# Spectral Clustering with Sparse Graph Construction Based on Markov Random Walk

**Jiangzhong Cao[1], Pei Chen[2], Bingo Wing-Kuen Ling[1], Zhijing Yang[1], Qingyun Dai [3]**

[1] Faculty of Information Engineering, Guangdong University of Technology,
Guangzhou, 510006, China.
[e-mail: cjz510@sdsc.edu, yongquanling@gdut.edu.cn, yzhj@gdut.edu.cn ]

[2] School of Information Science and Technology, Sun Yat-sen University,
Guangzhou, 510006, China.
[e-mail: chenpei@mail.sysu.edu.cn]

[3] Guangdong Polytechnic Normal University,
Guangzhou, 510006, China.
[e-mail: daiqy@gdut.edu.cn]

*Corresponding author: Jiangzhong Cao

*Received January 4, 2015; revised March 28, 2015; revised April 30, 2015; accepted June 9, 2015; published July 31, 2015*

## Abstract

Spectral clustering has become one of the most popular clustering approaches in recent years. Similarity graph constructed on the data is one of the key factors that influence the performance of spectral clustering. However, the similarity graphs constructed by existing methods usually contain some unreliable edges. To construct reliable similarity graph for spectral clustering, an efficient method based on Markov random walk (MRW) is proposed in this paper. In the proposed method, the MRW model is defined on the raw *k*-NN graph and the neighbors of each sample are determined by the probability of the MRW. Since the high order transition probabilities carry complex relationships among data, the neighbors in the graph determined by our proposed method are more reliable than those of the existing methods. Experiments are performed on the synthetic and real-world datasets for performance evaluation and comparison. The results show that the graph obtained by our proposed method reflects the structure of the data better than those of the state-of-the-art methods and can effectively improve the performance of spectral clustering.

***Keywords:*** *k*-NN graph, Markov random walk, spectral clustering

This research was supported by by the National Natural Science Foundation of China (no.61173083 and no. 61372173), the Natural Science Foundation of Guangdong Province (no. 2014A030310346) and the Guangdong Higher Education Engineering Technology Research Center for Big Data on Intellectual Property of Manufacturing.

**http://dx.doi.org/10.3837/tiis.2015.07.013** **ISSN : 1976-7277**

## 1. Introduction

Clustering, also called cluster analysis, is widely used in many research fields, including multimedia retrieval, pattern recognition and data mining [1−3]. Many clustering algorithms have been developed over the past decades. Among these algorithms, spectral clustering, as a promising method, has recently attracted considerable attention [4−7]. Spectral clustering reveals the cluster structures of data using eigenvectors of the Laplacian graph and can stably detect nonconvex patterns and nonlinearly separable clusters [5, 6]. It is considered as one of the most promising clustering techniques because of its superior performance compared to traditional clustering algorithms when utilized on certain challenging datasets [8].

Spectral clustering is a clustering method based on graph theory. Given a data set $\mathbf{V} = \{v_1, v_2 \cdots, v_n\} \subset \Re^l$, spectral clustering first constructs a similarity graph with the data as vertices and the similarity among the data as the weight of the edges. Then, the eigenvectors of a Laplacian matrix corresponding to the similarity graph is computed. Finally, clustering result is obtained by running a traditional clustering method (e.g., k-means) on the new data mapped by the eigenvectors. The objective of spectral clustering is to obtain new data that can be clustered more easily in the eigenspace constructed by the Laplacian matrix. Since the Laplacian matrix is computed in accordance with the similarity graph, the construction of the similarity graph plays a key role in spectral clustering. The most common similarity graph is a full graph in which each vertex is connected to all others. However, since the pairwise similarity between two faraway data points is less reliable than that between two neighbors, the full graph usually contains some unreliable edges that degrade the performance of spectral clustering. It has been shown that the sparse similarity graph can help to improve the performance of spectral clustering [5, 9]. Determining the neighbors is the core for constructing the sparse graph. Several methods for constructing sparse graphs currently have been proposed, with two most classical methods being $k$-nearest neighbor ($k$-NN) and $\varepsilon$-neighborhood [4]. In a $k$-NN graph a vertex is only connected to its $k$ nearest neighbors, whereas points whose pairwise distances (similarity) are smaller (larger) than ε are connected in ε-neighborhood graph. Classical sparse graphs can effectively reduce the effect caused by unreliable pairwise similarities from faraway vertices. However, in many cases, the pairwise similarities are not always accurate, especially for data with an elongated structure. Hence, classical sparse graphs often contain a number of unreliable relationships (an example is shown in **Fig. 1(b)**). A method based on the dominant set was subsequently proposed to overcome this difficulty [10]. In its graph, each vertex connects the vertices in the dominant set of its $k$-NN. For considering the affinities among the neighbors, this method is more robust to errors and outliers based on the pairwise similarities. However, the dominant neighborhood is inclined to make the local density samples condensing to an isolated clique from the non-uniform samples, as shown in **Fig. 2(c)**. In addition, several methods for constructing graphs have been proposed for other applications. Wang et al. [11] proposed a graph structure method named k-regular nearest neighbor graph and applied it to the manifold-ranking based retrieval framework. Li proposed the k-edge connected neighborhood graph, and presented two constructing methods named k-MST [12] and Min-k-ST [13, 14]. In [15], a method named b-matching was proposed to guarantee the number of edges of each vertex to be equal. Cheng et al. [16] proposed a method in which graphs are constructed using datum-wise sparse reconstructions of samples via $l_1$-norm minimization. The methods outlined above are designed for different tasks, but, to the best of our knowledge there is no evidence to show which graph is the best for spectral clustering. Consequently, a good graph construction method is still needed.

Markov random walk (MRW) can be useful for this purpose because its model has a close relation with graphs. Given a graph and a starting point, we can randomly select one of its neighbors and move to this neighbor; then, we can select this neighbor as the next starting point and repeat the above process. In this way, the sequence of points selected is a random walk on the graph. In fact, Markov chains can be viewed as random walks on a weighted graph [26]. Recently, the MRW model has been explored as a similarity measure and applied in many fields [17, 18, 21, 32, 33]. White and Smyth [27] explored average first-passage time as a similarity measure between vertices in the graph. Kondor and Lafferty [28] defined a graph regularization model that can be used to compute similarities between nodes. Newman [29] suggested a random-walk model to compute a “betweenness centrality” of a given node in a graph. Faloutsos et al. [30] extracted the “most relevant” connected subgraph relating two nodes of interest, based on electrical flows. Also, Palmer and Faloutsos [31] defined a similarity function based on random walks and electrical networks, and showed that this quantity provides a reasonably good measure for clustering and classifying categorical attributes. In [32], several other quantities, such as the average commute time and the pseudoinverse of the Laplacian matrix of the graph, were investigated to provide similarities between any pair of nodes. In addition, the transition probabilities of the MRW were used to define a metric among the points [17, 18]. In particular, the points are said to be more close if they have more similar probability distributions over all the points. The transition matrix was used as a smoothing kernel to improve the similarity metric [19]. Random-walk models on a graph have been proved effective in exploring the global relationships among the vertices.

