Bayesian model utilizes the benefits of shared parts to help object co-localization throughout the dataset. Through sharing a common set of features, the same semantic objects can be found simultaneously in each class. Through parts sharing, a few training images can make robust predictions of the objects. Meanwhile, with a small amount of training data and feature sharing, our model can save a great deal of computational resources.

(3) We define a constraint to distinguish between noisy images and clean images. Noisy images can be found by measuring the rate of transferring information of shared parts in each category. Furthermore, to illustrate the effectiveness of our model, we present the experiments performed on two challenging datasets, which represent the difficulties of intra-class variation and inter-class diversity. The results demonstrate that our method is robust in object discovery and localization.

The remaining sections are organized as follows. Related work is presented in Section 2. Section 3 presents the proposed model and parameters learning algorithm. Section 4 experimentally verifies the proposed method. Finally, Section 5 discusses and concludes the work.

2. Related Work

Weakly Supervised Object Localization. WSOL is a task of simultaneously locating objects in the images and learning their appearances, only using the weak labels indicating the presence/absence of the objects of interest. Due to the limitations of fully-annotated objects, WSOL has been attracting increasing attention. Many studies address this problem as multiple instance learning (MIL) [3,4,7,10,13], which alternates between selecting positive instances and learning object detectors. It, therefore, leads to a local minimum problem. Although many efforts have been made to overcome this problem by seeking better initialization models and optimization strategies [15], the localization performance is still limited.

Co-localization is a special WSOL to explore three types of cues existing in object localization tasks: the first cue is object saliency, which describes a region containing an object looking different from the other regions of an image. Alexe et al. [12] initialized object positions using the objectness method, which generates thousands of subwindows as candidate regions for saliency discovering. Most of the discriminative methods [2,5,7,10–15] for saliency regions are following this method. For instance, Pandey et al. [2] introduced deformable part-based models (DPM) by exploiting latent SVMs to enhance the supervision of bounding boxes. The second cue is similarities in intra-class, which describes a region containing an object that looks similar to other regions. These regions represent different images but contain similar objects in the same category [11]. The third cue is the difference in inter-class, which describes regions of interest that looks different to any other regions without the object of interest. To the best of our knowledge, Deselaers et al. [8] is the first one who employed a conditional random field (CRF) algorithm and generic prior knowledge to combine these three cues for WSOL. Tang et al. [11] presented a joint image-box formulation for discovering the similar objects in intra-class and turned the co-localization problem into a convex quadratic program. All these methods are discriminative and tend to return a bounding box to reflect the position of an object.

Recent work modifies CNN architectures for image classification as well as learn to localize objects by convolutional layers [8–10]. Other methods, like transfer learning [16,17] and self-taught learning [18], also provide effective cues for WSOL.

In contrast to these localization studies based on discriminative models, our method does not require a series of candidate bounding boxes. Instead, we take advantage of the generative models which are flexible to utilize object similarities and to cope with class variations and diversity.

Topic models for object localization. Topic models were originally developed for text analysis and have been successfully transferred to image classification and localization tasks [19,20]. Based on bag of words (BoW) model, topic models are used to cluster a set of discriminative features for classes of interest. However, features in topic models have no spatial information, thus, it cannot explicitly model the intra- and inter-class structure of feature distributions.

Recently, improved topic models with detailed structural information [21–24] have been proposed for object localization. Both unsupervised and supervised learning methods can be formulated to topic models. As unsupervised LDA and PLSA models have limited ability for image applications, supervised extension of these two models are more popular to incorporate class label variables. The classLDA (cLDA) [25] and supervised LDA (sLDA) [26] are classic models of this family. Via the weakly supervised class label information, Wang et al. [21] proposed a latent category learning method based on the extent to which PLSA could select the most discriminative category to localize objects in a clutter background. Sudderth et al. [23] developed a rich hierarchical topic model to capture features of the spatial structures with fixed classes. In addition, prior information is another unique advantage of topic models for specified vision tasks. Shi et al. [22] integrated geometry prior to topic models and proposed a novel Bayesian joint model for weakly supervised object localization. Niu et al. [24] proposed a novel knowledge-based topic model, named LDA with a mixture of Dirichlet trees to incorporate the must-links into topic modeling for object discovery. Object co-localization was explored in weakly supervised setups with multiple object categories in topic models [22–24]. Unlike the existing discriminative methods that treat different classes as unrelated entities, ref. [22] shared background similarities in the class. To those bounding boxes that were drawn around objects lack some strong supervision, ref. [23] learned both a hierarchical structure for visual appearance and its correspondence geometry information. We take advantage of [22–24] sharing some common features within and between-class to cope with class variations and diversity. In addition, ref. [10,11] focused on the co-localization problem with optimization and CNN respectively, and [11] also to pay attention to noisy images.

3. Methods

We introduced a novel topic model to localize a common set of features shared among classes. Object, parts, and features were modeled jointly in a hierarchical structure. Although objects are different in viewpoint, scale and size, etc., they are supposed to have some common features in each class. Conjugate priors were chosen for parameters learning. Gibbs sampling was adapted to update the features posterior distribution. Each part encoded a distribution over the features after learning. In addition, a constraint was defined to distinguish between noisy and clean images to co-localize object accurately.

