Proceedings Paper:
Huang, Y., Zhu, F., Shao, L. et al. (2016) Color object recognition via cross-domain learning on RGB-D images. In: Proceedings - IEEE International Conference on Robotics and Automation. 2016 IEEE International Conference on Robotics and Automation (ICRA), May 16th - 21st 2016, Stockholm, Sweden. IEEE, pp. 1672-1677. ISBN: 9781467380263. ISSN: 1050-4729.

https://doi.org/10.1109/ICRA.2016.7487308

Color Object Recognition via Cross-Domain Learning on RGB-D Images

Yawen Huang, Fan Zhu, *Member, IEEE,* Ling Shao, *Senior Member, IEEE*, Alejandro F. Frangi, *Fellow, IEEE*

***Abstract*—This paper addresses the object recognition problem using multiple-domain inputs. We present a novel approach that utilizes labeled RGB-D data in the training stage, where depth features are extracted for enhancing the discriminative capability of the original learning system that only relies on RGB images. The highly dissimilar source and target domain data are mapped into a unified feature space through transfer at both feature and classifier levels. In order to alleviate cross-domain discrepancy, we employ a state-of-the-art domain-adaptive dictionary learning algorithm that updates image representations in both domains and the classifier parameters simultaneously. The proposed method is trained on a RGB-D Object dataset and evaluated on the Caltech-256 dataset. Experimental results suggest that our approach can lead to significant performance gain over the state-of-the-art methods.**

I. Introduction

Previous research of color object recognition focused on RGB images. With the rapid development of sensors, depth information becomes a mainstream way to support some advanced recognition techniques. Particularly, the idea of using additional depth information to enhance the performance of a learning system has been presented in [1]. Most existing works consider that RGB and depth data share the same distribution. A major challenge in real-world applications is that desired data cannot always stay in the same feature space as the training data. In this scenario, conventional machine learning-based recognition algorithms are very likely to fail [3], because of the cross-domain feature mismatch problem. In order to deal with this issue, transfer learning techniques are proposed and widely applied, where typical examples can be found in [2, 4, 5, 8].

In this paper, we present a novel technique for recognizing color objects by using labeled RGB-D images, where the additional depth information is used as the auxiliary data to enhance the discriminative power of data representations in the original source domain. The motivation behind our proposed method is as follows: depth images contain useful discriminative information, which presents a different feature distribution when compared to the corresponding RGB image domain. A joint learning is then considered by combining depth data with RGB information into one model to enhance the discriminative capability of the original learning system. On the other hand, the cross-domain dictionary learning is proposed by inputting both RGB and depth images in the training stage, to maximize data inter-class distances while minimizing data intra-class distances. After that, a bridge can be established for each pair of depth image and RGB image at the feature representation level. We further learn the dictionary and the classifier simultaneously to optimize the learned dictionary for classification tasks (where knowledge transfer is conducted at the classification level). The idea of our proposed method is shown in Fig. 1, which represents a strategy to better handle the typical cross-domain object recognition problem. The major pipeline of our proposed work is provided in Fig. 2. Through learning RGB features with corresponding depth information, we expand the original inter-class diversity of training data, and also enhance the discriminative capability of the original recognition system.

We summarize the contributions of this work as follows: (1) we propose a novel discriminative cross-domain dictionary learning-based object recognition framework; (2) depth images are considered as the auxiliary data and jointly learned with RGB images; (3) the presented approach learns a domain-adaptive dictionary pair and classifier parameters in data representation level and classification level respectively, where the loss function is formulated according to the reconstruction error, discriminative capability, and cross-domain discrepancy, so as to avoid the feature distribution mismatch problem.

II. Related Work

Our algorithm is mostly related to the methods in [2, 6, 18]. We review these works from both dictionary learning and transfer learning (*a.k.a.*, domain adaption, domain transfer or knowledge transfer) aspects. Learning an over-complete dictionary for sparse coding has been applied to various areas in computer vision and artificial intelligence, for instance, image restoration [9], image denoising [11], and action recognition [10]. The K-Singular Value Decomposition (K-SVD) [6] method, as a classical solution to l_0–based dictionary learning, focuses on the reconstruction capability of the learned dictionary. Label Consistent K-SVD (LC-KSVD) [18] further explores the discriminative information for dictionary learning while learning the classification model simultaneously to avoid the problem of suboptimal dictionary for classification.

However, traditional dictionary learning techniques cannot deal with the domain mismatch problem because they assume

Y. Huang, and A. F. Frangi are with the Center for Computational Imaging and Simulation Technologies in Biomedicine (CISTIB), Department of Electronic and Electrical Engineering, The University of Sheffield, Sheffield, United Kingdom (e-mail: yhuang36@sheffield.ac.uk, a.frangi@sheffield.ac.uk).

