

Robust Path-Based Spectral Clustering

Hong Chang

Xerox Research Centre Europe

6 chemin de Maupertuis, 38240 Meylan, France

Dit-Yan Yeung

Department of Computer Science and Engineering

Hong Kong University of Science and Technology

Clear Water Bay, Kowloon, Hong Kong

March 2, 2007

Abstract

Spectral clustering and path-based clustering are two recently developed clustering approaches that have delivered impressive results in a number of challenging clustering tasks. However, they are not robust enough against noise and outliers in the data. In this paper, based on M-estimation from robust statistics, we develop a robust path-based spectral clustering method by defining a robust path-based similarity measure for spectral clustering under both unsupervised and semi-supervised settings. Our proposed method is significantly more robust than spectral clustering and path-based clustering. We have performed experiments based on both synthetic and real-world data, comparing our method with some other methods. In

particular, color images from the Berkeley Segmentation Dataset and Benchmark are used in the image segmentation experiments. Experimental results show that our method consistently outperforms other methods due to its higher robustness.

Keywords: path-based clustering, spectral clustering, robust statistics, unsupervised learning, semi-supervised learning, image segmentation.

1 Introduction

Clustering has been among the most active research topics in machine learning and pattern recognition. While many traditional clustering algorithms have been developed over the past few decades [6, 16], some new clustering algorithms emerged over the last few years give very promising results on some challenging tasks. Among them are *spectral clustering* [20, 24, 25, 28] and *path-based clustering* [8, 9, 10], which have demonstrated excellent performance on some clustering tasks involving highly nonlinear and elongated clusters in addition to compact clusters.

Despite the promising performance of these algorithms demonstrated on some difficult data sets, there exist some other situations when these algorithms do not perform well. Consider some examples in Figure 1. Although spectral clustering works perfectly well on the 2-circle data set (Figure 1(a)), it gives very poor result on the 3-spiral data set (Figure 1(b)). The poor clustering result is due mainly to the particular choice of the affinity matrix, which is usually defined in a way similar to the Gaussian kernel based on inter-point Euclidean distance in the input space. However, if path-based criteria from path-based clustering are used to define the (dis)similarity between points to form the affinity matrix before spectral clustering is applied, the three clusters in the 3-spiral data set can be found correctly, as shown in Figure 1(c).

While the combined use of path-based clustering and spectral clustering, referred to as *path-based spectral clustering* here, seems to be very effective, we will show later in the

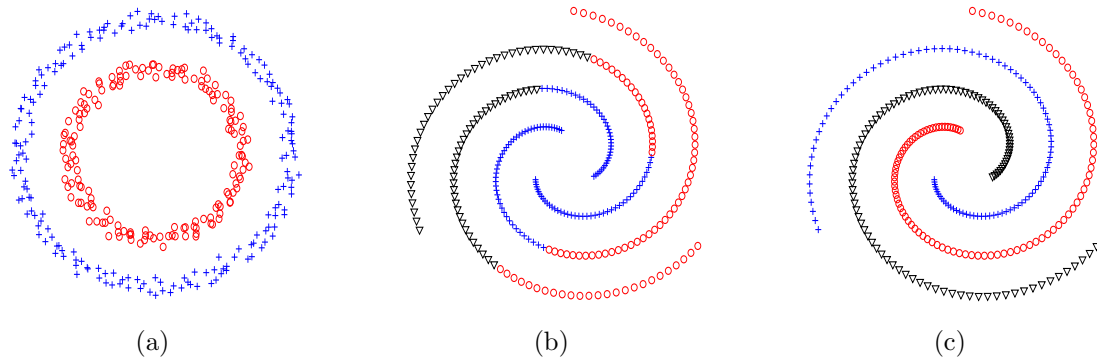


Figure 1: Simple examples of spectral clustering and path-based clustering: (a) spectral clustering result for 2-circle data set (same result given by path-based clustering); (b) spectral clustering result for 3-spiral data set; and (c) path-based spectral clustering result for 3-spiral data set.

paper that this combined method, like the separate use of spectral clustering or path-based clustering, is not robust enough against noise and outliers which commonly exist in real-world data.

In this paper, based on robust statistical techniques [15], we propose a novel scheme to make path-based (spectral) clustering more robust. Our work is built upon the recent work of Fischer et al. [8, 9, 10]. We devise an M-estimator and use it to define a robust path-based similarity measure which takes into account the existence of noise and outliers in the data and hence brings about robustness in the method.

The rest of this paper is organized as follows. Some related work is briefly reviewed in Section 1.1. In Section 2, we propose a robust path-based similarity measure based on robust statistics, with which a robust path-based spectral clustering algorithm can be devised. In Section 3, we extend this method for semi-supervised clustering with the supervisory information in the form of pairwise similarity and dissimilarity constraints. Experimental results on synthetic data as well as color image segmentation are presented in Sections 4 and 6, respectively, comparing our method with non-robust methods. Finally,

some concluding remarks are given in the last section.

1.1 Related Work

Some related clustering methods have been proposed in the literature. Besides spectral clustering and path-based clustering, there also exist some other methods that are capable of finding elongated structures. Hierarchical agglomerative clustering based on the single linkage algorithm [16] is efficient and has been widely used, but it is well known that this method is highly sensitive to outliers. Mean shift clustering [5] and spectral clustering have shown good performance in some clustering tasks. However, both of them are sensitive to the preselection of the kernel bandwidth parameter.

Fischer and Buhmann have a more recent paper which uses bagging (or *bootstrap aggregating*) with path-based clustering to address the robustness issue [7]. Like other bagging applications, data resampling is used to reduce data variance so that the results are less sensitive to noise. However, since many bootstrap samples are involved, efficiency has to be sacrificed in exchange for effectiveness of the method. Moreover, some bootstrap samples of the input data may still contain outliers which can affect the clustering results. Another robust variant of path-based distance measure is proposed in [4], where the authors “soften” the max operation to reduce the influence by outliers.

Zelnik-Manor and Perona proposed to use local scale for spectral clustering [31]. The local scaling parameter for each data point is defined as $\sigma_i = d(\mathbf{x}_i, \mathbf{x}_K)$, where K has to be chosen in advance. In [31], setting $K = 7$ can give good results even for high-dimensional data. However, this way of setting the parameter is heuristic in nature. In our method to be described in this paper, we determine the weight for each data point based on robust statistical techniques which require no local parameters to be set manually.

2 Robust Path-Based Clustering

2.1 Path-Based Similarity Measure

The path-based dissimilarity measure was originally proposed in [10]. Following their formulation, we define a path-based similarity measure which will be extended to a robust version in the next two subsections.

We denote the data set of n points by $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$. The data points can be represented as a fully connected graph with n vertices corresponding to the n points. Each edge (i, j) in the graph is assigned a weight s'_{ij} reflecting the original similarity between \mathbf{x}_i and \mathbf{x}_j analogous to the Gaussian kernel:

$$s'_{ij} = \begin{cases} \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}\right) & \text{for } i \neq j \\ 0 & \text{for } i = j, \end{cases} \quad (1)$$

except that s'_{ii} is equal to 0 rather than 1 for $i = 1, 2, \dots, n$. Here, the scaling parameter σ controls how fast s'_{ij} falls off with the Euclidean distance between \mathbf{x}_i and \mathbf{x}_j . Recall that this is actually one common way of defining the affinity matrix in spectral clustering [20], where σ is a somewhat sensitive parameter. While this parameter is usually prespecified, Ng et al. [20] proposed a method for choosing σ automatically.

Let \mathcal{P}_{ij} denote the set of all paths from vertex i to vertex j through the graph. For each path $p \in \mathcal{P}_{ij}$, the *effective similarity* s_{ij}^p between vertices i and j (or the corresponding data points \mathbf{x}_i and \mathbf{x}_j) is the minimum edge weight along the path. We define the total similarity s_{ij} between vertices i and j as the maximum of all path-based effective similarities s_{ij}^p 's for paths p 's in \mathcal{P}_{ij} :

$$s_{ij} = \max_{p \in \mathcal{P}_{ij}} \left\{ \min_{1 \leq h < |p|} s'_{p[h]p[h+1]} \right\},$$

where $p[h]$ denotes the h th vertex along the path from vertex i to vertex j .