However, these paths with many steps are not helpful to determine neighbors. As shown in **Fig. 1 (f)**, the neighbors determined by paths with 100 steps are not reliable. In view of this fact, this paper proposes to determine the neighbors of each vertex using only the paths in its neighborhood. In our proposed method, the probability of MRW after performing suitable steps is utilized to determine the neighbors in the sparse graph. We assume that $v_i$ is closer to $v_j$ if the *t*-step probability of the MRW from $v_\mathrm{i}$ to $v_j$ is higher. Also, the *t*-step probabilities of the MRW are adopted to determine the $k$-NN in this work. Due to the robustness of the transition probabilities of the MRW, our proposed method can not only effectively remove the unreliable neighbors, but also avoid, to an extent, the condensing degradation. It should be noted that our proposed method is different from the method presented in [20] that uses the MRW to improve spectral clustering. In our method, the probability of the MRW after performing suitable steps is utilized to determine the neighbors in sparse graphs, whereas in [20] the stable state of the MRW is used as an affinity matrix. As shown in **Fig. 1**, the probability of the MRW carries different information about neighbors after different steps. The probability after performing suitable steps can reflect the good relations of the neighborhood and the stable state carries a small amount of information about neighbors. There is also an inherent difference between our proposed method and the method presented in [21]. In [21], the MRW is utilized to intuitively interpret spectral clustering. By contrast, in our method, the MRW is used to measure the relationships among data.

The rest of this paper is organized as follows. The backgrounds of MRW and spectral clustering are reviewed in Section 2. In Section 3, we propose the construction of *k*-NN graph based on MRW. Experiment results on some datasets are presented in Section 4. Finally, some concluding remarks are given in Section 5.

## 2. Spectral Clustering and Markov Random Walk

### 2.1 Spectral clustering

Spectral clustering is a class of methods that is based on eigen-decomposition of graph affinity matrices. Given a set of data points $S=\{x_1,\cdots,x_n\}$, a weighted graph $G=(V,E)$ is first constructed in which every vertex corresponds to a point in $S$ and each edge is weighted by the similarity between the connected points. The Laplacian graph $L$ [22] is then derived from the adjacency matrix of $G$, and the eigenvectors of $L$ are computed. Finally, the traditional $K$-means method is applied to the low-dimensional representations of the original data. Many spectral clustering algorithms based on the above procedures have been proposed [3−7].

In this paper, the Ng-Jordan-Weiss (NJW) algorithm [6] is briefly reviewed and summarized in Algorithm 1 for completeness. It is noted that the proposed graph can also be used in any other spectral clustering algorithm.

**Algorithm 1:** NJW spectral clustering algorithm

**Input:** Dataset $S=\{x_1,\cdots,x_n\}$ in $\Re^l$ and the number of clusters $k$.

**Output:** $k$-way partition of the input data.

**Step 1:** Construct an affinity matrix A using the following Gaussian kernel function:

$$A_{ij}=\begin{cases}\exp\left(\dfrac{-\|x_i-x_j\|^2}{\delta^2}\right) & \text{for } i\neq j,\\ 0 & \text{for } i=j,\end{cases} \quad (1)$$

where δ is a scale parameter to control how fast the similarity attenuates with the distance between the data points $x_i$ and $x_j$.

**Step 2:** Compute the normalized affinity matrix $L=D^{-1/2}A\,D^{-1/2}$, where $D$ is the diagonal matrix with $D_{ii}=\sum_{j=1}^{n}A_{ij}$.

**Step 3:** Compute the $k$ eigenvectors of $L$, $v_1, v_2, \cdots, v_k$, which are associated with the $k$ largest eigenvalues, and form the matrix $X=[v_1\ v_2\ \cdots\ v_k]$.

**Step 4:** Renormalize each row to form a new matrix $Y\in\Re^{n\times k}$ with $Y_{ij}=X_{ij}\big/(\sum_j X_{ij}^2)^{1/2}$, so that each row of $Y$ has a unit magnitude.

**Step 5:** Treat each row of $Y$ as a point in $\Re^k$ and partition the n points (n rows) into $k$ clusters via a general clustering algorithm, such as the $K$-means algorithm.

**Step 6:** Assign the original point $x_i$ to the cluster $c$ if and only if the corresponding row $i$ of the matrix $Y$ is assigned to the cluster $c$.

## 2.2 Markov Random Walk

Markov chain is a discrete random process with memorylessness, in which the next state depends only on the current state and not on the sequence of events that preceded it. It has many applications, such as statistical models of real-world processes. Markov random walk can be seen as a Markov chain. More details about Markov random walk can be found in [23].

Given a sequence of random variables $X=\{X(n),n=0,1,2,\cdots\}$ and its state space $E=\{1,\cdots,N\}$, for $n_1,n_2,\cdots n_m(0\le n_1<n_2\cdots<n_m)$, $i_1,i_2,\cdots,i_m,j\in E$ and $t$, if the probabilities follow:

$$\begin{aligned} &P\{X(n_m+k)=j\mid X(n_1)=i_1,X(n_2)=i_2,\cdots,X(n_m=i_m) \\ &=P(X(n_m+k)=j\mid X(n_m)=i_m\} \end{aligned} \tag{2}$$

then, the sequence $X$ is called a Markov chain. The above equation formulates the Markov property: memorylessness. In (2), $P\{X(n+t)=j\mid X(n)=i\},t\ge 1$ is called the transition probabilities of the Markov chain going from the state $i$ at time $n$ to the state $j$ at time $n+t$, which is usually denoted by $p_{ij}(n,n+t)$. A Markov chain is said to be time-homogeneous if the value of $p_{ij}(n,n+t)$ is independent of the start time $n$. So, in a time-homogeneous Markov chain, the transition probability between two states depends only on the time interval, $t$, which is also called the transition step. Markov chain is assumed to be time-homogeneous in this paper. The probability of Markov chain from state $i$ to state $j$ after $t$ steps, $p_{ij}(t)$, is denoted by $p_{ij}^{(t)}$. If the number of steps is one, then $p_{ij}^{(1)}$ is abbreviated to $p_{ij}$. The transition matrix of a Markov chain can be constructed as follows:

$$P=\begin{bmatrix} p_{11} & p_{12} & \cdots & \cdots \\ p_{21} & p_{21} & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & p_{NN} \end{bmatrix} \tag{3}$$