3.1. Symbol Description

Given an image $j \in \{1, 2, ...J\}$ with N_j features, K-means algorithm is used to partition image features into K clusters. We define these clusters as a codebook W of visual words. Thus, the image j is represented as the appearance w_j and correspondence Gaussian position v_j, respectively.

3.2. The Joint Topic Model for Objects

Our topic model is a type of statistical model, specifying a joint probability distribution over observations. Given a set of J training images known with category information, the process of our generative Bayesian model is shown in Table 1.

Table 1. The generative process of our joint model.

```
for each image $j \in 1, ..., J$
  sample a topic distribution $\pi \sim Dir(\alpha)$
  sample a class label $l \sim Multi(O)$
    for each class $k \in 1, ..., K$
         sample a topic $z_i \sim \Pi_j$, $Z_i \in T = \{1, ..., K\}$
             for each observation $i \in 1, ..., N_j$
                  sample a visual word $w_{ji} \sim Multi(\eta_k; z_{ji}, l_j)$
                  sample the correspondence location $v_{ji} \sim N(\mu_{Z_{ji}}, \Lambda_{Z_{ji}})$
             end for
    end for
end for
```

Where Multi, Dir and N indicate Multinomial, Dirichlet and Normal distributions with specified parameters. These priors are chosen because they are conjugates to model parameters and hence, can enable efficient inference. For this joint model, shared parts are formalized as groups of features that are spatially clustered. Each object is a mixture of multinomial topic distributions and reuses shared parts in different proportions. We correlate class label L with appearance variable w, as shown in Figure 2, where parts are formalized as groups of features that are spatially clustered and have predictable appearances. This correlation makes conditional distributions of topics dependent on classes. Furthermore, as a larger set of topics are available per class, the model has greater ability to model inter-class structure.

To capture the intra structure of object, we define K distinct parts, which generate features with different typical appearance w_{ji} and location v_{ji}. The particular part z_{ji} associated with each feature is independently sampled from a category-specific multinomial distribution $Z_{ji} \sim \Pi_{lj}$. When learning model from training data, we assign a conjugate Dirichlet prior $\pi \sim Dir(\alpha)$ to these parts association probabilities, where α is a hyperparameter. It reflects prior knowledge of presence of each object class. Each part (w_{ji}, v_{ji}) is then defined by a multinomial distribution η_k on the discrete set of w_{ji} appearance descriptor, and a Gaussian distribution $N(\mu_k, \Lambda_k)$, described as $w_{ji} \sim Multi(\eta_k)$, $v_{ji} \sim N(\mu_{Z_{ji}}, \Lambda_{Z_{ji}})$. The joint distribution of all variables in the model is therefore:

$$\begin{aligned} & p(\boldsymbol{w}, \boldsymbol{v}, \boldsymbol{z}, \boldsymbol{\pi}, l, \mu, \nu | H, O, \eta, \alpha) \\ & = \prod_K p(\mu, \nu | H) \prod_J p(\boldsymbol{\pi} | \alpha) \prod_J p(l|O) \prod_{N_j} p(\mathbf{z}|\boldsymbol{\pi}) \prod_{N_j} p(\mathbf{w}|\mathbf{z}, l, \eta) \prod_{N_j} p(\mathbf{v}|\mathbf{z}, \mu, \nu) \end{aligned} \tag{1}$$

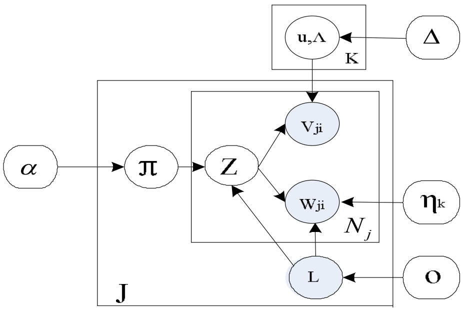


Figure 2. Our joint topic model proposed for describing visual sharing parts. The shade nodes are the observed data and the rounded rectangle are hyperparameters. Let w_j and v_j denote the appearance and two–dimensional position, respectively, with N_j features in image j. Using the visual codebook, the ith feature in image j is described by its discrete appearance w_{ji} and corresponding position v_{ji}.

3.3. Parameters Learning

To learn the parameters defining the parts of the model, we employ a Gibbs sampling algorithm. From the joint model's Markov properties, the posterior distribution over part assignments is as follows:

$$p(z_{ji} \middle| z_{\backslash ji}, \mathbf{w}, \mathbf{v}, l) \propto p(z_{ji} \middle| z_{\backslash ji}, l_j) p(w_{ji} \middle| w_{\backslash ji}, z) p(w_{ji} \middle| w_{\backslash ji}, l_j) p(v_{ji} \middle| v_{\backslash ji}, z) \tag{2}$$

Let N_{lk}^{-i} represents the number of features in classes of object l assigned to part k, and w the appearance descriptor assigned to part k by $z_{\backslash ji}$. C_{kw}^{-i} indicates the number of times of visual word. Using standard expression with Dirichlet priors to predict likelihood function, the first two terms of posterior are written as:

$$p(z_{ji} = k | z_{\backslash ji}, l) = \frac{N_{lk}^{-i} + \alpha/K}{\sum_{k'} N_{lk'}^{-i} + \alpha} \tag{3}$$

$$p(w_{ji} = w | z_{ji} = k, z_{\backslash ji}, w_{\backslash ji}) = \frac{C_{kw}^{-i} + \lambda/W}{\sum_{w'} N_{kw'}^{-i} + \lambda} \tag{4}$$

