2009 IEEE International Conference on Data Mining Workshops

Cross-domain Web Image Annotation

Si Si
Department of Computer Science
University of Hong Kong
Hong Kong
Email: ssi@cs.hku.hk

Dacheng Tao
School of Computer Engineering
Nanyang Technological University
Singapore 639798
Email: dctao@ntu.edu.sg

Kwok-Ping Chan
Department of Computer Science
University of Hong Kong
Hong Kong
Email: kpchan@cs.hku.hk

***Abstract*—In recent years, cross-domain learning algorithms have attracted much attention to solve labeled data insufficient problem. However, these cross-domain learning algorithms cannot be applied for subspace learning, which plays a key role in multimedia, e.g., web image annotation. This paper envisions the cross-domain discriminative subspace learning and provides an effective solution to cross-domain subspace learning. In particular, we propose the cross-domain discriminative Hessian Eigenmaps or CDHE for short. CDHE connects the training and the testing samples by minimizing the quadratic distance between the distribution of the training samples and that of the testing samples. Therefore, a common subspace for data representation can be preserved. We basically expect the discriminative information to separate the concepts in the training set can be shared to separate the concepts in the testing set as well and thus we have a chance to address above cross-domain problem duly. The margin maximization is duly adopted in CDHE so the discriminative information for separating different classes can be well preserved. Finally, CDHE encodes the local geometry of each training class in the local tangent space which is locally isometric to the data manifold and thus can locally preserve the intra-class local geometry. Experimental evidence on real world image datasets demonstrates the effectiveness of CDHE for cross-domain web image annotation.**

I. Introduction

With the rapid expansion of World Wide Web, the amount of pictorial data has been growing enormously. Annotating these images absolutely requires much expensive and time-consuming human labor, leading to the rise of interest in techniques relevant to image annotation [14]. However typical learning-based image annotation techniques need a large number of labeled training samples, and thus classifiers trained with only a limited number of labeled samples (e.g., one labeled image for each concept) are usually not robust for the real-world application. This critical problem can be well solved by recently proposed cross-domain learning [2] methods that has attracted more and more attentions for data analysis problems in multimedia information processing [13]. Cross-domain learning or transfer learning is specifically designed to deal with the situation where the labeled data in the training set is insufficient but sufficient labeling information can be obtained from other different but relevant domain(s).

A dozen of practical problems fall in cross-domain setting, because human annotation for obtaining training set is a very expensive and labor-intensive process. A natural concern is the possibility of utilizing the discriminative information for separating the training concepts to classify the concepts in the testing domain. Although they are of different types, they may share some common discriminative features. Therefore, it is possible to apply some popular cross-domain learning algorithms to solve the above problem.

A key role for image annotation is the distance or similarity between samples which can be solved via subspace learning [8][6], as subspace learning performs the annotation by enlarging the similarity among the intra-class samples and maximizing the difference among the inter-class samples in a subspace rather than the original feature space. However, the common assumption of the subspace learning algorithms is that both the training and the testing samples are drawn from an identical domain, or in a strict sense, they are independent and identically distributed (i.i.d.). Therefore, existing subspace learning algorithms cannot perform well when the training and the testing samples are drawn from different domains, and thus they can not utilize the information from other auxiliary domains to assist annotation when the data or information from the testing concept or class is limited.

In this paper, we tackle this problem by finding a shared subspace wherein the training and the testing samples are distributed in a similar way. In particular, the quadratic distance between the distributions of the training and testing domains is minimized in this subspace to solve the above distribution bias problem. However, this subspace could not be optimal for classifying samples from different classes. This is because we consider neither the manifold structure of intra-class samples nor the discriminative information of inter-class samples.

Under the patch alignment framework [9], we can model both conveniently. Specifically, for every sample associated with a patch (the neighbours of the sample), the following two aspects are taken into account: 1) to preserve the intra-class manifold structure, a local tangent space which is locally isometric to the manifold of the intra-class neighbour samples in the patch will help to preserve the local geometry information and thus can locally preserve the within class manifold structure; and 2) to preserve the inter-class discriminative information, the margin between the sample and its neighbours from different classes are maximized

978-0-7695-3902-7/09 $26.00 © 2009 IEEE
DOI 10.1109/ICDMW.2009.47

wherein the margin is measured by the average difference between the intra-class and inter-class distances. Because the intuition used for local geometry modelling is identical to that of the Hessian Eigenmaps, we term the proposed cross-domain subspace learning algorithm as the cross-domain discriminative Hessian Eigenmaps or CDHE for short.