From the definition, the similarity between points \mathbf{x}_i and \mathbf{x}_j should be large when the

two points belong to the same cluster, and small when they belong to different clusters. However, in the latter case, if there exist some outliers between the two clusters, the similarity between points residing in different clusters may become much larger than it should be. In other words, the path-based similarity measure is sensitive to noise and outliers.

2.2 Robust Estimator Based on M-Estimation

Robust statistics have recently emerged as a family of theories and techniques to deal with the parameter estimation problems where the data are contaminated by outliers. Here, we use the idea of M-estimation in robust statistics to achieve more accurate similarity measure in the presence of outliers. In this subsection, we devise a robust estimator based on the idea of M-estimation [15].

Let us regard the neighbors of a data point \mathbf{x}_i as realizations from an estimator of \mathbf{x}_i . Then the squared residual error e_{ij}^2 of \mathbf{x}_i for neighbor \mathbf{x}_j can be defined based on the distance between \mathbf{x}_i and \mathbf{x}_j :

$$e_{ij}^2 = \frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{|\mathcal{N}_i|},$$

where \mathcal{N}_i denotes the neighborhood of \mathbf{x}_i and $|\mathcal{N}_i|$ the number of neighbors it contains. The total squared residual error of the estimators on \mathbf{x}_i can be expressed as $\sum_{\mathbf{x}_j \in \mathcal{N}_i} \|\mathbf{x}_i - \mathbf{x}_j\|^2 / |\mathcal{N}_i|$. Therefore, the total squared error E for all data points is given by

$$E = \sum_{i=1}^n \sum_{\mathbf{x}_j \in \mathcal{N}_i} e_{ij}^2 = \sum_{i=1}^n \sum_{\mathbf{x}_j \in \mathcal{N}_i} \frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{|\mathcal{N}_i|}.$$

The standard least squares method tries to minimize the error E . In this paper, we make use of robust estimation techniques by replacing the least squares estimator with a robust estimator that minimizes

$$E_\rho = \sum_{i=1}^n \sum_{\mathbf{x}_j \in \mathcal{N}_i} \rho(e_{ij}),$$

where $\rho(\cdot)$ is usually selected to grow more slowly than the quadratic function, such that the influence of the outliers is reduced.

Using robust estimation techniques, E_ρ can be minimized by solving an *iterative reweighted least squares* problem. More specially, we express the optimization problem as minimizing the total weighted squared error

$$E_w = \sum_{i=1}^n \sum_{\mathbf{x}_j \in \mathcal{N}_i} a_{ij} e_{ij}^2.$$

In iteration t , the weight $a_{ij} = a(e_{ij}^{(t-1)})$ is recomputed based on the residual error estimated in the previous iteration. Using gradient descent to minimize E_ρ and E_w , it is easy to see the following relationship between $a(e_{ij})$ and $\rho(e_{ij})$:

$$a(e_{ij}) = \frac{\rho'(e_{ij})}{e_{ij}}.$$

We use a common M-estimator, *Welsch function*, to define $\rho(\cdot)$ as

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

where $c > 0$ is some positive parameter. Then the weight function can be expressed as

$$a(e_{ij}) = \frac{\rho'(e_{ij})}{e_{ij}} = \exp\left(-\left(\frac{e_{ij}}{c}\right)^2\right).$$

Therefore, in our problem, the weight a_{ij} for the j th neighbor of \mathbf{x}_i is

$$a_{ij} = a(e_{ij}) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{|\mathcal{N}_i| c^2}\right).$$

2.3 Robust Path-Based Similarity Measure

By summing up the weights a_{ij} of all neighbors $\mathbf{x}_j \in \mathcal{N}_i$ and setting $c = \sqrt{\frac{2}{|\mathcal{N}_i|}} \sigma$, we obtain a weight w'_i for each point \mathbf{x}_i which can be expressed solely based on the original similarity values:

$$w'_i = \sum_{\mathbf{x}_j \in \mathcal{N}_i} a_{ij} = \sum_{\mathbf{x}_j \in \mathcal{N}_i} \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}\right) = \sum_{\mathbf{x}_j \in \mathcal{N}_i} s'_{ij}.$$

The weight of a data point is large if many data points are in its vicinity, and small if very few other points are close to it. Neighborhood can be defined in a number of different ways. Although some methods may be better than others, we keep it simple in this paper by using the same neighborhood size for all points in a data set, so that the neighborhood is just large enough to include at least two neighbors in each neighborhood. Note that the justification for using local neighborhoods is based on the argument that the weights should depend on the local geometry rather than the global structure of the whole data set.

To make the range of w'_i not affected by σ , we normalize each weight as $w_i = w'_i / \max_{\mathbf{x}_i \in \mathcal{X}} w'_i$, so that all the weights are in $(0, 1]$. While a large value for w_i indicates that \mathbf{x}_i is likely to be inside a compact cluster, a small value indicates that \mathbf{x}_i is an outlier.

Therefore, the robust path-based similarity measure can be expressed as:

$$s_{ij} = \max_{p \in \mathcal{P}_{ij}} \left\{ \min_{1 \leq h < |p|} w_{p[h]} w_{p[h+1]} s'_{p[h]p[h+1]} \right\}. \quad (2)$$

This measure can reflect the genuine similarity between \mathbf{x}_i and \mathbf{x}_j even when outliers exist. If there exists a path from \mathbf{x}_i to \mathbf{x}_j going through only points with high values of $w_{p[h]}$'s and $s'_{p[h]p[h+1]}$'s, then the total similarity should be high and hence \mathbf{x}_i and \mathbf{x}_j are likely to belong to the same cluster. On the other hand, if all paths between \mathbf{x}_i and \mathbf{x}_j contain at least one low value of either $w_{p[h]}$ or $s'_{p[h]p[h+1]}$, then the total similarity should be relatively low, implying that \mathbf{x}_i and \mathbf{x}_j either belong to different clusters or are themselves outliers.

We define the robust similarity matrix $\mathbf{S} = [s_{ij}]$ and use it in place of the affinity matrix $\mathbf{A} = [s'_{ij}]$ commonly used in spectral clustering [20]. Besides its robustness property, \mathbf{S} is no longer sensitive to the parameter σ as pointed out above and to be illustrated further through sensitivity analysis in Section 4.3. Our robust path-based spectral clustering algorithm is simply the ordinary spectral clustering algorithm using the robust similarity measure described above to define the affinity matrix. This seemingly minor modification is in fact very significant as dramatically improved clustering results can be obtained.

3 Extension for Semi-Supervised Clustering

The robust path-based similarity measure is a connectivity measure between two data points in the input space by taking into account the possible existence of outliers. When two points belong to two different natural groups that, for some reasons, should be put into the same cluster or class, the measure can no longer reflect this correctly as its value is expected to be small. In such cases, additional background knowledge is needed. One type of background knowledge is in the form of pairwise similarity or dissimilarity information, which has been used directly for improving the clustering results [26, 27, 17]. Another approach is to use the background information to learn a better metric first and then perform clustering with the new metric to get improved clustering results [29, 1, 3, 32]. In this section, we show that it is actually quite straightforward to extend our robust path-based similarity measure to handle such semi-supervised clustering problems.

Let us denote the set of pairwise similarity constraints available as background knowledge or side information by \mathcal{S} , and the set of pairwise dissimilarity constraints by \mathcal{D} . Based on the side information in \mathcal{S} and \mathcal{D} , the original similarity measure s'_{ij} is modified as follows:

$$s'_{ij} = \begin{cases} \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}\right) & \text{if } (\mathbf{x}_i, \mathbf{x}_j) \notin \mathcal{S} \cup \mathcal{D} \text{ and } i \neq j \\ 0 & \text{if } (\mathbf{x}_i, \mathbf{x}_j) \notin \mathcal{S} \cup \mathcal{D} \text{ and } i = j \\ s_{max} & \text{if } (\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{S} \\ s_{min} & \text{if } (\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{D} \end{cases}$$

where $s_{max} = \max_{i \neq j} \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}\right)$ is the maximum value of the original similarity and $s_{min} = \min_{i \neq j} \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}\right)$ is the minimum value of the original similarity excluding the zero diagonal terms s'_{ii} .¹

Based on the new s'_{ij} , the total similarity s_{ij} is computed as before to obtain the robust similarity matrix \mathbf{S} for the subsequent spectral clustering step. This extension is simple

¹There are other ways to define s_{max} and s_{min} , e.g., $s_{min} = 0$, but that will not influence the performance of the robust path-based similarity measure and the subsequent clustering results.

and natural, but we will see appealing experimental results on semi-supervised clustering and interactive image segmentation tasks using this method.

