Multiclass Spectral Clustering

Stella X. Yu
Robotics Institute and CNBC
Carnegie Mellon University
Pittsburgh, PA 15213-3890

Jianbo Shi
Dept. of Computer and Information Science
University of Pennsylvania
Philadelphia, PA 19104-6389

Abstract

We propose a principled account on multiclass spectral clustering. Given a discrete clustering formulation, we first solve a relaxed continuous optimization problem by eigendecomposition. We clarify the role of eigenvectors as a generator of all optimal solutions through orthonormal transforms. We then solve an optimal discretization problem, which seeks a discrete solution closest to the continuous optima. The discretization is efficiently computed in an iterative fashion using singular value decomposition and non-maximum suppression. The resulting discrete solutions are nearly global-optimal. Our method is robust to random initialization and converges faster than other clustering methods. Experiments on real image segmentation are reported.

Spectral graph partitioning methods have been successfully applied to circuit layout [3, 1], load balancing [4] and image segmentation [10, 6]. As a discriminative approach, they do not make assumptions about the global structure of data. Instead, local evidence on how likely two data points belong to the same class is first collected and a global decision is then made to divide all data points into disjunct sets according to some criterion. Often, such a criterion can be interpreted in an embedding framework, where the grouping relationships among data points are preserved as much as possible in a lower-dimensional representation.

What makes spectral methods appealing is that their global-optima in the relaxed continuous domain are obtained by eigendecomposition. However, to get a discrete solution from eigenvectors often requires solving another clustering problem, albeit in a lower-dimensional space. That is, eigenvectors are treated as geometrical coordinates of a point set. Various clustering heuristics such as K-means [10, 9], transportation [2], dynamic programming [1], greedy pruning or exhaustive search [3, 10] are subsequently employed on the new point set to retrieve partitions.

We show that there is a principled way to recover a discrete optimum. This is based on a fact that the continuous optima consist not only of the eigenvectors, but of a whole family spanned by the eigenvectors through orthonormal transforms. The goal is to find the right orthonormal transform that leads to a discretization.

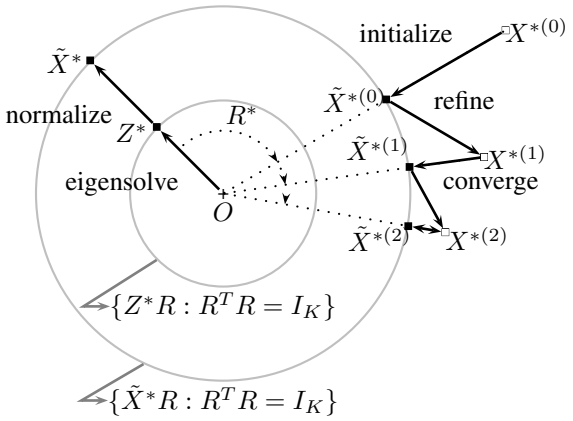

Figure 1. Schematic diagram of our algorithm. (1) We first obtain eigenvectors Z^*. Shown as the inner circle, Z^* generates the whole family of global optima through orthonormal transform R. After length normalization, each optimum corresponds to a partitioning solution in the continuous domain (the outer circle). (2) We then obtain a discrete solution closest to the continuous optima in an iterative fashion. Starting from discrete solution $X^{*(0)}$, we find $\tilde{X}^{*(0)}$ by computing R^* that brings $\tilde{X}^*$ closest to $X^{*(0)}$. Given the continuous optimum $\tilde{X}^{*(0)}$, we compute its closest discrete solution $\tilde{X}^{*(0)}$; so on and so forth. The algorithm converges at the solution pair $(X^{*(2)}, \tilde{X}^{*(2)})$, which are the closest to each other. The optimality of $\tilde{X}^{*(2)}$ guarantees that $X^{*(2)}$ is nearly global-optimal.

