Cross-Domain Object Recognition Using Object Alignment

Pengcheng Liu
pengcheng.liu@ia.ac.cn
Chong Wang
cwang@nlpr.ia.ac.cn
Peipei Yang
ppyang@nlpr.ia.ac.cn
Kaiqi Huang
kqhuang@nlpr.ia.ac.cn
Tieniu Tan
tnt@nlpr.ia.ac.cn

Center for Research on Intelligent
Perception and Computing,
National Laboratory of Pattern
Recognition,
Institute of Automation, Chinese
Academy of Sciences,
Beijing, China

Abstract

One popular solution to the problem of cross-domain object recognition is minimizing the difference between source and target distributions. Existing methods are devoted to minimizing that domain difference in a complex image space, which makes the problem hard to solve because of background influence. To discount the influence, we propose to minimize that difference using object alignment. We firstly present an algorithm to effectively align the object that appears in a set of images, and learn detectors for the aligned objects so that the detectors are robust to the influence of irrelevant background. Then we utilize the classification information from the image space to enhance our detectors. Finally, based on the detectors, we introduce a self-paced adaptation method to further reduce the domain difference. Experimental results demonstrate that the object alignment is effective to minimize the domain difference, and show the state-of-the-art recognition performance on several visual domain adaptation datasets.

1 Introduction

Cross-domain object recognition [27] has long been one of the challenging problems in computer vision. This problem typically arises when training (source domain) and test (target domain) samples are drawn from different distributions. In the problem of object recognition, this case is usually caused by the situation that training and test samples are acquired under different sets of background, lighting, view point, resolution conditions, etc. Taking the domain adaptation (DA) benchmark dataset for an example, it consists of four different domains, and some examples from each domain are shown in Figure 1. The images from Webcam and DSLR are obtained under different resolution conditions yet a similar background; the Amazon consists of product images from amazon.com with a clean background, while the images in Caltech are captured under some more complex environment conditions.

Pages 66.1-66.12
DOI: https://dx.doi.org/10.5244/C.29.66

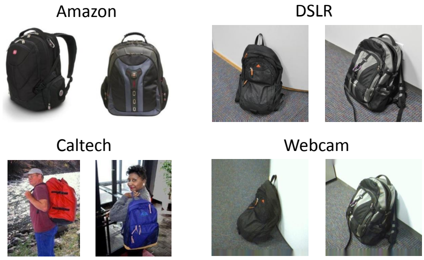
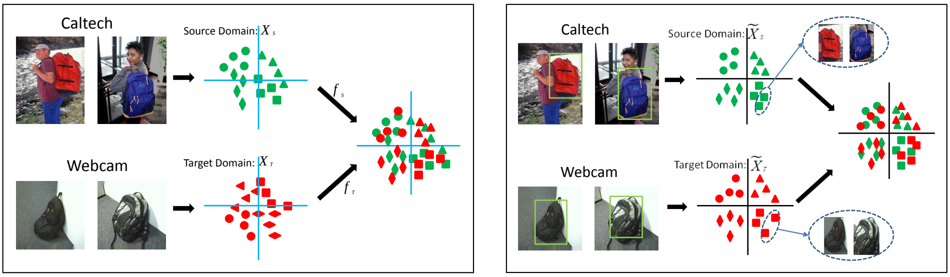


Figure 1: Sample images from four benchmark domains.



(a) (b)

Figure 2: Comparison of methods for a four-class problem on DA. (a) Existing methods minimize the domain difference based on two ambiguous image feature spaces: X_S and X_T. (b) We minimize the domain difference based on two more similar feature spaces consist of aligned objects: $\widetilde{X}_S$ and $\widetilde{X}_T$. Best viewed in color.

