Unsupervised Visual Domain Adaptation Using Auxiliary Information in Target Domain

Masaya Okamoto
Grad. School of Information Science and Technology
The University of Tokyo
Tokyo, Japan
Email: okamoto@nlab.ci.i.u-tokyo.ac.jp

Hideki Nakayama
Grad. School of Information Science and Technology
The University of Tokyo
Tokyo, Japan
Email: nakayama@ci.i.u-tokyo.ac.jp

***Abstract*—We propose a novel approach for unsupervised visual domain adaptation that exploits auxiliary information in a target domain. The key idea is to embed data in the target domain into a subspace where samples are better organized, expecting auxiliary information to serve as a somewhat semantically related signal. Specifically, we apply partial least squares(PLS) to RGB image features and corresponding depth features captured at the same time. Thus, we can improve the performance of domain adaptation without any help from manual annotation in the target domain. In experiments, we tested our approach with two state-of-the-art subspace based domain adaptation methods and show that, our method consistently improves the classification accuracy.**

***Keywords*-domain adaptation; transfer learning; multi-modal;**

I. Introduction

Visual object recognition is one of the most fundamental technologies in areas of multimedia. To successfully train a recognition system, we generally require a lot of supervised, i.e., manually annotated, training images specifically prepared for each target environment. However, hand labeling training datasets is quite time consuming and is regarded as a bottleneck in practical situations. Therefore, in recent years, visual domain adaptation, which was first proposed by Saenko et al. [1], has gathered more and more attention. The objective of domain adaptation is to transfer a classifier obtained from labeled examples in one domain to another domain. The domain where a classifier is trained is called the "source domain" and is expected to provide a lot of labeled data. The domain in which the classifier is actually tested is called the "target domain" and is assumed to have different characteristics in its nature, e.g., illumination and resolution, from the source domain. Figure 1 shows an example of the difference between two domains.

Among the many approaches to tackle this problem, unsupervised domain adaptation, where no labeled example is assumed in the target domain, is attractive for its practicality and has been intensively studied. Most previous work on unsupervised domain adaptation has typically used only visual information in the target domain, making the problem extremely challenging[1].

Figure 1. Difference between source and target domains

However, nowadays, we can easily get rich multimedia information in addition to RGB images such as GPS and gyroscopes owing to recent remarkable advances in wearable/sensing devices. Although this kind of auxiliary information has not been paid much attention in the context of visual domain adaptation, it is expected to work somewhat as semantically related signals and relax the extreme difficulty of unsupervised domain adaptation. In particular, considering that depth sensors are now a common technology, e.g., Kinect, distance information is thought to be the most promising and practical auxiliary information for supporting visual domain adaptation.

Motivated by these reasons, in this study, we propose a novel approach of unsupervised domain adaptation that exploits not only visual information but also distance information as an auxiliary signal to boost the performance. We implement our approach with two subspace based methods and demonstrate its effectiveness by using a real-world dataset for testing.

II. Related Work

The first work on domain adaptation for visual object recognition was proposed by Saenko et al. [1]. This method is based on information-theoretic metric learning (ITML) [2], which is a Mahalanobis distance metric learning method

[1]We might also assume a few labeled examples in a semi-supervised case, although this is not the scope of this work.

for letting the distance between samples be smaller if they belong to the same class, otherwise larger. This is a semi-supervised method where a large number of labeled examples is available in the source domain and also a few labeled ones are provided in the target domain.

Considering that our objective is to reduce the cost of manual labeling, an unsupervised setting, where no labeled example is used in the target domain, is the ultimate goal of domain adaptation, although it is essentially a quite difficult task. For unsupervised domain adaptation, the subspace based approach has been known to be a promising strategy, where multiple intermediate subspaces between source and target ones are generated as "virtual" domains that blend the properties of the source and target. This approach was first proposed by Gopalan et al. [3] as the geodesic flow sampling (GFS) method. This method creates intermediate subspaces by sampling points from the geodesic flow on the Grassmann manifold from the source to the target subspaces. One problem of GFS is the trade-off between performance and the dimensions of feature vectors that depend on a number of sampled intermediate subspaces. Namely, to improve the performance, we need to take more intermediate subspaces, but this results in higher computational costs. Some methods are proposed to relax this problem.

