# Graph Laplacian Kernels for Object Classification from a Single Example*

Hong Chang & Dit-Yan Yeung
Department of Computer Science, Hong Kong University of Science and Technology
{hongch,dyyeung}@cs.ust.hk

## Abstract

*Classification with only one labeled example per class is a challenging problem in machine learning and pattern recognition. While there have been some attempts to address this problem in the context of specific applications, very little work has been done so far on the problem under more general object classification settings. In this paper, we propose a graph-based approach to the problem. Based on a robust path-based similarity measure proposed recently, we construct a weighted graph using the robust path-based similarities as edge weights. A kernel matrix, called graph Laplacian kernel, is then defined based on the graph Laplacian. With the kernel matrix, in principle any kernel-based classifier can be used for classification. In particular, we demonstrate the use of a kernel nearest neighbor classifier on some synthetic data and real-world image sets, showing that our method can successfully solve some difficult classification tasks with only very few labeled examples.*

## 1. Introduction

Classification with only one labeled example per class is a challenging problem in machine learning and pattern recognition, since straightforward application of standard classification methods is infeasible due to overfitting.

Recently, some computer vision researchers proposed to study the *single training example per class problem* for face recognition tasks. The most common approach in this line of research is to augment the training information as much as possible from a single labeled face image. For example, some researchers proposed to extract various configural features from a single labeled face image [1]. Others proposed to obtain multiple training examples for each class by partitioning each labeled face image into a set of subimages or use the derived images to augment the training set [13, 4]. With this approach of augmenting the original training set which contains only one labeled example per person, traditional face recognition methods such as linear discriminant analysis (LDA) and its variants can then be used. The most recent work on face recognition from a single image under varying illumination uses a 3D spherical harmonic basis morphable model (SHBMM) [12] to eliminate the illumination effects and hence achieve high recognition rates. However, these methods which are devised for specific applications (face recognition in particular) cannot be used for general object classification tasks.

In the machine learning community, semi-supervised learning methods have emerged over the past few years as a promising approach to improving classification performance with the aid of unlabeled data. For example, Zhou et al. proposed an interesting graph-based method to propagate the labels from the limited labeled nodes to the unlabeled nodes in the graph based on the assumption of local and global consistency [14]. Fink proposed to use a nearest neighbor classifier based on class relevance metrics to discriminate between two target classes from a single example, where the class relevance metrics are learned from multiple labeled examples of other related classes [6]. Among these semi-supervised classification methods are some kernel-based methods that have been demonstrated to exhibit promising performance. Some of them are based on constructing appropriate kernels by transforming the spectrum of the graph over labeled and unlabeled data, such as cluster kernels [3] and nonparametric transformation of graph kernels [15]. Other methods include the connectivity kernels [7], which are directly induced from a so-called path-based dissimilarity measure defined on the weighted graph. However, despite the promising performance of these methods as demonstrated on some semi-supervised classification or clustering tasks, they are not robust enough in the presence of noise points or outliers.

In this paper, we propose a novel kernel-based method for general object classification applications with as few as only a single labeled example per class. More specifically, inspired by our recent work [2], we propose to construct a weighted graph based on some robust similarity measure and then define a kernel matrix based on the graph Laplacian for use in the subsequent kernel-based classification

*This research has been supported by Competitive Earmarked Research Grant HKUST621305 from the Research Grants Council of the Hong Kong Special Administrative Region, China.

algorithm. Not only does the proposed method make good use of unlabeled data to solve the classification problem in a semi-supervised setting, but it also exhibits high robustness against the data noise and insensitivity towards the model parameter.

The rest of this paper is organized as follows. In Section 2, we present the graph Laplacian kernels that incorporate a robust similarity measure. A semi-supervised classification algorithm using the graph Laplacian kernels is then presented in Section 3. Section 4 gives some experimental results on both synthetic and real-world data to demonstrate the effectiveness of our method. Finally, we give some concluding remarks in the last section.