The main contributions of this paper include: 1) To the best of our knowledge, CDHE is the first semi-supervised cross-domain subspace learning method. In contrast to the prior subspace learning methods, CDHE does not assume that the training and test data are drawn from the same domain or same distribution; 2) Several cross-domain learning algorithms directly transfer classifiers or models, and thus will heavily rely on the specific models whereas CDHE is general and flexible as it can be applied with any classification algorithms; 3) CDHE outperforms the state-of-the-art subspace learning and cross-domain learning methods on two real-world web image annotation databases: MSRA-MM and NUS-WIDE, demonstrating promising performance in real applications.

II. Cross-domain Discriminative Hessian Eigenmaps

Conventional subspace learning algorithms assume the training and the testing samples are drawn from an identical domain. In many practical applications, however, they are actually from different domains. Therefore, these algorithms cannot work well for these situations. This Section presents the cross-domain discriminative Hessian Eigenmaps or CDHE for short to solve the cross-domain classification tasks.

A. Modified Hessian Eigenmaps

Hessian Eigenmaps [11] can recover the underlying parameterization of a manifold M embedded in a high-dimensional space if the manifold M is locally isometric to an open and connected subset of R^d. Because the parameter space is not essentially convex in Hessian Eigenmaps, it can be applied to model a nonconvex manifold, e.g., a S-curve surface with a hole. Therefore, we adapt Hessian Eigenmaps in CDHE to preserve the local geometry for subspace learning.

Hessian Eignmaps finds the $d+1$ dimensional null-space of $H(f)$, wherein $H(f)$ is the Hessian matrix of a smooth mapping f, i.e., $f: M \mapsto R^d$. This $H(f)$ can be calculated by using $H(f) = \int_M \|H_f(x_i)\|_F^2 dx$ wherein $H_f(x_i)$ is the Hessian of f on the patch $X_{H(i)} = [x_i, x_{i^1}, \ldots, x_{i^{k_1}}]$ wherein $x_{i^1}, \ldots, x_{i^{k_1}}$, i.e., the k_1 nearest samples of x_i and the corresponding output in low dimensional space is $Y_{H(i)} = [y_i, y_{i^1}, \ldots, y_{i^{k_1}}]$. The tangent plane $T_{x_i}(M)$, a Euclidean space tangential to M at x_i, is an orthogonal coordinate system. In order to estimate $H_f(x_i)$, we calculate the local coordinate system of $X_{H(i)}$ and each sample in $X_{H(i)}$ has its own local coordinate Π_i on the tangent plane $T_{x_i}(M)$. Afterwards, this $H_f(x_i)$ can be estimated by using Π_i.

However, Hessian Eigenmaps cannot be applied to many practical applications, e.g., face recognition and image annotation because it requires that $k_1 > d$ where k_1 is the number of the neighbouring samples and d is the dimension of the subspace. It is difficult to guarantee this condition because we have a limited number of samples. Alternatively, we overcome this problem by performing PCA on M at x_i and orthnormalizing the d-dimensional representation to obtain the tangent coordinate in $T_{x_i}(M)$. The rest steps for the modified Hessian Eigenmaps are similar to those in Hessian Eigenmaps.

Under the patch alignment framework [9], the objective function for the modified Hessian Eigenmaps to preserve the local geometry on a local patch $Y_{H(i)}$ is given by

$$\begin{aligned} H(y_i) &= \mathrm{tr}\left(Y_{H(i)} H_f(x_i) H_f^T(x_i) Y_{H(i)}^T\right) \\ &= \mathrm{tr}\left(Y_{H(i)} L_{H(i)} Y_{H(i)}^T\right), \end{aligned} \tag{1}$$

where $L_{H(i)} = H_f(x_i) H_f^T(x_i)$ encodes the local geometry of the patch $X_{H(i)}$.

