Weakly Supervised Object Localization with Latent Category Learning

Chong Wang, Weiqiang Ren*, Kaiqi Huang, and Tieniu Tan

National Laboratory of Pattern Recognition
Institute of Automation, Chinese Academy of Sciences

Abstract. Localizing objects in cluttered backgrounds is a challenging task in weakly supervised localization. Due to large object variations in cluttered images, objects have large ambiguity with backgrounds. However, backgrounds contain useful latent information, *e.g.,* the sky for aeroplanes. If we can learn this latent information, object-background ambiguity can be reduced to suppress the background. In this paper, we propose the latent category learning (LCL), which is an unsupervised learning problem given only image-level class labels. Firstly, inspired by the latent semantic discovery, we use the typical probabilistic Latent Semantic Analysis (pLSA) to learn the latent categories, which can represent objects, object parts or backgrounds. Secondly, to determine which category contains the target object, we propose a category selection method evaluating each category's discrimination. We evaluate the method on the PASCAL VOC 2007 database and ILSVRC 2013 detection challenge. On VOC 2007, the proposed method yields the annotation accuracy of 48%, which outperforms previous results by 10%. More importantly, we achieve the detection average precision of 30.9%, which improves previous results by 8% and can be competitive with the supervised deformable part model (DPM) 5.0 baseline 33.7%. On ILSVRC 2013 detection, the method yields the precision of 6.0%, which is also competitive with the DPM 5.0.

Keywords: weakly supervised learning, object localization, category learning, latent semantic analysis.

1 Introduction

Weakly supervised localization is challenging in cluttered conditions. Different from the supervised task, the annotation of object location is not given. Though it requires less labeling, it is challenging because of large object variations in cluttered backgrounds. In recent years, many studies in weakly supervised learning have been proposed. They adopt a similar framework, as shown in Fig.1(a). They first use region proposals to extract candidate regions [1, 32], then the object regions (correct localizations) are selected among the candidate regions by

* Chong Wang and Weiqiang Ren contributed equally to this work.

D. Fleet et al. (Eds.): ECCV 2014, Part VI, LNCS 8694, pp. 431–445, 2014.

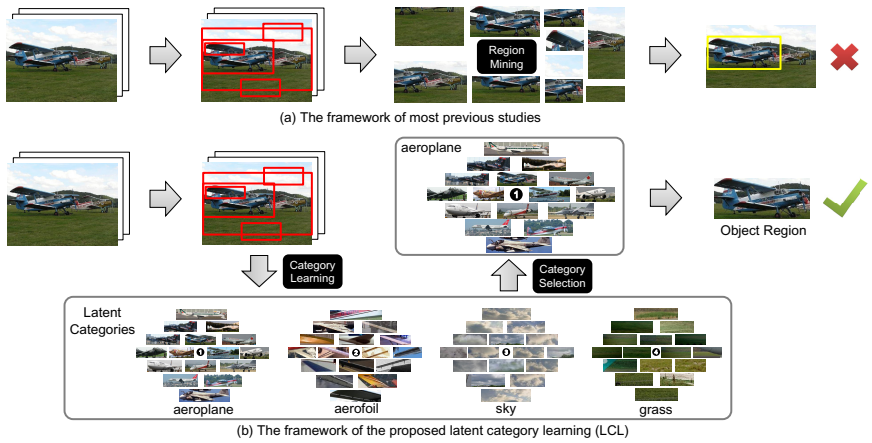

Fig. 1. The comparison of the framework of the proposed method and previous studies

region mining methods, *e.g.,* exhaustive search [20, 21], multiple instance learning [13, 28, 34], inter-intra-class modeling [5, 9, 23, 28, 29] and topic model [25, 30]. They have achieved promising results on object-centered conditions, in which objects occupy a large portion of an image [12]. However, on the cluttered condition such as the PASCAL VOC challenge [10], there still a long way to be competitive with the supervised approach [11].

In cluttered conditions, due to large object variations, objects usually have large ambiguity with backgrounds. Besides, in the weakly supervised task, only the image-level class labels are available, *e.g.,* the image has an aeroplane in Fig.1(a). However, a large number of candidate regions have a large background area, which makes it difficult to discover object regions in cluttered conditions, *e.g.,* the localization in Fig.1(a) contains too much background. However, backgrounds contain some latent information, *e.g.,* there is sky and grass in the image (Fig.1(a)). This latent information can be very useful because if it can be learned, the object-background ambiguity can be reduced to suppress the background, *e.g.,* the background area in Fig.1(b) is suppressed. Due to the unknown label of the candidate regions, learning these latent categories is an unsupervised learning problem. Many studies in unsupervised learning have attempted to discover the latent categories in object-centered conditions [4, 18, 19, 27, 30]. Inspired by them, we proposed to learn the latent categories in cluttered conditions.

In this paper, we propose the latent category learning (LCL) for weakly supervised localization. To learn the latent categories, we use the probabilistic Latent Semantic Analysis (pLSA) [16], which is a typical unsupervised learning method and achieves notable success in discovering latent semantics [30]. Fig.1(b) shows the framework of the proposed method. Compared to the previous studies in Fig.1(a), there are two main differences:

1) *Category Learning. Is it possible to learn meaningful latent categories in backgrounds?* We show that the typical unsupervised semantic analysis can

successfully learn the latent categories to represent objects, object parts and backgrounds, as shown in Fig.1(b).