## 2. Graph Laplacian Kernels

### 2.1. Path-Based Similarities on Graph

We denote a set of $n$ data points in a multidimensional Euclidean space by $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$. The data set can be represented as a fully connected undirected graph $G = (V, E)$, with vertex set $V = \{1, \ldots, n\}$ and edge set $E \subset V \times V$. For each edge $(i, j) \in E, i \neq j$, we assign a weight $w'_{ij} = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/(2\sigma^2))$ to represent the pairwise similarity between vertices $i$ and $j$ (corresponding to data points $\mathbf{x}_i$ and $\mathbf{x}_j$, respectively), with the scaling parameter $\sigma$ specifying the spread of the Gaussian window in the input space. We assume that there is no self-loop in the graph and hence we take $w'_{ii} = 0$ in the sequel.

Note that the pairwise similarities $w'_{ij}$ are solely determined by the Euclidean distances between the corresponding points in the input space, revealing no information about whether the points belong to the same class or not. To capture such information, one effective approach is to exploit the underlying manifold structure of the global data distribution so that points belonging to the same manifold (or class) have higher similarity while points belonging to different manifolds (or classes) have lower similarity. In general, the manifolds can be nonlinear and elongated in structure. Following the formulation of a path-based dissimilarity measure originally proposed in [8], we defined a path-based similarity measure which can implicitly convert elongated manifolds into compact ones [2].

The path-based similarity $w_{ij}$ between vertices $i$ and $j$ is defined as follows. For each path $p$ connecting vertices $i$ and $j$, the *effective similarity* $w^p_{ij}$ between the two vertices is the minimum edge weight along the path, i.e., $w^p_{ij} = \min_{1 \leq h < |p|} w'_{p[h]p[h+1]}$, where $p[h]$ denotes the $h$th vertex along path $p$ from vertex $i$ to vertex $j$ and $|p|$ denotes the number of vertices in $p$. Then, $w_{ij}$ is the maximum among all effective similarities corresponding to all paths from vertex $i$ to vertex $j$:

$$w_{ij} = \max_{p \in \mathcal{P}_{ij}} w^p_{ij} = \max_{p \in \mathcal{P}_{ij}} \left\{ \min_{1 \leq h < |p|} w'_{p[h]p[h+1]} \right\}. \quad (1)$$

However, as pointed out in [2], this path-based similarity measure is very sensitive to noise and outliers. In [2], this measure was further extended to a robust version, as to be analyzed in the next subsection.

### 2.2. Robust Adjacency Matrix

Robust statistical estimation techniques are estimation techniques which are insensitive to the presence of gross errors or outliers that do not fit to the stochastic model of parameter estimation [11]. Here we use the idea of M-estimation based on maximum likelihood considerations to devise a more robust similarity measure.

Consider the data points in the neighborhood of $\mathbf{x}_i$ as realizations from some estimator of $\mathbf{x}_i$. The squared residual error $e^2_{ij}$ of $\mathbf{x}_i$ for neighbor $\mathbf{x}_j$ can be defined based on the distance between $\mathbf{x}_i$ and $\mathbf{x}_j$: $e^2_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|^2/|\mathcal{N}_i|$, where $\mathcal{N}_i$ denotes the neighborhood set of $\mathbf{x}_i$ and $|\mathcal{N}_i|$ its cardinality. The standard least squares method tries to minimize $\sum_{i=1}^n \sum_{\mathbf{x}_j \in \mathcal{N}_i} e^2_{ij}$.

An M-estimator tries to reduce the influence of outliers through replacing the squared residual error by a convex function $\rho(\cdot)$. Solving the problem of $\min \sum_{i=1}^n \sum_{\mathbf{x}_j \in \mathcal{N}_i} \rho(e_{ij})$ with respect to some parameter to be estimated is equivalent to solving the following *iterative reweighted least squares* problem:

$$\min \sum_{i=1}^{n} \sum_{\mathbf{x}_j \in \mathcal{N}_i} a(e^{(t-1)}_{ij}) e^2_{ij}, \quad (2)$$

where $(t-1)$ denotes the iteration number and $a(\cdot)$ is the *weight function* for estimator $\rho(\cdot)$, which is defined as $a(e_{ij}) = \rho'(e_{ij})/e_{ij}$. During the robust estimation procedure, the weight function is recomputed after each iteration.

