Appears in: Big Data Meets Computer Vision: First International Workshop
on Large Scale Visual Recognition and Retrieval. In conjunction with NIPS 2012.

Overcoming Dataset Bias: An Unsupervised Domain Adaptation Approach

Boqing Gong
Dept. of Computer Science
U. of Southern California
Los Angeles, CA 90089
boqinggo@usc.edu

Fei Sha
Dept. of Computer Science
U. of Southern California
Los Angeles, CA 90089
feisha@usc.edu

Kristen Grauman
Dept. of Computer Science
U. of Texas at Austin
Austin, TX 78701
grauman@cs.utexas.edu

Abstract

Recent studies have shown that recognition datasets are biased. Paying no heed to those biases, learning algorithms often result in classifiers with poor cross-dataset generalization. We are developing domain adaptation techniques to overcome those biases and yield classifiers with significantly improved performance when generalized to new testing datasets. Our work enables us to continue to harvest the benefits of existing vision datasets for the time being. Moreover, it also sheds insights about how to construct new ones. In particular, domain adaptation raises the bar for collecting data — the most informative data are those which cannot be classified well by learning algorithms that *adapt* from existing datasets.

1 Introduction

Datasets are of paramount importance to visual recognition research. We use them extensively to train and evaluate learning algorithms and features, in the hope that they provide objective guidance for constructing robust classifiers.

This notion can no longer be taken for granted. Several recent studies have shown that instead of being objective, datasets are often *biased*—even when they appear to be neutrally composed of images from the same visual categories. The biases can be attributed to many exogenous factors in data collection, such as cameras, preferences over certain types of backgrounds, or annotator tendencies. Dataset biases adversely affect cross-dataset generalization; that is, the performance of a classifier trained on one dataset drops significantly when applied to another one [1, 2, 3]. Thus, instead of building classifiers that discriminate visual categories irrespective of dataset origins, our learning algorithms overfit on the datasets' idiosyncrasies and yield dataset-specific visual classifiers!

Given those discouraging results, it is only natural to doubt the value of biased datasets. In particular, *should we trust and continue to utilize existing datasets?* Our answer is a relieving and positive *yes*. Our goal is to *overcome the biases* so that existing datasets can still be instructive as training data in building robust classifiers with good generalization properties.

To this end, we develop powerful learning algorithms to reduce the idiosyncrasies in the training datasets. Concretely, we model the effect of biases as causing mismatches in datasets' distributions, and cast the problem as one of rectifying the mismatch between domains [4, 5, 6, 7, 8, 9, 10, 11, 12, 13]. Our key idea is to identify *domain invariant* features such that the training dataset (i.e., the source domain) and the testing dataset (i.e., the target domain) will be similar to each other. With such features, classifiers trained on the training dataset will also perform well on the target domain.

In this paper, we describe two different but closely related domain adaptation approaches exploring that idea. We will first describe our previous work of learning invariant features via the geodesic flow kernel (GFK) [14]. This method represents domains with low-dimensional subspaces and identifies different domains as points on a Grassmann manifold. It then models changes between domains as

geodesic paths between the points. Intuitively, the path encodes incremental changes in geometric and statistical properties for one domain "morphing" into the other. GFK then computes features that are insensitive to those changes, facilitating the adaptation between the source and target domains.

Our second approach advances the core idea behind GFK further. Instead of relying on a single kernel, we construct multiple kernels. Specifically, for the original pair of domains, we create multiple auxiliary pairs of domains. Adaptation between those auxiliary pairs is easier than in the original pair, as we construct them explicitly to have reduced distributional mismatch. The GFKs for those auxiliary pairs thus inform how to construct features for the adaptation task on the original pair. Specifically, we learn to combine their induced features discriminatively such that the final features are optimized on the original target domain.

