# DISCRIMINANT PAIRWISE LOCAL EMBEDDINGS

*Konstantinos Bozas and Ebroul Izquierdo*

School of EECS, Queen Mary University of London
{k.bozas, ebroul.izquierdo}@eecs.qmul.ac.uk

## ABSTRACT

This paper introduces Discriminant Pairwise Local Embeddings (DPLE) a supervised dimensionality reduction technique that generates structure preserving discriminant subspaces. This objective is achieved through a convex optimization formulation where Euclidean distances between data pairs that belong to the same class are minimized, while those of pairs belonging to different classes are maximized. These pairwise relations are encoded in two matrices and weighted with the data *affinity matrix* to ensure local structure preservation. The discriminant efficiency of our technique is demonstrated in two popular applications, face and sketch recognition, where DPLE outperforms competitive manifold learning algorithms. A kernelized version of DPLE, that further enhances recognition accuracy, is also explained.

***Index Terms*—** Dimensionality reduction, DPLE, manifold learning, sketch recognition, face recognition

## 1. INTRODUCTION

Dimensionality reduction or subspace learning is the transformation that maps data from a high-dimensional space into a meaningful low dimensional space. It has been widely used in recognition tasks to mitigate the inherent drawbacks of high-dimensional spaces. Real world data like images, videos and speech signals are by nature high-dimensional modalities. In order to efficiently process that data, we need to reduce its dimensionality. Furthermore, such real world data are accompanied by noise which affects the accuracy of classification algorithms. By exposing the intrinsic dimensionality of the input data, we can generate projection bases that are immune to noise. The benefits of dimensionality reduction include classification, visualization and compression of high-dimensional data [1].

One of the first and classic approaches to dimensionality reduction is the PCA algorithm that generates a subspace where data variance is maximized. PCA is an unsupervised technique, therefore does not produce discriminative subspaces. LDA [2] exploits the data labels and performs better in classification scenarios. PCA and LDA rely on assumptions on the data distributions which often do not hold for real world applications. LFDA [3] takes local structure of the data into account so multi-modal data can be embedded appropriately.

Manifold learning is the branch of dimensionality reduction that investigates the underlying manifold of data. Originated from ISOMAP [4], manifold learning techniques attempt to discover a low-dimensional manifold where the input data lie on. A famous example is the Swiss roll which is originally embedded in a three dimensional space, yet it easy to show by 'unfolding' it, that its points lie on a two dimensional manifold.

In the same spirit, LPP [5] and its variants [6, 7, 8, 9, 10] generate lower dimensional spaces that preserve the local neighborhood of the data, hence the restricting assumptions of PCA and LDA are avoided. LPP is an unsupervised technique, yet extensions have been published that make use of data labels. DLPP [9] incorporates in the optimization process the within and between scatter matrices to achieve class separability. ILPP [6], ARE [8] and max-margin MMP[10] are semi-supervised approaches obtaining label information from user feedback. ILPP updates its learned projection matrix according to user guidelines. MMP solves an eigenvalue problem that maximizes the margin between different labelled samples.

We present Discriminant Pairwise Local Embeddings (DPLE), a manifold learning algorithm inspired by LPP [5]. The main idea is to learn a discriminant subspace where the data will be better separated than in the original input space, without violating much its local neighbourhood. The latter ensures that the data will maintain their manifold structure in the learned subspace, so classification algorithms can generalize better. We form these goals in a convex optimization problem that can be efficiently solved through eigendecomposition. A kernelized version is also introduced to further enhance classification accuracy. Experiments on two datasets demonstrate the advantages of our technique.

DPLE's objective is similar to that of LDE[7]/ARE, yet our formulation is different and the superiority of our technique is attributed to the following factors: a) LDE does not exploit the importance of influential samples, i.e. samples with many proximate neighbours guaranteed not to be outliers. DPLE utilizes this information in its objective function. b) ARE employs a non-flexible encoding scheme for the relationships between data pairs. It weights equally every pair and does not take into account the distances of samples in the original space. This approach fails to alleviate the influence of noisy data pairs that belong to the same class but they are far away in the feature space. DPLE handles this problem by weighting these relationships with the affinity matrix.