2) *Category Selection.* After learning these categories, *which category contains the target object class?* We propose a category selection method by evaluating the discrimination of each category and select the most discriminative one. In this paper, we denote by "class" the given image-level object class and by "category" the latent category in an object class.

In the evaluation, we use the PASCAL VOC 2007 database [10] and ILSVRC 2013 detection challenge. For fair comparison with supervised methods, we use the complete dataset with only image-level class labels. On PASCAL VOC 2007, we obtain the annotation accuracy of 48%, which is 10% higher than the previous results [25, 28, 29]. More importantly, the LCL achieves the detection average precision of 30.9%, which outperforms previous results [23, 29] by 8% and can be competitive with the supervised deformable part model (DPM) 5.0 baseline 33.7% [15]. On ILSVRC 2013 detection challenge, we obtain the precision of 6.0% on the validation set, which is also competitive with the 8.8% by DPM.

2 Related Work

In recent years, many studies have been proposed in the weakly supervised localization, *e.g.,* exhaustive search [5,20,21,36], multiple instance learning [13,28,34], inter-intra-class modeling [9, 23, 29, 35] and topic model [25, 30]. Most of them adopt a similar framework, which has three main steps: (1) *Region Extraction*: region proposals extract candidate regions for each image; (2) *Region Representation*: feature representation is constructed for each region; (3) *Region Mining*: object regions (correct localizations) are discovered among the candidate regions. We review the main studies from these three parts.

Region Extraction. Nguyen *et al.* [20] and Pandy *et al.* [21] use dense regions in an initial bounding box as candidate regions, but the fixed size and shape make it difficult to generate enough object regions. To improve the quality of the candidate regions, various region proposals are used to extract regions based on object saliency. The one popularly used [9, 28, 29] is proposed by Alexe *et al.* [1], who present a generic objectness measure by combining multiple image cues in a Bayesian framework. Promising results have been obtained based on this proposal [1,9,28,29]. Recently, a segmentation based region proposal, named Selective Search [32], can generate regions with better objectness for its hierarchical segmentation and grouping strategies [6, 14, 32]. In this paper, we use the selective search for region extraction.

Region Representation. In [21], each candidate region is represented by the histogram of oriented gradients (HOG) descriptor [11]. With the additional viewpoint annotation, promising results are obtained on the subset of the PASCAL VOC 2007 challenge [10]. However, this gradient based low-level descriptor is sensitive to cluttered backgrounds. Many recent studies use the Bag-of-Words (BoW) feature for its mid-level object representation [3, 9, 23, 28, 29], and some

researchers combine the low-level and mid-level features for better discrimination [9]. Recently, the deep networks have achieved great success in large-scale and challenging object recognition tasks for its semantic object representation, especially the Convolutional Neural Network (CNN) [14, 17]. In this paper, we use the CNN for region representation.

Region Mining. In [21] and [20], object regions are obtained based on exhaustive search in an initial bounding box, which is usually determined based on object saliency [5,21]. However, the fixed size and shape make it difficult to collect enough object regions. To discover more objects, multiple instance learning considers inter-class relations by organizing the candidate regions as positive and negative bags [20, 22, 28, 34]. To improve the quality of the object regions, researchers further model intra-class relations to improve the similarity of the regions within the same object class [9, 23, 29, 35]. However, due to large object variations, localizations may have large background area. In fact, backgrounds contain useful latent categories, which can represent objects, object parts or backgrounds. They can be beneficial to reduce the object-background ambiguity and suppress the background area. Given only image-level class labels, learning these latent categories is an unsupervised learning problem. Some studies have attempted to learn them from large quantity of images in object-centered conditions [4, 18, 19, 27, 30]. Inspired by them, in this paper, we proposed to learn the latent categories in cluttered conditions.

3 Latent Category Learning

In this section, we present the latent category learning (LCL) for weakly supervised localization. We first introduce the extraction of the semantic candidate regions, then we elaborate how to learn the latent categories and discover the object regions among these categories. In this paper, we denote the object regions as correct localizations.

3.1 Region Extraction

Region proposal generates candidate regions for probable object locations. We use a segmentation based region proposal named Selective Search [32], which can generate regions with strong objectness [6, 14]. Compared to other region proposals [1], it is reported to have a higher overlap with ground truth bounding box but only with the comparable number of regions [32]. Fig.3(b) shows some examples on the training set of the PASCAL VOC 2007 database. Although objects vary a lot in size, illumination and occlusion, the selective search can extract object regions in most images.

After generating the candidate regions, the next step is to construct feature representation for them. In this paper, we use Convolutional Neural Network (CNN) to represent the regions. CNN has made a great breakthrough in many object recognition tasks [14,17]. It can construct semantic object representation for its deep hierarchical structure. As demonstrated in [14], the classification

results on ImageNet [8] can generalize well to the detection task in PASCAL VOC challenge. We train a CNN classification model on ILSVRC 2011 with the same setup to [14], which uses five convolutional layers and three fully-connected layers. We represent each candidate region by the *fc6* layer, which is the first fully-connected layer containing 4096 neurons. Therefore, the feature representation of each region has the dimension of 4096.