The key point of cross-domain object recognition is minimizing the distribution difference between source and target domains. This problem has received considerable attention in recent years. Some previous work [3, 16, 23, 24, 29] focus on learning a new domain-invariant feature representation by looking for a common projection space. Sample re-weighting, or selection methods [4, 13, 19] try to match the distributions of the source and target samples by assigning optimal weights to the source samples. To link the source and target domains, several subspace based DA methods [12, 14, 22] propose to learn a set of intermediate subspaces to capture the intrinsic domain shift between two domains.

Let X_S and X_T denote the samples of the source and target domains respectively, and $D(X_S, X_T)$ denote the domain distribution difference caused by different sets of background, lighting, view point, resolution conditions, etc. Based on the image feature space of X_S and X_T, the previous efforts are devoted to learning new transformations f_S and f_T to minimize $D(f_S(X_S), f_T(X_T))$, as shown in Figure 2 (a). For object recognition, a good object representation (e.g., deep learning feature [9]) is robust to some influence factors, such as lighting, view point, resolution conditions, etc. However, since the object and background are twisted in that image feature space, the discrepancy caused by background is difficult to eliminate, which makes it hard to learn optimal f_S and f_T for minimizing $D(f_S(X_S), f_T(X_T))$. If we can discount the influence from the ambiguous background before generating that feature space, the problem of cross-domain object recognition would be much easier.

In light of the above discussion, we propose a novel viewpoint on cross-domain object recognition. We minimize the domain difference by transferring to the feature space of aligned objects $\widetilde{X}_S$ and $\widetilde{X}_T$, but not the image feature space having background influence, as shown in Figure 2 (b). To explore the aligned objects, we first utilize the Probabilistic Latent Semantic Analysis (pLSA) [18] to discover the object that appears in a set of images. Then, based on the aligned objects, we learn detectors that are robust to the influence of the irrelevant background, which simplify the problem of cross-domain object recognition. In addition, we utilize the image classification information to enhance our detectors. Finally, to adapt source data to target data, we introduce a self-paced detector adaptation method that takes full advantage of the target domain to further reduce the domain difference. Experimental results demonstrate that the object alignment is beneficial to reduce the domain difference, and show the state-of-the-art performance on several visual domain adaptation datasets.

2 Related Work

We briefly review the relevant work on domain adaptation and object alignment as follows.

For cross-domain object recognition, one kind of approach is to learn a new domain-invariant feature representation by exploring a common projection space [3, 16, 23, 24, 28, 29]. Methods like manifold-alignment [32, 33] and low-rank reconstruction [20] are also introduced to learn that new representation. Another kind of approach, sample re-weighting [19] or selection [4, 13], is devoted to finding out a subset of source samples which are similar to the target samples. Moreover, to capture the intrinsic domain shift between source and target domains, a mapping function [10] between the source and target subspaces or a set of intermediate subspaces [12, 14, 17] is learned to link the two domains.

Object alignment based methods have been proposed for fine-grained recognition [2], pedestrian detector adaptation [35], etc. However, their detectors are learned in supervised way, which are not suitable for us since there is no ground-truth bounding box of the object available in the DA datasets.

In this paper, we present a new strategy for cross-domain object recognition. Compared with the existing methods that minimize the domain difference in an ambiguous image feature space, we minimize that difference based on the feature space of aligned objects, which makes two distributions more similar. This method, we believe and as suggested by the experiments, makes cross-domain object recognition much easier than before.

3 Proposed Method

The key insight of our approach is that the difference between the source and target distributions can be reduced by discounting the influence from the ambiguous background. We define the semantic object as the object that occurs in all the images of one class. To discount the background influence, our primary goal is to automatically localize the semantic object so that the irrelevant background is eliminated. Then based on the semantic object regions, we can learn an object detector that is robust to the influence of the irrelevant background and makes the cross-domain object recognition much easier than before. In addition, since our detectors are learned in a weakly supervised [34] way, we utilize the classification results from the image feature space to enhance our object detectors to avoid performance degradation. Finally, based on the detectors, a self-paced adaptation method is introduced to link the source and target data to further reduce the domain difference. The pipeline of our method is summarized in Figure 3. The rest of this section will describe our method step by step.