4 Experiments on Synthetic Data

To assess the efficacy of our algorithm for clustering tasks, we first perform some experiments on synthetic data sets for both unsupervised and semi-supervised settings.

4.1 Clustering Experiments

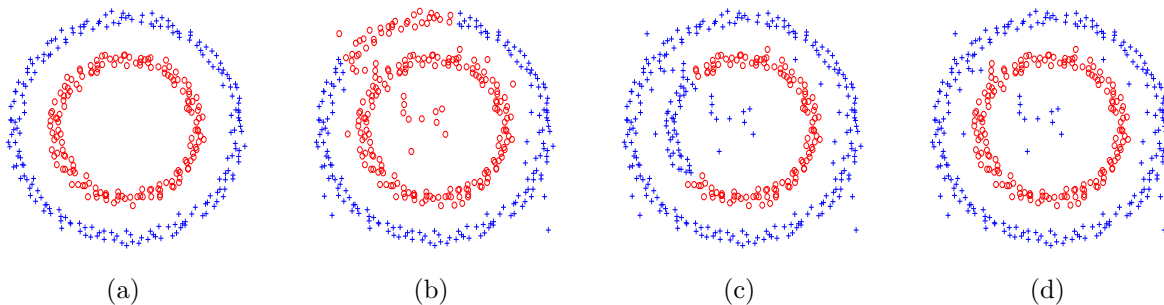


Figure 2: Clustering results for a noisy 2-circle data set: (a) original data set before noise points are added; (b) spectral clustering result; (c) path-based spectral clustering result; and (d) robust path-based spectral clustering result.

Although both standard spectral clustering and path-based clustering can find the two clusters in the 2-circle data set as shown in Figure 2(a), they are no longer robust enough to give satisfactory results if some noise points are added. Figure 2(b)–(d) compare these two methods with our robust path-based spectral clustering algorithm when 30 noise points are added. Note that some noise points located between the two circles end up connecting the two circular clusters. Figure 2(b) shows that spectral clustering cannot give good result

for this data set.² Due to the existence of noise points, the dissimilarity measure used by the original path-based clustering algorithm gives much lower dissimilarity values than they should be to point pairs residing in different circles. As a consequence, its result as shown in Figure 2(c) is not satisfactory either. Our method gives very satisfactory result, as shown in Figure 2(d), that agrees well with human judgement. This shows that the robust similarity measure is very effective in reducing the influence of the outliers.

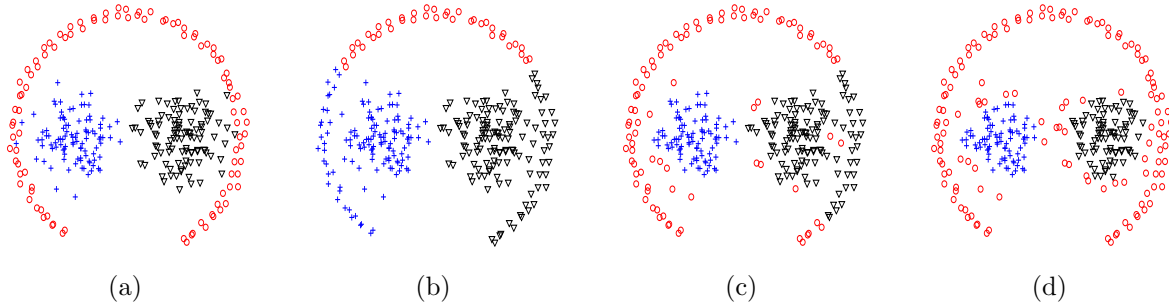


Figure 3: Clustering results for a 3-cluster data set: (a) original data set; (b) spectral clustering result; (c) path-based spectral clustering result; and (d) robust path-based spectral clustering result.

We also conduct some experiments on the 3-cluster data set as shown in Figure 3. The data set consists of a circular cluster with an opening near the bottom and two Gaussian distributed clusters inside. Each cluster contains 100 data points. Note that this data set is similar to the one shown in Figure 1(b) of [20], except that ours is more difficult. Unlike the 2-circle example, we do not add artificial noise points to this data set. However, the Gaussian clusters can be seen as having Gaussian noise points which tend to connect the clusters together. Figure 3(b) and (c) show that both spectral clustering and path-based clustering cannot find the three clusters correctly. Again, our robust path-based spectral clustering algorithm gives much more satisfactory result, as shown in Figure 3(d). However, some inter-cluster points are mis-clustered. Our method is successful in assigning lower weights to these points (and hence essentially detecting them as outliers). As a

²The noise level of this data set is higher than that of a similar data set in Figure 1(g) of [20].

result, they have relatively lower similarity values to other points within a cluster than those points closer to the Gaussian centers or on the incomplete circle. Spectral clustering will cluster such points incorrectly, as can also be seen in the 2-circle data set in Figure 2. If the robust path-based similarity measure is used with k -means clustering rather than spectral clustering, this problem actually does not happen and gives better result. However, in general, spectral clustering gives better results than k -means (e.g., the results shown in [20]) and hence is generally a better choice.

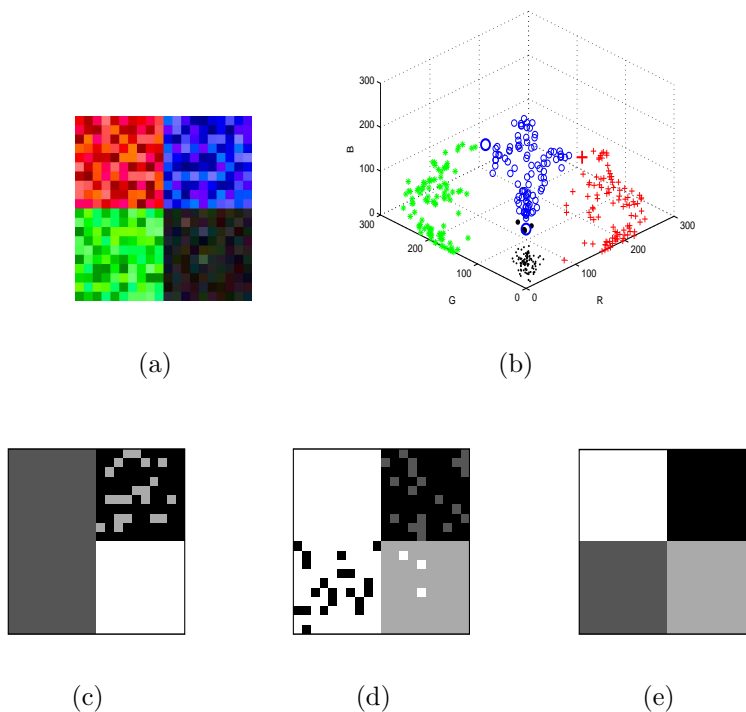


Figure 4: Segmentation results for a synthetic image: (a) image with some noise points; (b) feature vectors in RGB color space; (c) spectral clustering result; (d) path-based spectral clustering result; and (e) robust path-based spectral clustering result.

We further conduct experiments on a synthetic data set that can be seen as a simplified version of the image segmentation problem to be studied in the next section. Figure 4(a) shows a square image consisting of four regions, contaminated with some random noise points shown as small color squares of four different hues. Each small square (image

patch) corresponds to a point in the RGB color space, as shown in Figure 4(b). The clustering results based on different methods are shown in Figure 4(c)–(e). This again shows that our method is superior to other methods.

4.2 Semi-Supervised Clustering Experiments

While our robust path-based spectral clustering algorithm can handle many difficult clustering tasks even at the presence of outliers or for slighted coupled clusters, the extended algorithm incorporating pairwise side information can handle even more challenging cases, as discussed in Section 3. In this subsection, we demonstrate this empirically through some experiments on semi-supervised clustering tasks.