These probabilities are achieved through the pseudo-counts contributed by Dirichlet priors. To the third item of Equation (2), it represents object l assigned to words w is endowed with its own topic simplex.

$$p(w_{ji} | w_{\backslash ji}, l_j) = \int p(w_{ji} | z_{ji}, l_j) p(z_{ji} | l_j) = \sum_z p(w_{ji} | z_{ji}, l_j) p(z_{ji} | l_j) = \sum_z \eta_{zw}^l \pi_z^l \tag{5}$$

where $\eta_{1:K}^l$ denotes the parameters of mixture components ($\sum \eta_{zw}^l = 1$) and $\pi_{1:K}^l$ the mixing probabilities ($\sum \pi_z^l = 1$). Both of the parameters are defined as Dirichlet distributions:

$$\eta_k^l \sim Dir(C_{k1} + {}^{\lambda}/_{W}, ..., C_{kw} + {}^{\lambda}/_{W}) \tag{6}$$

$$\pi_k^l \sim Dir(N_{l1} + {}^{\alpha}/_{W}, ..., N_{lk} + {}^{\alpha}/_{W}) \tag{7}$$

As to the last item of the position likelihood, it depends on the visual words that are a set of features assigned to the same part by $z_{\backslash ji}$.

$$p(v_{ji} | z_{ji} = k, z_{\backslash ji}, v_{\backslash ji}) = p(v_{ji} | \{v_{j'i'} \Big| z_{j'i'} = k, (j', i') \neq (j, i)\}) \approx N(v_{ji}; \mu_k, \Lambda_k) \tag{8}$$

To the above equation, the mean μ_k and covariance Λ_k are given by regularized moment-matching of the features attributed to the part. Therefore, the Gaussian likelihood can make an accurate approximation for position.

$$\mu_k = \frac{1}{M_k} \sum_{j=1}^{J} \sum_{i|z_{ji}=k}^{N_i} v_{ji} \tag{9}$$

$$\Lambda_k = \delta \left(\Delta + \sum_{j=1}^{J} \sum_{i|z_{ji}=k}^{N_i} (v_{ji} - \mu_k)(v_{ji} - \mu_k)^T \right), \delta = \frac{M_k + 1}{M_k(M_k - d - 1)} \tag{10}$$

where M_k indicates the total number of features of any appearance assigned to part K. d represents features dimension and Δ is an inverse-Wishart hyperparameter. All of these priors are conjugate to the posterior of position distributions. Combine above equations, each of K candidate assignments z_{ji} can be evaluated or a new part of that feature can be sampled.

3.4. Supervision via Class Label Constraint on both of Appearances and Topics

In our weakly supervised topic model, l is used to encode the object supervision from class labels. As the topic simplex is smaller than the word simplex, it is limited to simultaneously model rich intra-class structure and to locate the objects separately. However, our goal is to discover and co-localize images that each contain a common object by exchanging and sharing information within and between-class, which puts more emphasis on class similarity. Therefore, we correlate class label l to topics and visual words together to form strong supervision on object position. Finally, the model can harness the benefits of both topics of supervision, as each topic is learned under the class supervision and topic discovery, as several topics are discovered per class.

3.5. Probabilistic Parts Sharing

To specify the WSOL task, we factor the images into combinations of K shared parts. All the position densities of shared parts depicted in Gaussian mixture models are pre-computed to a dedicated table for fast lookup. Therefore, each shared part corresponds to one object class, which can be learned with both a distribution over the size of appearance, and the spatial location of these appearance within an image. Objects are discovered and localized by means of transferring different proportions of shared parts.

3.6. Object Localization

Object transfers shared parts to compose its own parts in each class. In some special cases, it exploits parts from other classes, depending on which better explains the object. After the learning process, each latent part will be given both a distribution over the sized appearance vocabulary of each feature, as well as over the spatial location of these appearances within each image. In this way, objects are represented as a collection of spatially constrained parts. Then, a bounding box for topic K in image j can be obtained directly from the parts distribution of Gaussian. We do it through aligning a window to the max of four directions depicted by standard deviation ellipses. Final localization is the most standing-out region at the end.

To the noisy images, we use the information entropy to sort parts in the class, and then extend to the whole dataset to formalize global shared parts matrix. Those objects that share few similar characters with the entropy of sharing parts are considered as noisy images. A constraint is defined to distinguish between noisy and clean images as follows:

$$\begin{aligned}
&\sum_{k=1}^{K} I[\max(j_{m\times K}), H(k)] \leq 5 \\
&\text{where,} \\
&I(a,b) = \begin{cases} 1, \text{if } a = b \\ 0, \text{otherwise} \end{cases} \\
&H(k) = -\sum_{k=1}^{K} p(z_k)\ln(p(z_k))
\end{aligned} \tag{11}$$

where $j_{m\times K}$ represents image composed of different proportions of shared parts. m indicates the total number of appearance assigned to parts K in each image. $H(k)$ is the formula of information entropy. If $I \leq 5$, image j is considered as a noisy image.