3.2 Category Learning

With the candidate regions extracted, in this part, we learn the latent categories from them. Due to the unknown object class label of these regions, learning the latent category is an unsupervised learning problem. In this paper, we use the typical pLSA for latent category learning.

We use positive images in an object class for category learning. Suppose we have N candidate regions in positive images, and the CNN representation of each region is d_j. In document analysis, the pLSA usually takes the histogram of occurrence frequency on visual words as input, while the CNN region representation satisfies this histogram input for two reasons. Firstly, due to the Rectified Linear Units [14], all the region representation is non-negative. Secondly, we consider each neuron in the *fc6* layer as a visual word, and the CNN representation is the occurrence confidence on these words. The larger confidence leads to the larger occurrence probability of a word (neuron). If a hard threshold function ($d_j >$T 1;else 0) is used on the CNN representation, it will turn into the 0,1 value, thus the representation is the same to the histogram of occurrence frequency; while if the threshold function is not used, the CNN representation is not the strict frequency but the soft version. Therefore, this CNN region representation can fit well in the framework of topic modeling.

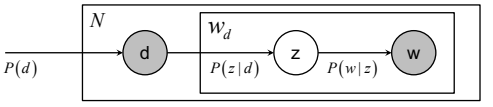

Fig. 2. The graphical model of the probabilistic Latent Semantic Analysis (pLSA) [30]

We denote each word (neuron) as w_i, thus the occurrence frequency of region d_j on w_i is the i-th dimension of d_j. In addition, there is a hidden topic variable z_k associated with all the visual words. We treat each topic z_k as a latent category in an object class. The pLSA optimizes the joint probability $P(w_i, d_j, z_k)$, which has the form of the graphical model shown in Fig.2 [30]. Marginalizing over the latent category z_k determines the conditional probability $P(w_i|d_j)$:

$$P(w_i|d_j) = \sum_{k=1}^{K} P(z_k|d_j)\, P(w_i|z_k), \tag{1}$$

where $P(z_k|d_j)$ is the probability of category z_k occurring in region d_j. Based on this term, each region has K probabilities for K latent categories. We consider that if region d_j has the maximum probability on category z_k, then d_j only

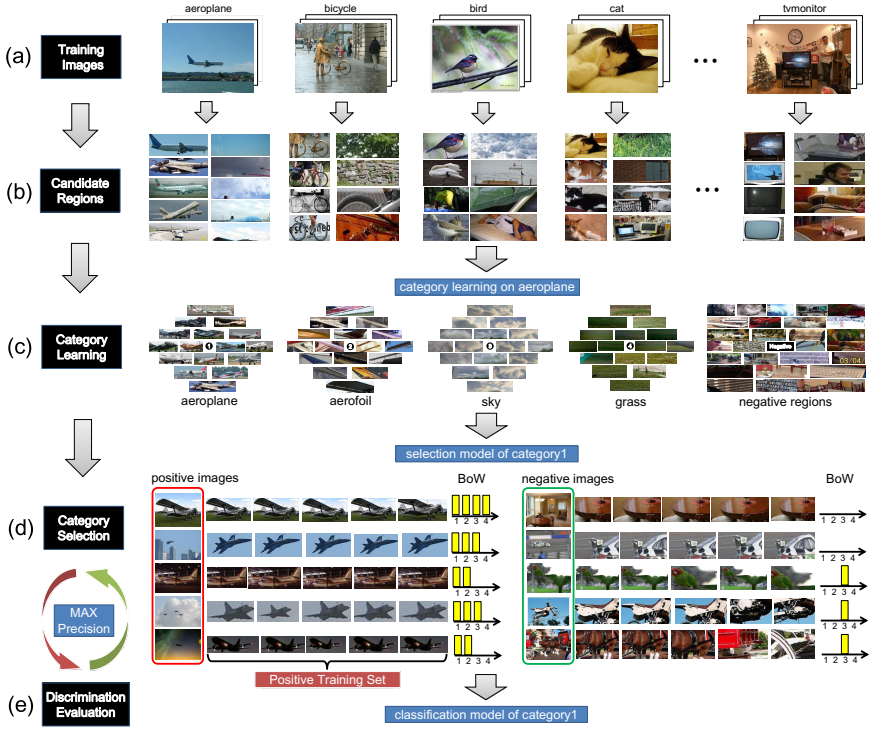

Fig. 3. The flowchart of the proposed latent category learning (LCL) for weakly supervised localization. (a) Original images on the PASCAL VOC 2007 training set. (b) Selective search and CNN extract semantic candidate regions. (c) Probabilistic Latent Semantic Analysis (pLSA) learns latent categories. (d) A selection model is trained for each latent category. (e) The discrimination of each category is evaluated by the classification model constructed in the manner of bag-of-words.

belongs to z_k. In this way, all regions are divided into K sets, each of which contains the regions with similar semantic meaning. Fig.3(c) shows some learned latent categories of the aeroplane class. These categories have strong semantic meanings, *e.g.,* category 1 represents the aeroplane, category 2 is the aerofoil, while others contain backgrounds such as sky and grass. The categories in each object class are learned separately to avoid a large memory cost.