We generate a data set which consists of four natural groups, but they belong to only two clusters or classes. As shown in Figure 5(a), one cluster consists of two smaller (inner) ellipses with 80 data points each, while the other cluster consists of two larger (outer) ellipses with 100 data points each. In addition to the data points, two pairs of similarity constraints are also provided and they are shown as two dotted lines. We apply both our original and the extended robust path-based spectral clustering methods to this data set. While the original robust algorithm (Figure 5(b)) cannot find the correct clusters, the extended algorithm (Figure 5(c)) can find the correct clusters with the aid of just two pairs of similarity constraints. The corresponding similarity matrices are shown in Figures 5(e) and (f), respectively. For comparison, we also show in Figure 5(d) the affinity matrix used by standard spectral clustering. In the similarity matrices, we order the data points along the inner ellipses first and then the outer ones. The darker the color, the lower the similarity. Note that only the extended robust path-based similarity measure incorporating side information can reflect the true cluster connectivity of the input data. Therefore, only the extended method can find the clusters correctly.

When a cluster consists of multiple natural groups, our extended algorithm can incorpo-

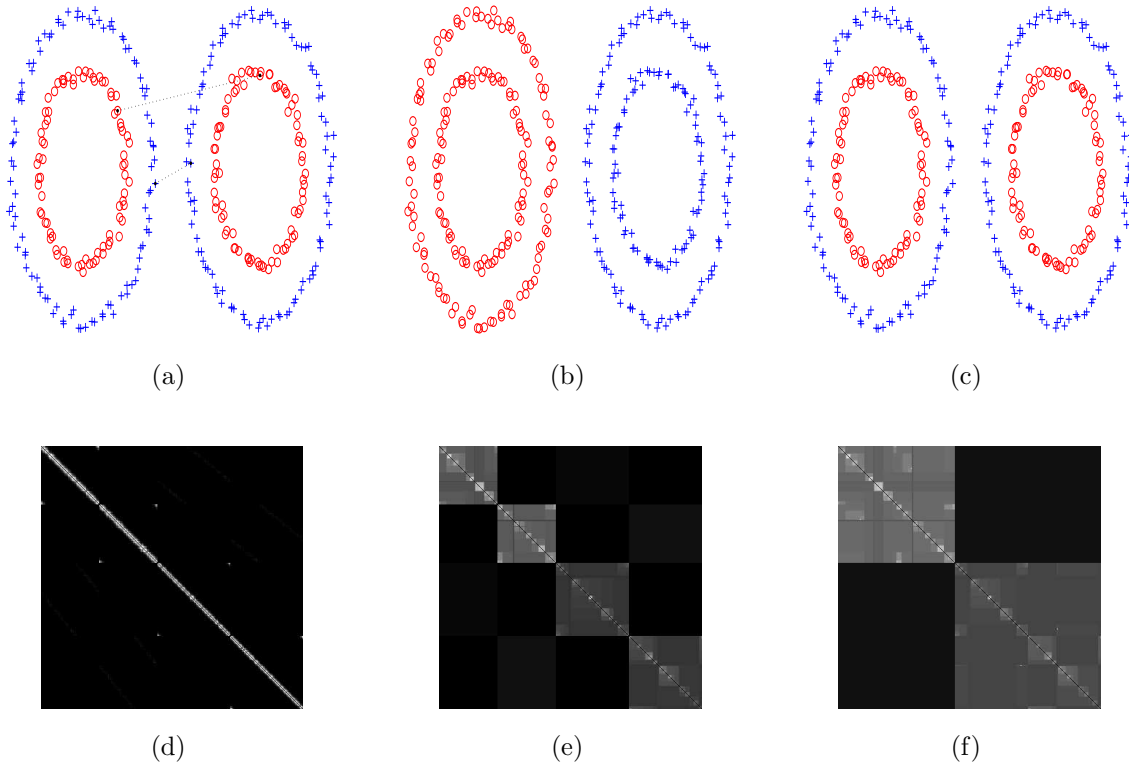


Figure 5: Semi-supervised clustering results: (a) data set with four natural groups belonging to two clusters and two pairs of similarity constraints shown as dotted lines; (b) robust path-based spectral clustering result without using side information; (c) extended robust path-based spectral clustering result using side information; (d) affinity matrix used in standard spectral clustering; (e) robust similarity matrix without using side information; (f) extended robust similarity matrix using side information.

rate background knowledge in the form of pairwise similarity or dissimilarity constraints to deliver very promising results even on fairly complex data sets. Although the example shown in Figure 5 appears somewhat hypothetical and unrealistic, there do exist some real-world problems where such semi-supervised learning setting is actually rather natural. The interactive image segmentation application to be described in Section 6.4 is a good example that performs image segmentation with human interaction.

4.3 Sensitivity Analysis of Parameters

Similar to standard spectral clustering, our method also requires choosing the scaling parameter σ . We argued above that our method is not sensitive to the setting of σ while the standard spectral clustering method is. In this subsection, we elaborate this point by performing sensitivity analysis to see how the clustering result varies with σ .

For simplicity, we only report results for the 2-circle data set, which are representative of the general case. Figure 6(a)–(c) show three 2-circle data sets of varying noise levels, with no noise in (a) (same as that in Figure 2(a)), 20 noise points in (b) and 30 noise points in (c) (same as that in Figure 2(b)–(d)). We compare robust path-based spectral clustering with standard spectral clustering and path-based spectral clustering on a range of different σ values. The clustering quality is measured using the Rand index [22] with respect to the ground truth clustering partition.

We can see that the standard spectral clustering algorithm is fairly sensitive to the choice of σ . With $\sigma^2 \in [0.018, 0.096]$, this method can correctly find the two clusters in the noise-free data set, as the Rand index curve in Figure 6(d) shows. When noise points are added, the reliable range of σ decreases with the increase of noise points (Figure 6(e) and (f)). For the noisy data set with 30 noise points, while there exists a very small range of σ values in which standard spectral clustering can find the correct clusters, this range is so narrow that it is very hard to locate and hence the method generally fails. As for

path-based spectral clustering, which can give good results (with a wide range of σ values) for clean data, it is very difficult to get satisfactory results for noisy data sets. On the other hand, our method can find the correct clusters for a very wide range of σ values on all three data sets (Figure 6(d)–(f)). It fails only when σ gets too small. This shows that our method is easier to use and is more stable than standard spectral clustering.

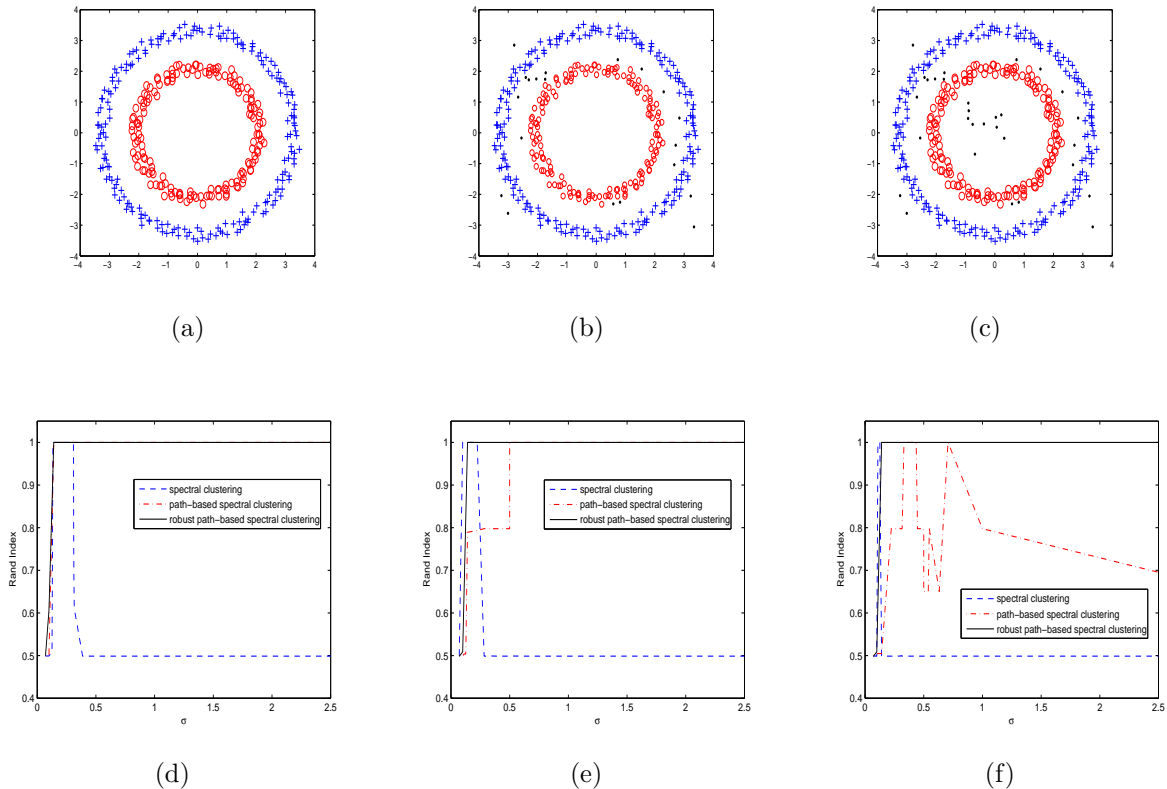


Figure 6: Effect of σ on clustering results: (a) 2-circle data set without noise points; (b) 2-circle data set with 20 noise points; (c) 2-circle data set with 30 noise points; (d)–(f) spectral clustering results (blue dash line), path-based spectral clustering results (red dash-dot line) and robust path-based spectral clustering results (black solid line) on (a)–(c).