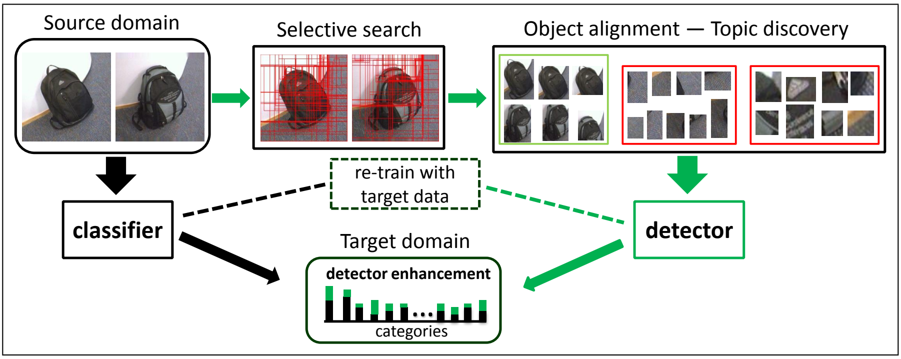


Figure 3: Pipeline of our method. Best viewed in color.

3.1 Object representation

In order to localize the semantic object region, we first generate sufficient category-independent region proposals of each image, which contains the semantic object with a high recall. A lot of recent studies offer solutions for the region proposals, such as objectness [1], constrained parametric min-cuts [8], selective search [31], etc. Among them, the selective search [31] shows higher recall on a generic object detection task (PASCAL VOC 2012). Thus we use selective search to generate about 1000 region proposals per image.

In order to distinguish the semantic object regions from other regions, it is necessary to describe each region proposal with a powerful descriptor, which would have tolerance to a certain degree of viewpoint, lighting, resolution and style change. In recent years, the convolutional neural network (CNN) based image representation has shown great success in the field of image recognition. The surprising image recognition accuracy on the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [9, 25] shows that CNN feature descriptor has great tolerance to intra-class variations. Thus, we compute a CNN based feature for each region proposal.

3.2 Object alignment for learning detectors

Object alignment: Given large quantities of region proposals, what we only know are their image category labels, instead of the labels indicating if they are semantic objects. Thus, we need to discover the semantic object that occurs in all the images of one class firstly. This is identical to semantic topic mining in statistical text analysis. The widely-used two semantic topic discovery models are probabilistic Latent Semantic Analysis(pLSA) [18] and Latent Dirichlet Allocation (LDA) [6]. Since the pLSA is simple and fast, we apply pLSA to discover the semantic object in the region proposals.

We will describe the model using the original terms "documents" and "words" as used in the text literature. In our case, documents correspond to region proposals, and the CNN representation of a document can correspond to the occurrence frequency of words for two reasons. Firstly, all the region representation is non-negative because of the Rectified Linear Units [11]. Secondly, we regard each neuron in the full connection layer as a visual word, and the CNN representation as the soft version of the occurrence frequency on these words.

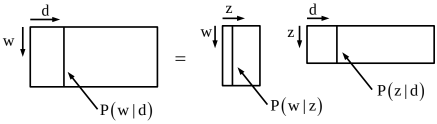


Figure 4: Latent variable model.

Suppose there are N documents, each of which is represented by a CNN feature of dimension M. The feature matrix of size M by N, namely the co-occurrence table in text analysis, where $n(w_i, d_j)$ stores the number of occurrences of a word (neuron) w_i in a document d_j.

The pLSA model introduces a latent topic variable z_k associated with each occurrence of a word w_i in a document d_j. Marginalizing over topics z_k, the conditional probability $P(w_i|d_j)$ can be formulated as following [26, 30],

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

where $P(z_k|d_j)$ is the probability of topic z_k occurring in document d_j, and $P(w_i|z_k)$ is the probability of a word w_i occurring in a particular topic z_k.