When the number of steps is larger than one, the transition matrix after $t$ steps is called a high-order transition matrix, which is represented as follows:

$$P^{(t)}=\begin{bmatrix} p_{11}^{(t)} & p_{12}^{(t)} & \cdots & \cdots \\ p_{21}^{(t)} & p_{21}^{(t)} & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & p_{NN}^{(t)} \end{bmatrix} \tag{4}$$

The transition matrix after $t$ steps can be computed using the $t$-th power of the one-step transition matrix as follows [23]:

$$P^{(t)} = P^t = \overbrace{P \cdot P \cdots\cdot P}^{n\ Ps} \tag{6}$$

From the equation (6), the $t$-step transition matrix $P^{(t)}$ can be easily computed once the one-step transition matrix is obtained. The elements of $P^{(t)}$, denoted as $P_{ij}^{(t)}$, represent the transition probability of the MRW going from state $i$ to $j$ after $t$ steps. The probability represents different meanings in different applications. In this paper, it is interpreted as the extent of connection among the data, as discussed in Section 3.

# 3. MRW-*K*-NN GRAPH

The transition probabilities of the MRW reflect the complex relations among vertices. If there are more paths with high weight between two vertices, the transition probability between them will be larger. Thus, the relationships among states carried by high-order transition probabilities are more reliable than those in one-step transition matrix. This observation motivates us to employ it to determine the neighbors of the sparse graph in spectral clustering. It is assumed that $v_i$ is closer to $v_j$ if the $t$-step probability of the MRW from $v_{\mathrm{i}}$ to $v_j$ is higher. The proposed method comprises two main steps. First, MRW is defined on the classical $k$-NN graph, which usually contains many unreliable neighbors. Then, the new sparse graph is constructed based on the $t$-step transition probabilities of MRW.

The robustness of the transition probabilities of MRW enables the proposed method to not only effectively remove unreliable neighbors, but also, to an extent, avoid the condensing degradation. The constructed sparse graph is called the MRW-*k*-NN graph. Details of the proposed method are given as follows.

## 3.1 Definition of MRW on the common k-NN graph

Given a data set $\mathbf{V} = \{v_1, v_2 \cdots, v_n\}$ and a similarity measure $f : \mathbf{V} \times \mathbf{V} \to \Re$, the common $k$-NN graph, denoted as the C-*k*-NNG, is constructed as follows. Each data point corresponds to a vertex in the graph. For each data $v_i$, it only connects to the $k$ nearest points according to the similarity measure *f*. The edge between two connected vertices is weighted by their similarity. The affinity matrix corresponding to the C-$k$-NNG is defined as:

$$\mathbf{W}_{ij} = \begin{cases} s_{ij} & \text{if } v_j \in N_k(v_i) \text{ } or \text{ } v_i \in N_k(v_j) \\ 0 & \text{otherwise} \end{cases}, \tag{7}$$

where $s_{ij}$ is the similarity between $v_i$ and $v_j$ (i.e., $f(v_i, v_j)$). Here, $N_k(v_i)$ is the set of $k$ nearest neighbors of $v_i$.

Next, define a MRW on the C-$k$-NNG [4], assign the states of the random walk to its vertices, and define the transition probabilities between the states as functions of the weighted edges. The one-step transition probability matrix $\mathbf{P}$ of the MRW can be obtained directly from the affinity matrix $\mathbf{W}$ as follows:

$$\mathbf{P}_{ij} = \frac{\mathbf{W}_{ij}}{\sum_k \mathbf{W}_{ik}}, \tag{8}$$

where, $\mathbf{P}_{ij}$ represents the one-step transition probability of the MRW going from $v_i$ to $v_j$, and

$$\sum_{j=1}^{n} \mathbf{P}_{ij} = 1. \tag{9}$$

Given the one-step transition matrix, the $t$-step probability matrix $\mathbf{P}^{(t)}$ can be easily obtained by computing the power of $\mathbf{P}$ using equation (6). The element of $\mathbf{P}^{(t)}$, denoted as $\mathbf{P}_{ij}^{(t)}$, represents the probability of the MRW going from $v_i$ to $v_j$ in $t$ steps.

The one-step transition matrix of the MRW only corresponds to the first order neighborhood structure of the graph. On the other hand, the $t$-step transition probability of the MRW corresponds to the high-order neighborhood structure of the graph. Hence, $\mathbf{P}_{ij}^{(t)}$ can effectively reflect the connection between two vertices, $v_i$ and $v_j$. $\mathbf{P}_{ij}^{(t)}$ is larger if there exist more paths with larger weighted edges between $v_i$ and $v_j$, and vice versa.

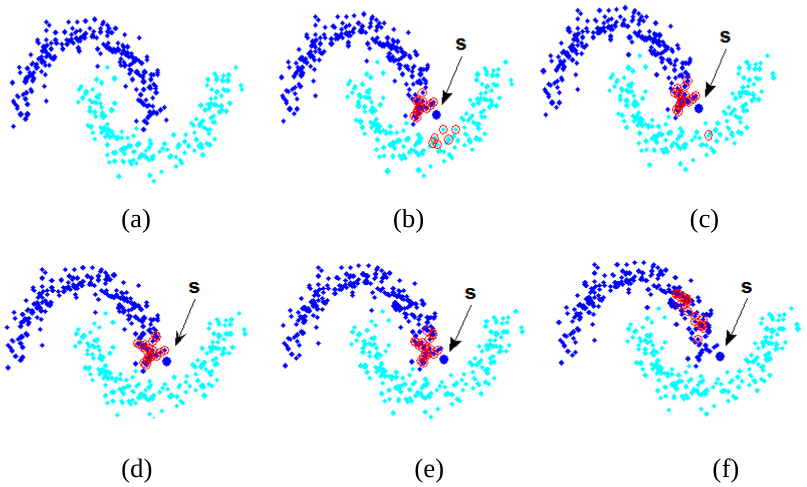
### 3.2 Construction of the MRW-k-NN graph