3.7. Complexity

To the complexity of our model, supposing an image j is composed of K topics with N_j features during training process, a Gibbs sampling algorithm update each feature assignment requiring $O(K \cdot j \cdot N)$ operations. These assignments also account for posterior distributions of topic parameters, which is logically to be inferred.

4. Experiments

We perform experiments on two datasets, the PASCAL VOC 2007 dataset [27] and Object Discovery dataset [28]. Following [29] in weakly supervised localization, we use CorLoc to evaluate performance. CorLoc is defined as the percentage of images correctly localized according to the PASCAL-criterion: $area(B_p \cap B_{gt})/area(B_p \cup B_{gt}) > 0.5$, where B_p is the predicted box and B_{gt} is the groundtruth box. All CorLoc results are given in percentages.

4.1. Experimental Settings

For each category with 21 training images, maximally stable extremal regions (MSER) is considered to represent interest regions as it favors larger, and more homogeneous image regions. As SIFT descriptor is invariant to lighting, pose and scale changes, it is used to characterize the appearance of interest regions. K-means clustering is used to identify a finite dictionary of W appearance patterns, where each feature is mapped to a disjoint set of visual words. We set the size of the dictionary with W = 300. In addition, hyperparameters used in our model are set as follows: The clustering rate $\alpha = 5$; the inverse-Wishart prior Δ = 6; the class information O = 1/number of classes is endowed with a uniform distribution; the multinomial appearance distributions are assigned symmetric Dirichlet priors $\eta_k = Dir(\lambda)$ and $\lambda = 0.1$; the number of shared parts are setting as K = 6 per each class. Lastly, our model updates the parts with Gibbs iterations 200 times.

4.2. PASCAL VOC 2007 6 × 2

PASCAL VOC 2007 6 × 2 is a subset of PASCAL VOC 2007. This subset consists of 6 classes images (airplane, bicycle, boat, bus, horse and motorbike) from the left and right point of view. Each class contains between 21 and 50 images for a total of 478 images, so that 21 images are used for training in each class. We test our method on this subset and compare with previous methods.

Table 2 shows CorLoc results for three terms combinations on PASCAL VOC 2007 6 × 2. We perform our method with heuristic pattern on each operation. Firstly, with no left or right information on each category, 6 classes are adopted to verify experimental performance of the original method, referring to the third row of Table 2. As the average of CorLoc results is not satisfactory, we turn to include the constraint to cope with noisy images. In this particular case, the results are improved nearly 15%. Lastly, we do the full experiment on the whole dataset with 12 classes labels included. The results have been significantly improved with 17% comparing with the second experiment. The reason is that our model has greater ability to model inter-class structure and allows a much larger set of topic distributions possible per class.

Table 2. CorLoc results for three terms combinations on PASCAL VOC 2007 6 × 2. Original method means 6 classes without left and right annotations. Improved method bases on the 1st experiment and adds the constraint to deal with noisy images. Full method means using 12 classes instead of 6 classes in the 2nd settings to complete the experiment.

Method	Airplane		Bicycle		Boat		Bus		Horse		Motorbike		Avg.
	Left	Right	Left	Right	Left	Right	Left	Right	Left	Right	Left	Right	
Our (original)	14.55	—	14.29	—	16.095	—	13.67	—	14.90	—	12.29	—	14.30
Our (improved)	27.935	—	26.54	—	29.89	—	36.44	—	27.67	—	26.06	—	29.09
Our (full)	46.51	43.59	43.75	42.00	45.45	46.51	57.14	56.52	43.75	45.65	43.59	44.12	46.55

In addition to the CorLoc results, we calculate the distance of each class to analyze its similarity. The similarity in intra-class enables model shares some parts for co-localization effectively but the similarity in inter-class cause interference when the appearances of objects look the same in different classes. Figure 3 shows the distribution of different object parts produced by multidimensional scaling. Except for boats, motorbikes and horses, the remaining classes seem to closely match the labels of category similarity. The reason for the deviation may be due to the fact that some learned parts are too similar to distinguish among classes. Figure 3 also indicates that sharing the common features can allow us to introduce prior correlations between parameters of nearby classes in the dataset.

Table 3 shows our results outperforming the previous methods that do not use noisy images in PASCAL VOC 2007 6 × 2.

Table 3. CorLoc result compared with previous methods on PASCAL VOC 2007 6 × 2.

Method	Average
Russell et al. [30]	22
Chum and Zisserman [31]	33
Deselaers et al. [29]	37
Tang et al. [11]	39
Our method	46