3.3 Category Selection

After learning the latent categories, a problem is to decide *which one contains the object regions of the target object class?* In this part, we propose a category selection strategy to discover the object regions. The idea is based on the fact that the latent categories have different semantic meanings, thus they have different discrimination to the target object class. We exploit the different discrimination to find out the most discriminative category. To evaluate the discrimination, it is observed that in each latent category, the regions of positive and negative images

have different occurrence frequencies on all the learned categories. For example, in category 1, regions of positive images have a high occurrence frequency on aeroplane but much lower frequency on others, while it is the opposite for the regions of negative images, as shown in Fig.3(d). Combined with image-level class labels, we select the category with the frequency which best differentiates the target object class and backgrounds.

Fig.3 is used for an illustration. To construct the frequency for each category, we first have to select the regions which can represent the category. We train a selection model to select them. For any target category (category 1), we consider the regions in it as positive regions, while the negative regions consist of two parts: the ones in other categories (category 2-4) and the ones from negative images (negative). Therefore, a selection model of the target category (category 1) can be trained. Secondly, we use the selection model to select the top T scored regions in each positive and negative image. We observe that the occurrence frequencies of the T selected regions is the BoW representation, as shown in Fig.3(d). Based on these regions, we construct the BoW image representation for each positive and negative image. Finally, with the BoW representation, a classification model of the target latent category (category 1) is trained on the training set with the image-level class label, and the discrimination of the model is evaluated by the classification precision on the validation set. By evaluating all categories, the one with the highest precision is considered as the most discriminative one, and its corresponding top T regions in positive images constitute the positive training set. Fig.3(d) shows the selection process and the positive training set on the aeroplane class.

In constructing the BoW representation, there are three steps: (1) *Codebook Generation*. We quantify each latent category by averaging the regions in it. Let $\mathbf{Z} = [\mathbf{z_1}, ..., \mathbf{z_K}]^T \in \Re^{M \times K}$ denote the codebook with K categories. We use the average to quantify the category for two reasons: one is that the regions in a category look very similar, and it is reasonable to use the center; another is that the regions in the correct category overlap heavily with the target object, thus averaging them is beneficial to suppress backgrounds. (2) *Feature Encoding*. In each image, suppose the T selected regions are denoted as $[\mathbf{d_1}, ..., \mathbf{d_T}]^T \in \Re^{M \times T}$, we encode each region by the Super Vector Coding [37]:

$$\left[\underbrace{0, ..., 0}_{(j-1)*M\ dim.}, \overbrace{\mathbf{d_i} - \mathbf{z_j}}^{M\ \text{dim.}}, \underbrace{0, ..., 0}_{(K-j)*M\ dim.} \right]. \quad (2)$$
$$s.t.\ \mathbf{z_j} = \arg\min_{\mathbf{z_k}} \|\mathbf{d_i} - \mathbf{z_k}\|_2$$

(3) *Feature Pooling*. After the encoding, average pooling [37] is used on the encoding of all the T regions to construct the BoW image representation, as shown in Fig.3(d).

4 Experimental Evaluation

In this section, we evaluate the proposed method on the PASCAL VOC 2007 dataset and ILSVRC 2013 detection challenge. We use the complete dataset with only image-level class labels for fair comparison with the supervised approach. The detailed setup is given as follows.

Region Extraction: In selective search, we use the source code released by Uijlings *et al.* [32]. We run the "fast" option to generate about 2000 candidate regions for an image. Then, to represent the regions, we train a convolutional neural network (CNN) on ILSVRC 2011 with five convolutional and three fully-connected layers, which has the same architecture to [14, 17]. We use the fc6 layer for representation with the dimensionality of 4096.

Category Learning and Selection: For each object class, all the regions from positive images are used for category learning. The number of the latent categories (K) is determined by the highest classification precision on different number, while K is around 30 for most classes. In training selection models in category selection, the number of the top selected regions (T) in each positive image is set up to be 10 to guarantee the quality of the predicted locations.

Training and Testing: In training the classification models and final object detectors, the stochastic dual coordinate ascent [24] in VLFeat [33] is adopted for high efficiency. In testing, we first select the regions with the score larger than -1, then the Non Maximum Suppression (NMS) [11] with the threshold of 0.5 is used to obtain final localizations.

4.1 Automated Annotation Results

Table 1 shows the annotation accuracy of the proposed LCL and the previous studies on the trainval set. The accuracy is measured by the percentage of training images in which an instance is correctly localized according to the PASCAL criterion, which requires the overlap of larger than 0.5 between the object region and the ground truth. We also use k-means in category learning as a baseline for comparison with pLSA. It is observed that LCL yields an annotation accuracy of 48.5%, which outperforms the previous best result by 10%. LCL improves most classes, and the improvement is quite promising on some difficult ones, *e.g.,* 18% on chair and 22% on plant. Besides, LCL-pLSA outperforms LCL-kmeans by a small margin, which shows that pLSA is slightly better in learning latent category, but it is much better than LCL-kmeans in the detection results, as shown below in Sec.4.2. Fig.4 shows some successful and failed difficult localizations by LCL on the trainval set. Although objects vary a lot in size, occlusion and illumination, LCL correctly localizes most difficult samples.