The model in Eq. (1) amounts to a matrix decomposition with the constraint that each document is expressed as a convex combination of K latent topic vectors, as shown in Figure 4. This latent variable model[1] can be fitted using the Expectation Maximization (EM) algorithm as described in [18].

To ensure that the visually similar objects (i.e., semantic object) are aligned into one topic, a region proposal d_j is assigned to topic z_k if $P(z_k|d_j)$ is the maximum probability among distributions $P(z|d_j)$. As shown in Figure 3, we discover 3 topics in the backpack category. One topic consists of the semantic object regions, i.e., backpack, and the other two topics consist of region proposals from background and some small object parts respectively.

Learning detectors: Given the topics of one image category, we aim to learn a detector for the semantic object. Although the semantic object has been aligned into one topic, we still do not know which topic contains the object of interest. For most images, the object of interest is the semantic object occurring in the original image. Therefore, if a detector (one-versus-all linear SVM) is trained by treating samples from the semantic object topic as positive samples and samples from the rest topics as negative samples, it would show the maximum response on the original image feature space. Following this intuitive approach, we learn one detector for each topic where positive samples are from this topic and negative samples are from the rest topics. Then all the detectors are tested on the original input image space, and the experimental results validate our intuition.

In addition, there may be some noisy region proposals in the semantic object topic. In order to learn a robust detector, all the proposals in the semantic object topic are sorted by the value of $P(z_k|d_j)$. Then part of the top-ranking proposals in the semantic object topic are selected as positive samples to train an object detector. According to this, our learned detectors are robust to the noise samples and irrelevant background, which would improve the performance of cross-domain object recognition.

[1] The code is available online.

3.3 Detector enhancement

Because the ground truth bounding box of the object is unavailable during training process, our learned detectors may localize some objects inaccurately. In addition, when the samples from different domains are captured under the same background, that background information would be useful to cross-domain object recognition. To avoid deterioration of recognition performance due to error-detection or missing the useful background, we utilize the classification results of the original image feature space to enhance our detectors. On one hand, we train a one-versus-all linear SVM on feature descriptors extracted over original images in source domain. On the other hand, given a test image from the target domain, we run selective search to generate region proposals. Since the aligned semantic object is more robust to the domain difference, the detectors learned in the source domain are used in the target domain to detect the object. Then we utilize the responses from the object detectors and the image classifiers jointly to determine the final category of an test image. The responses are linearly combined in the following form:

$$\vec{s} = \alpha * \vec{d} + (1-\alpha) * \vec{c}. \tag{2}$$

where $\vec{d}$ is the response of the object detectors, and $\vec{c}$ is the response of the image classifiers. Finally, the category of an image is determined based on the final score $\vec{s}$. In this paper, the parameter α is set by cross-validation.

3.4 Self-paced adaptation

Based on the learned detectors, we use a self-paced adaptation method to link the source and target domains. The algorithm we presented above has not utilized the information from the target domain yet. In order to take full advantage of the target domain to further reduce the domain difference, it will be helpful to add some unlabeled samples from target domain to source domain while training the models. To this end, all the samples in the target domain are scored by our learned models. Then the top-ranking target samples and their region proposals are selected as training data and added to source domain. The selected samples would be regarded as an intermediate domain linking the source and target domains, and then the model is re-trained. We can do this iteratively to ensure a high performance.

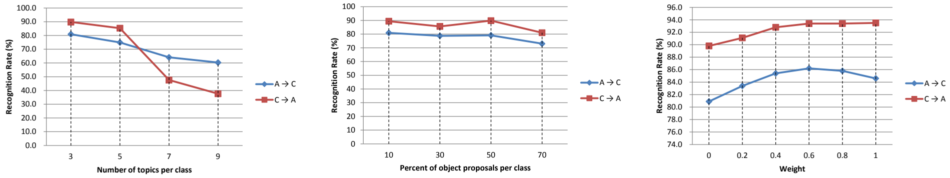
4 Experiments

4.1 Experiment on Office-Caltech dataset

In this section, we evaluate our method on the widely-used Office-Caltech benchmark [3, 4, 10, 12, 13, 14] for cross-domain object recognition. We compare the proposed method with several competitive ones. Experimental results show that our method is effective for cross-domain object recognition, and we achieve the state-of-the-art performance.