Usually, the robust estimator $\rho(\cdot)$ is selected to grow more slowly than the quadratic function. One example is the Welsch function used in [2]:

$$\rho(e_{ij}) = \frac{c^2}{2}[1 - \exp(-(e_{ij}/c)^2)], \quad (3)$$

where $c$ is some positive parameter. The corresponding weight function is $a(e_{ij}) = \rho'(e_{ij})/e_{ij} = \exp(-(e_{ij}/c)^2) = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/(|\mathcal{N}_i|\, c^2))$.

Based on the robust estimator, we can obtain a weight $\alpha'_i$ for each data point $\mathbf{x}_i$ by summing up the weight function values $a(e_{ij})$ of all its neighbors. It is worth noting that by setting $c = \sqrt{\frac{2}{|\mathcal{N}_i|}}\,\sigma$, the weight can be expressed solely in terms of the original similarities:

$$\alpha'_i = \sum_{\mathbf{x}_j \in \mathcal{N}_i} a(e_{ij}) = \sum_{\mathbf{x}_j \in \mathcal{N}_i} w'_{ij}. \quad (4)$$

Making use of the normalized weights, $\alpha_i = \alpha'_i / \max_{\mathbf{x}_i \in \mathcal{X}} \alpha'_i$, the robust path-based similarity measure is expressed as [2]:

$$w_{ij} = \max_{p \in \mathcal{P}_{ij}} \left\{ \min_{1 \le h < |p|} \alpha_{p[h]} \alpha_{p[h+1]} w'_{p[h]p[h+1]} \right\}. \quad (5)$$

This measure can reflect not only the possibility for $\mathbf{x}_i$ and $\mathbf{x}_j$ to belong to the same class, but also the genuine similarity even in the presence of noise points or outliers. Moreover, this measure is not sensitive to the choice of the Gaussian parameter $\sigma$.

From (5), we can define a robust adjacency matrix as $\mathbf{W} = [w_{ij}]$. The weighted graph $G$ with the robust adjacency matrix $\mathbf{W}$ will then be used to construct a graph Laplacian kernel.

### 2.3. Constructing Graph Laplacian Kernels

The *graph Laplacian* $\mathbf{L}$ is defined in terms of the adjacency matrix $\mathbf{W}$ as: $\mathbf{L} = \mathbf{D} - \mathbf{W}$, where $\mathbf{D}$ is a diagonal matrix with the $i$th diagonal entry defined as $D_{ii} = \sum_{j=1}^{n} w_{ij}$. The *normalized graph Laplacian* $\tilde{\mathbf{L}}$ is defined as: $\tilde{\mathbf{L}} = \mathbf{D}^{-\frac{1}{2}} \mathbf{L} \mathbf{D}^{-\frac{1}{2}}$ [5].

Let $\mathcal{R}(G)$ denote the linear space of real-valued functions defined on $G$ and $\{\lambda_i, \mathbf{u}_i\}_{i=1}^{n}$ denote an eigensystem of $\tilde{\mathbf{L}}$, where $\lambda_1 = \cdots = \lambda_r = 0$ and $0 < \lambda_{r+1} \le \cdots \le \lambda_n$. We define a Hilbert space of functions on $G$, $\mathcal{H}(G) = \{\mathbf{g} \mid \mathbf{g}^T \mathbf{u}_i = 0, i = 1, \ldots, r\}$, which is a linear subspace of $\mathcal{R}(G)$ orthogonal to the eigenvectors of $\tilde{\mathbf{L}}$ with zero eigenvalue. Similar to [10] for $\mathbf{L}$, we can prove that the pseudoinverse of $\tilde{\mathbf{L}}$ is the reproducing kernel of $\mathcal{H}(G)$:

**Theorem 1** *The pseudoinverse of the normalized graph Laplacian, $\tilde{\mathbf{L}}^+$, is the reproducing kernel of $\mathcal{H}(G)$.*