B. Margin maximization

Our main objective is the cross-domain classification, so it is insufficient to only retain the local geometry of intra-class samples. Therefore, the discriminative information should be exploited in the obtained subspace as well. Similar to the definition of the local geometry, we define a new margin maximization based scheme for discriminative information preservation over patches. For a patch $X_{M(i)} = \left[x_i, x_{i^1}, \ldots, x_{i^{k_1}}, x_{i_1}, \ldots, x_{i_{k_2}}\right]$ wherein $x_{i^1}, \ldots, x_{i^{k_1}}$, i.e., the k_1 nearest samples of x_i, are from the same class as x_i, and $x_{i_1}, \ldots, x_{i_{k_2}}$, i.e., the other k_2 nearest samples of x_i, are from different classes against x_i, the margin for the low dimensional corresponding patch $Y_{M(i)} = \left[y_i, y_{i^1}, \ldots, y_{i^{k_1}}, y_{i_1}, \ldots, y_{i_{k_2}}\right]$ is the average difference between the intra-class and inter-class distances, i.e.,

$$\begin{aligned} M(y_i) &= \sum_{j=1}^{k_1} \|y_i - y_{i^j}\|^2 \frac{1}{k_1} - \sum_{p=1}^{k_2} \|y_i - y_{i_p}\|^2 \frac{1}{k_2} \\ &= \mathrm{tr}\left(Y_{M(i)} L_{M(i)} Y_{M(i)}^T\right). \end{aligned} \tag{2}$$

In (2) we define $L_{M(i)}$ as

$$\begin{aligned} L_{M(i)} &= \begin{bmatrix} -e_{k_1+k_2}^T \\ I_{k_1+k_2} \end{bmatrix} \mathrm{diag}(w_i)\left[-e_{k_1+k_2}, I_{k_1+k_2}\right] \\ &= \begin{bmatrix} \sum\limits_{j=1}^{k_1+k_2} (w_i)_j & -w_i^T \\ -w_i & \mathrm{diag}(w_i) \end{bmatrix}. \end{aligned} \tag{3}$$

where $w_i = \left[\overbrace{1/k_1, \ldots, 1/k_1}^{k_1}, \overbrace{-1/\mathrm{k}_2, \ldots, -1/\mathrm{k}_2}^{k_2} \right]^T$; $I_{k_1+k_2}$ is the $(k_1+k_2) \times (k_1+k_2)$ identity matrix; $e_{k_1+k_2} = [1, \ldots, 1]^T \in R^{k_1+k_2}$. $L_{M(i)}$ encloses the local discriminative information in $Y_{M(i)}$. $M(y_i)$ can be viewed as the margin information representation.

C. Cross-domain parser

If samples from the training and the testing domains are independent and identically distributed, both the local geometry and the discriminative information can be well parsed from the training domain to the testing domain. However, in the cross-domain setting, the training and the testing samples are distributed differently in the original high dimensional space. Therefore, it is essential to find a subspace so that 1) the training and the testing samples are distributed similarly and 2) the local geometry and the discriminative information obtained from the training domain can be parsed to the testing domain.

The subspace can be obtained by minimizing a distance between the distribution of the training samples P_L and that of the testing samples P_U. Given a dataset $X = [x_1, x_2, \ldots, x_l, x_{l+1}, \ldots, x_{l+u}]$, suppose the first l samples are from the training set and the rest u samples are from the testing set. The corresponding low dimensional representation is $Y = [y_1, y_2, \ldots, y_l, y_{l+1}, \ldots, y_{l+u}]$. To provide a computationally tractable method to measure the distance between $p_L(y)$ the distribution of training samples in the low dimensional subspace and $p_U(y)$ the distribution of testing samples in the low dimensional subspace, the quadratic distance is applied here

$$\begin{aligned} Q_W(P_L \| P_U) &= \int (p_L(y) - p_U(y))^2 dy \\ &= \int (p_L(y)^2 - 2p_L(y)p_U(y) + p_U(y)^2) dy. \end{aligned} \tag{4}$$