Our approaches are advantageous both in terms of the visual recognition application and unsupervised domain adaptation in general. We do not assume labeled examples in the target domain (though our approaches can be extended easily to use them). The algorithms are virtually free of parameter-tuning, reducing the need for cross-validation and associated computation costs. Computing the proposed kernels is also scalable to very large datasets, requiring only matrix eigendecompositions. Our methods achieve the state-of-the-art performance for unsupervised domain adaptation, and they are sometimes even superior to methods that require labeled examples in the target domain.

2 Proposed Approaches for Unsupervised Domain Adaptation

Intuitively, datasets are biased because features encode information not only intrinsic to visual categories but also relevant to dataset-specific exogenous factors. Due to those factors, features are distributed differently across datasets. Correspondingly, classifiers optimized under one distribution will generalize poorly in other different distributions. Correcting the distribution mismatch is known as ***domain adaptation*** in the literature of statistics and machine learning [4, 5, 6, 7].

Thus, we cast overcoming dataset biases as an instance of domain adaptation, where one dataset is the source domain and the other is the target domain [8, 9, 10, 11, 12, 13]. For example, the source domain could be a benchmark dataset from the recognition community, while the target domain could consist of novel images taken on a mobile phone application. We focus on ***unsupervised domain adaptation*** where the target domain does not provide labels. The key challenge is then to extract domain-invariant features so as to reduce the mismatch between the two domains.

We have developed two approaches to address this challenge. The core idea is to derive kernels (which implicitly define feature mappings) with desirable properties. We start by describing our previous work of the geodesic flow kernel (GFK) [14]. We then describe how to improve the performance of GFK-based domain adaptation using discriminative training of multiple kernels.

2.1 Geodesic flow kernel (GFK)

The GFK technique models each domain with a d-dimensional linear subspace and embeds them onto a Grassmann manifold. Specifically, let $\boldsymbol{P}_S, \boldsymbol{P}_T \in \mathbb{R}^{\mathsf{D}\times\mathsf{d}}$ denote the basis of the PCA subspaces for each of the two domains, respectively. The Grassmann manifold $\mathbb{G}(\mathsf{d}, \mathsf{D})$ is the collection of all d-dimensional subspaces of the feature vector space $\mathbb{R}^{\mathsf{D}}$.

The geodesic flow $\{\boldsymbol{\Phi}(t) : t \in [0,\ 1]\}$ between $\boldsymbol{P}_S$ and $\boldsymbol{P}_T$ on the manifold parameterizes a path connecting the two subspaces. Every point on the flow is a basis of a d-dimensional subspace. In the beginning of the flow, the subspace is similar to $\boldsymbol{P}_S = \boldsymbol{\Phi}(0)$ and in the end of the flow, the subspace is similar to $\boldsymbol{P}_T = \boldsymbol{\Phi}(1)$. Thus, the flow can be seen as a collection of infinitely many subspaces varying gradually from the source to the target domain. The original feature $\boldsymbol{x}$ is projected into these subspaces and forms a feature vector of infinite dimensions: $\boldsymbol{z}^\infty = \{\boldsymbol{\Phi}(t)^{\mathsf{T}}\boldsymbol{x} : t \in [0,\ 1]\}$.

Using the new feature representation for learning will force the classifiers to be less sensitive to domain differences and to use domain-invariant features. In particular, the inner products of the new features give rise to a positive semidefinite kernel defined on the original features:

$$G(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle \boldsymbol{z}_i^\infty, \boldsymbol{z}_j^\infty \rangle = \boldsymbol{x}_i^{\mathsf{T}} \int_0^1 \boldsymbol{\Phi}(t)\boldsymbol{\Phi}(t)^{\mathsf{T}}\, dt\, \boldsymbol{x}_j = \boldsymbol{x}_i^{\mathsf{T}} \boldsymbol{G} \boldsymbol{x}_j. \tag{1}$$

The matrix $\boldsymbol{G}$ can be computed efficiently using singular value decomposition on $\boldsymbol{P}_S^{\mathsf{T}}\boldsymbol{P}_T$. Moreover, computing the kernel does not require any labeled data. The only free parameter is the dimensionality d of the subspace, which we show how to infer automatically. Details are in [14].