**Proof:** We can express $\tilde{\mathbf{L}}$ and $\tilde{\mathbf{L}}^+$ as $\tilde{\mathbf{L}} = \sum_{i=r+1}^{n} \lambda_i \mathbf{u}_i \mathbf{u}_i^T$ and $\tilde{\mathbf{L}}^+ = \sum_{i=r+1}^{n} \lambda_i^{-1} \mathbf{u}_i \mathbf{u}_i^T$. Let $\mathbf{K} = \tilde{\mathbf{L}}^+$ and $\mathbf{K}_i$ be the $i$th column of $\mathbf{K}$. Hence, $\tilde{\mathbf{L}}^+ \tilde{\mathbf{L}} = \sum_{i=r+1}^{n} \mathbf{u}_i \mathbf{u}_i^T = \mathbf{I} - \sum_{i=1}^{r} \mathbf{u}_i \mathbf{u}_i^T$. For any $\mathbf{g} = (g_1, \ldots, g_n)^T \in \mathcal{H}(G)$, we have $\tilde{\mathbf{L}}^+ \tilde{\mathbf{L}} \mathbf{g} = \mathbf{g} - \sum_{i=1}^{r} \mathbf{u}_i \mathbf{u}_i^T \mathbf{g} = \mathbf{g}$ since the second term is zero for all $\mathbf{g} \in \mathcal{H}(G)$. Let $\mathbf{e}_i$ be an $n$-dimensional indicator vector with all entries equal to 0 except the $i$th entry which is 1. So, $g_i = \mathbf{e}_i^T \mathbf{g} = \mathbf{e}_i^T \tilde{\mathbf{L}}^+ \tilde{\mathbf{L}} \mathbf{g} = \mathbf{e}_i^T \mathbf{K} \tilde{\mathbf{L}} \mathbf{g} = \mathbf{K}_i^T \tilde{\mathbf{L}} \mathbf{g} = \langle \mathbf{K}_i, \mathbf{g} \rangle$, which shows that the reproducing property holds. Hence $\mathbf{K} = \tilde{\mathbf{L}}^+$ is the reproducing kernel of $\mathcal{H}(G)$. □

We call the kernel matrix $\mathbf{K} = \tilde{\mathbf{L}}^+$ the *graph Laplacian kernel*. As opposed to using generic, data-independent kernels such as the polynomial kernel and the Gaussian RBF kernel, this approach results in a data-dependent kernel which captures the structure in the data. Specifically, in our case, it exploits the robust adjacency matrix defined above to construct a data-dependent graph Laplacian kernel that can be used subsequently by any kernel-based method.

## 3. Object Classification: A Kernel Approach

Let $c$ be the number of classes in the data set $\mathcal{X}$. In our setting, each class has exactly one labeled data point. For a labeled data point $\mathbf{x}_i$, we denote its class label as $y_i \in \{1, \ldots, c\}$. A labeled data point can also be represented more completely as $(\mathbf{x}_i, y_i)$. Without loss of generality, we assume that the first $c$ data points are labeled as: $\{(\mathbf{x}_1, 1), \ldots, (\mathbf{x}_c, c)\}$. The remaining $n-c$ data points, $\{\mathbf{x}_{c+1}, \ldots, \mathbf{x}_n\}$, are unlabeled.

With the graph Laplacian kernel $\mathbf{K}$ computed from both the labeled and unlabeled data points, we implicitly convert the elongated clusters into compact ones. The similarity between two data points belonging to the same class tends to have larger value than that between two data points from different classes. Therefore, even a simple kernel nearest neighbor classifier is expected to achieve impressive performance with a single training example per class. More specifically, the following classification is performed by comparing the squared Euclidean distance between the labeled and unlabeled data points in the kernel-induced feature space with $\phi$ being the implicit feature map: $\|\phi(\mathbf{x}_i) - \phi(\mathbf{x}_j)\|^2 = K_{ii} + K_{jj} - 2K_{ij}$, with $i = 1, \ldots, c$ and $j = c+1, \ldots, n$. Note that the term $K_{jj}$ is irrelevant to the classification of an example $\mathbf{x}_j$ to one of the $c$ classes. The overall classification algorithm based on the graph Laplacian kernel is summarized in Figure 1 below.