4.1.1 Dataset and data preparation

The widely-used Office-Caltech dataset for cross-domain image recognition is composed of four domains: Amazon (denoted by **A**), DSLR (denoted by **D**), Webcam (denoted by **W**) and Caltech (denoted by **C**). The first three domains are from the office dataset [27]. The



(a) (b) (c)

Figure 5: Recognition performance under different parameter settings: (a) number of topics for object alignment, (b) percent of object proposals for learning detectors, (c) weight for detector enhancement.

Caltech domain is introduced in [15]. There are 10 common classes in four domains. The number of images per class ranges from 8 to 151, and there are 2533 images in total.

The previous studies represented an image by SURF features encoded with a visual dictionary of 800 words, which were computed via K-means on a subset of Amazon images. In our experiments, we extract deep convolutional features based on an ImageNet pre-trained CNN. The CNN setup is identical to the one presented in [21]. We define **fc**6 as the output of the first fully connection layer. We use **fc**6 to represent the images and region proposals.

4.1.2 Parameter settings

In our method, there are 3 main parameters: topic number K for object alignment, top N percent of semantic object proposals for learning detectors and weight α for detector enhancement. We take two representative domains **A** (with clean background) and **C** (with complex background) for example, and set the three parameters in turn by 5-fold cross-validation. The corresponding experimental results on two pairs ($\mathbf{A} \rightarrow \mathbf{C}$ and $\mathbf{C} \rightarrow \mathbf{A}$) of cross-domain object recognition are shown in Figure 5. Based on these results, it is reasonable to set $K = 3$, $N = 10$ and $\alpha = 0.6$ for object recognition across **A** and **C**. For simplicity, we set $K = 3$, $N = 10$ and $\alpha = 0.6$ for all the domain pairs in Office-Caltech dataset.

4.1.3 Results of object alignment and detector enhancement

Since there are only a few samples in DSLR, it is not used as a source domain in most of the previous work. In this paper, we also focus on the remaining 9 pairs of source (**S**) and target (**T**) domains. We denote a cross-domain image recognition problem by the notation $\mathbf{S} \rightarrow \mathbf{T}$. Table 1 reports two sets of results base on SURF and CNN features respectively.

The first set of results in Table 1 are given based on SURF features. We quote these results directly from their papers.

The second set of results in Table 1 are given based on CNN features. For the baseline method, we use the source data only and the CNN features extracted from the whole image to train classifiers (linear SVM). Compared to the previous state of the art, our baseline validates that a good representation is helpful to reduce the difference of the source and target distributions. The two most popular DA methods (SA [10] and GFK [12]) show a slight performance increment over the baseline. Compared to the popular methods, although our detectors are learned using the source data only, it shows significant performance improvement except for the cases that $\mathbf{A} \rightarrow \mathbf{C}, \mathbf{D}, \mathbf{W}$ and $\mathbf{W} \rightarrow \mathbf{D}$. This is to be expected since the image in Amazon is object centered with clean background and the samples with the same