Illustrated in Fig. 1, our method has two steps. (1) We solve a relaxed continuous optimization problem. The global optima are given by some eigenvectors subject to arbitrary orthonormal transforms. (2) We iteratively solve for a discrete solution that is closest to the continuous optima using an alternating optimization procedure. We alternate the following: the continuous optimum closest to a discrete

Proceedings of the Ninth IEEE International Conference on Computer Vision (ICCV'03)
0-7695-1950-4/03 $ 17.00 © 2003 IEEE

solution is located by computing the best orthonormal transform, and the discrete solution closest to a continuous one is located by non-maximum suppression. Such iterations monotonously decrease the distance between a discrete solution and the continuous optima. After convergence, we obtain a nearly global-optimal partitioning. We apply our method to real image segmentation.

1 Multiclass normalized cuts

A weighted graph is specified by $\mathbb{G} = (\mathbb{V}, \mathbb{E}, W)$, where $\mathbb{V}$ is the set of all nodes; $\mathbb{E}$ is the set of edges connecting the nodes; W is an affinity matrix, with weights characterizing how likely two nodes belonging in the same group. W is assumed nonnegative and symmetric.

Let $[n]$ denote the set of integers between 1 and n: $[n] = \{1, 2, \cdots, n\}$. Let $\mathbb{V} = [N]$ denote the set of all elements (data points or pixels) to be grouped. To cluster N points into K groups is to decompose $\mathbb{V}$ into K disjoint sets, i.e., $\mathbb{V} = \cup_{l=1}^{K} \mathbb{V}_l$ and $\mathbb{V}_k \cap \mathbb{V}_l = \varnothing$, $\forall k \neq l$. We denote this K-way partitioning by $\Gamma_{\mathbb{V}}^{K} = \{\mathbb{V}_1, \cdots, \mathbb{V}_K\}$.

1.1 Multiclass partitioning criteria

Let $\mathbb{A}$, $\mathbb{B} \subset \mathbb{V}$. We define $\text{links}(\mathbb{A}, \mathbb{B})$ to be the total weighted connections from $\mathbb{A}$ to $\mathbb{B}$:

$$\text{links}(\mathbb{A}, \mathbb{B}) = \sum_{i \in \mathbb{A}, j \in \mathbb{B}} W(i, j). \tag{1}$$

The *degree* of a set is simply the total links to all the nodes:

$$\text{degree}(\mathbb{A}) = \text{links}(\mathbb{A}, \mathbb{V}). \tag{2}$$

Using the degree as a normalization term, we define

$$\text{linkratio}(\mathbb{A}, \mathbb{B}) = \frac{\text{links}(\mathbb{A}, \mathbb{B})}{\text{degree}(\mathbb{A})}, \tag{3}$$

i.e., the *proportion* of the links with $\mathbb{B}$ among those $\mathbb{A}$ has.

Two special linkratios are of particular interest. One is $\text{linkratio}(\mathbb{A}, \mathbb{A})$, which measures how many links *stay* within $\mathbb{A}$ itself. The other is its complement $\text{linkratio}(\mathbb{A}, \mathbb{V} \setminus \mathbb{A})$, which measures how many links *escape* from $\mathbb{A}$. A good clustering desires both tight connections within partitions and loose connections between partitions. These two goals are captured in the K-way *normalized associations* and *normalized cuts* criteria:

$$\text{knassoc}(\Gamma_{\mathbb{V}}^{K}) = \frac{1}{K} \sum_{l=1}^{K} \text{linkratio}(\mathbb{V}_l, \mathbb{V}_l) \tag{4}$$

$$\text{kncuts}(\Gamma_{\mathbb{V}}^{K}) = \frac{1}{K} \sum_{l=1}^{K} \text{linkratio}(\mathbb{V}_l, \mathbb{V} \setminus \mathbb{V}_l) \tag{5}$$