Though LCL shows promising improvements, it fails on some classes such as boat and table. Based on our observation, there are two main reasons for this: (1) Too much object variation. For example in boat, the size and appearance vary too much. Some images have small sailboats while some have large ships, which

Table 1. The comparison of annotation accuracy on PASCAL VOC 2007 trainval set

Method	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow	
Joint Learning [20]	30.7	16.5	23	14.9	4.9	29.6	26.5	35.3	7.2	23.4	
MIL-SVM [2]	37.8	17.7	26.7	13.8	4.9	34.4	33.7	46.6	5.4	29.8	
Drift Detect [29]	45.8	21.8	30.9	20.4	5.3	37.6	40.8	51.6	7	29.8	
MIL-Negative [28]	42.4	46.5	18.2	8.8	2.9	40.9	73.2	44.8	5.4	29.8	
Transfer Learning [26]	54.7	22.7	33.7	24.5	4.6	33.9	42.5	57	7.3	39.1	
Beyasian Topic [25]	67.3	54.4	34.3	17.8	1.3	46.6	60.7	68.9	2.5	32.4	
Multifold MIL [7]	56.6	58.3	28.4	20.7	6.8	54.9	69.1	20.8	9.2	50.5	
LCL-kmeans	74.9	61.7	49.6	13.5	17.0	57.4	73.3	44.0	27.5	70.0	
LCL-pLSA	80.1	63.9	51.5	14.9	21.0	55.7	74.2	43.5	26.2	53.4	
Method	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	Accuracy
Joint Learning [20]	20.5	32.1	24.4	33.1	17.2	12.2	20.8	28.8	40.6	7	22.4
MIL-SVM [2]	14.5	32.8	34.8	41.6	19.9	11.4	25	23.6	45.2	8.6	25.4
Drift Detect [29]	27.5	41.3	41.8	47.3	24.1	12.2	28.1	32.8	48.7	9.4	30.2
MIL-Negative [28]	14.5	32.8	34.8	41.6	19.9	11.4	25	23.6	45.2	8.6	30.4
Transfer Learning [26]	24.1	43.3	41.3	51.5	25.3	13.3	28	29.5	54.6	11.8	32.1
Beyasian Topic [25]	16.2	58.9	51.5	64.6	18.2	3.1	20.9	34.7	63.4	5.9	36.2
Multifold MIL [7]	10.2	29.0	58.0	64.9	36.7	18.7	56.5	13.2	54.9	59.4	38.8
LCL-kmeans	16.3	56.3	55.3	69.5	13.6	40.0	60.3	46.2	45.5	61.9	**47.7**
LCL-pLSA	16.3	56.7	58.3	69.5	14.1	38.3	58.8	47.2	49.1	60.9	**48.5**

Fig. 4. Some successful and failed difficult localizations on the trainval set

makes it difficult to learn meaningful latent categories under the limited number of positive images. (2) Similar co-occurrent classes. For example the table, it always co-exists with chairs. They look very similar in most cases, *e.g.,* both the table and chair have a flat area with several legs, which makes it difficult to learn two different latent categories. Therefore, under the cases of too much variations and similar co-occurrent classes, it is challenging for LCL to generate good localizations.

4.2 Detection Results

Table 2 shows the detection mean average precision (mAP) of the proposed LCL, the previous studies and the supervised approaches on the PASCAL VOC 2007 test set. It is observed that LCL-pLSA yields a detection mAP of 30.9%, which improves the previous best result by 8% and improves most classes by a large margin, *e.g.,* 21% on aeroplane, 13% on cow, 10% on motorbike and 15% on sofa. We also make a breakthrough on the classes which are almost zero in previous results, *e.g.,* the improvement is about 11% on chair. More importantly, compared to the supervised approach, the 30.9% obtained by LCL-pLSA can be

Table 2. The comparison of the detection mAP on PASCAL VOC 2007 test set

Method	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow
Drift-Detect [29]	13.4	44.0	3.1	3.1	0.0	31.2	43.9	7.1	0.1	9.3
Object-Centric [23]	-	-	-	-	-	-	-	-	-	-
Multifold MIL [7]	35.8	40.6	8.1	7.6	3.1	35.9	41.8	16.8	1.4	23.0
Latent SVM [31]	27.6	41.9	19.7	9.1	10.4	35.8	39.1	33.6	0.6	20.9
LCL-kmeans	41.5	29.7	24.9	12.0	10.7	30.3	40.9	31.8	10.5	21.8
LCL-pLSA	48.8	41.0	23.6	12.1	11.1	42.7	40.9	35.5	11.1	36.6
DPM 5.0 [11]	33.2	60.3	10.2	16.1	27.3	54.3	58.2	23.0	20.0	24.1
CNN Supervise [14]	68.1	72.8	56.8	43.0	36.8	66.3	74.2	67.6	34.4	63.5