The $t$-step transition probabilities are employed in this paper because they can reflect more reliable relationships among data. Vertex $v_i$ is only connected to the $k$ neighbors with the largest $t$-step transition probabilities going from $v_i$. The edges are weighted by $\mathbf{W}_{ij}$. The corresponding affinity matrix is denoted as $\mathbf{W}^{(t)}$ and defined as follows:

$$\mathbf{W}_{ij}^{(t)} = \begin{cases} \mathbf{W}_{ij} & \text{if } \mathbf{P}_{ij}^{(t)} \geq I_k(\mathbf{P}_{i\cdot}^{(t)}) \\ 0 & \text{otherwise} \end{cases}, \tag{10}$$

where $\mathbf{P}_{i\cdot}^{(t)}$ denotes the set containing $\mathbf{P}_{ij}^{(t)}$ for $j = 1, 2, \cdots, n$ and $I_k(\mathbf{P}_{i\cdot}^{(t)})$ denotes the $k$*-th* largest element in $\mathbf{P}_{i\cdot}^{(t)}$. It is worth noting that the edges in the MRW-$k$-NN graph are still weighted by $\mathbf{W}$, and $\mathbf{P}^{(t)}$ is only used to determine the connected vertices. This is because the value of $\mathbf{P}_{ij}^{(t)}$ varies with respect to the value of $t$ and the degree of vertex $v_i$. **Fig. 1** illustrates the advantages of using the MRW-$k$-NN in the construction of the graph. In the figure, the solid circle denotes the point $s$ and the red circles denote its neighbors. The 30 nearest neighbors of point $s$ determined by the transition probability with different values of $t$ are shown in **Figs. 1(b)–(f)**. In **Fig. 1(b)**, the neighbors of the point $s$ are determined by $\mathbf{P}$, which is the same as the result obtained by the common $k$-NN method. It can be seen that its neighbors contain some points from other manifolds. This shows that some of the neighbors determined by the common $k$-NN method are unreliable. However, the neighbors become increasingly more reliable as the value of $t$ increases. It is worth noting that the neighbors of the point $s$ are unreliable when $t$ becomes very large, as illustrated in **Fig. 1(f)**. This is because when $t$ becomes very large, the probabilities converge to a unique stationary distribution. Hence, they have little distinguishable information. As a result, the value of $t$ plays an important role in the MRW based methods. In [17], it is suggested that the value of $t$

should be chosen according to the task. In this paper, a heuristic method is proposed for spectral clustering, and the value of $t$ is searched from 1 to 20. The value of $t$ minimizing the normalized cut [5] of the resultant clusters is chosen.



(a) (b) (c)

(d) (e) (f)

**Fig. 1.** The 30 nearest neighbors of the point s determined by the transition probability at different values of *t*. (a) Original data. (b) *t*=1. (c) *t*=2. (d) *t*=4. (e) *t*=6. (f) *t*=100.

### 3.3 Spectral clustering with the MRW-k-NN graph

The MRW-$k$-NN graph can be constructed by following the above steps. It is easy to apply the proposed graph to spectral clustering by substituting the affinity matrix in Algorithm 1 with the best $\mathbf{W}^{(t)}$ that minimizes the normalized cut. Spectral clustering with the MRW-$k$-NN graph is described in Algorithm 2.

**Algorithm 2:** Spectral clustering with the MRW-$k$-NN graph

**Input:** Dataset $S=\{x_1,\cdots,x_n\}$ in $\Re^l$, the number of clusters, $k$, and the maximum order , $N$.
**Output:** $k$-way partition of the input data.
**Step 1:** Construct the common $k$-NN graph as in (7), and define the MRW on it as in (8).
**Step 2:** Compute the transition matrices $\mathbf{P}^{(t)}$ $(t\in\{1,2,\cdots,N\})$, and obtain $N$ corresponding affinity matrices $\mathbf{W}^{(t)}$ $(t\in\{1,2,\cdots,N\})$ as in (10).
**Step 3:** For each $\mathbf{W}^{(t)}$, run the NJW algorithm with $\mathbf{W}^{(t)}$ by substituting the affinity matrix A in the step 1 of Algorithm 1, and obtain $N$ clustering results $C^t$ $(t\in\{1,2,\cdots,N\})$.
**Step 4:** Evaluate the normalized cut of $C^t$ $(t\in\{1,2,\cdots,N\})$ and select the clustering result, $C^*$, corresponding to the minimum cut as the return.

### 3.4 Complexity analysis of spectral clustering with the MRW-k-NN graph

The difference between the new spectral clustering and the NJW algorithm (Algorithm 1) lies in their sparse graphs. In the proposed method, there exists a process to select the suitable order for the MRW. As described in Algorithm 2, we search for the best order by choosing the minimum cut among the $N$ partitions. Suppose that we divide $n$ data points into $k$ clusters, the computing complexity of the NJW algorithm is $O(n^3)$ and the computing complexity of Algorithm 2 is $O(Nn^3)$. In Section 4, we will show the time-consuming evaluation for the proposed method. In addition, the space complexity of Algorithm 2 is $O(n^2)$, which is the same as that of the NJW algorithm.

## 4. Experimental Results and Analysis

To illustrate the effectiveness of our proposed method, we conduct the experiments on several synthetic and real datasets. We compared three existing methods for constructing the $k$-NN graphs: the classical method [4], the method based on dominant set [10] and $k$-regular nearest neighbor graph [11]. They are denoted as the C-$k$-NN, DN-$k$-NN and R-$k$-NN, respectively. Here, the similarity measure function is defined as:

$$f(v_i, v_j) = \exp\left(\frac{-\| v_i - v_j \|^2}{\delta^2}\right), \tag{11}$$

where ||•|| denotes the Euclidean distance and $\delta$ is a scale parameter that controls the rate of change of the similarity with the distance among the data points. There are two parameters to be set: the value of $k$ in (7) and in (10), as well as the value of $\delta$ in (11), which will be discussed in the following subsections.

### 4.1 Synthetic dataset

The synthetic dataset contains the *two-half-moon* data and the *circle-Gaussian* data, as shown in **Figs. 1(a)** and **3(a)**, respectively. The *two-half-moon* data consist of two clusters that are distributed with nonlinear elongated structures. There are some noises between the two clusters. The *circle-Gaussian* data comprise three clusters: two clusters in the middle with Gaussian distribution, and one outer cluster with a circular shape. It is worth noting that the *circle-Gaussian* data are not well separated. The points around the Gaussian clusters tend to connect all these three clusters.

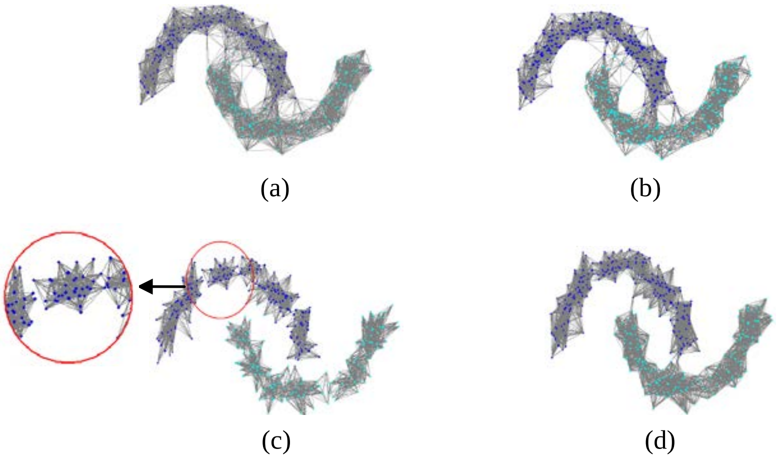


**Fig. 2.** The sparse graphs on the *two-half-moon* data constructed by different methods. (a) The C-k-NN method [4]. (b) The R-k-NN method [11]. (c) The DN-k-NN method [10]. (d) Our proposed method.