Method	A→C	A→D	A→W	C→A	C→D	C→W	W→A	W→C	W→D
Baseline	41.2	38.2	34.9	49.5	42.0	38.0	35.0	32.8	83.4
TCA [23]	35.0	36.3	27.8	41.4	45.2	32.5	24.2	22.5	80.2
GFS [14]	39.2	36.3	33.6	43.6	40.8	36.3	33.5	30.9	75.7
GFK [12]	42.2	42.7	40.7	44.5	43.3	44.7	31.8	30.8	75.6
SCL [7]	42.3	36.9	34.9	49.3	42.0	39.3	34.7	32.5	83.4
SA [10]	39.9	38.8	39.6	46.1	39.4	38.9	39.3	31.8	77.9
LMS [13]	45.5	47.1	46.1	56.7	57.3	49.5	40.2	35.4	75.2
DIP [3]	47.2	49.0	47.8	58.7	61.2	58.0	40.9	37.2	91.7
SIE [4]	48.2	49.1	48.1	56.7	61.2	58.0	42.7	38.6	93.0
Our baseline	80.8	78.4	77.3	85.4	76.6	72.0	70.1	64.8	95.1
SA [10]	79.9	78.5	77.7	88.6	79.6	77.2	71.1	64.1	95.5
GFK [12]	79.1	80.4	78.5	82.8	79.7	74.5	70.0	67.3	96.1
Our detector	80.9	71.3	72.9	89.8	81.5	78.6	73.4	70.7	90.4
Our DE	**85.4**	**80.9**	**81.0**	**93.4**	**85.4**	**80.0**	**79.6**	**72.1**	**97.5**

Table 1: Recognition accuracies on 9 pairs of cross-domain object recognition. The first set of results are obtained by using SURF features. The second set of results are obtained by using CNN features.

class label in Webcam and DSLR domains are captured under the same background conditions. This evidences that our object alignment is beneficial to reduce the difference between domains with different backgrounds, but it will miss the useful background information of some domain pairs. To avoid losing recognition performance, our detector enhancement (DE) utilizes the classification information from the image space to enhance our detectors, which performs the best on all pairs.

4.1.4 Results on self-paced adaptation

It is worth noting that all our results are given by the models trained with data from the source domain only. Compared with other existing DA methods, we have not utilized the information from the target domain yet. As a consequence, in this experiment, our self-paced adaptation method retrains our models 4 times by setting top N as $[5, 10, 15, 20]$ respectively. For each time, we select the top N target samples in each class by the models learned in the previous time as training samples added to the source domain. In this process, our detectors are retrained according to the top 5 region proposals of each selected sample. The recognition accuracies are reported in Table 2. Compared with the baseline method which is learned in a similar way but without the detectors, the performance of our self-paced adaptation is gradually improved and is much better than the baseline. It demonstrates that the source domain is effectively adapted to the target domain according to our object alignment.

4.2 Experiment on Bing-Caltech dataset

In this section, we evaluate our method on a larger and more complicated Bing-Caltech dataset created by [5].

4.2.1 Experimental set-up

The Bing-Caltech dataset contains all 256 categories from Caltech dataset. Each category in Bing-Caltech is augmented with 300 web images which collected through textual search

Method	top N	A→C	A→D	A→W	C→A	C→D	C→W	W→A	W→C	W→D
Baseline	5	**84.7**	82.2	80.0	93.5	79.0	77.6	**77.7**	64.6	96.8
	10	84.2	84.7	83.1	93.9	80.9	78.0	**77.7**	**65.3**	**97.5**
	15	83.3	**86.0**	85.4	**94.1**	**82.2**	79.0	73.6	62.8	**97.5**
	20	83.2	**86.0**	**87.5**	93.9	**82.2**	**80.7**	74.1	60.7	**97.5**
Ours	5	86.2	87.9	84.1	93.7	87.3	84.1	86.2	76.9	98.1
	10	86.5	89.2	89.8	93.9	88.5	87.8	90.1	80.4	98.7
	15	86.5	**97.5**	90.8	**94.4**	94.3	90.5	92.0	82.1	**100.0**
	20	**86.7**	**97.5**	**91.5**	**94.4**	**96.8**	**93.6**	**92.8**	**83.0**	**100.0**

Table 2: Recognition accuracies of self-paced adaptation on 9 pairs of domains using the deep features.