We conduct more experiments to analyze the effect of σ on the clustering results for the synthetic image shown in Figure 4. The feature vectors of the small patches in the

image form four regions with different densities and random noise points, as shown in Figure 7(a). The analysis results are shown in Figure 7(b).

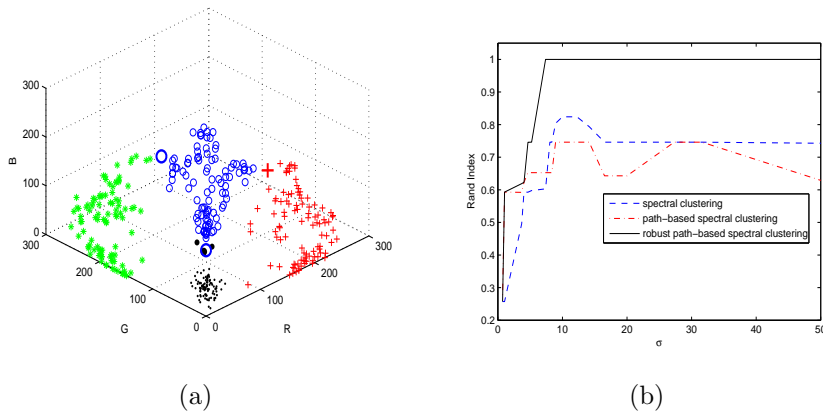


Figure 7: Effect of σ on clustering results: (a) feature vectors in RGB color space; (b) spectral clustering results (blue dash line), path-based spectral clustering results (red dash-dot line) and robust path-based spectral clustering results (black solid line) on (a).

Another parameter setting of our method that may affect the clustering results is the way of defining the neighborhoods \mathcal{N}_i 's. In our experiments, we use the same neighborhood size for all points in a data set so that the neighborhood is just large enough to include at least two neighbors in each neighborhood. This is not the only way to define the neighborhoods though. For example, we may use the mean edge length of the neighborhood graph derived from a data set to determine the radius of each neighborhood. The neighborhood graph can be constructed by connecting \mathbf{x}_i and \mathbf{x}_j if \mathbf{x}_i is one of the K nearest neighbors of \mathbf{x}_j or if \mathbf{x}_j is one of the K nearest neighbors of \mathbf{x}_i based on the Euclidean distance. Under this alternative setting, which we have also tried, our method gives very similar results to those shown above.

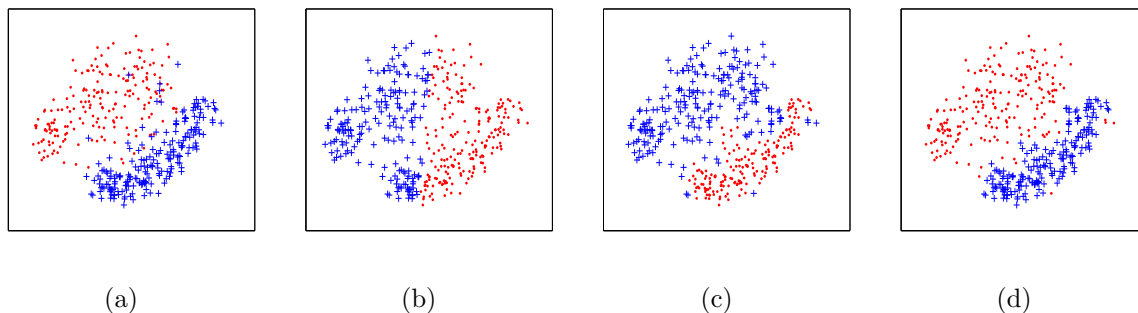


Figure 8: Clustering results for digits “8” and “9”: (a) digit images plotted based on the two leading principal components; (b) spectral clustering result; (c) path-based spectral clustering result; and (d) robust path-based spectral clustering result.

5 Experiments on Real-World Data

5.1 MNIST Digits

We further perform experiments on handwritten digits from the well-known MNIST database.³ Unlike the synthetic data, this data set is of much higher dimensionality. The digits in the database have been size-normalized and centered to 28×28 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 8 shows the clustering 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 8(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. The clustering results using different methods are shown in Figure 8(b), (c) and (d). It

³<http://yann.lecun.com/exdb/mnist/>

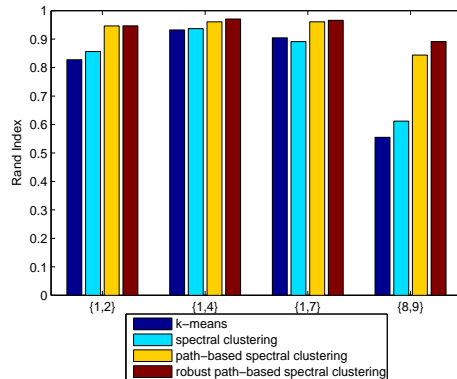


Figure 9: Clustering results on several subsets of the MNIST digit database.

can be seen that our method gives the best clustering result.

We conduct more experiments on different digit subsets under the same experimental settings. We use k -means clustering for baseline comparison. The clustering results are summarized in Figure 9, where the four groups of bars, from left to right, represent the results for subsets 1,2, 1,4, 1,7 and 8,9. We use Rand Index [22] to quantify the clustering accuracy. From the results, we can see that our proposed method outperforms all other clustering methods for all data sets.

5.2 UMIST Face Images

The UMIST face database [13] consists of 564 gray-level images from 20 persons. Images of each person contain a range of poses from profile to frontal views. We down-sample pre-cropped images to smaller ones of size 56×46 . 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}_{new}) is the weighted average of two randomly selected face images (\mathbf{f}_i and \mathbf{f}_j) from two different



Figure 10: Examples of synthesized “noise” face images.

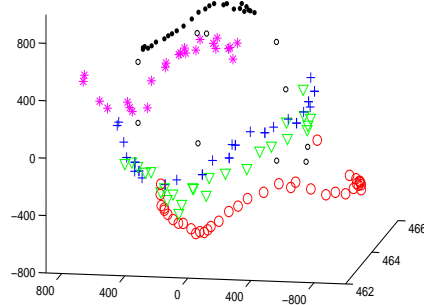


Figure 11: 3D embeddings of face images from five persons. Ten “noise” faces are marked as black circles.

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 10. These artificially created images act as outliers in the subsequent experiments.

Instead of computing the leading principal components, we perform clustering 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 11. 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.

The clustering results using different methods are shown in Figure 12. The left group of bars contains the clustering results for the clean image set, while the right group contains

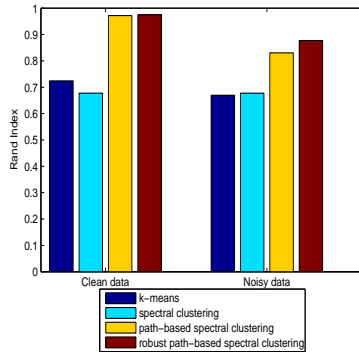


Figure 12: Clustering results on a subset of the UMIST face database.

the results when “noise” images are added. For both settings, our proposed robust path-based spectral clustering method achieves the highest accuracy.

6 Experiments on Color Image Segmentation

While the results presented in the previous section are very promising, we would like to test our robust path-based spectral clustering algorithm on real-world clustering tasks too. In particular, we study color image segmentation in this section.

6.1 Color Image Segmentation

Image segmentation tries to parse natural images into homogeneous tiles corresponding to different surfaces and objects. Homogeneity is usually defined as pairwise similarity between pixels or regions. Recent work on pairwise clustering, including spectral clustering [24, 25, 28] and path-based clustering [8, 7, 10], has demonstrated impressive results for image segmentation.