In order to have a fair comparison, the value of $k$ is chosen as 30 in all methods. The value of $\delta$ is chosen as the mean distance between the vertices to their *tenth* neighbor defined, according to the Euclidean distance. The sparse graphs of these two synthetic data constructed by four different methods are shown in **Fig. 2** and **Fig. 3**, respectively. From **Fig. 2** and **Fig. 3**, it is clear that the graphs of these two datasets have common characteristics. There are some erroneous neighbors on the C-$k$-NN graph. The $k$ nearest neighbors of some points are in other clusters. This violates the principle requiring that the similarities among the points in the same manifold should be larger than those in different manifolds. Compared with the C-$k$-NN graph, there are fewer unreliable neighbors in R-$k$-NN graphs. However, the performances of the R-$k$-NN graphs are worse than those of the DN-$k$-NN and MRW-$k$-NN graphs. The DN-$k$-NN graph determines the nearest neighbors well. However, the vertices in the same manifold are separated into several parts. This causes the DN-$k$-NN graphs to be unsuitable for applications such as spectral clustering. This deficiency can be illustrated as follows. For the MRW-$k$-NN graph, the edges connecting various vertices distributed in different manifolds have been significantly reduced. Also, the vertices in the same manifold are not disconnected. Hence, the graphs constructed by our proposed method outperform those constructed by the other methods.

## 4.2 MNIST dataset

We also conducted experiments on the MNIST dataset (http://yann.lecun.com/exdb/mnist/). This dataset consists of 10 digits with a total of 70,000 examples. Every example is a 28×28 gray level image. These 784 pixels are used as a feature vector in the simulations, and principal component analysis is applied for pre-processing before computing the Euclidean distances. In the experiments, each subset is constructed by selecting the first 200 examples from each digit. To illustrate the significance of these sparse graphs, the full graph (FULL) is also included for the comparisons.

The value of $\delta$ is chosen as the local scale parameter defined in [24] and $k$ is chosen as 10. The normalized mutual information (NMI) is employed as a performance index of the clustering, which is widely employed to evaluate clustering results [9, 25]. Suppose X and Y denote two random variables, then the NMI is defined as:

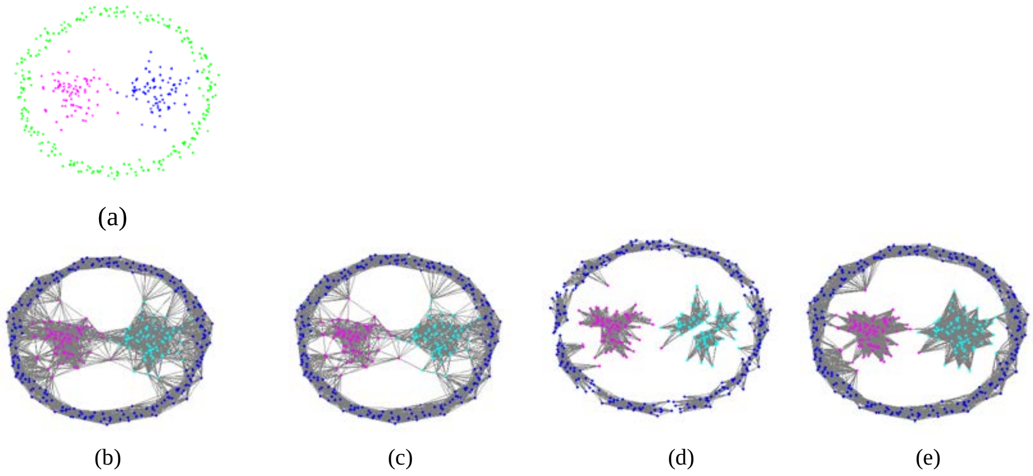


**Fig. 3.** The sparse graphs on the *circle-Gaussian* data constructed by different methods. (a) Circle-gauss data. (b) The C-*k*-NN method [4]. (c) The R-*k*-NN method [11]. (d) The DN-*k*-NN method [10]. (e) Our proposed method.

$$\mathrm{NMI}(\mathbf{X},\mathbf{Y}) = \frac{I(\mathbf{X},\mathbf{Y})}{\sqrt{H(\mathbf{X})H(\mathbf{Y})}} \tag{12}$$

where $I(\mathrm{X},\mathrm{Y})$ is the mutual information between **X** and **Y**, and $H(\mathrm{X})$ is the entropy of **X**. When evaluating the performance of clustering algorithms, the NMI is computed by regarding the clustering result and true class label as two random variables. Specifically, given a dataset with n points, the true classes and clustered classes are represented by $\{C_1^l, C_2^l, \cdots, C_k^l\}$ and $\{C_1^h, C_2^h, \cdots, C_{k'}^h\}$, respectively, where *k* and *k'* denote the number of the classes and the number of the true classes, respectively. The NMI is computed as:

$$\mathrm{NMI} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{k'} n_{ij}\log\left(\frac{nn_{ij}}{n_i^l n_j^h}\right)}{\sqrt{\left(\sum_{i=1}^{k} n_i^l \log\frac{n_i^l}{n}\right)\left(\sum_{j=1}^{k'} n_j^h \log\left(\frac{n_j^h}{n}\right)\right)}} \tag{13}$$

where $n_i^l$ and $n_j^h$ denote the number of points in $C_i^l$ and $C_j^h$, and $n_{ij}$ denotes the number of points belonging to $C_i^l$ and $C_j^h$. The range of NMI is from 0 to 1 and larger NMI implies better clustering results.

First, we constructed several subsets containing the digital groups {1, 2, 3}, {0, 5, 6}, {1, 4, 7, 9} and {0, … , 9}. The results obtained for these subsets are shown in **Fig. 4**. It can be seen that the sparse graphs obtained by our proposed method achieved better NMIs for all the subsets than those obtained by other methods. Moreover, these sparse graphs of the existing methods achieved better results compared to the full graph (FULL), which confirms that sparse graphs are important.