F. Zhu is with the Multimedia and Visual Computing Lab, Department of Electrical and Computer Engineering, New York University Abu Dhabi, Abu Dhabi (e-mail: fan.zhu@nyu.edu).

L. Shao is with the Computer Vision and Artificial Intelligence Group, Department of Computer Science and Digital Technologies, Northumbria University, Newcastle upon Tyne, United Kingdom (e-mail: ling.shao@ieee.org).

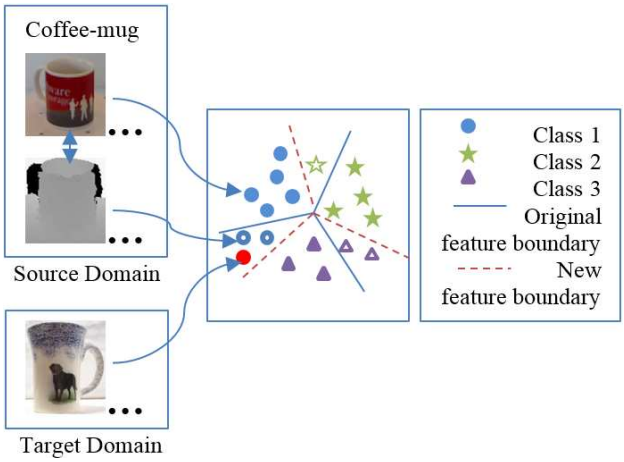


Fig. 1: Illustration of the cross-domain object recognition problem. Through jointly training with both RGB and depth image features, we aim to recognize images in the target RGB domain with the help of the source domain depth images. Examples from three classes are used to describe the difference between the original decision boundaries and the new decision boundaries which are obtained by adding depth features as auxiliary data.

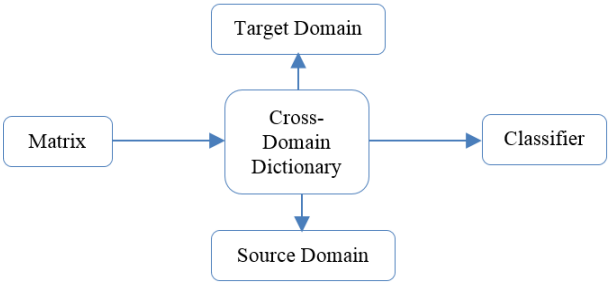


Fig. 2: Cross-domain dictionary learning flowchart. A transformation matrix is constructed to establish virtual correspondences between the source domain and the target domain data. The cross-domain dictionary learning is then performed through learning a discriminative dictionary pair and the corresponding classifier simultaneously.

that data in source and target domains share the same distribution. In this situation, transfer learning algorithms are developed to alleviate the discrepancies between mismatched cross-domain features. For example, domain-adaption from multi-view to single-view (DA_M2S) method [2] attempts to seek an optimal projection matrix to map samples from two different domains into a common feature space. Our approach aims to learn a discriminative cross-domain dictionary [12], in which both dictionary learning and classification tasks are unified into a single learning process.

III. Method

A. Dictionary Learning

Suppose there is a projection dictionary $D = \{d_1, d_2, \dots, d_K\} \in \mathbb{R}^{n\times K}$, and sparse vector $x \in \mathbb{R}^K$ in a K-dimensional feature space, thus the input signal can be represented as: $y = Dx$, where $y \in \mathbb{R}^n$ is an n-dimensional signal. The objective function for learning a reconstructive dictionary can be formulated as:

$$\min_x \|x\|_0 \quad s.t.\ y = Dx\ (or\ \|y - Dx\|_p \le \epsilon) \qquad (1)$$

where $\|\cdot\|_0$ denotes the l_0-norm sparse constraint, which fixes the number of non-zero elements in sparse representation x. Given a set of input signals $Y = \{y_1, y_2, \dots, y_N\}$, the sparse model follows the formulation of $Y = DX$, where $Y \in \mathbb{R}^{n\times N}$, $X \in \mathbb{R}^{K\times N}$, and $D \in \mathbb{R}^{n\times K}$. Engan et al. [16] presented a method of optimal directions (MOD) as a general approach to update the dictionary atoms sequentially. The error $E = \{e_1, e_2, \dots, e_i\}$ is computed by:

$$\|E\|_F^2 = \|Y - DX\|_F^2 \qquad (2)$$

Then, the objective function of K-SVD can be formulated by minimizing error E,

$$\min_{D,X} \|Y - DX\|_F^2 \quad s.t. \forall i, \|x_i\|_0 \le T \qquad (3)$$

where T is the sparsity constraint, which represents the number of non-zero elements. In our implementation, the reconstruction error term $\|Y - DX\|_F^2$ is solved by computing $\sum_{i=1}^N \|y - Dx\|_2^2$.