In this section, we compare our robust path-based spectral clustering algorithm with

several non-robust methods for image segmentation based on color and spatial features from the images.

6.2 Experimental Settings

The images used in our experiments are from the Berkeley Segmentation Dataset and Benchmark [19]. As in [8, 7, 10], we formulate image segmentation as a data clustering problem based on sparse proximity data. For each image of 154401 (321×481) pixels, we consider 1855 (35×53) overlapping patches with each patch of size equal to 13×13 pixels.

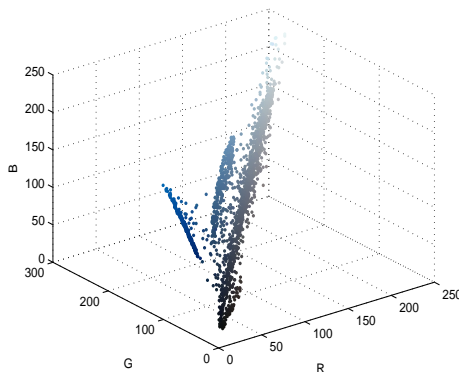


Figure 13: Image patches from the bottommost image of Figure 14(a) shown as feature vectors in the RGB color space.

Similarity between patches is computed based on color and spatial features. For simplicity, we use a relatively simple feature representation scheme in our experiments. For each image, the average color of each patch is represented as a point or feature vector $\mathbf{x}_i^c, i = 1, \dots, 1855$ in the RGB color space. Figure 13 shows all the feature vectors for the bottommost image of Figure 14(a). We can see that the sky, lake and mountain form three manifolds with elongated structures. To segment the three regions correctly, a

clustering algorithm should not only find the elongated structures but should also be sufficiently robust even at the presence of outliers.

Besides color features, we also include some spatial features as in [5]. The spatial features are simply the two coordinates of each patch in the image lattice, denoted as $\mathbf{x}_i^s, i = 1, \dots, 1855$. The overall feature vector \mathbf{x}_i for patch i is the concatenation of the vectors for color and spatial features, i.e., $\mathbf{x}_i = ((\mathbf{x}_i^c)^T, \lambda(\mathbf{x}_i^s)^T)^T$, where the superscript T denotes vector/matrix transpose and $\lambda > 0$ specifies the relative importance of the spatial information for defining the similarity measure. It is set to 0.5 in our experiments.

6.3 Image Segmentation Results

Figure 14 shows the segmentation results for four images on natural scenes. From left to right, the five columns show the input color images, segmentation results based on k -means clustering, spectral clustering, path-based spectral clustering, and robust path-based spectral clustering, respectively. For each image, the number of clusters for the four clustering methods is user-defined. The segmentation results are shown with different gray levels representing different segments. In these images, color changes in the sky, water, grass land and mountain form elongated manifold structures, such as the example shown in Figure 13. It can be seen that the robust path-based spectral clustering algorithm outperforms all other methods. Although path-based clustering can also cluster elongated structures, it is not robust enough against noise and outliers. Specifically, the existence of outlier points between manifolds in the feature space ends up connecting the corresponding color segments together.

For more quantitative comparison of the different methods, we make use of the global consistency measure (GCE) and local consistency measure (LCE) proposed by [19]. These two measures quantify the consistency between different segmentation results and have been validated through extensive experiments. With these measures, deviation or refine-

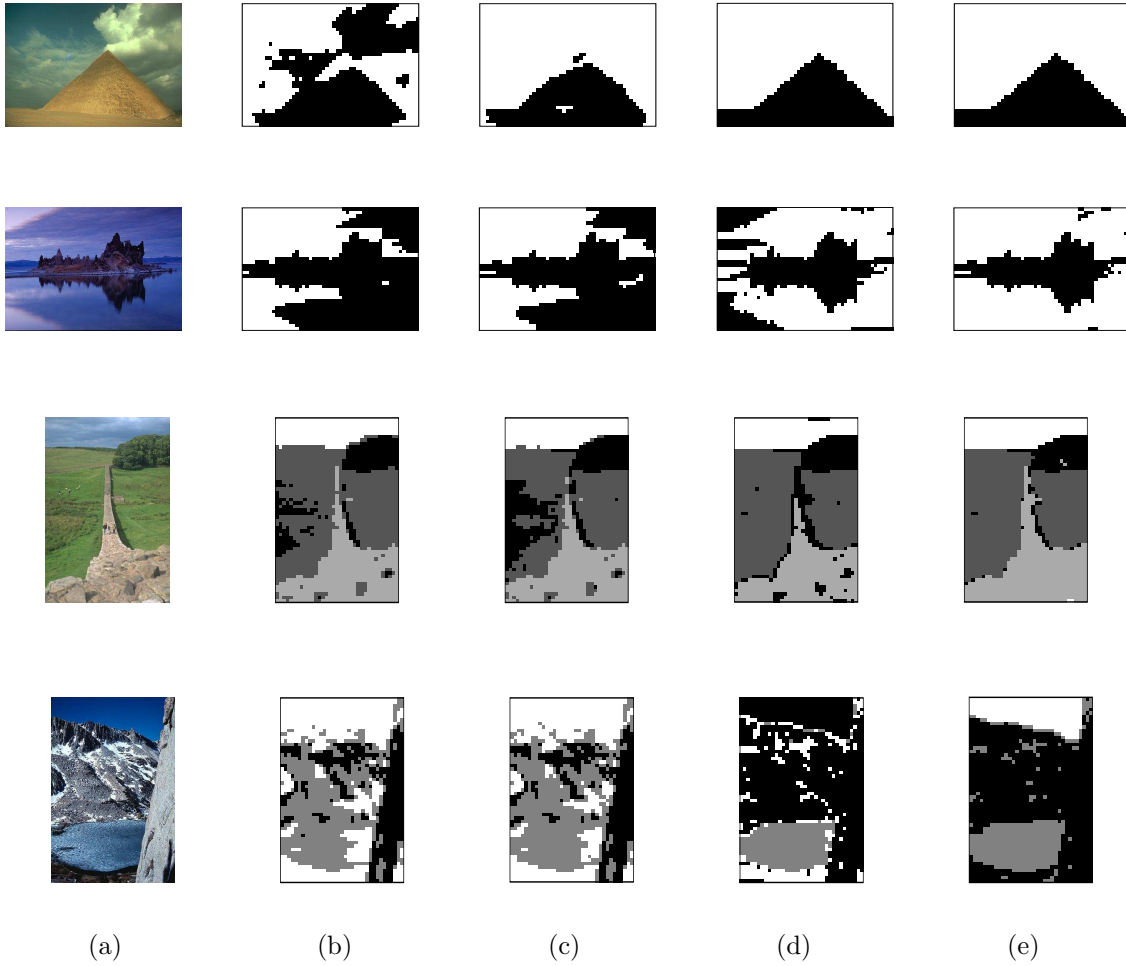
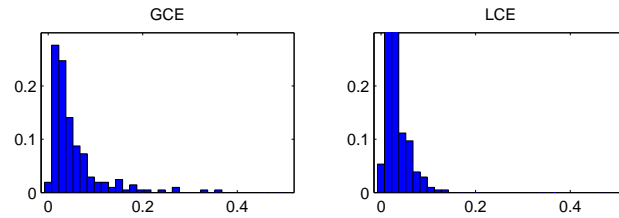
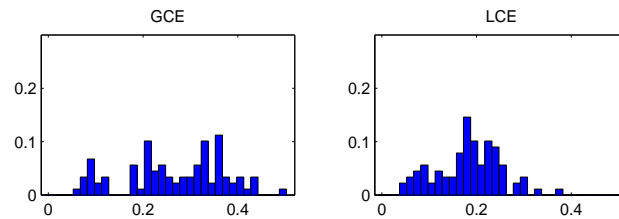


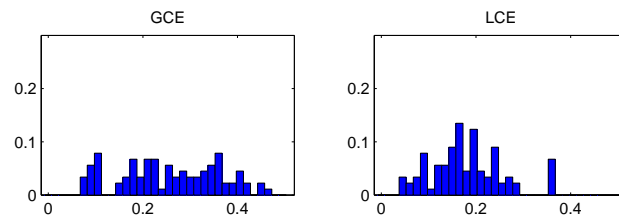
Figure 14: Color image segmentation results: (a) input images; (b) k -means clustering results; (c) spectral clustering results; (d) path-based spectral clustering results; and (e) robust path-based spectral clustering results.