The geodesic flow kernel (GFK) [4] is based on an analytic solution of what GFS has done in the sampling based approach. As an another example of a subspace based method, Fernando et al. proposed the unsupervised visual domain adaptation using subspace alignment (SA) method [5]. It shows that a transformation matrix that best matches the source subspace to the target one can be obtained in a simple closed form, leading to an extremely fast algorithm.

The subspace based approach is probably the current most successful approach for the unsupervised domain adaptation problem. In this framework, it is important source and target subspaces are constructed for adaptation. It has been shown that applying PLS [6] analysis to build the subspace of the source domain improves the final classification accuracy more than applying PCA instead. This is probably because PLS can make a better subspace by using corresponding semantic information (category label in their case), and thus, lead to a better knowledge transfer. However, since labels are provided in the source domain only, there has been no choice but to apply PCA for the target domain. In this study, we propose a novel approach of applying PLS in the target domain by using visual information and corresponding auxiliary information instead of labels. This is a fundamental concept and can be broadly applied to any subspace based domain adaptation methods.

III. Proposed Method

A. Concept

In previous work on unsupervised domain adaptation, it has been typically assumed that we have a lot of unlabeled images in the target domain. In the near feature, it will not be difficult to assume that auxiliary signals are provided naturally thorough multimedia devices, as we discussed in the introduction.

We improve the subspace based visual domain adaptation methods by applying PLS analysis over visual information and auxiliary information in the target domain. Figure 2 shows the difference between our method and the previous work.

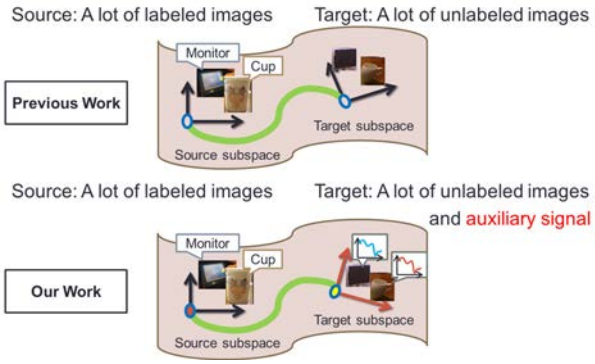

Figure 2. Approaches of our approach and previous work

B. Partial Least Squares

Partial least squares (PLS) was proposed by Wold et al. [6] in the field of chemometrics. It is a statistical multivariate analysis method for finding a latent subspace that bears some relation to principal component analysis.

While PCA finds a subspace that preserves the variance of one observed variable (visual features), PLS is used to generate a subspace that has relationships between two variables, i.e., visual features and labels, or visual features and auxilary information. More specifically, PLS maximizes the covariance of two variables projected into the latent subspace. Therefore, a subspace obtained via PLS is expected to capture the essential latent structure that bridges the two observed variables.

C. Process Flow

Figure 3 shows the process flow of our method. It consists of five parts as follows.

1) Extract the image features from RGB images in both domains.
2) Apply PLS analysis in source domain to obtain the source subspace by using label information.
   Note that the idea of applying PLS is proposed in the previous work and is not new itself. In the experiments (Section IV), we also test PCA for building source subspaces and investigate their performances.
3) Extract the distance features from depth images.
4) Apply PLS analysis in the target domain to obtain the target subspace by using distance features. This is the most important process of our method.

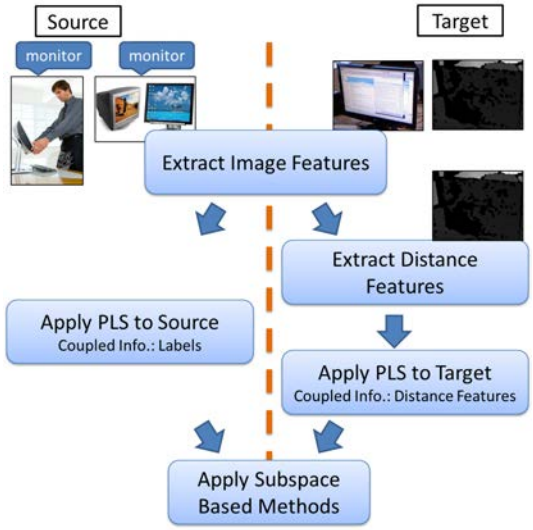

Figure 3. Process flow of our method. Image and label pairs are provided in the source domain, while image and depth image pairs are provided in the target domain.