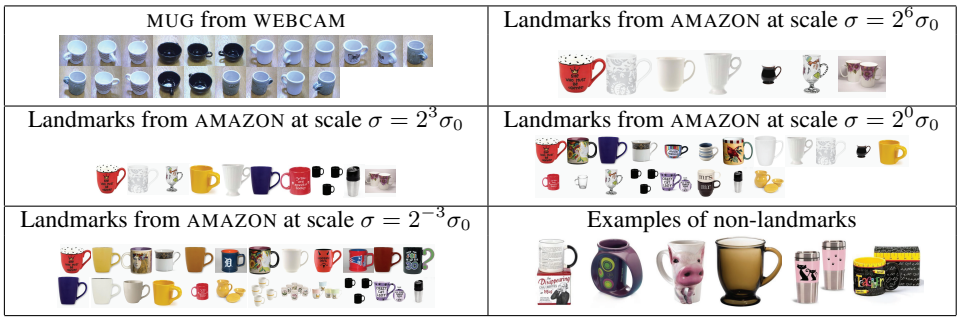

Figure 1: Landmarks selected from the source domain AMAZON for the target WEBCAM, as well as non-landmarks. As the scale decreases, images with greater variance in appearance are selected.

Our previous study has demonstrated the significant advantage of geodesic flow kernels over other competing methods; section 3 gives a snapshot of those empirical results. In what follows, we describe a new approach to advance the core idea behind GFK further.

2.2 Discriminative learning of multiple cross-domain kernels

The strength of the GFK — requiring no labeled data from the target domain — can sometimes also be perceived as its shortcoming. It is not clear from the construction of the GFK whether the learned domain-invariant features aim directly to minimize classification error on the target domain. *Yet, how can we learn **discriminative** domain-invariant features for **unsupervised** domain adaptation?*

Our first insight to answer this seemingly oxymoronic question is that in the source domain, there are data points we call *landmarks* that are distributed in such a way that they look like they could be sampled from the target domain. Fig. 1 displays several discovered landmark images for the datasets we use in this work. Our intuition is to discriminatively optimize the performance of the adapted classifiers on these *labeled* landmarks as a proxy to the true errors on the target domain.

Our second insight is to exploit those landmarks further but *without* their labels to construct multiple auxiliary domain adaptation tasks. Those auxiliary tasks are easier to solve as we purposely use the landmarks to bridge the source and the target domains in those tasks. Each one of those tasks gives rise to a GFK kernel that implies a domain-invariant feature mapping. We then discriminatively combine those mappings in the framework of multiple kernel learning.

In the following, we summarize several key steps, with details to appear in a longer version.

Identifying landmarks We use a variant of maximum mean discrepancy (MMD) [15] to select samples from the source domain to match the distribution of the target domain. Let $\mathcal{D}_S = \{(\boldsymbol{x}_m, y_m)\}_{m=1}^{\mathsf{M}}$ denote M data points and their labels from the source domain and likewise $\mathcal{D}_T = \{\boldsymbol{x}_n\}_{n=1}^{\mathsf{N}}$ for the target domain. We use $\boldsymbol{\alpha} = \{\alpha_m \in \{0,1\}\}$ to denote M indicator variables, one for each data point in the source domain. If $\alpha_m = 1$, then $\boldsymbol{x}_m$ is regarded as a landmark. We identify α_m by minimizing the MMD metric, defined with a kernel mapping function $\phi(\boldsymbol{x})$,

$$\min_{\boldsymbol{\alpha}} \left\| \frac{1}{\sum_m \alpha_m} \sum_m \alpha_m \phi(\boldsymbol{x}_m) - \frac{1}{\mathsf{N}} \sum_n \phi(\boldsymbol{x}_n) \right\|_{\mathcal{H}}^2 \quad \text{s.t.} \quad \frac{1}{\sum_m \alpha_m} \sum_m \alpha_m y_{mc} = \frac{1}{\mathsf{M}} \sum_m y_{mc}, \quad (2)$$

where y_{mc} denotes the indicator variable for $y_m = c$. Note that the right-hand-side of the constraint is simply the prior probability of the class c, estimated from the source domain. The constraint is used to avoid the case that some categories dominate the selected landmarks. We solve the intractable eq. (2) with linear relaxation; details are omitted for brevity.