We apply the kernel density estimation (KDE) technique to estimate $p_L(y)$ and $p_U(y)$, i.e., $p(y) = (1/n)\sum_{i=1}^{n} G_{\sum}(y - y_i)$. Here, n is the number of samples, and $G_{\sum}(y)$ is the d-dimensional Gaussian kernel with the covariance matrix $\sum$. If we introduce estimated distributions based on KDE to (4), we have

$$Q_W(P_L \| P_U) = \frac{1}{l^2} \sum_{s=1}^{l} \sum_{t=1}^{l} G_{\sum_{11}}(y_t - y_s) \tag{5}$$
$$+ \frac{1}{u^2} \sum_{s,t=l+u}^{l+1} G_{\sum_{22}}(y_t - y_s) - \frac{2}{lu} \sum_{s=1}^{l} \sum_{t=l+1}^{l+\mathrm{u}} G_{\sum_{12}}(y_t - y_s),$$

where $\sum_{11} = \sum_1 + \sum_1$, $\sum_{12} = \sum_1 + \sum_2$ and $\sum_{22} = \sum_2 + \sum_2$.The quadratic distance $Q_W(P_L \| P_U)$ serves as a bridge to parse the local geometry and the discriminative information from the training domain to the testing domain.

D. Optimization framework

By using the results obtained from the above subsections, we can obtain the optimization framework to learn the projection matrix W, which can parse both the local geometry and the discriminative information from the training domain to the testing domain. Because the margin maximization $M(y_i)$ and the local geometry representation $H(y_i)$ are defined over patches, each patch has its own coordinate system. The alignment strategy is adopted to build a global coordinate for all patches defined for the training samples. As a consequence, the objective function to solve the cross-domain subspace learning is given by

$$\begin{aligned} W = \underset{W \in R^{D \times d}}{\arg\min} & \sum_{i=1}^{l} (M(y_i) + \beta H(y_i)) \\ & + \lambda Q_W(P_L \| P_U), \end{aligned} \tag{6}$$

where λ and β are two tuning parameters. If we define two selection matrixes $S_{H(i)}$ and $S_{M(i)}$, which select samples in the i^{th} patch from all the training samples $Y_L = [y_1, y_2, \cdots, y_l]$ for constructing $M(y_i)$ and $H(y_i)$, respectively. Therefore, $Y_{H(i)} = Y_L S_{H(i)}$ and $Y_{M(i)} = Y_L S_{M(i)}$. According to (1) and (2) and letting $Y_L = W^T X_L$, the first part of objective function defined in (6) can be rewritten as

$$\begin{aligned} & \sum_{i=1}^{l} (M(y_i) + \beta H(y_i)) \\ & = \sum_{i=1}^{l} \left(\begin{array}{c} \mathrm{tr}\left(Y_L S_{M(i)} L_{M(i)} \left(Y_L S_{M(i)}\right)^T\right) \\ +\beta \mathrm{tr}\left(Y_L S_{H(i)} L_{H(i)} \left(Y_L S_{H(i)}\right)^T\right) \end{array} \right) \\ & = \mathrm{tr}\left(W^T X_L L X_L^T W\right), \end{aligned} \tag{7}$$

where $L = \sum_{i=1}^{l} \left(S_{M(i)} L_{M(i)} S_{M(i)}^T + \beta S_{H(i)} L_{H(i)} S_{H(i)}^T \right)$ is the alignment matrix [9]. X_L is the high-dimensional representation of all the training samples.

As a consequence, based on (7), the objective function in (6) can be further changed into

$$\begin{aligned} W = \underset{W \in R^{D \times d}}{\arg\min} & \, \mathrm{tr}\left(W^T X_L L X_L^T W\right) \\ & + \lambda Q_W(P_L \| P_U). \end{aligned} \tag{8}$$

To solve the above optimization problem, in this paper, the gradient descent technique is applied to obtain the optimal linear projection matrix W.

III. Experiments

In this Section, we apply the proposed CDHE on two real-world web image annotation databases: MSRA-MM [12] and NUS-WIDE [3] respectively. Because there are no public web image databases for cross-domain annotation, we design six datasets based on MSRA-MM and NUS-WIDE databases. Because we are working on cross-domain setting,

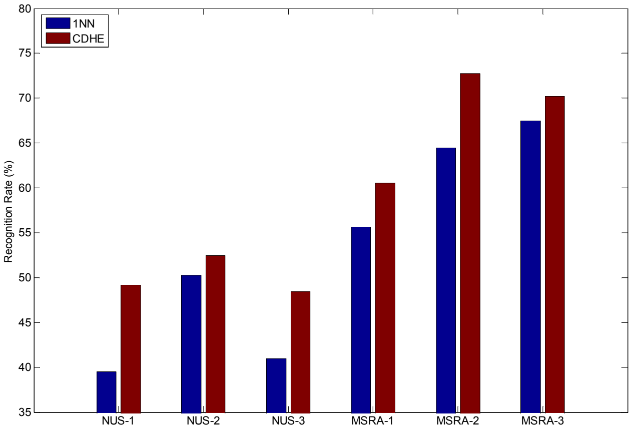

Figure 1. Recognition rates over 1NN and CDHE on six datasets, i.e., NUS-1, NUS-2, NUS-3, MSRA-1, MSRA-2, and MSRA-3, respectively.