We also conducted simulations on the subsets comprising one pair of digits—a total of 45 subsets (pairs). **Fig. 5** shows these results. In most cases, our proposed method outperformed other methods. In particular, the spectral clustering with the graphs constructed via the C-$k$-NN method, the DN-$k$-NN method, the R-$k$-NN method and our proposed method achieved the best results in 1, 11, 6, 6, 21 cases, respectively. The mean and the standard deviation of the values of NMIs of different methods on these 45 pairs are shown in **Table 1**. It can be seen from Table 1 that our proposed method obtained the highest mean and the lowest standard deviation of the values of NMIs. The larger mean value refers to the better spectral clustering performance, while the smaller standard deviation refers to the more robustness of the method. Thus, from **Table 1** it is clear that our proposed method comprehensively outperforms the other three methods.

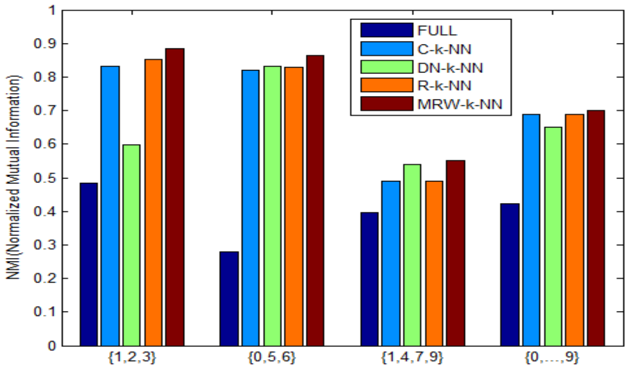


**Fig. 4.** The results on several subsets of the MNIST dataset.

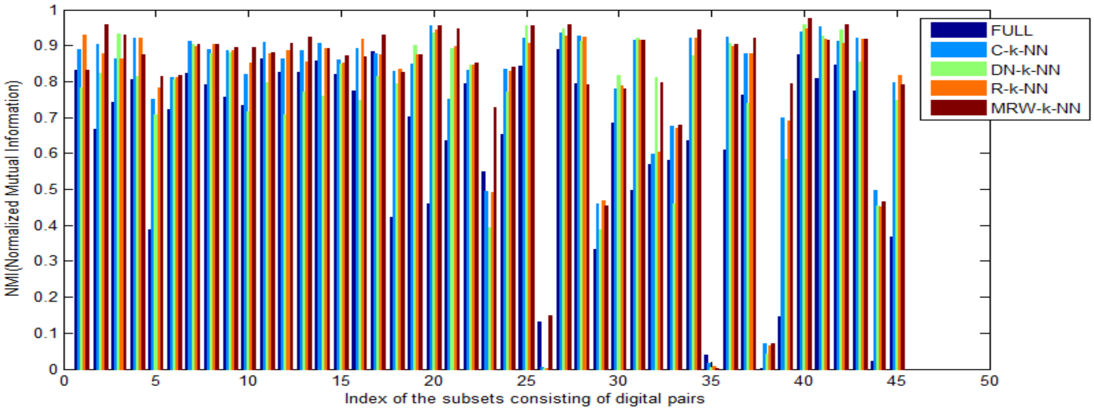


**Fig. 5.** Results for 45 subsets of the MNIST dataset

**Table 1.** Means and standard deviations of the NMIS of the spectral clustering for the different graphs performed on the 45 tests.

| | FULL | C-*k*-NN | DN-*k*-NN | R-*k*-NN | MRW-*k*-NN |
|---|---|---|---|---|---|
| **mean** | 0.6338 | 0.7806 | 0.7438 | 0.7827 | **0.8055** |
| **standard deviation** | 0.2504 | 0.2359 | 0.2431 | 0.2379 | **0.2262** |

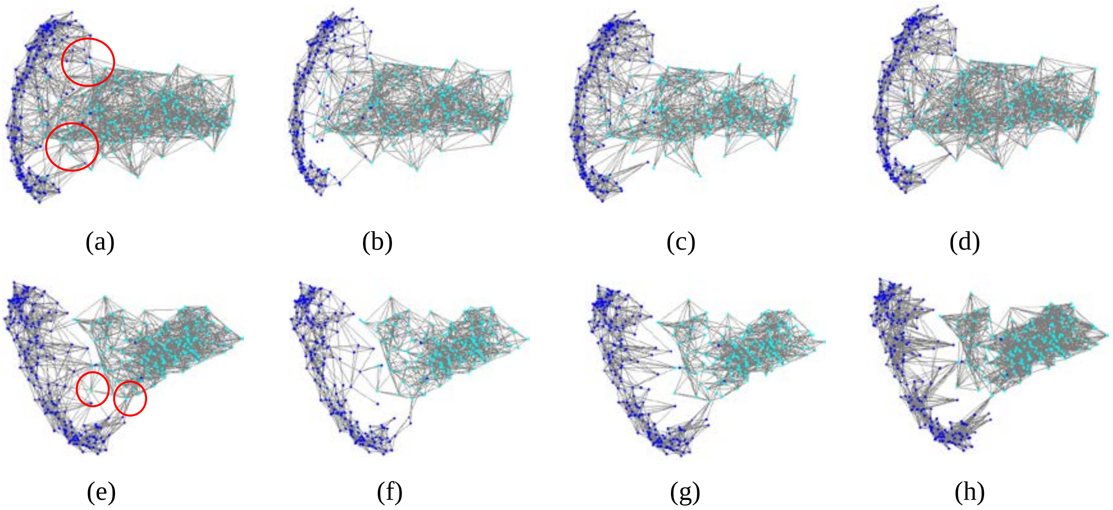


**Fig. 6.** Sparse graphs constructed on the pairwise digit subsets by various methods. Graphs (a)-(d) are constructed on the subset {1, 2} by C-*k*-NN method [4], R-*k*-NN method [11], DN-*k*-NN method [10] and our proposed method respectively. Graphs (e)-(h) are constructed on the subset {1, 9} by C-*k*-NN method [4], R-*k*-NN method [11], DN-*k*-NN method [10] and our proposed method respectively. The red circles in (a) and (e) are the notable regions.

To further illustrate the effectiveness of the proposed method, we also show the sparse graphs on the real-world dataset. It is difficult to draw the actual graph because the samples are in high-dimension space; therefore, we applied principal component analysis and obtained graphs in the two principal dimensions, as shown in **Fig. 6**. In the figure, the vertices with the same color denote the same kind of digits, of which there are two kinds in the graphs. From the red circles indicating the notable regions, it can be seen that: 1) there exist many edges connected vertices from different digits in C-*k*-NN graph; 2) the other three methods, namely R-*k*-NN, DN-*k*-NN and the proposed method, can reduce the wrong edges to an extent, which can be seen in the notable regions; 3) compared with the graphs by R-*k*-NN and DN-*k*-NN, especially in the notable regions, more right edges and less error are found in the graph based on our proposed method, In general, the graph based on our proposed method is more suitable for spectral clustering.