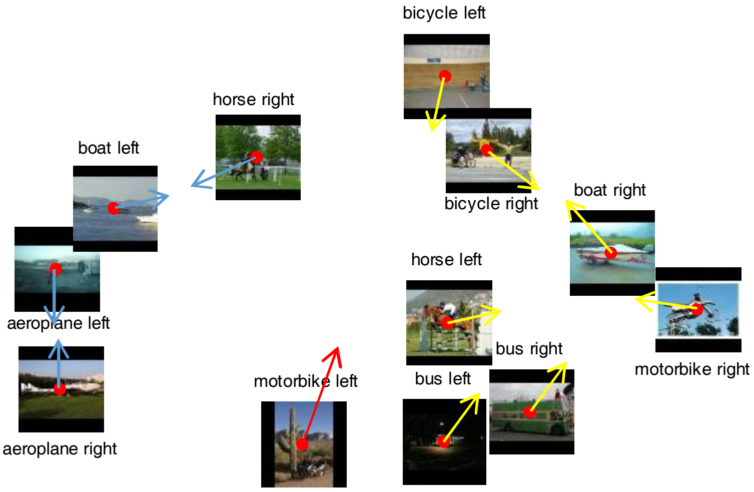


Figure 3. *Cont.*

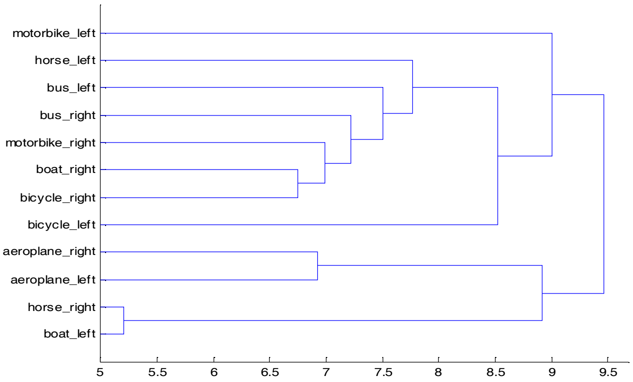
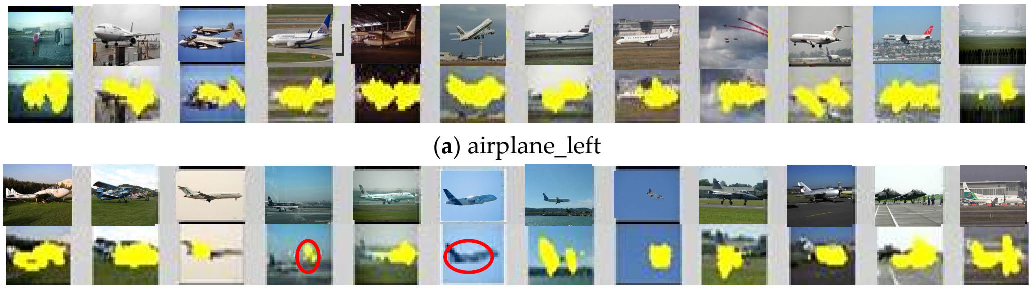


Figure 3. Visualizations of learned parts distribution of our model. Top: Category distance embedding computed by multidimensional scaling, where coordinates for each object category are chosen to approximate pairwise KL distances. The left is clustered with blue arrows, the middle with a red arrow and the right with yellow arrows. Bottom: Category distance dendrogram illustrating a hierarchical clustering, where branch lengths are proportional to inter-category distances. This dendrogram describes a more detailed relationship of classes corresponding to category distance embedding.

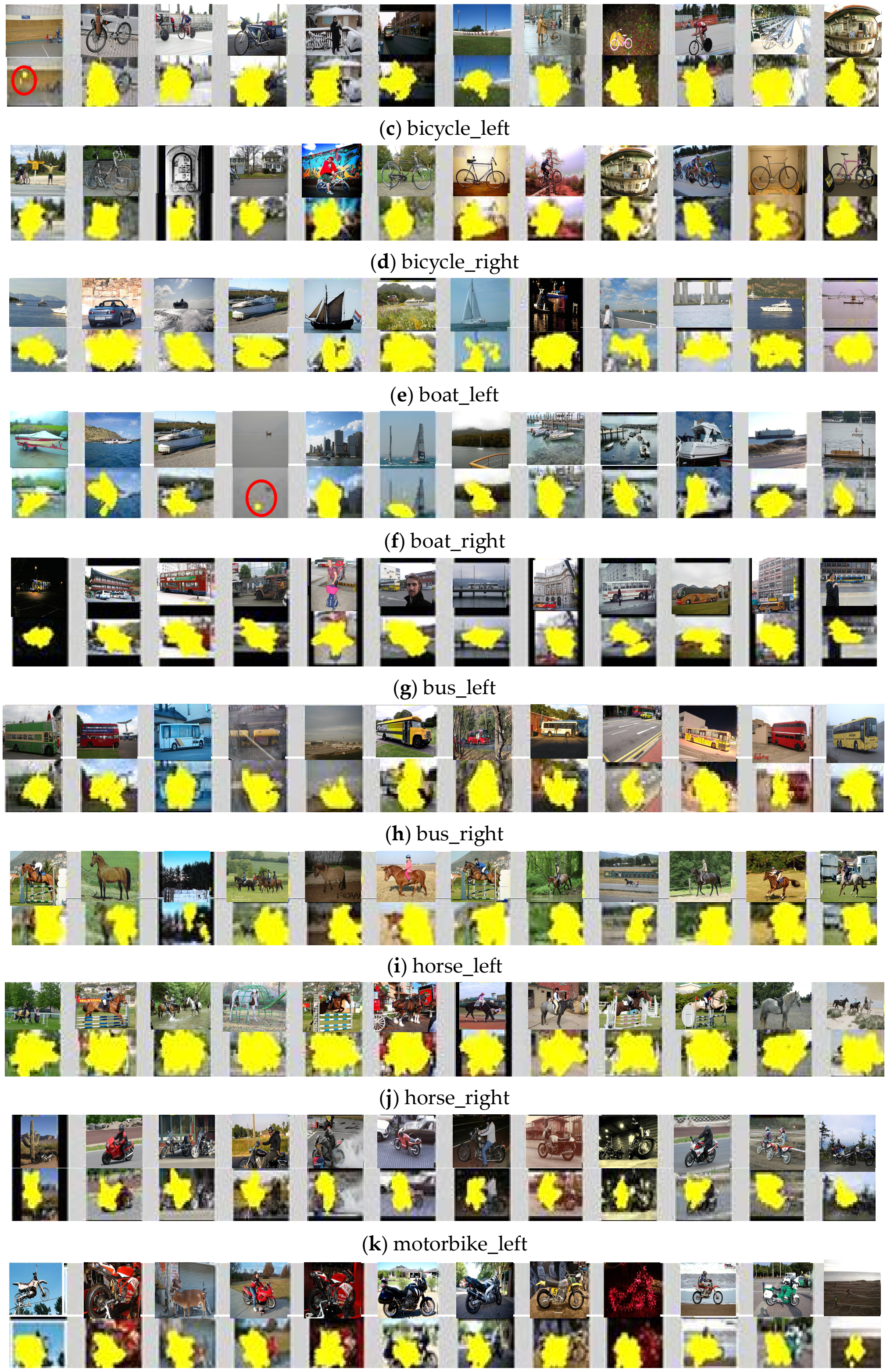
In Figure 4, 12 images are chosen per class to visualize the object co-localization. Our model simultaneously introduces the appearance and position of the image features. The appearances which have the maximum posterior are selected and the correspondence position use ellipses to represent the saliency parts. From Figure 4, we can see the model demonstrates objects localization eliminating clutters in the background and transferring attention to objects themselves. In the bicycle_left class, the bicycle in the first image is not found. It is the same to the fourth image in (b) airplane_right and (f) boat_right in their class. The reason for this phenomenon is that few common features are detected to formalize an effective part of sharing. These unfound objects appear throughout the dataset and thus considered as noisy images. However, whether or not an image will be considered as noisy image depends on the number of matching parts that we have defined in Equation (11). Furthermore, this visualization also demonstrates that our model depends on the most expressive component of the parts to achieve object discrimination.