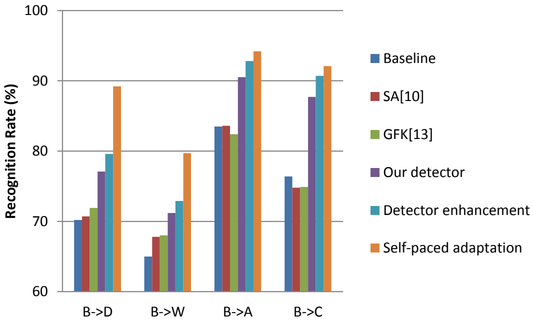


Figure 6: Performance of different methods on Bing-Caltech based cross-domain object recognition. This figure is best viewed in color.

using Bing. Thus, in Bing-Caltech, the image content include object and background is more complicated than the one in Office-Caltech dataset. However, we can see that there are many error labeled images in Bing-Caltech due to the textual search. In order to ensure the reliability of recognition accuracy, the Bing-Caltech cannot be considered as a test dataset. As a consequence, we can just take the Bing-Caltech as the source domain for cross-domain object recognition. Here, we take the domains in Office-Caltech as the target domains.

Our experiments are based on the data from the 10 common classes among Bing-Caltech (denoted by **B**) and the four domains in Office-Caltech. Similar to the previous one, all the images and region proposals are represented by the deep convolutional features. For Bing-Caltech, we set $K = 9$, $N = 50$ and $\alpha = 0.6$ for all domain pairs based on the cross-validation introduced before.

4.2.2 Recognition results

Figure 6 shows a comparison of the results of different methods on 4 source-target pairs. For simplicity, we set top N as 5 for our final self-paced adaptation method. As can be seen, our object alignment based object detector shows significant performance improvement compared to the two representative methods for domain adaptation. It proves that discounting the influence from the ambiguous background is an effective way for reducing the domain difference. Moreover, the recognition performance is further improved base on our detector enhancement. After that, by utilizing some unlabeled samples from target domain, our final self-paced adaptation method achieves the best performance over all domain pairs.

5 Conclusion

We have presented a novel solution for the problem of cross-domain object recognition. We have reduced the domain difference by discounting the influence from the ambiguous background using object alignment. In particular, we have presented an effective algorithm to eliminate the irrelevant background by aligning the semantic object, and learned detectors which simplify the problem of cross-domain object recognition. We have also utilized the classification information to enhance our detectors. Finally, based on the detectors, we have introduced a self-paced adaptation method to further reduce the domain difference. Our experimental results have demonstrated the benefits of our object alignment on minimizing domain difference. We have showed that our final approach achieved the state-of-the-art performance on several visual domain adaptation datasets.

Acknowledgment

This work is funded by the National Basic Research Program of China (Grant No. 2012CB316302), National Natural Science Foundation of China (Grant No. 61135002), and the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant XDA06040102).

References

[1] B. Alexe, T. Deselaers, and V. Ferrari. What is an object? In *CVPR*, 2010.

[2] A. Angelova and Shenghuo Zhu. Efficient object detection and segmentation for fine-grained recognition. In *CVPR*, 2013.

[3] Mahsa Baktashmotlagh, Mehrtash T. Harandi, Brain C. Lovell, and Mathieu Salzmann. Unsupervised domain adaptation by domain invariant projection. In *ICCV*, 2013.

[4] Mahsa Baktashmotlagh, Mehrtash T. Harandi, Brain C. Lovell, and Mathieu Salzmann. Domain adaptation on the statistical manifold. In *CVPR*, 2014.

[5] Alessandro Bergamo and Lorenzo Torresani. Exploiting weakly-labeled web images to improve object classification: a domain adaptation approach. In *NIPS*, 2010.

[6] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[7] John Blitzer, Ryan McDonald, and Fernando Pereira. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, 2006.

[8] J. Carreira and C. Sminchisescu. Constrained parametric min-cuts for automatic object segmentation. In *CVPR*, 2010.