Additionally, we also examined the time consumption of these methods. **Fig. 7** shows the computational time required for the simulations of these five methods. As expected, the time taken for the spectral clustering with MRW-k-NN is always approximately N (maximum order) times that of the spectral clustering with C-k-NN, the same as those with FULL and DN-k-NN. However, our proposed method takes less time than the R-k-NN method. The main reason more time was required for our proposed method is that the spectral clustering algorithm has to run repeatedly to look for the best order of the MRW.

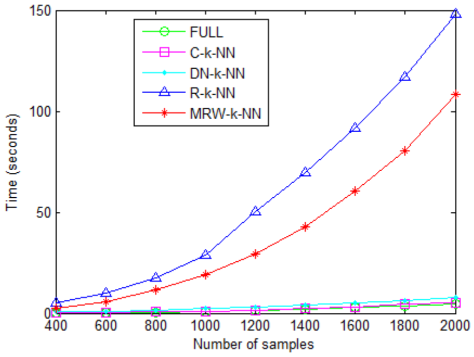


**Fig. 7.** Comparison of running-time for five algorithms on the MNIST dataset.

## 5. CONCLUSIONS

To improve the performance of spectral clustering, the MRW based method for constructing reliable k-NN graphs is proposed in this paper. The high-order transition probability of MRW is adopted to determine the neighbors in the proposed method. The proposed method consists of two steps. In the first step, an MRW model on the common $k$ -NN graph constructed by the data is introduced. Then, in the second step, the probability of the MRW is computed and utilized to determine the neighbors. Because of the advantage that the high-order transition probabilities characterize the complex relationships among data, the neighbors selected in the proposed graph are more reliable than those of existing methods. The results of experiments show that the graph constructed by our proposed method performs better and can efficiently improve the performance of spectral clustering.

In future work, we plan to explore a more efficient algorithm to determine the order of the MRW and further verify the effectiveness of the proposed graph in other applications.

## References

[1] K. Jain, “Data clustering: 50 years beyond K-means, Pattern Recognition Letters,” vol. 31, pp. 651-666, 2010. Article (CrossRef Link)

[2] F. Elijorde, S. Kim, J. Lee, “A Wind Turbine Fault Detection Approach Based on Cluster Analysis and Frequent Pattern Mining,” *KSII Transactions on Internet and Information Systems (TIIS)*, vol. 8, no. 2, pp. 664-677, 2014. Article (CrossRef Link)

[3] Wang, L., Dong, M., “Multi-level Low-rank Approximation-based Spectral Clustering for image segmentation,” *Pattern Recognition Lett*, vol. 33, pp. 2206-2215, 2012. Article (CrossRef Link)

[4] Luxburg, U, “A tutorial on spectral clustering. Statistics and Computing,” vol. 17, pp. 395-416, 2007. Article (CrossRef Link)

[5] Shi, J., Malik, J., “Normalized cuts and image segmentation,” *IEEE Trans. Pattern Anal. Machine Intell.* vol. 22, pp. 888-905, 2000. Article (CrossRef Link)

[6] Ng, A., Jordan, M., Weiss. Y., “On spectral clustering: analysis and an algorithm,” in *Proc of. Advances in Neural Information Processing Systems (NIPS)*, vol. 15, MIT Press, 2002. http://citeseerx.ist. psu.edu/viewdoc/summary?doi=10.1.1.19.8100

[7] J. Zhou, S. Gao, Z. Jin, "A New Connected Coherence Tree Algorithm For Image Segmentation," *KSII Transactions On Internet And Information Systems*, vol. 6, no. 4, pp.1118-1202, 2012. Article (CrossRef Link)

[8] Verma, D., Meila, M., "A comparison of spectral clustering algorithms," *Technical Report 03-05-01*, University of Washington Department of Computer Science, 2003. http://citeseerx. ist.psu.edu/viewdoc/summary?doi=10.1.1.57.6424

[9] J. Z. Cao, P. Chen, Q. Y. Dai, Wing-Kuen Ling, "Local information-based fast approximate spectral clustering, Pattern Recognition Letters," vol. 38, pp. 63-69, 2014. Article (CrossRef Link)

[10] X. Yang, L. J. Latecki, "Affinity learning on a tensor product graph with applications to shape and image retrieval," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 2369-2376, 2011. Article (CrossRef Link)

[11] Wang B, Pan F, Hu K M, Paul J C, "Manifold-ranking based retrieval using k-regular nearest neighbor graph," *Pattern Recognition*, vol. 45, pp. 1569-1577, 2012. doi>10.1016/j.patcog.2011.09.006

[12] Li Y, "K-edge connected neighborhood graph for geodesic distance estimation and nonlinear data projection," in *Proc. of The International Conference on Pattern Recognition*, 2004. Article (CrossRef Link)

[13] Li Y, "Building k-edge-connected neighborhood graph for distance-based data projection," *Pattern Recognition Letters*, vol. 26, no. 13, pp. 2015-2021, 2005. Article (CrossRef Link)

[14] Li Y, "Building k edge-disjoint spanning trees of minimum total length for isometric data embedding," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 10, pp. 1680-1683, 2005. Article (CrossRef Link)

[15] Jebara T, Wang J, Chang S F, "Graph construction and b-matching for semi-supervised learning," in *Proc. of International Conferenceon Machine Learning*, 2009. Article (CrossRef Link)

[16] Cheng B, Yang J, Yan S, Fu Y. Huang T S. Learning with l1-graph for image analysis. IEEE Transactionson Image Processing, 19: 858-866, 2010. Article (CrossRef Link)

[17] M. Szummer, T. Jaakkola, "Partially labeled classification with Markov random walks," *Advances in Neural Information Processing Systems (NIPS)*, pp.945-952, 2001. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.25.1955

[18] S. Lafon, A. B. Lee, "Diffusion maps and coarse-graining: a unified framework for dimensionality reduction," *graph partitioning, and data set parameterization, IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 9, pp. 1393-1403, 2006. Article (CrossRef Link)

[19] J. Jiang, B. Wang, Z. Tu, "Unsupervised metric learning by self-smoothing operator," in *Proc. IEEE Conf. on Computer Vision (ICCV)*, pp. 794-801, 2011. Article (CrossRef Link)

[20] Xianchao ZHANG, Quanzeng YOU, "An improved spectral clustering algorithm based on random walk [J]," *Frontiers of Computer Science in China*, vol. 3, pp. 268-278, 2012. Article (CrossRef Link)

[21] M Meila, J Shi, "A Random Walks View of spectral segmentation," in *Proc. of 8th Inernational Workshop on Artificial intelligence and Statistics*, 2001. http://citeseer.ist.psu.edu/viewdoc/ summary?doi=10.1.1.33.1501

[22] Chung, F. R. K, "Spectral Graph Theory, Regional Conference Series in Mathematics," *AMS. 92*, 1997.