it is essential to require the training and testing domains to share some common properties. Otherwise there would be nothing useful to transfer from the training domain to the testing domain. As a consequence, the design strategy is that we first select some relevant concepts from the database and then split these concepts into two disjoint parts with different concepts, one for training and the other for test. Because the training and test data belong to different concepts, they are from different domains. The composition of six datasets are shown in Table I. To reduce the computational burden, we sample 100 training and 100 test examples for each concept.

To demonstrate the effectiveness of CDHE, we compare it against five popular subspace learning algorithms, which are Fisher's linear discriminant analysis (FLDA) [4], locality preserving projections with supervised setting (LPP) [5], discriminative locality alignment (DLA) [10], semi-supervised discriminate analysis (SDA) [1], and the maximum mean discrepancy embedding (MMDE) [7]. FLDA is a conventional supervised learning method. DLA and LPP are discriminative manifold learning based subspace learning. Both achieve top level performance in many computer vision tasks. SDA is a semi-supervised method and it assumes the training and the testing samples are drawn from an identical manifold. MMDE is a cross-domain learning method and has been identified to be effective for cross-domain learning problems. All of these methods are subspace learning algorithms with the same training and test data and testing strategy, and thus the comparison is fair.

To further verify the effectiveness of CDHE as a subspace learning algorithm, we compare CDHE(the subspace dimension is set to 50) with nearest-neighbour classifier (1NN) by Euclidean distance (i.e., without subspace learning) in six datasets in Figure 1. Since both CDHE and 1NN use only one labeling image from the test concepts, the comparison is also fair. Figure 1 shows that CDHE can achieve more satisfactory performance than 1NN, which reflects CDHE can better discover the similarity between samples than 1NN.

Table I
THE COMPOSITION OF SIX DATA SETS.

Date Set		Concept
NUS-1	Train	six concepts from 12 animals: bear,..., zebra
	Test	the remaining six concepts from 12 animals
NUS-2	Train	eight concepts from 16 locations: airport,..., town
	Test	the remaining eight concepts from 16 locations
NUS-3	Train	seven concepts from 14 scenes: snow,...,rainbow
	Test	the remaining seven concepts from 14 scenes
MSRA-1	Train	five concepts from 10 animals: cat,...,cow
	Test	the remaining five concepts from 10 animals
MSRA-2	Train	tree, waterpark, baseball, party, and military
	Test	plant, hotel, football, medical, and war
MSRA-3	Train	boy, baby, flower, Disney, and earth
	Test	cowboy, children, rose, cartoon, and star

A. MSRA-MM Test

MSRA-MM database [12] consists of 65,443 labeled web images with 68 concepts (classes) collected from the Internet by using Microsoft Live Search. Example web images from the MSRA-MM database are shown in Figure 2. For representing images in the MSRA-MM, the dimension of features is 899-D, including seven kinds of features, i.e., HSV color histogram and wavelet texture. In this experiment, we evaluate the effectiveness of CDHE for cross-domain image annotation on these three datasets: MSRA-1, MSRA-2, and MSRA-3, respectively.

In the annotation stage, we select one reference image from each concept and then apply the nearest-neighbour rule to predict labels of the rest testing images in the selected subspace W. In the training stage, the labelling information from the reference images is blind to all the subspace learning algorithms.

Figure 4 compares CDHE against the other five subspace learning algorithms on MSRA-MM database under 4 different dimensions. It uses the boxplot to describe the comparison results. It has four groups, each of which stands for one dimension, i.e., 5, 10, 20, and 50. Each group contains six boxes, where boxes from left to right are the average accuracies of FLDA, LPP, DLA, SDA, MMDE, and CDHE, respectively. The figure shows that CDHE consistently and significantly outperform other subspace learning algorithms.