Since $\text{knassoc}(\Gamma_{\mathbb{V}}^{K}) + \text{kncuts}(\Gamma_{\mathbb{V}}^{K}) = 1$, maximizing the associations and minimizing the cuts are achieved simultaneously. Among the numerous criteria such as minimum cuts and various definitions of average cuts, only minimum cuts and normalized cuts have this duality property. However, minimum cuts [5] are noise-sensitive, i.e., a few isolated nodes could easily draw the cuts away from a global partitioning, whereas normalized cuts are robust to weight perturbation [10]. Since knassoc and kncuts are equivalent, we make no distinct further and denote our K-way normalized cuts objective as:

$$\varepsilon(\Gamma_{\mathbb{V}}^{K}) = \text{knassoc}(\Gamma_{\mathbb{V}}^{K}). \tag{6}$$

ε is a unit-less value between 0 and 1 regardless of K.

For any K-way partitioning criterion, we need to examine its performance over K's. For example, how does it change with K? Can it produce refinement of partitioning when K increases? The definitions in Eqn. (4) and (5) do not lend themselves an obvious answer to these questions. However, we will show that an upperbound of ε decreases monotonically with increasing K. While our criterion does not take into account the requirement of hierarchical refinement over the number of classes, a consistent optimal partitioning can often be obtained with little extra cost.

1.2 Representation

We use $N \times K$ *partition matrix* X to represent $\Gamma_{\mathbb{V}}^{K}$. Let $X = [X_1, \cdots, X_K]$, where X_l is a binary indicator for $\mathbb{V}_l$:

$$X(i, l) = \langle i \in \mathbb{V}_l \rangle, \quad i \in \mathbb{V},\ l \in [K], \tag{7}$$

where $\langle \cdot \rangle$ is 1 if the argument is true and 0 otherwise. Since a node is assigned to one and only one partition, there is an exclusion constraint between the columns of X: $X\, 1_K = 1_N$, where 1_d denotes the $d \times 1$ vector of all 1's. We define the *degree matrix* for the symmetric weight matrix W to be:

$$D = \text{Diag}(W 1_N), \tag{8}$$

where $\text{Diag}(\cdot)$ denotes a diagonal matrix formed from its vector argument. We can rewrite *links* and *degree* as:

$$\text{links}(\mathbb{V}_l, \mathbb{V}_l) = X_l^T W X_l \tag{9}$$

$$\text{degree}(\mathbb{V}_l) = X_l^T D X_l. \tag{10}$$

The K-way normalized cuts criterion is expressed in an optimization program of variable X, called program *PNCX*:

$$\text{maximize} \quad \varepsilon(X) = \frac{1}{K} \sum_{l=1}^{K} \frac{X_l^T W X_l}{X_l^T D X_l} \tag{11}$$

$$\text{subject to} \quad X \in \{0, 1\}^{N \times K} \tag{12}$$

$$X\, 1_K = 1_N. \tag{13}$$

This problem is NP-complete even for $K = 2$ and even when the graph is planar [10]. We will develop a fast and principled algorithm to find its near-global optima.

Proceedings of the Ninth IEEE International Conference on Computer Vision (ICCV'03)
0-7695-1950-4/03 $ 17.00 © 2003 IEEE

2 Solving K-way normalized cuts

We solve program *PNCX* in two steps. We first relax a transformed formulation into an eigenvalue problem. We show that its global optimum is not unique, and a special solution is the generalized eigenvectors of the matrix pair (W, D). Transforming the eigenvectors to the space of partition matrices, we get a set of continuous global optima. We then solve a discretization problem, where the discrete partition matrix closest to the continuous optima is sought. Such a discrete solution is thus near global-optimal.