[23] P. Bertsekas , N. Tsitsiklis, "Introduction to Probability," *AMS*, 1997. http://vfu.bg/en/e-Learning/Math--Bertsekas_Tsitsiklis_Introduction_to_probability.pdf

[24] L. Zelnik-Manor, P. Perona, "Self-tuning spectral clustering," *Advances in Neural Information Processing Systems (NIPS)*, pp. 1601-1608, 2004. http://www.vision.caltech.edu/lihi/Demos/SelfTuningClustering.html

[25] Shang, F., Jiao, L. C., Shi, J., Gong, M., Shang, R. H, "Fast density-weighted low-rank approximation spectral clustering," *Data Min Knowl Disc*, vol. 23, pp.345-378, 2011. http://link.springer.com/article/10.1007/s10618-010-0207-5

[26] Lovasz L, "Random Walks on Graphs: A Survey," *Combinatorics*, vol. 80, no. 2, pp. 1-46, 1993. www.cs.elte.hu/~lovasz/erdos.pdf

[27] S. White, P. Smyth, “Algorithms for Estimating Relative Importance in Networks,” in *Proc. of Ninth ACM SIGKDD Int’l Conf. Knowledge Discovery and Data Mining*, pp. 266-275, 2003. Article (CrossRef Link)

[28] R.I. Kondor, J. Lafferty, “Diffusion Kernels on Graphs and Other Discrete Structures,” in *Proc. of 19th Int’l Conf. Machine Learning*, pp. 315-322, 2002. Article (CrossRef Link)

[29] M. Newman, “A Measure of Betweenness Centrality Based on Random Walks,” *Social Networks*, vol. 27, no. 1, pp. 39-54, 2005. Article (CrossRef Link)

[30] C. Faloutsos, K. McCurley, A. Tomkins, “Fast Discovery of Connection Subgraphs,” in *Proc. of 10th ACM SIGKDD Int’l Conf. Knowledge Discovery and Data Mining*, pp. 118-127, 2004. Article (CrossRef Link)

[31] C. Palmer, C. Faloutsos, “Electricity Based External Similarity of Categorical Attributes,” in *Proc. of Seventh Pacific-Asia Conf. Knowledge Discovery and Data Mining (PAKDD ’03)*, pp. 486-500, 2003. Article (CrossRef Link)

[32] F. Fouss, A. Pirotte, J. Renders, M. Saerens, “Random-walk computation of similarities between nodes of a graph, with application to collaborative recommendation,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, no. 3, 2007. Article (CrossRef Link)

[33] R. Lambiotte, J. Delvenne, M. Barahona, “Random Walks, Markov Processes and the Multiscale Modular Organization of Complex Networks,” IE*EE Transactions on Network Science and Engineering*, vol. 1, no. 2, pp. 76-90, 2014. Article (CrossRef Link)

**Jiangzhong Cao** received his Ph.D. degree in communication and information system from School of Information Science and Technology, Sun Yat-sen University, China, in 2013. He is currently a Lecturer in the School of Information Engineering, Guangdong University of Technology, Guangzhou, China. His research interests include computer vision and pattern recognition.

**Pei Chen** received two Ph.D. degrees in wavelets and computer vision from Shanghai Jiaotong University, Shanghai, China, and Monash University, Melbourne, Australia, in 2001 and 2004, respectively. He is currently a Professor in the School of Information Science and Technology, Sun Yat-sen University, Guangzhou, China. He has been a Post-Doctoral Researcher with Monash University, a Senior Research Engineer with Motorola Labs, Shanghai, and a Research Professor with SIAT/CAS, Shenzhen, China. His current research interests include topics in computer vision and machine learning.

**Wing-KuenLing** received the B.Eng.(Hons) and M.Phil. degrees from the Department of Electronic and Computer Engineering, Hong Kong University of Science and Technology, in1997 and 2000, respectively, and the Ph.D. degree in the Department of Electronic and Information Engineering of the Hong degree in the Department of Electronic and Information Engineering of the HongKong Polytechnic University in 2003. In 2004, he joined the King's College London as a Lecturer. In 2010, he joined the University of Lincon as a Principal Lecturer and promoted to a Reader in 2011. In 2012, he joined Guangdong University of Technology as a Full Professor. He is a Fellow of IET, a senior member of IEEE, and a China National Young Thousand-People-Plan Distinguished Professor. He currently serves in the technical committees of the Non-linear Circuits and Systems Group, the digital signal processing group and the power electronics and systems group of the IEEE Circuits and Systems Community. He was awarded the best reviewer prizes from the IEEE Instrumentation and Measurement Society in 2008 and 2012. He has also served as the guest editor-in-chief of several special issues of highly rated international journals, such as the Circuits, Systems and Signal Processing, the American Journal of Engineering and Applied Sciences, the International Journal of Digital Content Technology and Its Applications, and the Mediterranean Journal of Electronics and Communications. He is currently an associate editor of several international journals, including International Journal of Bifurcation and Chaos, the Circuits, Systems and Signal Processing, Journal of the Franklin Institute, American Journal of Engineering and Applied Sciences, International Journal of Engineering and Emerging Technologies, International Journal of Applied Mathematics and Modeling, and Open Journal Advanced Engineering Techniques. He has published an undergraduate textbook, are search monograph, four book chapters, a book review published in an IEEE journal, more than 75 papers in international leading journals and more than 55 papers in highly rated international conference. He has held 7 visiting positions and delivered more than 20 seminars in the fields of image processing, time frequency analysis, optimization theory, symbolic dynamics, as well as fuzzy and impulsive control theory.

**Zhijing Yang** received the B.S. and Ph.D. degrees from the Faculty of Mathematics and Computing Science, Sun Yat-Sen University, Guangzhou, China, in 2003 and 2008, respectively. Then, he became a Lecturer in the same university. He was a Visiting Research Scholar in the School of Computing, Informatics and Media, University of Bradford, U.K., between June 2009 and December 2009, and a Research Fellow in the School of Engineering, University of Lincoln, between January 2011 and January 2013. He is currently an Associate Professor in the Faculty of Information Engineering, Guangdong University of Technology, China. He has been a project leader for three projects on signal processing from National Natural Science Foundation of China. His current research interests include time-frequency analysis, signal processing, machine learning, and pattern recognition.

**Qingyun Dai** received the Ph.D. degree in communication and information systems from South China University of Technology, Guangzhou, China. She is currently a Professor in the School of Information Engineering, Guangdong University of Technology, Guangzhou, China. She has also been a Post-Doctoral Researcher at Sun Yat-sen University, Guangzhou, China. Her current research interests include Image Processing, RFID and Internet of Things.