## 2. DISCRIMINANT PAIRWISE LOCAL EMBEDDING

This section describes Discriminant Pairwise Local Embeddings (DPLE), a novel supervised dimensionality reduction technique and its kernelized variant via the *kernel trick* [11].

### 2.1. Linear DPLE

Let $n$ pairs of data samples and its associated labels $(\mathbf{x}_i, y_i)$, $i = \{1, 2, \ldots, n\}$, where $\mathbf{x}_i \in \mathbf{R}^d$ represents a data sample and $y_i \in \{1, 2, \ldots, |C|\}$ is the label of the $i$-th sample. $|C|$ is the total number of classes. Let $\mathbf{X} \in \mathbf{R}^{d\times n}$ be the matrix of all samples. The $i$-th column of $\mathbf{X}$ is $\mathbf{x}_i$. Let $\mathbf{z}_i \in \mathbf{R}^p (1 \leq p \leq d)$ be an embedded sample and $p$ the dimension of the embedding space. Since we investigate dimensionality reduction scenarios, we usually require $p \ll d$.

Linear dimensionality reduction is performed via the transformation matrix $\mathbf{W} \in \mathbf{R}^{d\times p}$:

$$\mathbf{z}_i = \mathbf{W}^\top \mathbf{x}_i \tag{1}$$

The structure information of the data set is represented in the *affinity matrix* $\mathbf{A}$. The matrix $\mathbf{A}$ captures similarities between data pairs and is defined as:

$$\mathbf{A}_{i,j} = \begin{cases} e^{-\|\mathbf{x}_i - \mathbf{x}_j\|^2/2\sigma^2}, & \text{if } \mathbf{x}_i \in \mathcal{N}_k(\mathbf{x}_j) \\ & \text{or } \mathbf{x}_j \in \mathcal{N}_k(\mathbf{x}_i) \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

where $\mathcal{N}_k(\mathbf{x})$ represents the set of $k$-nearest neighbours of $\mathbf{x}$. A simpler alternative to (2) is to set $\mathbf{A}_{i,j} = 1$ if $\mathbf{x}_i$ is a nearest neighbor of $\mathbf{x}_j$ or vice versa; otherwise $\mathbf{A}_{i,j} = 0$. In both cases, a high value of $\mathbf{A}_{i,j}$ indicates that $\mathbf{x}_i$ and $\mathbf{x}_j$ lie close in the defined metric space and a low value that they lie apart.

Based on the label information, we define two pairwise relation matrices. The *same-label matrix* $\mathbf{A}^{(s)}$ representing all the sample pairs that share the same label and the *different-label matrix* $\mathbf{A}^{(d)}$ representing all the sample pairs with different labels:

$$\mathbf{A}^{(s)}_{i,j} = \begin{cases} \mathbf{A}_{i,j}, & \text{if } y_i = y_j \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

$$\mathbf{A}^{(d)}_{i,j} = \begin{cases} \mathbf{A}_{i,j}, & \text{if } y_i \neq y_j \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

We observe from (3) and (4) that matrices $\mathbf{A}^{(s)}$ and $\mathbf{A}^{(d)}$ are weighted with the affinity matrix $\mathbf{A}$. If we assign a constant value to similar and dissimilar pairs as in [8]; for instance if $\mathbf{A}^{(s)}_{i,j} = 1$ when $y_i = y_j$ and $\mathbf{A}^{(d)}_{i,j} = 1$ when $y_i \neq y_j$, then all the sample pairs will have equal weights resulting in loss of structure information. Instead, by employing the affinity matrix we assign an 'importance' value to each pair. Samples that lie close in the original input space are more significant and are imposed to lie close in the embedding space. On the other hand, pairs that are apart in the original space are either ignored or slightly contribute to the optimal solution. This idea is similar to the local variant of LDA [3], yet employed in a different learning framework.