**Input:** Labeled data set $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_c, y_c)\}$
Unlabeled data set $\{\mathbf{x}_{c+1}, \ldots, \mathbf{x}_n\}$
**Offline kernel construction:**
Construct fully connected graph $G$ over $\mathcal{X}$
Compute robust adjacency matrix $\mathbf{W}$ (Equation (5))
Compute normalized graph Laplacian $\tilde{\mathbf{L}}$ from $\mathbf{W}$
Compute graph Laplacian kernel $\mathbf{K} = \tilde{\mathbf{L}}^+$
**Online classification:**
**for** $j = c+1, \ldots, n$ **do**
$y_j = \arg\min_{i=1,\ldots,c} \{K_{ii} - 2K_{ij}\}$
**end**
**Output:** Labels $y_j, j = c+1, \ldots, n$.

Figure 1. Classification algorithm based on graph Laplacian kernels

## 4. Experiments

In this section, we empirically evaluate the semi-supervised classification algorithm described above on both synthetic and real-world data sets.

### 4.1. Experimental Setup

We compare our kernel-based classification method described in Section 3 with two related methods proposed previously by other researchers. The first method is a kernel nearest neighbor classifier based on the connectivity kernel [7]. In their work, the connectivity kernel is induced from effective dissimilarities over the weighted graph, which are related to the robust path-based similarities used in our

method. The authors apply the connectivity kernel to some clustering tasks. Although their method can deal with some challenging clustering problems with elongated data structures, the effective dissimilarity measure (hence the connectivity kernel) does not possess the robustness property, as will be shown in our experimental results below. Another method is the method proposed by Zhou et al. [14]. Their semi-supervised classification method is not a kernel method. Based on enforcing local and global consistency, their method has shown very promising results for some difficult classification tasks with very limited labeled data. Besides these two methods, we also include an ordinary nearest neighbor classifier for baseline comparison.[1] In summary, the following four classification methods are included in our comparative study:

| | |
|---|---|
| 1-NN | 1-nearest neighbor classifier |
| Connectivity kernel | Kernel 1-NN using connectivity kernel |
| LLGC | Local and global consistency method |
| Graph Laplacian kernel | Kernel 1-NN using our proposed graph Laplacian kernel |

We perform experiments on a synthetic data set and two real-world image data sets (MNIST digits and UMIST face images). For the image data sets, we use the classification accuracy on the unlabeled data to quantify the classification performance of different methods. For each data set, we randomly generate 10 different training sets, with each training set containing one example per class. The average classification results over the 10 runs are reported.

### 4.2. Synthetic Data

We first demonstrate the power of our proposed method on a synthetic data set.

Figure 2(a) shows a noisy 2-moon data set. Data points with the same mark and color belong to the same class. This synthetic data set is commonly used in some recent semi-supervised learning work. However, the difference is that we also add some noise points (black dots) to the otherwise clean data set.

In our classification experiments, we choose one example for each class as training examples, which are shown with larger marks in the figures. If the 2-moon data set contains no noise points, both the connectivity kernel and LLGC can classify the clean data very well even with only one labeled example per class. This has been empirically verified by us for the connectivity kernel method. As for LLGC, please refer to [14] for illustration. However, neither of the two methods can give good classification results when there exist noise points in the data, as shown in Figure 2(b) and (c). Note that some noise points located between the two moons end up connecting the two classes. Due to the existence of noise points, the dissimilarity measure, from which the connectivity kernel is induced, gives much lower dissimilarity values than they should be to point pairs residing in different classes. As a consequence, its classification result is not satisfactory. Similarly, the propagation procedure of the LLGC algorithm is also seriously affected by the noise points. On the other hand, our classification method based on a data-dependent graph Laplacian kernel gives fairly good result, as shown in Figure 2(d). This shows that the robust path-based similarity measure is indeed very effective in reducing the influence of the outliers.