Method	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	**mAP**
Drift-Detect [29]	9.9	1.5	29.4	38.3	4.6	0.1	0.4	3.8	34.2	0.0	13.9
Object-Centric [23]	-	-	-	-	-	-	-	-	-	-	15.0
Multifold MIL [7]	4.9	14.1	31.9	41.9	19.3	11.1	27.6	12.1	31.0	40.6	22.4
Latent SVM [31]	10.0	27.7	29.4	39.2	9.1	19.3	20.5	17.1	35.6	7.1	22.7
LCL-kmeans	15.4	29.4	24.3	37.8	19.1	14.7	33.1	24.1	36.2	43.0	**26.6**
LCL-pLSA	18.4	35.3	34.8	51.3	17.2	17.4	26.8	32.8	35.1	45.6	**30.9**
DPM 5.0 [11]	26.7	12.7	58.1	48.2	43.2	12.0	21.1	36.1	46.0	43.5	33.7
CNN Supervise [14]	54.5	61.2	69.1	68.6	58.7	33.4	62.9	51.1	62.5	64.8	58.5

Fig. 5. Some successful and failed difficult localizations on the test set

competitive with the deformable part model 5.0 released baseline 33.7%. The precision on most classes is comparable to DPM 5.0, and some classes show better precision, *e.g.,* the improvement is about 15% on aeroplane, 12% on bird, cat and cow, and 23% on dog. This result is very encouraging because without the tedious and ambiguous annotation of object locations, the weakly supervised localization yields the comparable detection precision to the supervised methods in cluttered image conditions. Some successful and failed difficult detections on the test set are shown in Fig.5, in which LCL correctly localizes most objects under large variations of size, occlusion and illumination.

Table 3 shows the detection mean average precision (mAP) of the proposed LCL and the DPM 5.0 baseline on the validation set of the ILSVRC 2013 detection challenge (200 object classes). For higher efficiency, we use the k-means in category learning instead of pLSA, and the number of latent categories (K) is fixed to be 30. It is observed that the proposed LCL yields the detection mAP of 6.0%, which can be competitive with DPM 5.0 baseline 8.8%. This result demonstrates that LCL can be effective in large-scale image conditions.

Table 3. The comparison of the detection mAP of the LCL-kmeans and DPM 5.0 baseline on the validation set of the ILSVRC 2013 detection challenge

ILSVRC 2013 detection challenge	mAP (Validation)
LCL-kmeans	6.0%
DPM 5.0 (without context)	8.8%

Table 4. The detection mAP of the proposed LCL by incorporating object structure and inter-class relation

Method	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow	
Drift-Detect [29]	13.4	44.0	3.1	3.1	0.0	31.2	43.9	7.1	0.1	9.3	
LCL-pLSA	48.8	41.0	23.6	12.1	11.1	42.7	40.9	35.5	11.1	36.6	
LCL+DPM	30.2	46.9	10.4	4.6	11.1	47.0	44.9	14.7	5.6	17.4	
LCL+Context	48.9	42.3	26.1	11.3	11.9	41.3	40.9	34.7	10.8	34.7	
Method	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
Drift-Detect [29]	9.9	1.5	29.4	38.3	4.6	0.1	0.4	3.8	34.2	0.0	13.9
LCL-pLSA	18.4	35.3	34.8	51.3	17.2	17.4	26.8	32.8	35.1	45.6	30.9
LCL+DPM	4.6	15.0	38.6	41.8	13.9	10.6	19.3	31.8	16.3	37.9	23.1
LCL+Context	18.8	34.4	35.4	52.7	19.1	17.4	35.9	33.3	34.8	46.5	**31.6**

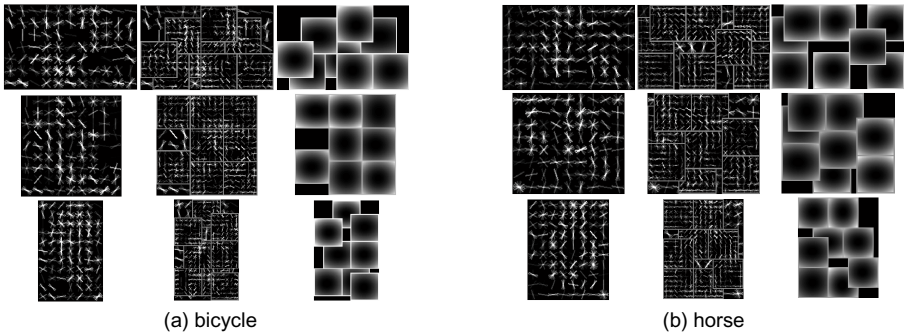

Fig. 6. The visualization of the detection model by using the LCL localizations as ground truth. Each detection model is trained with three components.

Though LCL has achieved comparable performance to DPM 5.0, the precision on some classes is relatively low, *e.g.,* bicycle, car, horse and person. We observe that for the classes which DPM beats LCL, most of them are the classes of rigid objects, *e.g.,* bicycle, boat, bottle, chair and table. Under this condition, object structures provide good representations because rigid objects do not change much. Combined with the HOG representation, the DPM achieves better results.

4.3 DPM and Context Embedding

To incorporate object structure and inter-class relations, we consider DPM and context in LCL for further enhancement. In DPM, we use the LCL annotations as ground truth, and the same setup to [11] is used, *i.e.,* 8 object parts and 3 object

components. For the context, similar to the contextual operation in [11], we concatenate the region score, region location and the detection score of each class to the CNN region representation, thus the feature dimension of each candidate region is $4096 + 25 = 4121$.