We suggest the following optimization problem:

$$\begin{aligned} &\arg\max_{\mathbf{W}} \frac{1}{2} \sum_{i,j}^{n} \|\mathbf{W}^\top \mathbf{x}_i - \mathbf{W}^\top \mathbf{x}_j\|^2 \left(\mathbf{A}^{(d)}_{i,j} - \gamma \mathbf{A}^{(s)}_{i,j}\right) \\ &\text{subject to: } \mathbf{W}^\top \mathbf{X}\mathbf{D}\mathbf{X}^\top \mathbf{W} = \mathbf{I} \end{aligned} \tag{5}$$

where $\mathbf{D}_{i,i} = \sum_{j=1}^{n} \mathbf{A}_{i,j}$ is a diagonal matrix consisted of the row sums of $\mathbf{A}$ and $\gamma$ is a scalar to compensate for any imbalances occurred by different number of pair samples between $\mathbf{A}^{(d)}$ and $\mathbf{A}^{(s)}$.

The above formulation minimizes the Euclidean distances between all sample pairs that belong to the same category through matrix $\mathbf{A}^{(s)}$ and in the same time maximizes those between pairs belonging to different classes through matrix $\mathbf{A}^{(d)}$. We have previously seen that each pair relationship is weighted by the affinity matrix $\mathbf{A}$, therefore the intrinsic structure of data is maintained. The constrain $\mathbf{W}^\top \mathbf{X}\mathbf{D}\mathbf{X}^\top \mathbf{W} = \mathbf{I}$ is imposed to avoid the trivial solution $\mathbf{W} = \mathbf{0}$ and each entry $\mathbf{D}_{i,i}$ provides a measure of importance to the embedded sample $\mathbf{z}_i = \mathbf{W}^\top \mathbf{x}_i$.

The objective function in (5) can be rewritten as follows using linear algebra properties:

$$\begin{aligned} &\arg\max_{\mathbf{W}} J(\mathbf{W}) = \mathbf{W}^\top \mathbf{X} \left(\mathbf{L}^{(d)} - \gamma \mathbf{L}^{(s)}\right) \mathbf{X}^\top \mathbf{W} \\ &\text{subject to: } \mathbf{W}^\top \mathbf{X}\mathbf{D}\mathbf{X}^\top \mathbf{W} = \mathbf{I} \end{aligned} \tag{6}$$

where $\mathbf{L}^{(s)} = \mathbf{D}^{(s)} - \mathbf{A}^{(s)}$ and $\mathbf{L}^{(d)} = \mathbf{D}^{(d)} - \mathbf{A}^{(d)}$ are the Laplacian matrices of $\mathbf{A}^{(s)}$ and $\mathbf{A}^{(d)}$ respectively.

We apply the Lagrange multipliers to the above problem and the set the derivative with respect to $\mathbf{W}$ to zero.

$$\mathbf{X} \left[\mathbf{L}^{(d)} - \mathbf{L}^{(s)}\right] \mathbf{X}^\top \mathbf{w} = \bar{\lambda} \mathbf{X}\mathbf{D}\mathbf{X}^\top \mathbf{w} \tag{7}$$

The result is a generalized eigenvalue problem and since $\mathbf{L}^{(s)}$, $\mathbf{L}^{(d)}$ and $\mathbf{D}$ are symmetric semi-definite matrices all the eigenvalues are real positive numbers.

The optimal projection matrix $\mathbf{W}_{DPLE}$ is given by:

$$\mathbf{W}_{DPLE} = \left(\sqrt{\bar{\lambda}_1}\mathbf{w}_1 \mid \sqrt{\bar{\lambda}_2}\mathbf{w}_2 \mid \cdots \mid \sqrt{\bar{\lambda}_p}\mathbf{w}_p\right) \tag{8}$$

where $\{\mathbf{w}\}_{i=1}^{p}$ are the generalized eigenvectors associated with the $p$ largest eigenvalues $\bar{\lambda}_1 \geq \bar{\lambda}_2 \geq \cdots \geq \bar{\lambda}_p$ of (7).