### 4.3. MNIST Digits

We further perform experiments on handwritten digits from the well-known MNIST database.[2] Unlike the synthetic data, this data set is of much higher dimensionality. The digits in the database have been size-normalized and centered to $28 \times 28$ gray-level images, so the dimensionality of the digit space is 784. In our experiments, we randomly choose 200 images for each digit from a total of 60,000 digit images in the MNIST training set.

Figure 3 shows the classification results on the data set containing digits "8" and "9". The digit images are plotted based on the two leading principal components estimated from the data, as shown in Figure 3(a), where "8" and "9" are represented by (red) dots and (blue) crosses, respectively. As we can see, the data points form relatively compact clusters in the 2D space with some outliers located between them or even inside the other clusters. Two labeled examples are denoted by larger marks. The classification results using different methods are shown in Figure 3(b), (c) and (d). It can be seen that our method outperforms the connectivity kernel and LLGC methods.

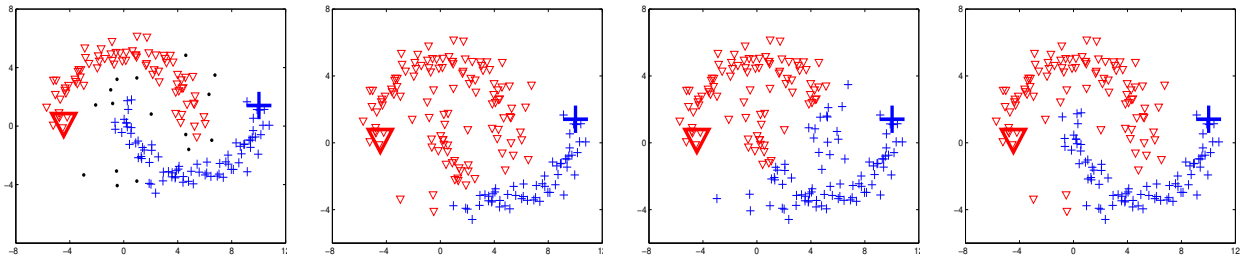
We conduct more experiments on different digit subsets under the same experimental settings. The classification results are summarized in Table 1. There are two result values for each classification method and each data set. The upper value is the test accuracy averaged over 10 random trials and the lower value represents the standard deviation. From the results, we can see that the connectivity kernel generally can improve the classification performance but LLGC leads to poor results. On the other hand, our proposed method outperforms all other classification methods for all data sets.

### 4.4. UMIST Face Images

The UMIST face database [9] consists of 564 gray-level images from 20 persons. Images of each person contain a range of poses from profile to frontal views. The original pre-cropped images are of size $112 \times 92$. We down-sample

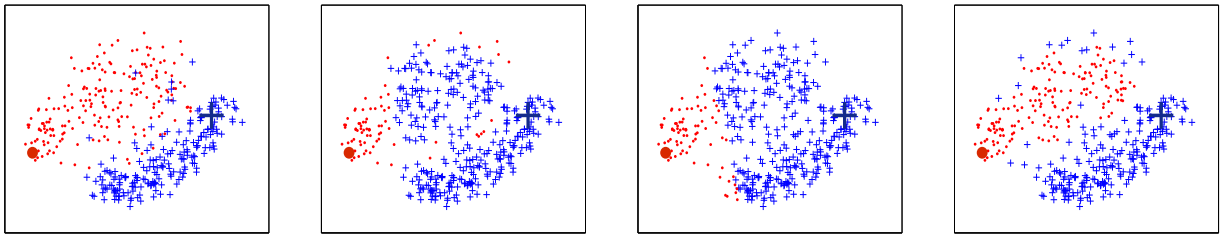
[1] Note that a kernel nearest neighbor classifier with the RBF kernel gives the same result as an ordinary nearest neighbor classifier, since the RBF kernel does not change the relative ordering of the distances in the input space.