B. NUS-WIDE Test

NUS-WIDE database [3] contains 269,648 well labelled web images with 81 concepts (classes). The features used in the experiment for NUS-WIDE are 500-D bag of visual words. Example web images from the NUS-WIDE database are shown in Figure 3. Analogy to MSRA-MM database, NUS-WIDE database is also not intuitionally designed under cross-domain setting, and thus in order to perform cross-domain learning, we build three sub-databases (i.e., NUS-1, NUS-2, and NUS-3) based on the concepts within it as shown in Table I. In this experiment, we evaluate the effectiveness of CDHE for cross-domain image annotation

Figure 2. Sample images under the MSRA-2 dataset.

Figure 3. Sample images under the NUS-1 dataset.

on these three datasets: NUS-1, NUS-2, and NUS-3 respectively. The testing stage in NUS-WIDE is similar to the testing in the MSRA-MM database.

Figure 4 compares CDHE against the other five subspace learning algorithms on NUS-WIDE database under 4 different dimensions. In this figure, we have four groups, which indicate 5, 10, 20, and 50 dimensions. Each group contains six boxes, where boxes from left to right show the annotation accuracies of FLDA, LPP, DLA, SDA, MMDE, and CDHE, respectively. This figure shows that CDHE consistently and significantly outperforms other subspace learning algorithms.

IV. Conclusion

This paper has exploited the subspace learning under the cross-domain setting. We have proposed the cross-domain discriminative Hessian Eigenmaps (CDHE). CDHE parses both the local geometry and the discriminative information from the training domain to the testing domain. Therefore, for CDHE, it is not necessary to assume that training and testing samples are distributed independently and identically. This assumption is a high barrier between conventional subspace learning algorithms and many practical applications, e.g., cross-domain classification tasks. Thorough experimental results demonstrate that CDHE can significantly reduce the human labelling efforts for web images annotation problems. There are several directions for future work including the development of new distribution distance measurement to reduce the computation cost of the CDHE. In addition, we plan to apply Graph-based algorithms to address the scalability of CDHE for large-scale applications.

V. Acknowledge

The authors would like to thank Dr. M. Wang for helpful discussions on web image annotation. Si Si and Kwok-Ping Chan thank the support from HKU-SPF Grant (under project number 10400016). Dacheng Tao thanks the support from Nanyang SUG Grant (under project number M58020010) and Microsoft Operations PTE LTD-NTU Joint R&D (under project number M48020065). We also thank the anonymous reviewers for their valuable comments.

References

[1] Cai, D. et al.: Semi-supervised Discriminant Analysis. In: Proc. IEEE ICCV, pp. 1–8 (2007)

[2] Caruana, R. et al.: Multitask Learning. Machine Learning, pp. 41–75 (1997)

[3] Chua, T.S. et al.: NUS-WIDE: A Real-World Web Image Database from National University of Singapore. In: Proc. CIVR (2009)

[4] Fisher, R.A. et al.: The Use of Multiple Measurements in Taxonomic Problems. Annals of Eugenics, vol.7, pp. 179–188 (1936)

[5] He, X. et al.: Locality Preserving Projections. Advances in Neural Information Processing Systems, vol. 16 (2004)

[6] Liu, W. et al.: Transductive Component Analysis. In Proc. ICDM, pp. 433–442 (2008)

[7] Pan, S.J. et al.: Transfer Learning via Dimensionality Reduction. In: Proc. AAAI, pp. 677–682 (2008)

[8] Tao, D. et al.: Geometric Mean for Subspace Selection. IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 31, no. 2, pp. 260–274 (2009)

[9] Zhang, T. et al.: Patch Alignment for Dimensionality Reduction. IEEE Trans. Knowledge and Data Engineering, vol. 21, no. 9, pp. 1299–1313 (2009)

[10] Zhang, T. et al.: Discriminative Locality Alignment. In Proc. ECCV, pp. 725–738 (2008)

[11] Donoho, D. L. et al.: Hessian eigenmaps: locally linear embedding techniques for high-dimensional data. In: Proc. NAAS, pp. 5591–5596 (2003)

[12] Li, H. et al.: MSRA-MM 2.0: A Large-Scale Web Multimedia Dataset. In International Workshop on Internet Multimedia Mining, in association with ICDM 2009.