Thereby, the optimal dictionary D and sparse vector x can be acquired via iteratively minimizing errors. Since solving eq. (3) is generally NP-hard under the l_0-norm constraint, an alternative solution [17] is proposed to approximate the objective function with a higher order l_1-norm regularization [17, 20] for a near-optimum solution. The new objective function is then written as:

$$E(x) = \|Y - DX\|_2^2 \quad s.t. \|x\|_1 \le T \qquad (4)$$

The above function is a convex function which can be solved in polynomial time.

B. Discriminative Dictionary Learning

Taking a step further, discriminative dictionary learning provides a turning point by learning a dictionary and a classification model simultaneously for each class. Thus, it differs from most conventional dictionary learning approaches. As mentioned above, we suppose that Y represents a set of n-dimensional input signals, *i.e.*, $Y = \{y_i\}_{i=1}^N \in \mathbb{R}^{n\times N}$, where $i = 1, \dots, N$, D denotes the over-complete dictionary with K-dimensional dictionary atoms, and X indicates the sparse coefficients. The sparse representation of signals Y can be obtained by solving the problem of (4). In order to guarantee the learned cross-domain dictionary is over-complete, a strict constraint is added as $K > n$. Here, D can be constructed by iteratively minimizing reconstruction errors while subjecting to l_1–norm regularization. Also, the number of items in each signal is less than T in its decomposition.

The sparse representation x_i for each element can be directly used as feature in the classification level with a classifier $f(x_i)$, and $f(x_i)$ can be obtained by satisfying:

$$W = arg \min_W \mathcal{L}\{h_i, f(x_i, W)\} + \lambda \|W\|_F^2 \qquad (5)$$

where W is the classifier parameters ($W \in \mathbb{R}^{m\times K}$), $\mathcal{L}$ denotes the classification loss function, h_i is the label of x_i, and λ represents a regularization parameter which is used for preventing overfitting. Considering two separate procedures (*i.e.*, dictionary learning and classification) might lead to a suboptimal dictionary. A unified learning function is proposed

by jointly learning the dictionary and the classification model as in [12, 18]:

$$< D, W, X > = \arg\min_{D,W,X} \|Y - DX\|_F^2 + \sum_I \mathcal{L}\{h_i, f(x_i, W)\} + \lambda\|W\|_F^2 \quad s.t. \forall i, \|x_i\|_0 \le T \tag{6}$$

C. Cross-Domain Dictionary Learning

Since the discriminative information is included in the learning process, learning the dictionary and its corresponding classifier simultaneously achieves significant improvement over the classical K-SVD algorithm. However, if the target domain has a different data distribution with the source domain, discriminative dictionary learning can only ensure the local data smoothness. Therefore, the extended cross-domain discriminative dictionary learning is explored to handle such a weakness, as in [12], which redefines a reconstructive dictionary pair and gives the optimization function as:

$$\min_{D_d, X_d} \|Y_d - D_d X_d\|_F^2 + \lambda \sum_{i=1}^{N_d} \|x_{d_i}\|_1 \tag{7}$$

where $D_d = D_t \; or \; D_s = \{\mathcal{D}_{d_i}\}_{i=1}^{K} \in \mathbb{R}^{n \times K}$ (d is the name of the domain (t: target domain, s: source domain)) represents a target domain dictionary or a source domain dictionary. The term $K > n$ is set to make the dictionary over-complete. Similarly, $X_d = X_t \; or \; X_s = \{x_{d_i}\}_{i=1}^{N_d} \in \mathbb{R}^{K \times N_d}$ indicates a set of sparse codes for the source domain or the target domain. The notation λ is a tradeoff parameter. The sparse representation X_d still remains in the original data space for two separate domains while performing eq. (7) to obtain the corresponding dictionary. In order to avoid the data mismatch problem, the global data smoothness is explored via the combination of objective functions from two domains:

$$\min_{D_t, D_s, X_t, X_s} \{\|Y_t - D_t X_t\|_F^2 + \|Y_s - D_s X_s\|_F^2 + \Psi[X_t \; X_s]\} + \lambda \sum_{i=1}^{N_t} \|x_{t_i}\|_1 + \partial \sum_{i=1}^{N_s} \|x_{s_i}\|_1 \tag{8}$$