(**a**) airplane_left

(**b**) airplane_right

Figure 4. *Cont.*



(**c**) bicycle_left

(**d**) bicycle_right

(**e**) boat_left

(**f**) boat_right

(**g**) bus_left

(**h**) bus_right

(**i**) horse_left

(**j**) horse_right

(**k**) motorbike_left

(**l**) motorbike_right

Figure 4. Example of co-localization results for each class. In each category, the upper row represents the original images, which are in contrast to the visualized images below. The yellow region indicates the salient parts, which consist of dozens to hundreds of ellipses in visualizations. Due to sharing a few parts, the red ellipses indicate no objects discovered in (**b**) airplane_right, (**c**) bicycle_left and (**f**) boat_right classes, respectively.

4.3. Object Discovery Dataset

The object discovery dataset [28] is composed of three categories: airplane, car and horse. Through the query of the class name, the images were downloaded automatically by using the Bing search engine. Thus, there were some noisy images in the dataset without accurate query. This dataset was originally introduced for co-segmentation. For a better comparison with the previous methods of co-segmentation and co-localization results, we drew a tight bounding box around the ground truth segmentations. The experimental data settings were the same as the ones used in PASCAL VOC 2007 6 × 2 and 100 image subset was used to be consistent with [11,28].

Table 4 gives the Corloc results on three categories. The Corloc results are comparing with previous methods on the same dataset. Obviously, there were improvements in the airplane, but there were some decreased data in the car. This is because there were more noisy images that were found in the airplane category. Our results also demonstrated the defined constraint is an effective method to find noisy images.

Table 4. Corloc results on the 100 image subset of the Object Discovery dataset.

Method	Airplane	Car	Horse	Average
Kim et al. [32]	21.95	0	16.13	12.69
Joulin et al. [33]	32.93	66.29	54.84	51.35
Joulin et al. [34]	57.32	64.04	52.69	58.02
Rubinstein et al. [28]	74.39	87.64	63.44	75.16
Tang et al. [11]	71.95	93.26	64.52	76.58
Our Method	80.44	86.25	67.82	78.17

In Figure 5, we visualize some examples of co-localization results on the three categories. From these bounding boxes of objects, it can be seen that the objects are discovered and localized simultaneously, while ignoring a wide range of viewpoints, locations and background clutter. Object shared parts can be used to explain the reasons for this visualization.