[13] Yang, J. et al.: Cross-domain video concept detection using adaptive svms. In ACM Multimedia, pp. 188–197 (2007)

[14] Sebe, N. et al.: Toward Improved Ranking Metrics. IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 22, pp. 1132–1143 (2000)

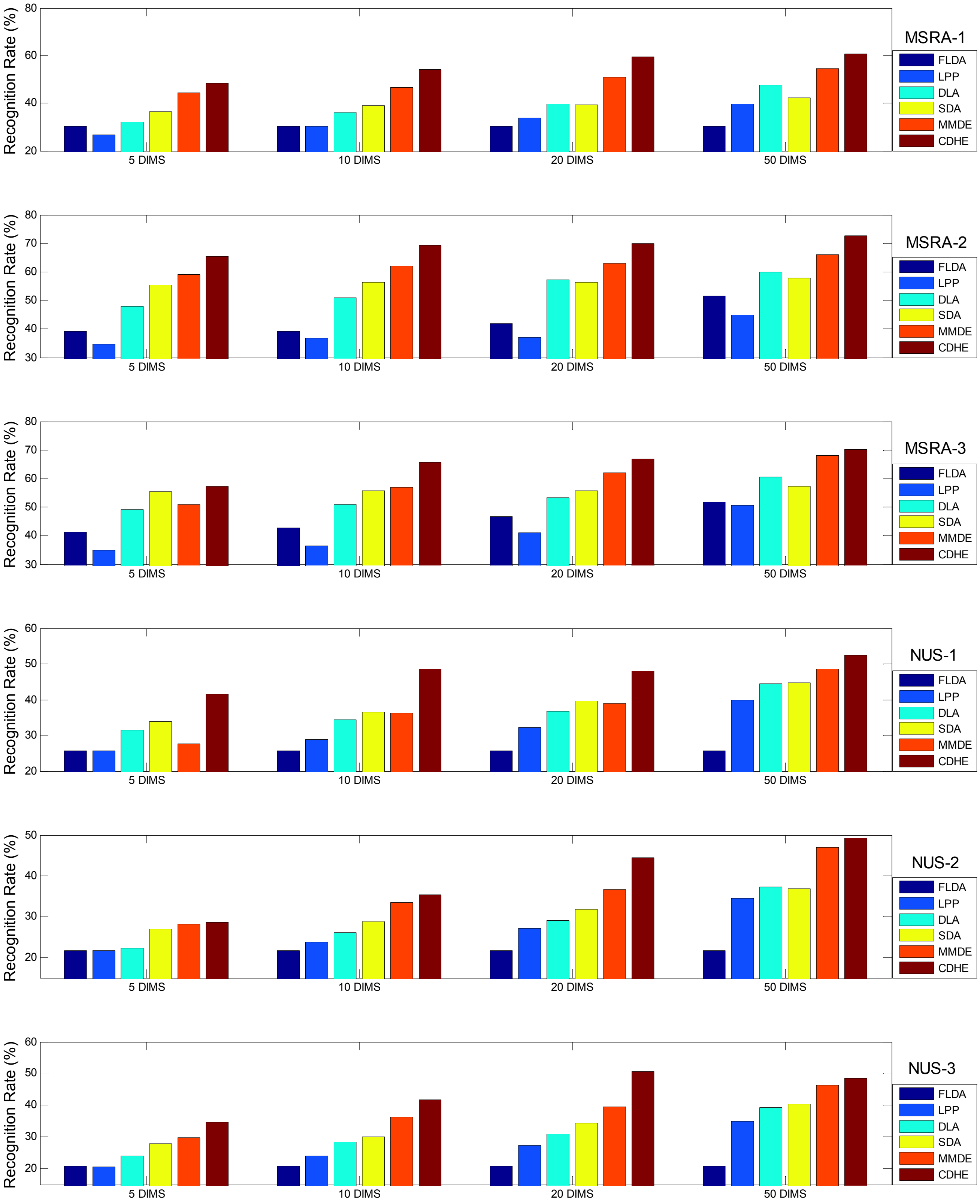

Figure 4. Recognition rates vs. different subspace learning algorithms under the 5, 10, 20, and 50 dimensions on six datasets, i.e., MSRA-1, MSRA-2, MSRA-3, NUS-1, NUS-2, and NUS-3, respectively.