We use the geodesic flow kernel computed between the source $\mathcal{D}_S$ and the target $\mathcal{D}_T$, as defined in eq. (1), to compose the kernel mapping function $\phi(\boldsymbol{x})$

$$\phi(\boldsymbol{x}_i)^{\mathrm{T}} \phi(\boldsymbol{x}_j) = K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\{-d^2_{\boldsymbol{G}}(\boldsymbol{x}_i, \boldsymbol{x}_j)/\sigma^2\} = \exp\{-(\boldsymbol{x}_i - \boldsymbol{x}_j)^{\mathrm{T}} \boldsymbol{G} (\boldsymbol{x}_i - \boldsymbol{x}_j)/\sigma^2\}. \quad (3)$$

Constructing auxiliary tasks The bandwidth σ in the kernel eq. (3) is a scaling factor for measuring similarities at different granularities. We use a set of factors $\{\sigma_q \in [\sigma_{min},\ \sigma_{max}]\}_{q=1}^{\mathsf{Q}}$. For each σ_q, we solve eq. (2) to obtain the corresponding landmarks $\mathcal{L}^q$ whose α_m is 1.

Table 1: Comparison of our newly proposed approach (MKL) to the baseline and existing methods for unsupervised domain adaptation. MKL performs the best on 8 out of 9 pairs, while our previous approach (GFK) [14] is the second best. C: CALTECH, A: AMAZON, W: WEBCAM, D: DSLR.

	A→C	A→D	A→W	C→A	C→D	C→W	W→A	W→C	W→D
SRC ONLY	26.0	25.5	29.8	23.7	25.5	25.8	23.0	19.9	59.2
GFS [13]	39.2	36.3	33.6	43.6	40.8	36.3	*33.5*	*30.9*	75.7
TCA [7]	35.0	36.3	27.8	41.4	*45.2*	32.5	24.2	22.5	**80.2**
GFK [14] (ours)	*42.2*	*42.7*	*40.7*	*44.5*	43.3	*44.7*	31.8	*30.8*	75.6
GFK+MKL (ours)	**45.5**	**47.1**	**46.1**	**56.7**	**57.3**	**49.5**	**40.2**	**35.4**	75.2

For each set of landmarks, we construct a new domain pair by moving the landmarks from the original source to the target domains, yielding the source domain $\mathcal{D}_S \setminus \mathcal{L}^q$ and the target domain $\mathcal{D}_T \bigcup \mathcal{L}^q$. Arguably, each auxiliary task is "easier" to adapt than the original pair $\mathcal{D}_S$ and $\mathcal{D}_T$, due to the increased distributional match between the new source and target domains.

Discriminative multiple kernel learning We learn the final kernel as a convex combination of all the kernels from the auxiliary tasks: $\boldsymbol{F} = \sum_q w_q \boldsymbol{G}_q$, where $\boldsymbol{G}_q$ is the GFK for the q-th auxiliary task. The coefficients w_q are optimized on a labeled training set $\mathcal{D}_{\text{TRAIN}} = \sum_q \mathcal{L}^q$, composed of all landmarks selected at different granularities. We use $\boldsymbol{F}$ in a support vector machine classifier whose accuracy is optimized with the standard multiple kernel learning algorithm to learn w_q [16]. Intuitively, since landmarks are distributed similarly to the target, we expect the classification error on $\mathcal{D}_{\text{TRAIN}}$ to be a good proxy to that of the target.