[9] J. Deng, A. Berg, S. Satheesh, A. Khosla H. Su, and L. Fei-Fei. Imagenet large scale visual recognition competition 2012 (ILSVRC2012). http://www.image-net.org/challenges/LSVRC/2012/.

[10] Basura Fernando, Amaury Habrard, Marc Sebban, and Tinne Tuytelaars. Unsupervised visual domain adaptation using subspace alignment. In *ICCV*, 2013.

[11] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014.

[12] B. Gong, Y. Shi, F. Sha, and K. Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *CVPR*, 2012.

[13] B Gong, K Grauman, and F Sha. Connecting the dots with landmarks: Discriminatively learning domain-invariant features for unsupervised domain adaptation. In *ICML*, 2013.

[14] Raghuraman Gopalan, Ruonan Li, and Rama Chellappa. Domain adaptation for object recognition: An unsupervised approach. In *ICCV*, 2011.

[15] G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. *Technical Report*, 2007.

[16] J Hoffman, E Rodner, J Donahue, K Saenko, and T Darrell. Efficient learning of domain-invariant image representations. In *International Conference on Learning Representations*, 2013.

[17] Judy Hoffman, Trevor Darrell, and Kate Saenko. Continues manifold based adaptation for evolving visual domains. In *CVPR*, 2014.

[18] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 43:177–196, 2001.

[19] Jiayuan Huang, Alexander J. Smola, Arthur Gretton, Karsten M. Borgwardt, and Bernhard Schölkopf. Correcting sample selection bias by unlabeled data. In *NIPS*, 2007.

[20] I.-H. Jhuo, D. Liu, D. T. Lee, and S.-F. Chang. Robust visual domain adaptation with low-rank reconstruction. In *CVPR*, 2012.

[21] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[22] Jie Ni, Qiang Qiu, and Rama Chellappa. Subspace interpolation via dictionary learning for unsupervised domain adaptation. In *CVPR*, 2013.

[23] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang. Domain adaptation via transfer component analysis. In *IJCAI*, 2009.

[24] Q. Qiu, V. Patel, P. Turage, and R. Chellappa. Domain adaptive dictionary learning. In *ECCV*, 2012.

[25] Olga Russakovsky, Sean Ma, Jonathan Krause, Jia Deng, Alex Berg, and L. Fei-Fei. Imagenet large scale visual recognition competition 2014 (ILSVRC2014). http://www.image-net.org/challenges/LSVRC/2014/.

[26] Bryan C. Russell, Alexei A. Efros, Josef Sivic, William T. Freeman, and Andrew Zisserman. Using multiple segmentations to discover objects and their extent in image collections. In *CVPR*, 2006.

[27] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *ECCV*, 2010.

[28] Sumit Shekhar, Vishal M. Patel, Hien V. Nguyen, and Rama Chellappa. Generalized domain-adaptive dictionaries. In *CVPR*, 2013.

[29] Y. Shi and F. Sha. Information-theoretical learning of discriminative clusters for unsupervised domain adaptation. In *ICML*, 2012.

[30] Josef Sivic, Bryan C. Russell, Alexei A. Efros, Andrew Zisserman, and William T. Freeman. Discovering objects and their location in images. In *ICCV*, 2005.

[31] Koen E. A. van de Sande, Jasper R. R. Uijlings, Theo Gevers, and Arnold W. M. Smeulders. Segmentation as selective search for object recognition. In *ICCV*, 2011.

[32] C. Wang and S. Mahadevan. Manifold alignment without correspondence. In *IJCAI*, 2009.

[33] C. Wang and S. Mahadevan. Heterogeneous domain adaptation using manifold alignment. In *IJCAI*, 2011.

[34] Chong Wang, Weiqiang Ren, Kaiqi Huang, and Tieniu Tan. Weakly supervised object localization with latent category learning. In *ECCV*, 2014.

[35] Jiaolong Xu, S. Ramos, D. Vazquez, and A.M. Lopez. Domain adaptation of deformable part-based models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36:2367–2380, 2014.