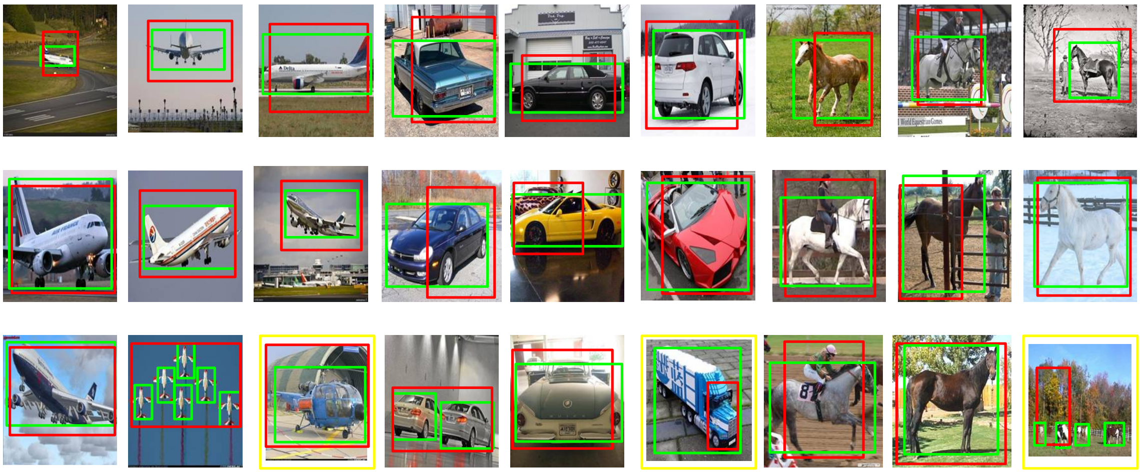


Figure 5. Example of co-localization results on object discovery dataset. Red boxes are our method and green boxes are ground truth localizations. Yellow boxes at the bottom rows represent the wrong localizations.

4.4. Time Complexity

Our experiments are conducted on a 2.5 GHz Intel Dual-core i5 processor. All images are cropped to the size of 128 × 128 implemented with MATLAB tools. Approximately 1.5 s is taken per image to extract SIFT features. With 21 images per class for training, our hierarchical model takes about 27 h and 8 h of PASCAL VOC 2007 6 × 2 and object discovery dataset, respectively. It requires about 6 min per image. However, during the testing process, it takes about 30 ms to localize the position of each object with Gaussian mixture algorithm.

5. Discussion and Conclusions

We describe the weakly supervised object co-localization task based on a new joint Bayesian model. Our work relies heavily on the learning of discriminative shared parts, which are the unique compositions of objects. In addition, our work also relies heavily on the efficient method for dealing with noisy images. To illustrate the effectiveness of our model, extensive experiments were conducted to evaluate the performance of our method. The results showed 7% and 1.5% improvement in the two datasets respectively, in comparison to the state-of-the-art approaches. Three possible explanations are presented for the current findings: (1) A hierarchical structure modeling for object, parts, and features; (2) object parts learned in its class as well as shared between classes; (3) a constraint proposed for determining noisy images. In future research, we would like to extend our work to handle multiple instances of objects on more challenge datasets.

Author Contributions: The authors contributed equally to the design of ideas, analysis of results, and writing of the article.

Acknowledgments: The authors would like to acknowledge funding from the National Science Foundation Committee (NSFC) of China (Grant No. 51675389 and No. 51475347) and the contributions of all collaborators within the projects mentioned.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Dalal, N.; Triggs, B. Histogram of Oriented Gradients for Human Detection. In Proceedings of the IEEE Computer Society Conference on CVPR, San Diego, CA, USA, 20–25 June 2005.
2. Pandey, M.; Lazebnik, S. Scene Recognition and Weakly Supervised Object Localization with Deformable Part-based Models. In Proceedings of the IEEE International Conference on ICCV, Barcelona, Spain, 6–13 November 2011; pp. 1307–1314.
3. Leibe, B.; Schindler, K.; Van Gool, L. Coupled Detection and Trajectory Estimation for Multi-Object Tracking. In Proceedings of the IEEE 11th International Conference on ICCV, Rio de Janeiro, Brazil, 14–21 October 2007.
4. Andriluka, M.; Roth, S.; Schiele, B. Monocular 3D Pose Estimation and Tracking by Detection. In Proceedings of the IEEE Conference on CVPR, San Francisco, CA, USA, 13–18 June 2010.
5. Cho, M.; Kwak, S.; Schmid, C.; Ponce, J. Unsupervised Object Discovery and Localization in the Wild: Part-Based Matching with Bottom-Up Region Proposals. *arXiv*, 2015, arXiv:1501.06170.
6. LeCun, Y.; Bengio, Y.; Hinton, G. Deep Learning. *Nature* **2015**, *521*, 436. [CrossRef] [PubMed]
7. Felzenszwalb, P.F.; Girshick, R.B.; McAllester, D.A.; Ramanan, D. Object Detection with Discriminatively Trained Part-based Models. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *32*, 1627–1645. [CrossRef] [PubMed]
8. Oquab, M.; Bottou, L.; Laptev, I.; Sivic, J. Is Object Localization for Free? Weakly-Supervised Learning with Convolutional Neural Networks. In Proceedings of the IEEE conference on CVPR, Boston, MA, USA, 7–12 June 2015.
9. Zhou, B.; Khosla, A.; Oliva, A.; Torralba, A. Learning Deep Features for Discriminative Localization. *arXiv*, **2016**, arXiv:1512.04150.
10. Zhu, Y.; Zhou, Y.; Ye, Q.; Qiu, Q.; Jiao, J. Soft Proposal Networks for Weakly Supervised Object Localization. *arXiv*, 2017, arXiv:1709.01829.
11. Tang, K.; Joulin, A.; Li, J.; Li, F.F. Co-Localization in Real World Images. In Proceedings of the IEEE Conference on CVPR, Columbus, OH, USA, 23–28 June 2014.
12. Alexe, B.; Deselaers, T.; Ferrari, V. Measuring the Objectness of Image Windows. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 2189–2202. [CrossRef] [PubMed]
13. Nguyen, M.H.; Torresani, L.; Torre, F.; Rother, C. Weakly Supervised Discriminative Localization and Classification: A Joint Learning Process. In Proceedings of the 12th IEEE International Conference on CVPR, Kyoto, Japan, 29 September–2 October 2009.