**Algorithm 1:** DPLE embedding

**Data**: $(\mathbf{x}_i, y_i)\ i \in \{1, 2, \dots, n\}, \gamma, p$
**Result**: Projection matrix: $\mathbf{W}_{DPLE}$

1 Compute affinity matrix $\mathbf{A}$ according to (2).
2 Compute matrices $\mathbf{A}^{(s)}$ and $\mathbf{A}^{(d)}$ from (3) and (4).
3 Solve the generalized eigenproblem of (7).
4 Form the columns of $\mathbf{W}_{DPLE}$ from the eigenvectors of (7) corresponding to the largest eigenvalues.

The steps of DPLE are summarized in Algorithm 1. DPLE exploits the labelled information encoded through the matrices $\mathbf{A}^{(s)}$ and $\mathbf{A}^{(d)}$ to generate discriminate projection bases without violating the intrinsic structure of the data. The latter is ensured by the leverage of the affinity matrix $\mathbf{A}$ which weights accordingly each sample pair. The embedded data lie on a discriminative semantic manifold which preserves local geometric relations. As a result classes become better separated in the learned subspace.

### 2.2. Kernel DPLE

In most real world applications, data in the original input space cannot be linearly separated, due to it is being generated from non-linear processes. In such cases, linear algorithms like DPLE fail to produce efficient embedding spaces. We show that by using the *kernel trick* [11], we can generate a non-linear map from the original high-dimensional feature space to a lower-dimensional manifold where non-linear data can be efficiently represented.

Let $\phi : \mathbf{R}^d \rightarrow \mathcal{H}$ be a non-linear map function, mapping the Euclidean space $\mathbf{R}^d$ to Hilbert space $\mathcal{H}$. In Hilbert space the eigenvector problem of (7) becomes:

$$\phi(\mathbf{X}) \left[\mathbf{L}^{(d)} - \mathbf{L}^{(s)}\right] \phi(\mathbf{X})^\top \mathbf{w} = \bar{\lambda} \phi(\mathbf{X}) \mathbf{D} \phi(\mathbf{X})^\top \mathbf{w} \quad (9)$$

There is no easy way to directly compute the mapping $\phi(\mathbf{X})$, yet we can employ inner products of mapped data to solve the problem. We define the inner products of the mapped data as:

$$\mathbf{K}(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^\top \phi(\mathbf{x}_j) \quad (10)$$

The eigenvectors of (9) are linear combinations of $\phi(\mathbf{x}_1), \phi(\mathbf{x}_2), \dots, \phi(\mathbf{x}_n)$, hence we can write:

$$\mathbf{w} = \sum_{i=1}^{n} \alpha_i \phi(\mathbf{x}_i) = \phi(\mathbf{X}) \alpha \quad (11)$$

where $\alpha = [\alpha_1, \alpha_1, \dots, \alpha_n]^\top \in \mathbf{R}^n$. Using (11) it is easy to obtain the *kernelized* eigenvalue problem:

$$\mathbf{K}\left[\mathbf{L}^{(d)} - \mathbf{L}^{(s)}\right] \mathbf{K}\alpha = \bar{\lambda} \mathbf{K} \mathbf{D} \mathbf{K} \alpha \quad (12)$$

As before, the optimal embedding is consisted from the $p$ eigenvectors corresponding to largest eigenvalues of (12).

## 3. EXPERIMENTS AND RESULTS

In this section, the classification efficiency of DPLE is demonstrated. Our method is applied to two popular learning tasks, face recognition and sketch recognition, and compared against various well-known discriminant subspace learning algorithms.