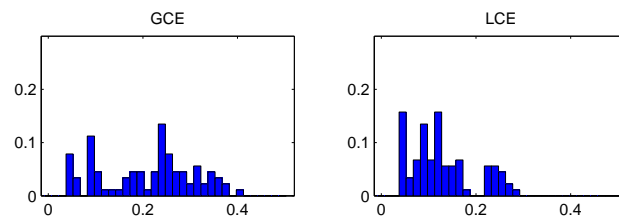
(a) Human vs. Human



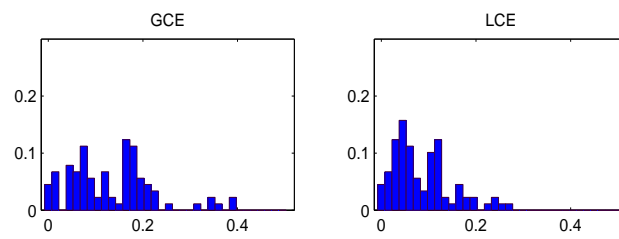
(b) Human vs. k -means



(c) Human vs. spectral clustering



(d) Human vs. path-based spectral clustering



(e) Human vs. robust path-based spectral clustering

Figure 15: Histograms of the distribution of errors (GCE and LCE) for different segmentation methods. Human segmentation results compared to results based on: (a) human segmentation; (b) k -means clustering; (c) standard spectral clustering; (d) path-based spectral clustering; and (e) robust path-based spectral clustering.

ment from a reference segmentation result (typically human segmentation result) can be quantified by an error measure, but not simply regarded as an incorrect result in the form of a binary decision. Between GCE and LCE, GCE has more stringent requirements since it requires all local refinements to be in the same direction while LCE allows refinements in different directions in different parts of the image.

We perform image segmentation experiments on a subset of 40 color images from the Berkeley image database.⁴ Each image has at least five human segmentation results available in the database. For each image, we compare the segmentation results based on different clustering methods with each human segmentation result. Figure 15 depicts the comparison results, where the distributions of GCE and LCE are shown as histograms. The horizontal axis of each histogram shows the range of GCE or LCE values, while the vertical axis indicates the percentage of comparisons. From the subfigures, we can see that human segmentation results for the same image are fairly consistent. The segmentation results using k -means clustering, spectral clustering and path-based clustering algorithms have much higher errors. On the other hand, the robust path-based clustering algorithm can give significantly lower errors. The mean and standard deviation of each histogram are summarized in Table 1.

Table 1: Mean and standard deviation of errors for different segmentation methods.

	GCE (%)	LCE (%)
HUMAN VS. HUMAN	5.27(\pm 5.97)	3.32(\pm 2.33)
HUMAN VS. k -MEANS CLUSTERING	26.65(\pm 10.52)	18.33(\pm 6.79)
HUMAN VS. SPECTRAL CLUSTERING	25.49(\pm 10.60)	18.02(\pm 7.60)
HUMAN VS. PATH-BASED SPECTRAL CLUSTERING	20.95(\pm 9.94)	12.85(\pm 6.75)
HUMAN VS. ROBUST PATH-BASED SPECTRAL CLUSTERING	13.32(\pm 8.92)	8.56(\pm 6.16)

⁴More segmentation results are available in <http://www.cs.ust.hk/~hongch/image-segmentation.htm>.

6.4 Interactive Image Segmentation Results

We have seen in the previous subsection that the robust path-based clustering algorithm can give promising image segmentation results. However, there do exist some situations when the algorithm (as well as other algorithms) in its fully automatic form cannot give satisfactory results, because image segmentation, like many computer vision and image processing tasks, is ill-posed in nature and can be ambiguous especially for complicated scenes. For example, with object occlusion caused by lighting or background clutter, the same object may appear as disconnected in an image. In the same spirit as some recent efforts in *interactive* computer vision and computer graphics, one possibility is to use interactive image segmentation techniques by incorporating human prior knowledge about desirable groupings through human-computer interaction. Some related recent work along this line includes region-based image segmentation methods that allow the user to provide hints by specifying the extent of different segments, such as the interactive graph cut method [2, 18] and segmentation with partial grouping constraints [30].

The semi-supervised extension of our algorithm can naturally be used for interactive image segmentation simply by requiring the user to provide the pairwise supervisory information through a well-designed user interface. Some proof-of-concept experiments are reported here. Two input images are shown in Figure 16(a). Each solid curve in an image specifies the extent of an object of interest provided by the user using a simple user interface. Multiple curves can appear in an image, with the same color referring to the same object of interest and different colors corresponding to different objects. All image patches passed through by a curve induce pairwise similarity constraints among themselves, and pairwise dissimilarity constraints with image patches passed through by another curve of a different color. Note that the information provided by this high-level user interaction process does not have to be very precise. In fact, the curves provided are quite far from the object boundaries. Figure 16(b) and (c) show the segmentation results for the two images without and with supervisory information from the user, respectively. While the

segmentation result for the first image without user interaction is not satisfactory due to shadow clutter, the interactive segmentation result gives more meaningful clusters corresponding to the ground, forest and horses. For the second image, user interaction can help to segment the animal from the background successfully, which is otherwise very difficult due to the small bald region on the grassland.

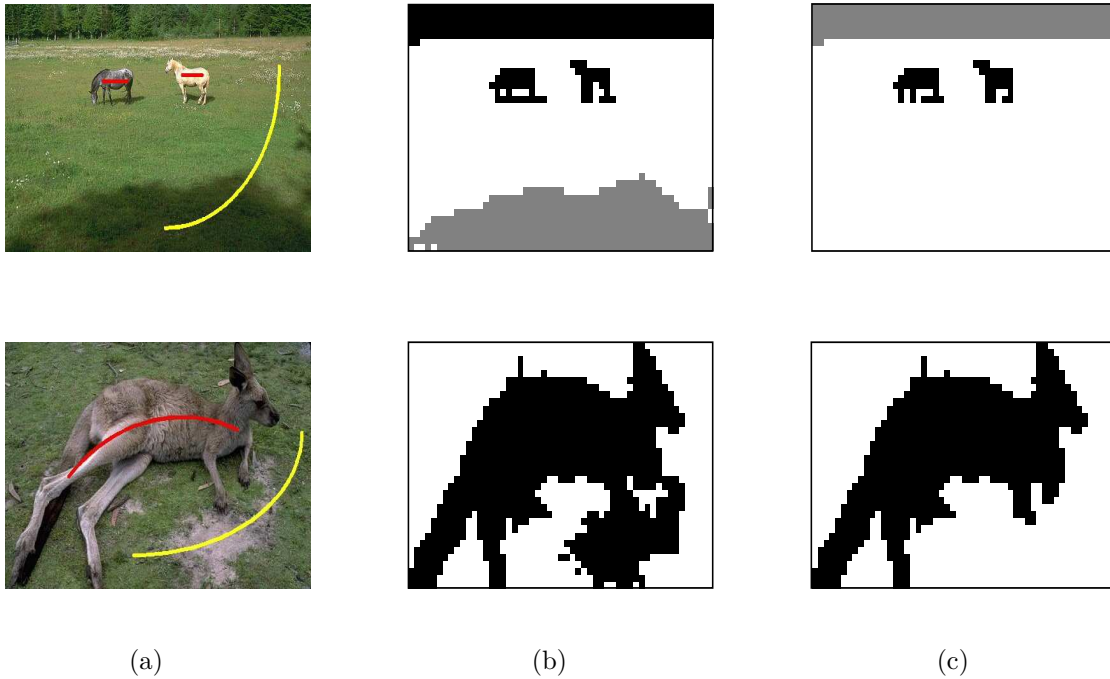


Figure 16: Interactive color image segmentation results: (a) input images, with supervisory information provided as solid curves; (b) robust path-based spectral clustering results; and (c) semi-supervised robust path-based spectral clustering results.

7 Concluding Remarks

In this paper, we have presented a robust path-based spectral clustering algorithm that makes the combined use of spectral clustering and path-based clustering more robust against noise and outliers. We summarize some major advantages of our method here.