where function $\Psi[\cdot]$ expresses the distance measure between similar instances across different domains for each category. A desired property is that the sparse codes, which possess the same class labels, are forced to be close to each other. In this case, function $\Psi[\cdot]$ only pursues data smoothness for the target domain. Therefore, rewriting the $\Psi[X_t \; X_s]$ term as $\|X_t - m(Y_t, Y_s)X_s\|_F^2$ rather than $\|X_t - m(Y_t, Y_s)X_s\|_F^2 + \|X_s - m(Y_s, Y_t)X_t\|_F^2$, where $m(\cdot)$ is designed to calculate the mapping of corresponding cross-domain samples. The smaller the value of $\|X_t - m(Y_t, Y_s)X_s\|_F^2$ is, the greater the possibility of sharing the same labels between similar points can be. Thus, the divergence is measured through a linear transformation mapping $T(\vec{x})$ within each category to construct virtual correspondences between two domains. A global optimization function is then formulated as:

$$\min_{D_t, D_s, X_t, X_s} \|Y_t - D_t X_t\|_F^2 + \|Y_s - D_s X_s\|_F^2 + \|X_t - T(\vec{x})X_s\|_F^2 + \lambda \sum_{i=1}^{N_t} \|x_{t_i}\|_1 + \partial \sum_{i=1}^{N_s} \|x_{s_i}\|_1 \tag{9}$$

Note that the transformation matrix is written as $T(\vec{x})$. Setting the maximum item of each column of $T(\vec{x})$ to 1 while setting remainders to 0, the overall transformation matrix $T(\vec{x})$ is then computed as a binary matrix. Assuming that $T(\vec{x})$ can result in a one-to-one mapping across X_t and X_s after encoding, each matched pair owns an identical representation for two domains, (*i.e.*, $\left\|{X_t}^T - T(\vec{x})'{X_s}^T\right\|_F^2 = 0$, and $\left\|{Y_t}^T - T(\vec{x})'{Y_s}^T\right\|_F^2 = 0$), we then obtain:

$$\min_{D_t, D_s, X_t} \|Y_t - D_t X_t\|_F^2 + \|Y_s T(\vec{x})^T - D_s X_t\|_F^2 + \lambda \sum_{i=1}^{N_t} \|x_{t_i}\|_1 \tag{10}$$

Further, the desired objective function can be acquired based on the method in [12] which gives the discriminative cross-domain dictionary learning as in eq. (10):

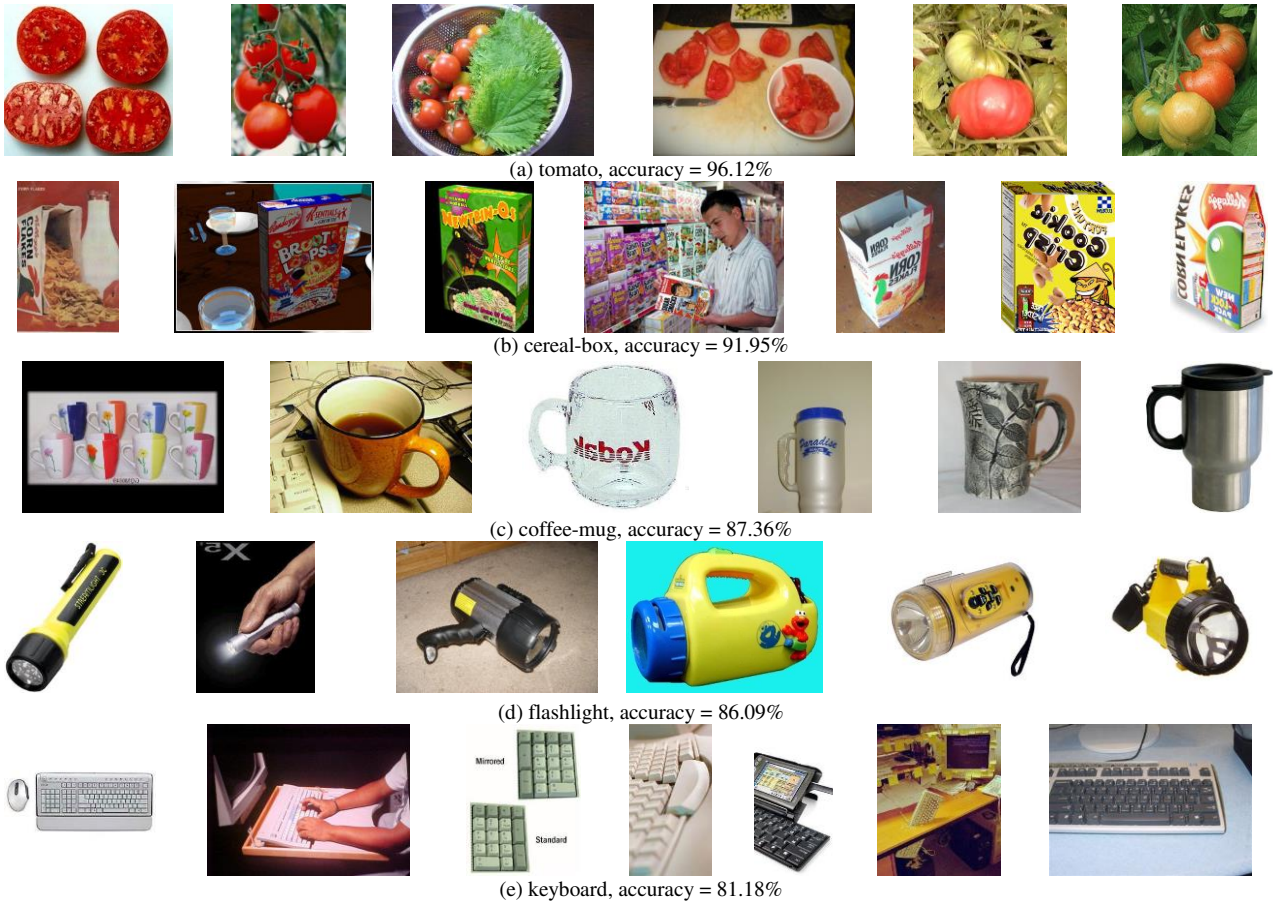
$$\min_{D_t, D_s, X_t, \mathcal{A}, \Theta} \|Y_t - D_t X_t\|_F^2 + \|Y_s T(\vec{x})^T - D_s X_t\|_F^2 + \beta \left\| \mathcal{H} - \{\arg\min_{\Theta} \sum_i \mathcal{L}\{h_i, f(x_i, W)\}\} X_t \right\|_F^2 + \alpha\|\mathcal{Q} - {X_t}'\|_F^2 + \lambda \sum_{i=1}^{N_t} \|x_{t_i}\|_1 \tag{11}$$