We assume that the features from depth images that interrelate with their classes serve as cues to obtain a better subspace for the target domain. More specifically, we expect that samples from the same hidden concept, i.e., class, are located near in this subspace.

5) Apply a subspace based domain adaptation method on top of the learned source and target subspaces. In experiments, we exploit two methods: GFK and SA. We exploit publicly available codes provided by the authors of each method.

Finally, we conduct classification by using the transferred representation obtained by our method. In the experiments, we use the nearest neighbor algorithm as it is the simplest classifier and suitable for investigating the pure performance of domain adaptation.

IV. Experiments

A. Dataset

We used ImageNet [7] as the source dataset and the RGB-D object dataset (B3DO) [8] as the target. The latter provides depth images corresponding to each RGB image. We chose six classes that commonly appear in both B3DO and ImageNet. We cropped the images and depth images of B3DO by using the provided bounding boxes of objects. Table I summarizes the details of our dataset. Figure 4 shows examples of dataset images.

B. Setup

The core contribution of our study is applying PLS to the target domain. The objective of our experiments is to prove its effectiveness, for which we compare the four combinations of source and target subspaces as follows.

Table I
Number of samples

Class	B3DO (target)	ImageNet (source)
bottle	238	920
bowl	142	919
cup	256	919
keyboard	129	1512
monitor	243	1134
sofa	109	982
SUM	1119	6386
AVG	186.5	1064.3

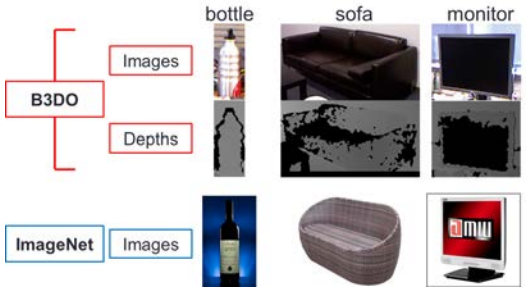

Figure 4. Examples of dataset images. For visibility, we adjusted the dynamic range of depth images in this figure. In the experiments, we use raw depth images of the B3DO dataset (best viewed in color).

1) OURS1 (Source: PCA Target: PLS)
2) Baseline1 (Source: PCA Target: PCA)
3) OURS2 (Source: PLS Target: PLS)
4) Baseline2 (Source: PLS Target: PCA)

The methods applying PLS analysis in the target domain correspond to the proposed methods, while the methods applying PCA on the target domain correspond to the baselines. In all cases, we tested two independent subspace based domain adaptation methods: GFK and SA.

The comparison of OURS1 and Baseline1 illustrates the effectiveness of our approach when PCA was used for building the source subspace. Similarly, OURS2 and Baseline2 are comparable when PLS was used in the source domain. We expected to observe the respective improvements in each case. We experimentally chose dimensions of subspaces among 10, 20, 30, 40, and 50 that maximize the classification accuracy for each case because fixed dimensions may bias a particular method to work better.

For visual features, we extracted dense SIFT features [9] and created a bag-of-words dictionary of 1000 words by using only source (ImageNet) images. We then obtained a 1000 dimensional image representation. As for distance features, we exploited the kernel descriptors [10] proposed by Bo et al. In their work, they proposed two types of features; one is based on both RGB and depth images, and the other is based on depth images only. In our experiment, we used the latter one, and obtained 14000-d depth representations. We used the public codes provided by the authors[2].

[2]http://www.cs.washington.edu/robotics/projects/kdes/

We changed the numbers of source samples from 20, 50, 100, and 300 to 500 per class (total: 120, 300, 600, 1800, and 3000).