3 Experimental Results

We evaluate the proposed methods on benchmark datasets extensively used in domain adaptation for object recognition [13, 11, 12]. We use four datasets: CALTECH [17], AMAZON, WEBCAM, and DSLR [11]. Each dataset is distinctly biased: objects in the CALTECH are mostly centered with clean backgrounds; AMAZON is collected from online catalogs, WEBCAM and DSLR were taken in office environments with different resolutions. We follow the same procedure in [11] to prepare our data.

Table 1 reports the adapted classifiers' accuracies on the target domains under 9 adaptation tasks. We contrast our methods (geodesic flow kernel (GFK) [14] and landmark-based multiple kernel learning (MKL)), to a baseline where there is no adaptation (SRC ONLY), as well as two leading methods for domain adaptation: geodesic flow sampling (GFS) [13] and transfer component analysis (TCA) [7].

In most cases, domain adaptation techniques improve over classifiers without being adapted. The best performing method is MKL, outperforming all others in 8 out of 9 pairs. Our previous approach GFK is the second best performing method. More results and details can be found at http://rainflower.usc.edu/projects/domainadaptation.

4 Conclusion

Despite the extensive effort in collecting data in both volume and diversity, dataset biases will remain as a challenging problem in computer vision for a long period of time, due to the combinatorial explosion of too many exogenous factors. Moreover, in many practical applications, we may want the classifiers to perform well on a specific target distribution, instead of on all possible distributions.

We have developed unsupervised domain adaptation techniques to overcome the biases. We show the empirical success of the proposed methods. We believe that this will be a fruitful direction for future research, complementing the effort of building large-scale unbiased datasets. In particular, our work raises the bar for collecting data—we should scrutinize and aim only for data which cannot be classified well by learning algorithms that *adapt* from existing datasets, as those data will be the most informative addition.

Acknowledgements

This work is partially supported by DARPA D11AP00278 and NSF IIS 1065243 (B.G. and F.S.), and NSF IIS 1065390 (K.G.).

References

[1] A. Torralba and A.A. Efros. Unbiased look at dataset bias. In *CVPR*, 2011.

[2] F. Perronnin, J. Senchez, and Y. Liu. Large-scale image categorization with explicit data embedding. In *CVPR*, 2010.

[3] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: A benchmark. In *CVPR*, 2009.

[4] S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira. Analysis of representations for domain adaptation. *NIPS*, 2007.

[5] J. Blitzer, R. McDonald, and F. Pereira. Domain adaptation with structural correspondence learning. In *EMNLP*, 2006.

[6] J. Blitzer, M. Dredze, and F. Pereira. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *ACL*, 2007.

[7] S.J. Pan, I.W. Tsang, J.T. Kwok, and Q. Yang. Domain adaptation via transfer component analysis. *IEEE Trans. NN*, (99):1–12, 2009.

[8] L. Duan, IW Tsang, D. Xu, and SJ Maybank. Domain transfer svm for video concept detection. In *CVPR*, 2009.

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

[10] L. Bruzzone and M. Marconcini. Domain adaptation problems: A DASVM classification technique and a circular validation strategy. *IEEE Trans. PAMI*, 32(5):770–787, 2010.

[11] K. Saenko, B. Kulis, M. Fritz, and T. Darrell. Adapting visual category models to new domains. In *ECCV*, 2010.

[12] B. Kulis, K. Saenko, and T. Darrell. What you saw is not what you get: Domain adaptation using asymmetric kernel transforms. In *CVPR*, 2011.

[13] R. Gopalan, R. Li, and R. Chellappa. Domain adaptation for object recognition: An unsupervised approach. In *ICCV*, 2011.

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

[15] A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola. A kernel method for the two-sample-problem. In *NIPS*. 2006.

[16] G. R. G. Lanckriet, N. Cristianini, P. Bartlett, L. El Ghaoui, and M. Jordan. Learning the kernel matrix with semidefinite programming. *JMLR*, 5:27–72, December 2004.

[17] G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. Technical report, California Institute of Technology, 2007.