Table.4 shows the detection mAP of LCL by incorporating DPM and context. LCL+DPM obtains a mAP of 23.1%, which is 9% higher than the Drift-Detect [29] which also trains DPM. However, compared to the LCL-pLSA, it decreases by 7% due to the inaccurate annotations of LCL, and the precision on most classes decreases a lot. But we see some promising improvements in detecting rigid objects, *e.g.,* the improvement over LCL-pLSA is about 6% in bicycle, 5% on bus and car, and 4% in horse. Fig.6 shows the detection model trained by LCL-pLSA with three components on the classes of bicycle and horse. The top two components describe the side views of the objects based on the different size, and the bottom component is more like the frontal or the rear view. These results show that incorporating object structures in latent category learning can be beneficial to detect rigid objects.

We see that by considering inter-class relations in LCL, performance can be further improved. LCL+Context achieves the mAP of 31.6%, which outperforms the LCL-pLSA baseline by 0.7%. The improvements on some classes are promising, *e.g.,* 9% on sheep, 3% on bird and 2% on person, but the average improvement is too small. The reason may be that the locations and scores of the detections are not accurate enough to provide meaningful co-occurrence information. As a result, this will hurt the detection precision, *e.g.,* the precision decreases about $1 \sim 2\%$ on boat, bus, cow and dog.

4.4 Category Selection

One key step in the latent category learning is to select the category containing the target object class. As elaborated in Sec.3.3, this category selection is based on each category's discrimination, which is evaluated by the classification precision on the validation set. In constructing the BoW image representation, we set the number of the top scored regions T to be 10. Fig.7 shows the Classification Mean Average Precision (Cls-mAP) of aeroplane and motorbike based on the number of latent categories of 20 and 30 respectively. The Maximum Average Best Overlap (MABO) [14] with ground truth is used to validate the correctness of the selection. It is observed that the highest Cls-mAP always corresponds to the highest MABO, *e.g.,* the $17th$ and $14th$ category in aeroplane and motorbike, which demonstrates the effectiveness of this selection strategy. However, we observe that the highest Cls-mAP does not have a large margin over the ones of other categories, *e.g.,* the $13th$ and $20th$ category of the aeroplane also has high precision. The reason is that the categories such as aerofoil and sky also contribute a lot to classify aeroplane. Although the margin is small, the most discriminative category can obtain the highest classification performance. In future, we will consider the more powerful Latent Dirichlet Allocation (LDA) to improve the discrimination.

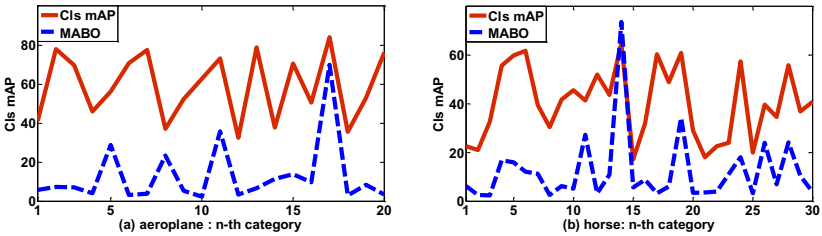

Fig. 7. The category selection on the classes of aeroplane and horse. The testing number of latent categories is set to be 20 and 30.

Another problem is how to set the number of latent categories for each object class. Due to the large variations of the number of positive images in different classes, using the appropriate number of latent categories is critical to learn meaningful ones. In our implementation, we initially set the number K to be $20 \sim 60$, then we use the above selection process to obtain the most discriminative category for each number. Finally, the number with the highest Cls-mAP is used. Table.5 shows the highest Cls-mAP and MABO of bicycle and cow under the different number of categories. We see the $K = 60$ is the best for both classes. If K is too small, the discriminative category contains many background regions; while if K is too large, object regions will be assigned to different latent categories which may not be discriminative to the target object class.

Table 5. The selection of the number of latent categories K on bicycle and cow

Cls-mAP						MABO					
K	20	30	40	50	60	K	20	30	40	50	60
bicycle	66.7	67.6	65.2	68.4	**69.6**	bicycle	57.2	68.2	67.3	62.2	**70.0**
cow	45.6	47.1	48.5	44.1	**51.2**	cow	66.7	47.8	60.9	60.1	**68.8**

5 Conclusion

In this paper, we have proposed the latent category learning (LCL) for weakly supervised object localization. We first use a segmentation based region proposal to generate semantic candidate regions, each of which is represented by the Convolutional Neural Network (CNN) trained on ILSVRC 2011. Then, based on the large number of candidate regions, the probabilistic Latent Semantic Analysis (pLSA) is used to learn the latent categories, from which the category containing target object class is selected by evaluating each latent category's discrimination. Evaluation on the challenging PASCAL VOC 2007 dataset and the large-scale ILSVRC 2013 detection competition shows encouraging results achieved by LCL, with state-of-the-art annotation and detection performance among the weakly supervised localization methods. More importantly, the results

are competitive with the supervised deformable part model 5.0 released baseline. In the future, we will improve the discrimination of the latent categories by LDA and design a category learning algorithm which automatically determine the number of latent categories for use in large-scale conditions.

Acknowledgement. This work is funded by the National Basic Research Program of China (Grant No. 2012CB316302), National Natural Science Foundation of China (Grant No. 61322209 and Grant No. 61175007), the National Key Technology R&D Program (Grant No. 2012BAH07B01).

