Supplementary Material for Connecting the Dots with Landmarks: Discriminatively Learning Domain-Invariant Features for Unsupervised Domain Adaptation

Boqin Gong BOQINGGO@USC.EDU
Department of Computer Science, University of Southern California, Los Angeles, CA 90089

Kristen Grauman GRAUMAN@CS.UTEXAS.EDU
Department of Computer Science, University of Texas at Austin, Austin, TX 78701

Fei Sha FEISHA@USC.EDU
Department of Computer Science, University of Southern California, Los Angeles, CA 90089

In this Supplementary Material, we provide extra details on the following:

- Sec. A: proof of Theorem 1 in the main text. We prove that the new target distribution is closer to the source distribution than the original one, after adding landmarks to the target domain.
- Sec. B: results on object recognition accuracies. Expanding what is reported in section 3.1 of the main text, we report additional results on a variation of GFK (Gong et al., 2012), choosing landmarks with random selection (with standard errors reported), and results from applying DASVM, a transudtive-style domain adaptation technique (Bruzzone & Marconcini, 2010).
- Sec. C: more detailed analysis. Expanding what is reported in section 3.1.2 of the main text, we provide additional analysis on the effect of selecting landmarks.

A. Proof of Theorem 1 in the main text

The proof is straightforward, appealing to the convexity of KL-divergence on its argument. Specifically,

$$\begin{aligned} KL(P_S(X)\|Q_T(X)) &= KL(P_S(X)\|(1-\mu)P_T(X)+\mu P_S(X)) && (1)\\ &\le (1-\mu)KL(P_S(X)\|P_T(X))+\mu KL(P_S(X)\|P_S(X)) && (2)\\ &= (1-\mu)KL(P_S(X)\|P_T(X)) && (3)\\ &\le KL(P_S(X)\|P_T(X)) && (4)\end{aligned}$$

The last step follows from the fact that $\mu \le 1$.

We can also assume a slightly more general model and prove similar results. Suppose the original target distribution is $P_T(X) = \alpha P_S(X) + (1-\alpha)P_O(X)$ where $P_O(X)$ is a mixture component that is unique to the target domain.

After adding the landmarks, suppose the new target distribution is $Q_T(X) = \beta P_S(X) + (1-\beta)P_O(X)$ with $\alpha \le \beta$. Then, similarly, we have

$$KL(P_S(X)\|Q_T(X)) \le KL(P_S(X)\|P_T(X)) \qquad (5)$$

The proof is as straightforward as the Theorem 1. In fact, we recognize $Q_T(X)$, a more skewed binary source, as a concatenation of a less skewed binary source $P_T(X)$ with a bit-flipping binary symmetric channel with transition probability $\epsilon = (\beta-\alpha)/(1-\alpha) \in [0,1]$. Namely,

$$Q_T(X) = \epsilon P_S(X) + (1-\epsilon)P_T(X) \qquad (6)$$

Applying Theorem 1, we arrive immediately at the last inequality.

B. Results on object recognition accuracies

Table 1 summarizes the classification accuracies on the target domain for 9 source-target pairs. The best result for each pair is in bold and red font. Comparing to the results reported in the main text (Table 1 and 2 there), we have added

- a variation of GFK (GFK+SVM). Originally, GFK (Gong et al., 2012) was used to perform kernelized 1-nearest neighbor classification using labeled source data. Alternatively, we take the square root of the kernel and transform features linearnly, $\sqrt{\boldsymbol{G}}\boldsymbol{x}$. We then train a linear support vector machine classifier to classify the transformed data in the target domain.
- the standard errors for RAND. SEL. (randomly selecting data points as landmarks). This is obtained by running experiments 10 times and computing the averaged accuracies as well as the standard errors of those 10 trials. Note that there are no standard errors in methods other than RAND. SEL., as we have used the whole source domain to select landmarks. Since the selection algorithm (eq. (3) in the main text) is a convex optimization, the landmarks are selected deterministically.

We can see that LANDMARK outperforms the state-of-the-arts, TCA (Pan et al., 2009), GFS (Gopalan et al., 2011), GFK (Gong et al., 2012) and its variation replacing 1-nearest neighbor with SVM, SCL (Blitzer et al., 2006), and KMM (Huang et al., 2007) to large margins. One exception is on WEBCAM→DSLR. As mentioned in the main text, WEBCAM and DSLR share the same set of object instances. Namely, for each particular object in DSLR there are image(s) of it in WEBCAM, and vice versa. As a result, our algorithm selects most images out of WEBCAM as landmarks and leaves probably too few samples to do model selection and validation. We leave this issue for future work.

The proposed method LANDMARK outperforms RAND. SEL. significantly (beyond the range of standard errors) on 7 out of 9 pairs, and works equally well as RAND. SEL. on the other two pairs of CALTECH → WEBCAM and WEBCAM → AMAZON.

In addition to the methods reported in Table 1, we have also tested domain adaptation SVM (DASVM) (Bruzzone & Marconcini, 2010). Since several parameters in DASVM cannot be cross-validated using the labeled data in the source domain, we report the range of its classification accuracies here. Changing the parameters in DASVM, we get accuracy ranges of 37.4–44.5%, 40.1–42.0%, 25.7–39.7%, 24.1–28.1%, and 48.4–68.2% on AMAZON→CALTECH, AMAZON→DSLR, WEBCAM→AMAZON, WEBCAM→CALTECH, and WEBCAM→DSLR, while LANDMARK's are 45.5%, 47.1%, 40.2%, 35.4%, and 75.2%, respectively. DASVM underperforms our LANDMARK in general.