14. Siva, P.; Russell, C.; Xiang, T.; Agapito, L. Looking Beyond the Image: Unsupervised Learning for Object Saliency and Detection. In Proceedings of the IEEE Conference on CVPR, Portland, OR, USA, 23–28 June 2013.
15. Bilen, H.; Pedersoli, M.; Tuytelaars, T. Weakly Supervised Object Detection with Convex Clustering. In Proceedings of the IEEE Conference on CVPR, Boston, MA, USA, 7–12 June 2015.
16. Shi, M.; Caesar, H.; Ferrari, V. Weakly Supervised Object Localization Using Things and Stuff Transfer. *arXiv*, **2017**, arXiv:1703.08000.
17. Rochan, M.; Wang, Y. Weakly Supervised Localization of Novel Objects Using Appearance Transfer. In Proceedings of the IEEE Conference on CVPR, Boston, MA, USA, 7–12 June 2015.
18. Jie, Z.; Wei, Y.; Jin, X.; Feng, J.; Liu, W. Deep Self-Taught Learning for Weakly Supervised Object Localization. *arXiv*, 2017, arXiv:1704.05188.
19. Rasiwasia, N.; Vasconcelos, N. Latent Dirichlet Allocation Models for Image Classification. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 2665–2679. [CrossRef] [PubMed]
20. Li, L.; Zhang, X.; Zhou, M.; Carin, L. Nested Dictionary Learning For Hierarchical Organization of Imagery and Text. In Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence, Catalina Island, CA, USA, 15–17 August 2012.
21. Wang, C.; Ren, W.; Zhang, J.; Maybank, S. Large-Scale Weakly Supervised Object Localization via Latent Category Learning. *IEEE Trans. Image Process.* **2015**, *24*, 1371–1385. [CrossRef] [PubMed]
22. Shi, Z.; Hospedales, T.M.; Xiang, T. Bayesian Joint Modeling for Object Localisation in Weakly Labeled Images. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1959–1972. [CrossRef] [PubMed]
23. Sudderth, E.; Torralba, A.; Freeman, W.; Willsky, A. Learning Hierarchical Models of Scenes, Objects, and Parts. In Proceedings of the IEEE Computer Society Conference on CVPR, San Diego, CA, USA, 20–25 June 2005.
24. Niu, Z.; Hua, G.; Wang, L.; Gao, X. Knowledge Based Topic Model for Unsupervised Object Discovery and Localization. *IEEE Trans. Image Process.* **2017**, *27*, 50–63. [CrossRef] [PubMed]
25. Li, F.F.; Perona, P. A Bayesian Hierarchical Model for Learning Natural Scene Categories. In Proceedings of the IEEE Computer Society Conference on CVPR, San Diego, CA, USA, 20–25 June 2005.
26. Wang, C.; Blei, D.; Li, F.F. Simultaneous Image Classification and Annotation. In Proceedings of the IEEE Computer Society Conference on CVPR, Miami, FL, USA, 20–25 June 2009.
27. Everingham, M.; Van Gool, L.; Williams, C.K.I.; Winn, J.; Zisserman, A. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. Available online: http://www.pascalnetwork.org/challenges/VOC/voc2007/workshop/index.html,2007 (accessed on 7 June 2007).
28. Rubinstein, M.; Joulin, A.; Kopf, J.; Liu, C. Unsupervised Joint Object Discovery and Segmentation in Internet Images. In Proceedings of the IEEE Conference on CVPR, Portland, OR, USA, 23–28 June 2013.
29. Deselaers, T.; Alexe, B.; Ferrari, V. Weakly Supervised Localization and Learning with Generic Knowledge. *Int. J. Comput. Vis.* **2012**, *100*, 275–293. [CrossRef]
30. Russell, B.C.; Efros, A.A.; Sivic, J.; Freeman, W.T.; Zisserman, A. Using Multiple Segmentations to Discovery Objects and Their Extent in Image Collections. In Proceedings of the IEEE Computer Society Conference on CVPR, New York, NY, USA, 17–22 June 2006.
31. Chum, O.; Zisserman, A. An Exemplar Model for Learning Objective Classes. In Proceedings of the IEEE Conference on CVPR, Minneapolis, MN, USA, 17–22 June 2007.
32. Kim, G.; Xing, E.P.; Li, F.F.; Kanade, T. Distributed Cosegmentation via Submodular Optimization on Anisotropic Diffusion. In Proceedings of the IEEE International Conference on ICCV, Barcelona, Spain, 6–13 November 2011.
33. Joulin, A.; Bach, F.; Ponce, J. Discriminative Clustering for Image Co-Segmentation. In Proceedings of the IEEE Conference on CVPR, San Francisco, CA, USA, 13–18 June 2010.
34. Joulin, A.; Bach, F.; Ponce, J. Multi-Class Cosegmentation. In Proceedings of the IEEE Conference on CVPR, Providence, RI, USA, 16–21 June 2012.