First, it is robust in the sense that Equation (2) can give a reliable measure of the inter-point similarity even at the presence of noise and outliers. Therefore, based on the robust similarity measure, our method is much more robust than standard spectral clustering and path-based clustering on noisy data sets and data sets involving slighted coupled clusters. Second, it is also robust in the sense that the algorithm is not sensitive to the scaling parameter σ while the standard spectral clustering algorithm is. The implication of this desirable property is that our method is easier to use and is more stable, making it a good candidate as a general clustering method. Third, it is easy for our method to be extended to the semi-supervised learning setting, which allows it to handle some even more challenging cases with the aid of supervisory side information. This makes our method even more powerful for solving some highly challenging clustering problems.

Despite its promising performance, there is still room for us to further enhance our method. To trade accuracy for efficiency, we perform our image segmentation experiments on image patches instead of pixels, as was done in [8, 7, 10]. As a consequence, the segmentation results are relatively crude compared with human segmentation results. Although some approximation methods have been proposed [9], we can look for even more efficient approximation methods for specific clustering tasks. Recently, Fowlkes et al. [11] proposed an approximation technique based on sampling to alleviate the computational burden of spectral grouping. The more recent work by Ouimet and Bengio [21] using greedy approximation shows even better performance than randomly sampling or simply using the Nyström method to generalize the sample points. Inspired by their ideas, one possible extension is to explore the possibility of solving the segmentation problem for a small subset of pixels first and then extrapolating this solution to the full set of pixels in the image. Moreover, image segmentation methods typically make use of combined features, including color, brightness and texture, in patch or gradient forms [12]. We will consider incorporating more feature types into our method to further improve its image segmentation accuracy.

Another interesting direction to extend the current research is to explore the relationship

between our robust path-based similarity measure and Mercer kernels [23]. One possible approach is like the connectivity kernel proposed by [9]. Another possibility is to make use of the graph Laplacian to give a kernel view of our method, as in [14]. We will pursue this research direction in our future work.

Acknowledgments

The research reported in this paper has been supported by Competitive Earmarked Research Grant (CERG) HKUST6174/04E from the Research Grants Council of the Hong Kong Special Administrative Region, China.

References

- [1] A. Bar-Hillel, T. Hertz, N. Sental, and D. Weinshall. Learning distance functions using equivalence relations. In *Proceedings of the Twentieth International Conference on Machine Learning*, pages 11–18, Washington, DC, USA, 21–24 August 2003.
- [2] Y.Y. Boykov and M.P. Jolly. Interactive graph cuts for optimal boundary & region segmentation of objects in N-D images. In *Proceedings of the Eighth IEEE International Conference on Computer Vision*, volume 1, pages 105–112, Vancouver, BC, Canada, 7–14 July 2001.
- [3] H. Chang and D.Y. Yeung. Locally linear metric adaptation for semi-supervised clustering. In *Proceedings of the Twenty-First International Conference on Machine Learning*, pages 153–160, Banff, Alberta, Canada, 4–8 July 2004.
- [4] O. Chapelle and A. Zien. Semi-supervised classification by low density separation. In *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, pages 57–64, Barbados, 6–8 January 2005.

- [5] D. Comaniciu and P. Meer. Mean shift: a robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):603–619, 2002.
- [6] R.O. Duda, P.E. Hart, and D.G. Stork. *Pattern Classification*. Wiley, New York, NY, USA, 2nd edition, 2001.
- [7] B. Fischer and J.M. Buhmann. Bagging for path-based clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(11):1411–1415, 2003.
- [8] B. Fischer and J.M. Buhmann. Path-based clustering for grouping of smooth curves and texture segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(4):513–518, 2003.
- [9] B. Fischer, V. Roth, and J.M. Buhmann. Clustering with the connectivity kernel. In S. Thrun, L. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, USA, 2004.
- [10] B. Fischer, T. Zöllner, and J.M. Buhmann. Path based pairwise data clustering with application to texture segmentation. In M.A.T. Figueiredo, J. Zerubia, and A.K. Jain, editors, *Proceedings of the Third International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition*, pages 235–250, Sophia Antipolis, France, 3–5 September 2001.
- [11] C. Fowlkes, S. Belongie, F. Chung, and J. Malik. Spectral grouping using the Nyström method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(2):214–225, 2004.
- [12] C. Fowlkes, D. Martin, and J. Malik. Learning affinity functions for image segmentation: combining patch-based and gradient-based approaches. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 54–61, Madison, WI, USA, 18–20 June 2003.

- [13] 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.
- [14] J. Ham, D.D. Lee, S. Mika, and B. Schölkopf. A kernel view of the dimensionality reduction of manifolds. In *Proceedings of the Twenty-First International Conference on Machine Learning*, pages 369–376, Banff, Alberta, Canada, 4–8 July 2004.
- [15] P.J. Huber. Robust regression: asymptotics, conjectures, and Monte Carlo. *Annals of Statistics*, 1(5):799–821, 1973.
- [16] A.K. Jain and R.C. Dubes. *Algorithms for Clustering Data*. Prentice Hall, Englewood Cliffs, NJ, USA, 1988.
- [17] D. Klein, S.D. Kamvar, and C.D. Manning. From instance-level constraints to space-level constraints: making the most of prior knowledge in data clustering. In *Proceedings of the Nineteenth International Conference on Machine Learning*, pages 307–314, Sydney, Australia, 8–12 July 2002.
- [18] Y. Li, J. Sun, C.K. Tang, and H.Y. Shum. Lazy snapping. *ACM Transactions on Graphics*, 23(3):303–308, 2004.
- [19] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings of the Eighth IEEE International Conference on Computer Vision*, volume 2, pages 416–423, Vancouver, BC, Canada, 7–14 July 2001.
- [20] A.Y. Ng, M.I. Jordan, and Y. Weiss. On spectral clustering: analysis and an algorithm. In T.G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, pages 849–856. MIT Press, Cambridge, MA, USA, 2002.

- [21] M. Ouimet and Y. Bengio. Greedy spectral embedding. In *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, pages 253–260, Barbados, 6–8 January 2005.
- [22] W.M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66:846–850, 1971.
- [23] B. Schölkopf and A.J. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, USA, 2002.
- [24] J. Shi and J. Malik. Normalized cuts and image segmentation. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 731–737, San Juan, Puerto Rico, 17–19 June 1997.
- [25] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.
- [26] K. Wagstaff and C. Cardie. Clustering with instance-level constraints. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 1103–1110, Stanford, CA, USA, 29 June – 2 July 2000.
- [27] K. Wagstaff, C. Cardie, S. Rogers, and S. Schroedl. Constrained k -means clustering with background knowledge. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 577–584, Williamstown, MA, USA, 28 June – 1 July 2001.
- [28] Y. Weiss. Segmentation using eigenvectors: a unifying view. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, volume 2, pages 975–982, Kerkyra, Greece, 20–27 September 1999.
- [29] E.P. Xing, A.Y. Ng, M.I. Jordan, and S. Russell. Distance metric learning, with application to clustering with side-information. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, pages 505–512. MIT Press, Cambridge, MA, USA, 2003.

- [30] S.X. Yu and J. Shi. Segmentation given partial grouping constraints. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(2):173–183, 2004.
- [31] L. Zelnik and P. Perona. Self-tuning spectral clustering. In *Advances in Neural Information Processing Systems 17*. 2005.
- [32] Z. Zhang. Learning metrics via discriminant kernels and multidimensional scaling: toward expected Euclidean representation. In *Proceedings of the Twentieth International Conference on Machine Learning*, pages 872–879, Washington, DC, USA, 21–24 August 2003.

About the Author – HONG CHANG received her Bachelor degree, M.Phil. degree and Ph.D. degree in Computer Science from Hebei University of Technology, Tianjin University, and Hong Kong University of Science and Technology, respectively. She is currently a Research Scientist in Xerox Research Centre Europe. Her main research interests include semi-supervised learning, nonlinear dimensionality reduction, and related applications.

About the Author – DIT-YAN YEUNG received his B.Eng. degree in Electrical Engineering and M.Phil. degree in Computer Science from the University of Hong Kong, and his Ph.D. degree in Computer Science from the University of Southern California in Los Angeles. He was an Assistant Professor at the Illinois Institute of Technology in Chicago before he joined the Department of Computer Science and Engineering of the Hong Kong University of Science and Technology, where he is currently an Associate Professor. His research interests are in machine learning and pattern recognition. He is currently serving on the editorial boards of Journal of Artificial Intelligence Research and Pattern Recognition.