[2] http://yann.lecun.com/exdb/mnist/



(a) (b) (c) (d)

Figure 2. Classification results for noisy 2-moon data set: (a) 2-moon data with some noise points (black dots); and classification results using (b) connectivity kernel; (c) LLGC; and (d) graph Laplacian kernel.



(a) (b) (c) (d)

Figure 3. Classification results for digits "8" and "9": (a) digit images plotted based on the two leading principal components; and classification results using (b) connectivity kernel; (c) LLGC; and (d) graph Laplacian kernel.

Figure 4. Examples of synthesized "noise" face images.

them to smaller ones of size $56 \times 46$. In our experiments, a subset of 149 face images belonging to the first five people are selected. The numbers of images from each person are 38, 35, 26, 24 and 26, respectively.

Besides these clean images, we synthesize 10 more images artificially. Each of them ($\mathbf{f}_{\text{new}}$) is the weighted average of two randomly selected face images ($\mathbf{f}_i$ and $\mathbf{f}_j$) from two different persons in the image set: $\mathbf{f}_{\text{new}} = 0.75 \times \mathbf{f}_i + 0.25 \times \mathbf{f}_j$. As a result, the new face image can be seen as a "noise" image of $\mathbf{f}_i$ with shade from another person ($\mathbf{f}_j$). Some examples of the "noise" face images for the first person are shown in Figure 4. These artificially created images act as outliers in the subsequent experiments.

Instead of computing the leading principal components, we perform classification directly on the high-dimensional feature vectors extracted from the face images. For visualization purpose, we use multidimensional scaling (MDS) to embed the face images into a 3-dimensional space, as shown in Figure 5. We can see that the face images of different persons form elongated manifolds. The "noise" faces marked as black circles scatter among the five manifolds.

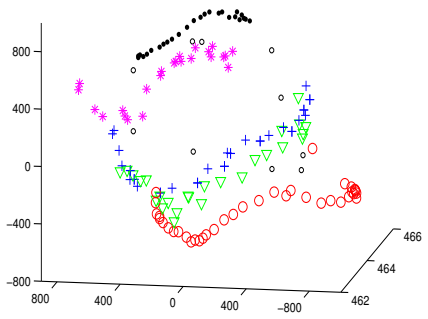



Figure 5. 3D embeddings of face images from five persons. Ten "noise" faces are marked as black circles.

The average test accuracy and the standard deviation for five different methods are summarized in Table 2. The left column contains the classification results for the clean image set, while the right column contains the results when "noise" images are added. For both settings, our proposed kernel-based method achieves the highest accuracy.

| Method | $\{1,4\}$ | $\{1,7\}$ | $\{3,6\}$ | $\{8,9\}$ | $\{0,6,8\}$ | $\{1,3,7\}$ |
|---|---|---|---|---|---|---|
| 1-NN | 0.9261 | 0.8920 | 0.8588 | 0.8201 | 0.7296 | 0.6197 |
| | $\pm$0.0013 | $\pm$0.0034 | $\pm$0.0154 | $\pm$0.0116 | $\pm$0.0317 | $\pm$0.0002 |
| Connectivity kernel | 0.9644 | 0.9298 | 0.9048 | 0.7573 | 0.6928 | **0.6372** |
| | $\pm$0.0001 | $\pm$0.0000 | $\pm$0.0000 | $\pm$0.0114 | $\pm$0.0506 | $\pm$0.0172 |
| LLGC | 0.8327 | 0.8332 | 0.7983 | 0.7181 | 0.6698 | 0.5801 |
| | $\pm$0.0103 | $\pm$0.0039 | $\pm$0.0068 | $\pm$0.0003 | $\pm$0.0269 | $\pm$0.0102 |
| Graph Laplacian kernel | **0.9739** | **0.9548** | **0.9457** | **0.9075** | **0.7394** | 0.6284 |
| | $\pm$0.0001 | $\pm$0.0001 | $\pm$0.0000 | $\pm$0.0163 | $\pm$0.0146 | $\pm$0.0257 |

Table 1. Classification results on several subsets of the MNIST digit database.