where ${X_t}' = \mathcal{A}X_t$, $\mathcal{A}$ is a linear transformation matrix that transforms the original sparse codes to be most discriminative in $\mathbb{R}^K$; $\mathcal{Q} = \{q_i\}_{i=1}^{N} \in \mathbb{R}^{K \times N}$ denotes a set of discriminative sparse codes of input signals Y_t in the target domain (*e.g.*, a set of discriminative sparse codes of y_i is represented as $q_i = \{q_{i_1}, q_{i_2}, \dots, q_{i_K}\} = \{0, \dots, 1, 1, \dots, 0\} \in \mathbb{R}^K$), and $\mathcal{H} = \{h_i\}_{i=1}^{N} \in \mathbb{R}^{C \times N}$ describes the class labels of Y_t, for example, a labelled vector of y_i is $h_i = \{h_{i_1}, h_{i_2}, \dots, h_{i_K}\} = \{0, \dots, 1, \dots, 0\} \in \mathbb{R}^C$. $\|\mathcal{Q} - {X_t}'\|_F^2$ term and $\|\mathcal{H} - \Theta X_t\|_F^2$ term are presented as the discriminative sparse representation errors and the classification errors respectively, in which scalars α and β are utilized to control the balance between two terms.

IV. Experiments

To illustrate the effectiveness of our approach for cross-domain object recognition, our experiments are conducted on two popular datasets. The RGB-D (Kinect) object dataset [14] is treated as the source domain which includes both color and corresponding depth images of 300 instances from 51 categories, and the Caltech-256 dataset [7] is served as the target domain which contains only color images. In this work, we randomly select 10 common categories[1] within each separate dataset, and uniformly choose images with an interval of twenty frames for each category from the RGB-D (Kinect) object dataset.

[1]Ten categories are: ball, calculator, cereal-box, coffee-mug, flashlight, keyboard, light bulb, mushroom, soda-can, and tomato.



(a) tomato, accuracy = 96.12%

(b) cereal-box, accuracy = 91.95%

(c) coffee-mug, accuracy = 87.36%

(d) flashlight, accuracy = 86.09%

(e) keyboard, accuracy = 81.18%

Fig. 3: 32 example images with high accuracy results from Caltech-256 dataset.

We obtain 1817 training samples in the source domain, and ensure that each color image corresponds to a depth image. On the other hand, the entire ten categories' color images (*i.e.* 1131 images) which come from the Caltech-256 dataset (shown in Fig. 3), are used in the target domain to test the performance of our proposed method. We take the factor of feature performance into consideration, and then choose OverFeat Feature Extractor[2] [26] to extract features for color images since its outstanding properties for representing RGB features. Considering depth images have the space-specific characteristic, we use Kernel Descriptors[3] (KDES) [23] to represent their features. Especially, Gradient Kernel Descriptor (GKD)-a subclass of KDES, shows better quality over the others, therefore, we choose it. Then, we extract depth features in accordance with the provision of [23]. In order to possess a consistent dimension before entering the discriminative cross-domain dictionary learning approach, the recorded depth features of each image are first aggregated to a uniform size by performing a primary dictionary learning with one level pyramid (*i.e.* 2×2) while fixing the size of dictionary as 1024.