C. Auxiliary tasks

The Value of auxiliary tasks The auxiliary tasks are domain adaptation problems over new pairs of source and target domains $\mathcal{D}_S^q \to \mathcal{D}_T^q$. As pointed by Theorem 1 in section 2.2, by incorporating landmarks in the augmented target domain, the domain adaptation becomes easier to solve. Fig. 1 provides strong empirical evidence.

In the figure, we show the object recognition accuracies on the original target domain as a result of solving those auxiliary tasks individually. Specifically, for each scale σ_q, we use the method of GFK to compute $\boldsymbol{G}_q$ for the pair $\mathcal{D}_S^q \to \mathcal{D}_T^q$ to extract invariant features then train a SVM classifier to minimize classification errors on the landmarks. We contrast to GFK+SVM reported in Table 1, where the only difference is to solve the original adaptation problem.

Clearly, the auxiliary tasks are easier to solve, resulting more effective adaptations such that the accuracies on the target domains are in general much better than GFK+SVM. This asserts firmly that landmarks *connect the dots* between the source and the target, and thus are an important adaptation mechanism to exploit.

The Benefits of multi-scale analysis and combining In Fig. 1, we also contrast results of individual tasks to the proposed method LANDMARK where the solutions of multiple auxiliary tasks are *combined* discriminatively. Combination clearly improves individual tasks. Moreover, we also marked in red color those individual tasks whose kernels have contributed to the final solution in eq. (7). Note that, the selected scales are indeed sparse. Both observations support our hypothesis that the data is modeled better with distances and similarities at multiple scales.

References

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

Bruzzone, L. and Marconcini, M. Domain adaptation problems: A DASVM classification technique and a circular validation strategy. *IEEE Trans. PAMI*, 32

Table 1. Recognition accuracies of the proposed method LANDMARK, baseline, and several variants of LANDMARK. Total 9 pairs of source/target domains are reported. C: CALTECH, A: AMAZON, W: WEBCAM, D: DSLR. The proposed method performs the best.

%	A→C	A→D	A→W	C→A	C→D	C→W	W→A	W→C	W→D
TCA (Pan et al., 2009)	35.0	36.3	27.8	41.4	45.2	32.5	24.2	22.5	80.2
GFS (Gopalan et al., 2011)	39.2	36.3	33.6	43.6	40.8	36.3	33.5	30.9	75.7
GFK (Gong et al., 2012)	42.2	42.7	40.7	44.5	43.3	44.7	31.8	30.8	75.6
GFK+SVM	38.8	43.3	37.3	50.2	40.1	45.1	39.1	34.5	67.5
SCL (Blitzer et al., 2006)	42.3	36.9	34.9	49.3	42.0	39.3	34.7	32.5	**83.4**
KMM (Huang et al., 2007)	42.2	42.7	42.4	48.3	53.5	45.8	31.9	29.0	72.0
LANDMARK	**45.5**	47.1	**46.1**	**56.7**	**57.3**	**49.5**	**40.2**	**35.4**	75.2
RAND. SEL.	44.5±0.3	44.5±0.9	41.9±0.9	53.8±0.4	49.9±0.8	**49.5**±1.0	**39.8**±0.8	34.1±0.5	74.2±0.5
SWAP	41.3	**47.8**	37.6	46.2	42.0	46.1	38.2	32.2	70.1
UNBALANCED	37.0	36.9	38.3	55.3	49.0	**50.1**	39.4	34.9	73.9
EUC. SEL.	44.5	44.0	41.0	50.2	40.1	45.1	39.1	34.5	67.5

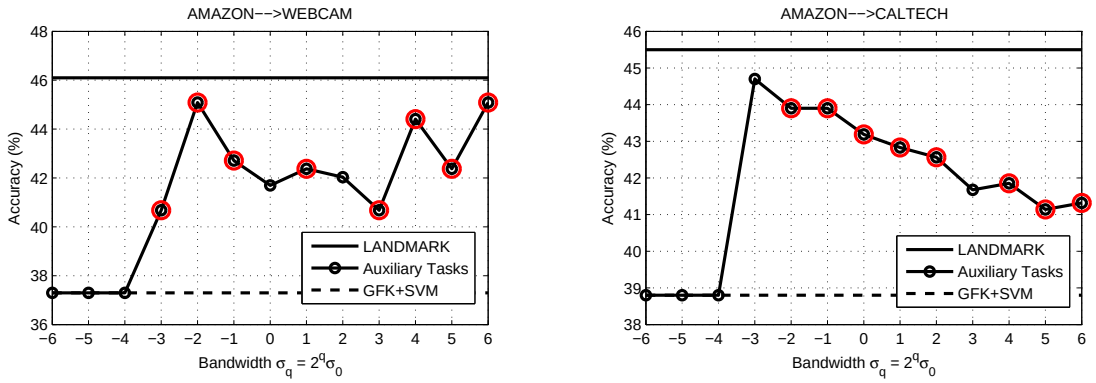

Figure 1. Performance of individual auxiliary tasks. The marked circle points on the curves show recognition accuracies on the original target domain $\mathcal{D}_T$, by using the kernel computed for the auxiliary task. Individual auxiliary tasks do not perform as well as LANDMARK. However, they all outperform BASELINE except when the scale is very small. In that case, all source domain data are selected as landmarks and auxiliary tasks are not defined. The red circles denote the auxiliary tasks whose kernels contribute to the final kernel $\boldsymbol{F}$ in eq. (7) after discriminative learning.

(5):770–787, 2010.

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

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

Huang, J., Smola, A.J., Gretton, A., Borgwardt, K.M., and Scholkopf, B. Correcting sample selection bias by unlabeled data. In *NIPS*, 2007.

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