References

1. Alexe, B., Deselaers, T., Ferrari, V.: What is an object? In: CVPR (2010)
2. Andrews, S., Tsochantaridis, I., Hofmann, T.: Support vector machines for multiple-instance learning. In: NIPS (2003)
3. Boureau, Y.L., Bach, F., LeCun, Y., Ponce, J.: Learning mid-level features for recognition. In: CVPR (2010)
4. Chen, X., Shrivastava, A., Gupta, A.: Neil: Extracting visual knowledge fromweb data. In: ICCV (2013)
5. Chum, O., Zisserman, A.: An exemplar model for learning object classes. In: CVPR (2007)
6. Cinbis, R.G., Verbeek, J., Schmid, C.: Segmentation driven object detection with fisher vectors. In: ICCV (2013)
7. Cinbis, R.G., Verbeek, J., Schmid, C.: Multi-fold mil training forweakly supervised object localization. In: CVPR (2014)
8. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR (2007)
9. Deselaers, T., Alexe, B., Ferrari, V.: Weakly supervised localization and learning with generic knowledge. IJCV 100(3) (2012)
10. Everingham, M., Gool, L.V., Williams, C.K.I., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. IJCV (2010)
11. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part based models. TPAMI 32(9) (2010)
12. Fergus, R., Perona, P., Zisserman, A.: Object class recognition by unsupervised scale-invariant learning. In: CVPR (2003)
13. Galleguillos, C., Babenko, B., Rabinovich, A., Belongie, S.: Weakly supervised object localization with stable segmentations. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part I. LNCS, vol. 5302, pp. 193–207. Springer, Heidelberg (2008)
14. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: CVPR (2014)
15. Girshick, R.B., Felzenszwalb, P.F., McAllester, D.A.: Discriminatively trained deformable part models, release 5, `http://people.cs.uchicago.edu/~rbg/latent-release5/`
16. Hofmann, T.: Probabilistic latent semantic indexing. In: SIGIR (1999)
17. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS (2012)

18. Kumar, M.P., Packer, B., Koller, D.: Modeling latent variable uncertainty for loss-based learning. In: ICML (2012)
19. Li, Q., Wu, J., Tu, Z.: Harvesting mid-level visual concepts from large-scale internet images. In: CVPR (2013)
20. Nguyen, M.H., Torresani, L., de la Torre, F., Rother, C.: Weakly supervised discriminative localization and classification: a joint learning process. In: ICCV (2009)
21. Pandey, M., Lazebnik, S.: Scene recognition and weakly supervised object localization with deformable part-based models. In: ICCV (2011)
22. Ren, W., Wang, C., Huang, K., Tan, T.: On automatic and efficient localization of objects with weak supervision. In: ACCV (2014)
23. Russakovsky, O., Lin, Y., Yu, K., Fei-Fei, L.: Object-centric spatial pooling for image classification. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part II. LNCS, vol. 7573, pp. 1–15. Springer, Heidelberg (2012)
24. Shalev-Shwartz, S., Zhang, T.: Stochastic dual coordinate ascent methods for regularized loss minimization. JMLR 14(1), 567–599 (2013)
25. Shi, Z., Hospedales, T.M., Xiang, T.: Bayesian joint topic modelling for weakly supervised object localisation. In: ICCV (2013)
26. Shi, Z., Siva, P., Xiang, T.: Transfer learning by ranking for weakly supervised object annotation. In: BMVC (2012)
27. Singh, S., Gupta, A., Efros, A.A.: Unsupervised discovery of mid-level discriminative patches. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part II. LNCS, vol. 7573, pp. 73–86. Springer, Heidelberg (2012)
28. Siva, P., Russell, C., Xiang, T.: In defence of negative mining for annotating weakly labelled data. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part III. LNCS, vol. 7574, pp. 594–608. Springer, Heidelberg (2012)
29. Siva, P., Xiang, T.: Weakly supervised object detector learning with model drift detection. In: ICCV (2011)
30. Sivic, J., Russell, B.C., Efros, A.A., Zisserman, A., Freeman, W.T.: Discovering objects and their location in images. In: ICCV (2005)
31. Song, H.O., Girshick, R., Jegelka, S., Mairal, J., Harchaoui, Z., Darrell, T.: On learning to localize objects with minimal supervision. arXiv:1403.1024 (2014)
32. Uijlings, J., van de Sande, K., Gevers, T., Smeulders, A.: Selective search for object recognition. IJCV (2013)
33. Vedaldi, A., Fulkerson, B.: VLFeat: An open and portable library of computer vision algorithms (2008), `http://www.vlfeat.org/`
34. Vijayanarasimhan, S., Grauman, K.: Keywords to visual categories: Multiple-instance learning for weakly supervised object categorization. In: CVPR (2008)
35. Wang, S., Joo, J., Wang, Y., Zhu, S.C.: Weakly supervised learning for attribute localization in outdoor scenes. In: CVPR (2013)
36. Zhang, Y., Chen, T.: Weakly supervised object recognition and localization with invariant high order features. In: BMVC (2010)
37. Zhou, X., Yu, K., Zhang, T., Huang, T.S.: Image classification using super-vector coding of local image descriptors. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part V. LNCS, vol. 6315, pp. 141–154. Springer, Heidelberg (2010)