We divide our experiments into two groups: first group includes several baseline approaches without domain adaption where the original input space is directly used without learning a new representation; second group involves transfer learning approaches that take into account transfer learning between diverged source domain data and target domain data and adapt them into a common space. Specifically, we evaluate single visual features and both visual and depth features respectively in our algorithm and demonstrate the experimental results for two groups.

V. Results

To illustrate the effectiveness of our algorithm, and the advantage of domain adaption in object recognition, we conduct two groups of experiments to compare our method with most relevant and state-of-the-art approaches. Experiments are set as follows:

- Group 1 (general object recognition using single RGB image-based features in source domain without domain adaption): LC_KSVD [18], Sparse codes Spatial Pyramid Matching (ScSPM) [15], No Adaption 1-Nearest-Neighbor (NA_NN) [13].
- Group 2 (object recognition methods with domain adaption by using both RGB and depth image-based features in source domain): Domain Adaption Subspace Alignment (DA_SA 1-Nearest-Neighbor (NN) or Support Vector Machine (SVM)) [13], and Geodesic Flow Kernel (GFK) [25], and our proposed method.

[2] http://cilvr.nyu.edu/doku.php?id=software:overfeat:start
[3] http://www.cs.washington.edu/robotics/projects/kdes/

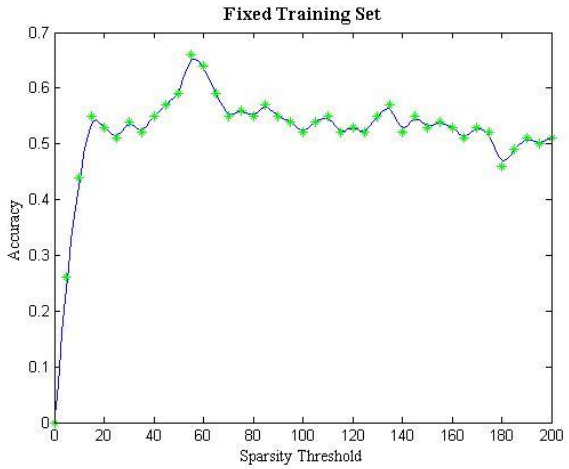


Fig. 4: Performance analysis on sparsity threshold while fixing the dictionary size as 1024.

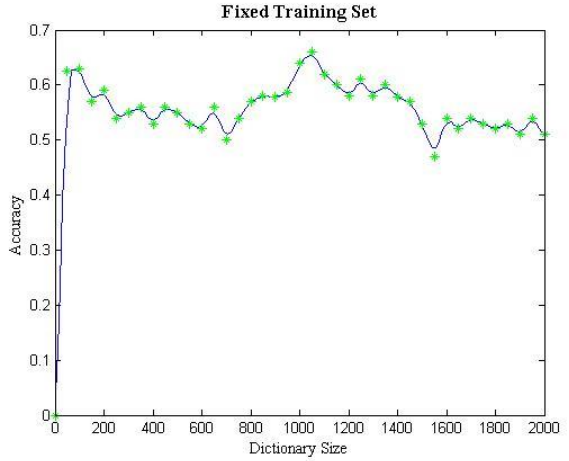


Fig. 5: Performance analysis on dictionary size while fixing the sparsity factor as 55.

TABLE I. PERFORMANCE COMPARISON OF BASELINE APPROACHES WITHOUT DOMAIN ADAPTION ON CALTECH-256 DATASET WHILE TRAINING ON RGB-D OBJECT DATASET.

Method	NA_NN	LC_KSVD2	LC_KSVD1	ScSPM
	56.32%	59.70%	60.30%	61.54%
(depth)	56.49%	60.10%	60.90%	62.51%

TABLE II. PERFORMANCE COMPARISON BETWEEN OUR METHOD AND STATE-OF-THE-ART TRANSFER LEARNING APPROACHES ON CALTECH-256 DATASET WHILE TRAINING ON RGB-D OBJECT DATASET.

Method	GFK	SA_NN	SA_SVM	OURS
	56.50%	59.33%	62.95%	**64.99%**
(depth)	57.59%	61.27%	62.16%	**65.98%**

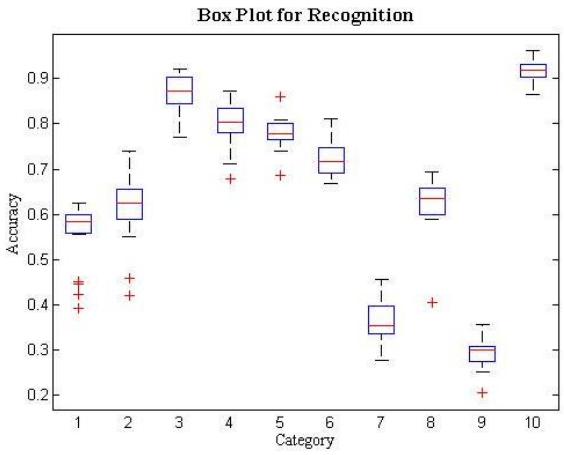


Fig. 6: Boxplot of recognition accuracy for each category in our RGB-D experiments.