### 3.1. Datasets and experimental setup

The two datasets used in our evaluation are the ORL face database [12] and the sketch recognition database (SKETCH) of [13]. The ORL dataset includes 40 subjects with 10 grayscale images per subject. Following the preprocessing of [7], we resize each image to $28 \times 23$ pixels and vectorize the outcome. We apply PCA to the image vectors and keep 98% of the information.

The SKETCH dataset of [13] encompasses 20,000 unique human drawn sketches evenly distributed over 250 object categories. Each image depicts a binary sketch of a single object. All sketches are rescaled to a fixed size and centred in the image canvas to accommodate scale and translation invariance. The human accuracy on the above database is 73% which highlights the challenge for machine classification. We observe *high inter-class* and *intra-class variability*. Some classes are easily recognized while others regularly misclassified to categories with similar visual appearance. Moreover, an object can be sketched quite differently by various individuals a fact that contributes to aforementioned intra-class variations. Each sketch is represented by an ensemble of local features that capture the main gradient orientations of a local sketch region. The data are publicly available from the authors' website and in this paper we use them as provided with no alternations.

We compare our method with the k-nn classifier in the original space denoted as (NN), the classic PCA and LDA algorithms and a collection of more sophisticated manifold learning techniques, namely LPP [5], LFDA [3] and LDE [7] along with kernelized versions for the last two. The recognition accuracy of the k-nn classifier in the learned subspace is reported. The parameters of each algorithm are empirically tuned for every dataset. In the kernelized version of the algorithms, we employ the rbf kernel with $\sigma = 1$. In the ORL database we perform 5-fold cross validation, whereas in SKETCH dataset we follow the protocol of [13] and perform 3-fold cross-validation with stratified sampling.

### 3.2. Results

The evaluation results are illustrated in Table 1. NN accuracy indicates that AT&T dataset is easy. We observe that the discriminant manifold learning algorithms perform better than PCA, LDA and the unsupervised LPP. DPLE and KDPLE achieves the highest recognition rates in this dataset.

| Method | ORL | SKETCH |
|---|---|---|
| | Linear | |
| NN | 97.5% | 45% |
| PCA | 98% ($p = 32$) | 41.97% ($p = 250$) |
| LDA | 98% ($p = 24$) | 41.2% ($p = 100$) |
| LPP | 96.25% ($p = 20$) | 41.74% ($p = 300$) |
| LFDA | 98.5% ($p = 12$) | 48.10% ($p = 120$) |
| LDE | 98.5% ($p = 21$) | 48.18% ($p = 120$) |
| *DPLE* | **99%** ($p = 23$) | **49.02%** ($p = 100$) |
| | Kernelized | |
| KLFDA | 99% ($p = 24$) | 48.93% ($p = 95$) |
| KLDE | 99.25% ($p = 21$) | 52.64% ($p = 200$) |
| *KDPLE* | **99.25%** ($p = 23$) | **53.70%** ($p = 253$) |

**Table 1**. Best recognition rates in the evaluation datasets.

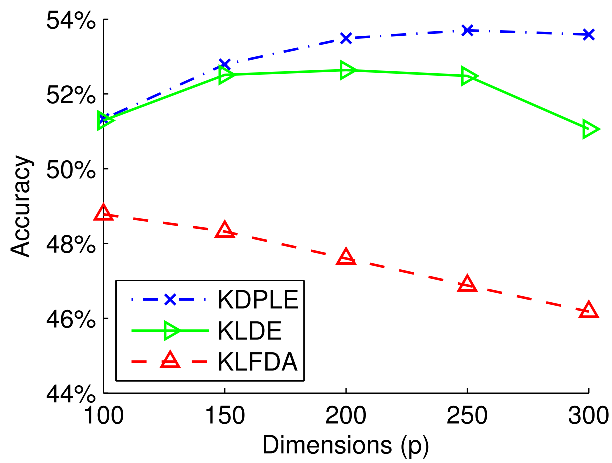


**Fig. 1**. Sketch recognition accuracy of kernelized algorithms across varying dimensionality using k-nn classification. KDPLE constantly outperforms the rest methods.