| Method | Clean data | Noisy data |
|---|---|---|
| 1-NN | 0.5361 | 0.5215 |
| | $\pm$0.0091 | $\pm$0.0063 |
| Connectivity kernel | 0.8924 | 0.6458 |
| | $\pm$0.0032 | $\pm$0.0054 |
| LLGC | 0.5646 | 0.5187 |
| | $\pm$0.0112 | $\pm$0.0062 |
| Graph Laplacian kernel | **0.9722** | **0.8326** |
| | $\pm$0.0156 | $\pm$0.0029 |

Table 2. Classification results on a subset of the UMIST database

So far we have only considered classification under the transductive setting. For new test data points, we can compute the robust path-based similarities between the test points and the labeled data points based on the constructed graph $G$ with the robust adjacency matrix $\mathbf{W}$. Nearest neighbor classification is then performed. Preliminary experiments show that our method is also effective under this out-of-sample setting.

## 5. Concluding Remarks

We have proposed a novel kernel-based method for object classification with a single labeled example per class. Experimental results on both synthetic and real-world data sets verify the effectiveness of the method.

In our future research, we will perform more experiments on the out-of-sample extension and apply our method to more real-world problems.

## References


[1] R. Brunelli and T. Poggio. Face recognition: features versus templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(10):1042–1062, 1993. 1

[2] H. Chang and D. Yeung. Robust path-based spectral clustering with application for image segmentation. In *Proceedings of the Tenth IEEE International Conference on Computer Vision*, 2005. 1, 2

[3] O. Chapelle, J. Weston, and B. Schölkopf. Cluster kernels for semi-supervised learning. In *Advances in Neural Information Processing Systems 15*. 2003. 1

[4] S. Chen, J. Liu, and Z.-H. Zhou. Making FLDA applicable to face recognition with one sample per person. *Pattern Recognition*, 37(7):1553–1555, 2004. 1

[5] F. Chung. *Spectral Graph Theory*, volume 92 of *Regional Conference Series in Mathematics*. American Mathematical Society, 1997. 3

[6] M. Fink. Object classification from a single example utilizing class relevance metrics. In *Advances in Neural Information Processing Systems 17*. 2004. 1

[7] B. Fischer, V. Roth, and J. M. Buhmann. Clustering with the connectivity kernel. In *Advances in Neural Information Processing Systems 16*. 2004. 1, 3

[8] B. Fischer, T. Zöller, and J. M. Buhmann. Path based pairwise data clustering with application to texture segmentation. *Energy Minimization Methods in Computer Vision and Pattern Recognition*, 2134:235–250, 2001. 2

[9] D. B. Graham and N. M. Allinson. Characterizing virtual eigensignatures for general purpose face recognition. In H. Wechsler, P. J. Phillips, V. Bruce, F. Fogelman-Soulie, and T. S. Huang, editors, *Face Recognition: From Theory to Applications, NATO ASI Series F, Computer and Systems Sciences*, volume 163, pages 446–456. 1998. 4

[10] M. Herbster, M. Pontil, and L. Wainer. Online learning over graphs. In *Proceedings of the Twenty-Second International Conference on Machine Learning*, Bonn, Germany, 2005. 3

[11] P. Huber. Robust regression: asymptotics, conjectures, and Monte Carlo. *Annals of Statistics*, 1(5):799–821, 1973. 2

[12] S. W. L. Zhang and D. Samaras. Face synthesis and recognition from a single image under arbitrary unknown lighting using a spherical harmonic basis morphable model. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2005. 1

[13] A. Martinez. Recognizing imprecisely localized, partially occluded, and expression variant faces from a single sample per class. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(6):748–763, 2002. 1

[14] D. Zhou, O. Bousquet, T. Lal, J. Weston, and B. Schölkopf. Learning with local and global consistency. In *Advances in Neural Information Processing Systems 16*. 2004. 1, 4

[15] X. Zhu, J. Kandola, Z. Ghahramani, and J. Lafferty. Nonparametric transforms of graph kernels for semi-supervised learning. In *Advances in Neural Information Processing Systems 17*. 2004. 1