TABLE III. NUMERICAL COMPARATION OF RECOGNITION ACCURACY FOR 20 TIMES OF TOTAL EXPERIMENTS PER CATEGOR.

	1	2	3	4	5	6	7	8	9	10	AVG.
MINIMUM (%)	39.31	42.00	77.01	67.82	68.70	66.88	27.71	40.59	20.69	86.41	**53.71**
MAXIMUM (%)	62.43	74.00	91.98	87.36	86.09	81.18	45.65	69.50	35.63	96.12	**72.99**
AVERAGE (%)	55.55	61.35	86.78	80.09	78.09	72.28	36.61	62.33	29.35	91.65	**65.40**

For all compared methods, we tune parameters according to the characteristics of each algorithm and demonstrate the best results among overall experiments. The sparsity factor and dictionary size are two impartment parameters in our algorithm as they are used for controlling the quality of our reconstructed dictionary. Here, we simply fix the dictionary size as 1024 and set the threshold value of sparsity as 55 (the descriptions of our selections are shown below). We carry out two sets of experiments to explain the reason of our setting for recognizing RGB images by jointly training RGB and depth images. In particular, although these two parameters are related, we analyze the behavior of one by assuming the other one is fixed. We collect some statistics about the setting of sparsity threshold and dictionary size through empirical experiments in many prior works (*e.g.*, [19, 21, 22, 24]), and found that the algorithm demonstrates its stable and better performance for sparsity threshold $\in$ [15, 200] and dictionary size $\in$ [50, 2000], in spite of the varying accuracy within this range. In order to guarantee the integrity of experiments, we extend the value of sparsity threshold and dictionary size to [0, 200] and [0, 2000], respectively. The effects on both parameters are illustrated in Fig. 4 and Fig. 5. As shown in Fig. 4, we provide the recognition results for sparsity between 0 and 200 while fixing the dictionary size as 1024. Although there are not many differences for the results using sparsity factors from 20 to 40 and 70 to 200, a detailed observation is made that sparsity from 45 to 65, particularly at 55, yields best performance among all results. Fig. 5 shows that how dictionary size influences the recognition results by setting the sparsity factor as 55. As expected, three dictionaries with sizes 1000, 1050, and 1100 achieve better performance than others. We therefore set the dictionary size as 1024 in all our experiments, which is same as most dictionary learning works.

Results are reported in Table I and Table II, where in each table, the first row indicates method name, the second row presents the recognition results by training on single RGB images, and the third row gives recognition rate by training on both RGB and depth images. The results of all comparative approaches in Table I and Table II are obtained by learning either RGB features or both RGB and depth features in the source domain while the methods in Table I do not apply domain adaption. All experimental results demonstrate that the additional depth data do improve the performance of algorithms. It is worth to point out that our method leads to the best results in both cases over the others by appending the depth features to enhance the diversity of intra-class, and also performing discriminative domain adaption dictionary learning to avoid the domain distribution mismatch problem. For more accurate statistics of our results, we record experimental results 20 times and show them in Fig. 6 and Table III, respectively. We also illustrate some samples of five categories with high recognition accuracies in Fig. 3. As shown in Fig. 6, the recognition rate for each category has a margin of error of $\pm 10\%$. For numerical analysis, Table III provides the minimum, maximum and average of recognition accuracies per category to further evaluate the variation of our proposed method. In this case, despite some outliers, our algorithm demenstrates the significant performance in almost all cases.

VI. Conclusion

In this paper, we have proposed an object recognition approach for recognizing RGB images in the target domain, where both RGB and auxiliary depth features are learned in the source domain. By performing cross-domain dictionary learning over both RGB and depth images in the training stage, we aim to span the intra-class diversities, so as to maximize the inter-class distances while minimizing the intra-class distances. Our method involves updating both image representations in source and target domains and the classifier parameters in a joint optimization process, so that the data distribution mismatch problem can be alleviated. We compare the proposed approach with other well established transfer learning approaches. Experimental results illustrate significant improvements over the state-of-the-art methods when incorporating the auxiliary depth images for enhancing the performance of cross-domain recognition.