SKETCH dataset is more challenging and exposes the disadvantages of each method. PCA and LDA output much lower rates than NN, highlighting the limitation induced by the Gaussian distribution assumption of these methods. LPP also fails to meet NN accuracy because of its unsupervised nature. LPP solely focuses on data structure preservation, hence generates non-discriminant projection bases. All the linear versions of the discriminant manifold algorithms, namely LFDA, LDE and DPLE, outperform the NN accuracy. DPLE demonstrates the higher rate among the linear techniques.

We further evaluate the recognition accuracy of the kernelized versions of the algorithms. KDPLE achieves again the best recognition rate and performs constantly better than KLFDA and KLDE under various dimensionality setups as Figure 1 shows. The superiority of KDPLE is accredited to the exploitation of the pairwise relationships between data pairs and the importance factor assigned to each train sample by matrix $\mathbf{D}$. We also note that the kernelized extension of DPLE offers a significant accuracy boost.

## 4. CONCLUSIONS

We have presented DPLE, a supervised manifold learning algorithm that generates discriminant embeddings with a convex optimization process based on pairwise relations between the data. A non-linear variant of the algorithm is also illustrated. We have demonstrated the superiority of DPLE over competitive dimensionality reduction techniques in two recognition datasets. Future work could be concentrated on the online updating of the projection matrix upon new sample arrival.

# Acknowledgments

This work was partially supported by EU project 3DLife under grant agreement FP7-247688.

## 5. REFERENCES


[1] L.J.P. van der Maaten, E. O. Postma, and H. J. van den Herik, "Dimensionality reduction: A comparative review," 2008.

[2] Keinosuke Fukunaga, *Introduction to statistical pattern recognition*, Academic Press Professional, Inc., San Diego, CA, USA, 1990.

[3] Masashi Sugiyama, "Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis," *J. Mach. Learn. Res.*, vol. 8, pp. 1027–1061, May 2007.

[4] J. B. Tenenbaum, V. Silva, and J. C. Langford, "A Global Geometric Framework for Nonlinear Dimensionality Reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[5] Xiaofei He and Partha Niyogi, "Locality preserving projections," in *In Advances in Neural Information Processing Systems*. 2003, MIT Press.

[6] Xiaofei He, "Incremental semi-supervised subspace learning for image retrieval," in *Proceedings of ACM Conference on Multimedia*, 2004.

[7] Hwann-Tzong Chen, Huang-Wei Chang, and Tyng-Luh Liu, "Local discriminant embedding and its variants," in *Proceedings IEEE Computer Vision and Pattern Recognition (CVPR'05) - Volume 2 - Volume 02*, 2005, CVPR '05, pp. 846–853.

[8] Yen-Yu Lin, Tyng-Luh Liu, and Hwann-Tzong Chen, "Semantic manifold learning for image retrieval," in *Proceedings of ACM Conference on Multimedia*. 2005, pp. 249–258, ACM.

[9] Weiwei Yu, Xiaolong Teng, and Chongqing Liu, "Face recognition using discriminant locality preserving projections," *Image and Vision Computing*, vol. 24, no. 3, pp. 239 – 248, 2006.

[10] Xiaofei He, Deng Cai, and Jiawei Han, "Learning a maximum margin subspace for image retrieval," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 20, no. 2, pp. 189 –201, feb. 2008.

[11] Bernhard Scholkopf and Alexander J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, MIT Press, Cambridge, MA, USA, 2001.

[12] F. S. Samaria and A. C. Harter, "Parameterisation of a stochastic model for human face identification," in *Applications of Computer Vision, 1994., Proceedings of the Second IEEE Workshop on*, pp. 138–142.

[13] Mathias Eitz, James Hays, and Marc Alexa, "How do humans sketch objects?," *ACM Transactions on Graphics (Proceedings SIGGRAPH)*, vol. 31, no. 4, pp. 44:1–44:10, 2012.