C. Results

Table II
RECOGNITION ACCURACY OF SOURCE SAMPLES EXPERIMENT

	OURS1	Baseline1	OURS2	Baseline2
120 (GFK)	28.33	**28.95**	**32.35**	31.64
300 (GFK)	29.31	**29.85**	**32.71**	31.55
600 (GFK)	**29.04**	28.60	**32.53**	28.87
1800 (GFK)	**32.17**	30.92	**34.32**	31.81
3000 (GFK)	**33.42**	31.72	**34.94**	33.92
Exec. time (GFK)	3.83	2.26s	135.17s	128.03s
120 (SA)	**34.05**	29.85	**34.23**	30.83
300 (SA)	**33.15**	30.21	**32.17**	31.90
600 (SA)	**33.78**	33.15	**33.33**	32.71
1800 (SA)	**33.15**	30.21	**32.17**	31.90
3000 (SA)	**34.85**	32.44	**33.69**	32.89
Exec. time (SA)	3.07s	0.98s	130.90s	120.30s

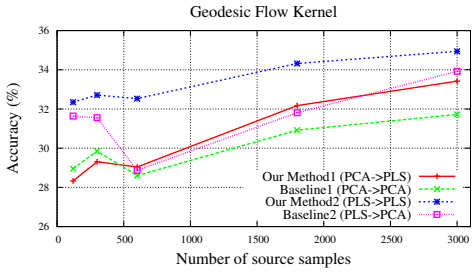

Figure 5. Result on geodesic flow kernel (best viewed in color)

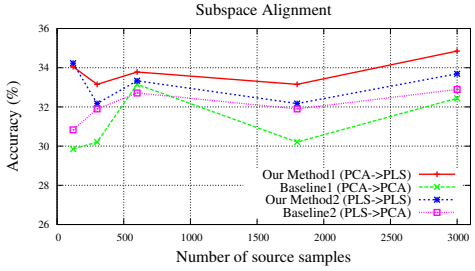

Figure 6. Result on subspace alignment

Table II shows the result with a different number of source samples, and the average execution times of each method on 120 source samples case. Figure 5 and Figure 6 show the results of our methods for GFK and SA, respectively. For both GFK and SA, our methods (OURS1 and OURS2) that applied PLS analysis in the target domain outperformed the baselines. It is notable that our approach improved the performance in all four test cases of the combinations of the domain adaptation method (GFK and SA) and source subspace method (PCA and PLS), indicating its consistent effectiveness.

V. CONCLUSION

We proposed a novel approach of unsupervised visual domain adaptation that improved subspace based methods with auxiliary information. In experiments, we showed that the proposed approach consistently outperformed the previous ones over two independent subspace based domain adaptation methods. We demonstrated the consistent effectiveness of our method in several situations in which the number of source samples was changed. To the best of our knowledge, we proposed the first visual domain adaptation method that utilizes auxiliary information in a target domain.

Although we used only distance features as auxiliary information, our method is generic and can be used with any weakly coupled auxiliary information in theory. Considering that other multimedia information such as sounds and GPS can now be easily obtained, we plan to evaluate whether they can also be used for domain adaptation. In addition, it would be interesting to exploit many types of multimedia information jointly to overcome domain gaps more effectively.

ACKNOWLEDGEMENTS

This work was supported by JST, CREST and the Okawa foundation research grant.

REFERENCES

[1] K. Saenko, B. Kulis, M. Fritz, and T. Darrell, "Adapting visual category models to new domains," in *Proc. of ECCV*, 2010.

[2] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, "Information-theoretic metric learning," in *Proc. of ICML*, 2007.

[3] R. Gopalan, R. Li, and R. Chellappa, "Domain adaptation for object recognition: an unsupervised approach," in *Proc. of ICCV*, 2011.

[4] B. Gong, Y. Shi, and F. Sha, "Geodesic flow kernel for unsupervised domain adaptation," in *Proc. of CVPR*, 2012.

[5] B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars, "Unsupervised visual domain adaptation using subspace alignment," in *Proc. of ICCV*, 2013.

[6] H. Wold, S. Kotz, and N. L. Johnson, "Partial least squares," in *Encyclopedia of Statistical Sciences*, 1985.

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: a large-scale hierarchical image database," in *Proc. of CVPR*, 2009.

[8] A. Janoch, S. Karayev, Y. Jia, J. Barron, M. Fritz, K. Saenko, and T. Darrell, "A category-level 3-d object dataset: putting the kinect to work," in *Proc. of ICCV Workshop on Consumer Depth Cameras in Computer Vision*, 2011.

[9] D. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[10] L. Bo, X. Ren, and D. Fox, "Depth kernel descriptors for object recognition," in *Proc. of IROS*, 2011.