References

[1] T. Huynh, R. Min, and J. L. Dugelay, "An efficient LBP-based descriptor for facial depth images applied to gender recognition using RGB-D face data," in *ACCV Workshop on Computer Vision with Local Binary Variants*, 2012, pp. 133-145.

[2] L. Chen, W. Li, and D. Xu, "Recognizing RGB images by learning from RGB-D data," in *IEEE International Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1418-1425.

[3] A. Torralba, A. Efros, "Unbiased look at dataset bias," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 1521-1528.

[4] L. Shao, F. Zhu, and X. Li, "Transfer learning for visual categorization: a survey," in *IEEE Trans. on Neural Networks and Learning Systems*, vol. 26, no. 5, pp. 1019-1034, Jul. 2014.

[5] L. Fei-Fei, "Knowledge transfer in learning to recognize visual objects classes," in *International Conference on Development and Learning*, 2006.

[6] M. Aharon, M. Elad, and A. Bruckstein, "K-svd: an algorithm for designing over-complete dictionaries for sparse representation," in *IEEE Trans. on Signal Process.*, vol. 54, no. 11, pp. 4311-4322, Nov. 2006.

[7] G. Griffin, A. Holub, and P. Perona, "The Caltech 256 dataset," Caltech Technical Report, California Institute of Technology, 2007.

[8] J. Liu, M. Shan, B. Kuipers, and S. Savarese, "Cross-view action recognition via view knowledge transfer," in *IEEE Conference Computer Vision and Pattern Recognit.*, Jun. 2011, pp. 3209-3216.

[9] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online dictionary learning for sparse coding," *International Conference on Machine Learning*, 2009, .

[10] F. Zhu and L. Shao, "Enhancing action recognition by cross-domain dictionary learning," *British Machine Vision Conference*, 2013.

[11] M. Elad, M. Aharo, "Image denoising via sparse and redundant representations over learned dictionaries," in *IEEE Trans. on Image Process.*, vol. 15, no. 12, pp. 3736-3745, 2006.

[12] F. Zhu, L. Shao, "Weakly-supervised cross-domain dictionary learning for visual recognition," *International Journal of Computer Vision*, vol. 109, no. 1, pp. 42-59, Mar. 2014.

[13] B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars, "Unsupervised visual domain adaption using subspace alignment," in *European Conference on Computer Vision*, 2013, pp. 2960-2967.

[14] K. Lai, L. Bo, X. Ren, and D. Fox, "A large-scale hierarchical multi-view RGB-D object dataset," in *IEEE International Conference on Robotics and Automation*, May 2011, pp. 1817-1824.

[15] J. Yang, K. Yu, Y. Gong, T. Huang, and B. Institute, "Linear spatial pyramid matching using sparse coding for image classification," in *IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2009, pp. 1794-1801.

[16] K. Engan, S. O. Aase, and J. H. Hakon-Husoy, "Method of optimal directions for frame design," in *IEEE International Conference on Acoust., Speech, Signal Process,* 1999.

[17] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," in *SIAM Journal on Scientific Computing*, vol. 43, no. 1, pp. 129–159, Aug. 2001.

[18] Z. Jiang, Z. Lin, and L. S. Davis, "Learning a discriminative dictionary for sparse coding via label consistent k-svd," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 2651-2664.

[19] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazaian. "Image denosing by sparse 3D tranform-domain collaborative filtering," *IEEE Trans. Image Process.*, vol. 16, no. 8, Aug. 2007.

[20] M. Zibulevsky, and M. Elad, "l1-l2 optimization in signal and image processing," *IEEE Signal Process. Mag.*, vol. 27, no. 3, pp. 76-88, May, 2010.

[21] W. Dong, L. Zhang, and G. Shi, "Centralized sparse representation for image restoration," in *IEEE International Conference on Computer Vision*, Nov. 2011, pp. 1259-1266.

[22] J. Yang, J. Wright, T.S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Trans. Image Process*., vol. 19, no. 11, pp. 2861-2873, Nov. 2010.

[23] L. Bo, X. Ren, and D. Fox, "Depth kernel descriptors for object recognition," in *IEEE International Conference on Intelligent Robot and Systems*, 2011.

[24] W. Dong, L. Zhang, G. Shi, and X. Li, "Nonlocally centralized sparse representation for image restoration," *IEEE Trans. Image Process.*, vol. 22, no. 4, pp. 1620-1630, Apr. 2013.

[25] B. Gong, Y. Shi, F. Sha, and K. Grauman, "Geodesic flow kernel for unsupervised domain adaption," in *IEEE International Conference on Computer Vision and Pattern Recognition*, Jun. 2012, pp. 2066-2073.

[26] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. L. Cun, "OverFeat: Integrated recognition, localization and detection using convolutional networks," in *International Conference on Learning Representations*, 2014.