2.1 Finding optimal relaxed solutions

We simplify Eqn. (11) as $\varepsilon(X) = \frac{1}{K}\,\mathrm{tr}(Z^T W Z)$, where tr denotes the trace of a matrix, and Z is a *scaled partition matrix* [3]:

$$Z = X(X^T DX)^{-\frac{1}{2}}. \tag{14}$$

Since $X^T DX$ is diagonal, the columns of Z are those of X scaled by the inverse square root of the degrees of partitions. A natural constraint on Z is:

$$Z^T DZ = (X^T DX)^{-\frac{1}{2}} X^T DX (X^T DX)^{-\frac{1}{2}} = I_K,$$

where I_K denotes the $K \times K$ identity matrix. Ignoring the constraints in *PNCX*, we derive a new program of variable Z and call it *PNCZ*:

$$\text{maximize} \quad \varepsilon(Z) = \frac{1}{K}\,\mathrm{tr}(Z^T W Z) \tag{15}$$

$$\text{subject to} \quad Z^T DZ = I_K. \tag{16}$$

Relaxing Z into the continuous domain turns the discrete problem into a tractable continuous optimization problem. This program has a special property stated below, which can be proved trivially using $\mathrm{tr}(AB) = \mathrm{tr}(BA)$.

Proposition 1 (Orthonormal Invariance). *Let R be a $K \times K$ matrix. If Z is a feasible solution to* PNCZ*, so is $\{ZR : R^T R = I_K\}$. Furthermore, they have the same objective value: $\varepsilon(ZR) = \varepsilon(Z)$.*

Therefore, a feasible solution remains equally good with arbitrary rotation and reflection. Program *PNCZ* is a Rayleigh quotient optimization problem that has been addressed in Rayleigh-Ritz theorem and its extensions. Proposition 2 rephrases the theorem in our problem setting. It can also be proved directly using Lagrangian relaxation. The proposition shows that among all the optima are the eigenvectors of (W, D), or equivalently those of *normalized weight matrix* P:

$$P = D^{-1} W. \tag{17}$$

Since P is a stochastic matrix [8], it is easy to verify that 1_N is a trivial eigenvector of P and it corresponds to the largest eigenvalue of 1.

Proposition 2 (Optimal Eigensolution). *Let (V, S) be the eigendecomposition of P: $PV = VS$, where $V = [V_1, \ldots, V_N]$ and $S = \mathrm{Diag}(s)$ with eigenvalues ordered nonincreasingly: $s_1 \geq \ldots \geq s_N$. (V, S) is obtained from the orthonormal eigensolution $(\bar{V}, S)$ of the symmetric matrix $D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$, where*

$$V = D^{-\frac{1}{2}} \bar{V}, \tag{18}$$

$$D^{-\frac{1}{2}} W D^{-\frac{1}{2}}\, \bar{V} = \bar{V} S, \quad \bar{V}^T \bar{V} = I_N. \tag{19}$$

Therefore, V and S are all real and any K distinct eigenvectors form a local optimum candidate to PNCZ*, with*

$$\varepsilon([V_{\pi_1}, \ldots, V_{\pi_K}]) = \frac{1}{K} \sum_{l=1}^{K} s_{\pi_l}, \tag{20}$$

where π is an index vector of K distinct integers from $[N]$. The global optimum of PNCZ *is thus achieved when $\pi = [1, \ldots, K]$:*

$$Z^* = [V_1, \ldots, V_K], \tag{21}$$

$$\Lambda^* = \mathrm{Diag}([s_1, \ldots, s_K]), \tag{22}$$

$$\varepsilon(Z^*) = \frac{1}{K}\,\mathrm{tr}(\Lambda^*) = \max_{Z^T DZ = I_K} \varepsilon(Z). \tag{23}$$

To summarize, the global optimum of *PNCZ* is not unique. It is a subspace spanned by the first K largest eigenvectors of P through orthonormal matrices:

$$\{Z^* R :\, R^T R = I_K,\, PZ^* = Z^* \Lambda^*\}. \tag{24}$$

Unless the eigenvalues are all the same, $Z^* R$ are no longer the eigenvectors of P. All these solutions have the optimal objective value, which provides a nonincreasing upperbound to *PNCX*.

Corollary 1 (Upperbound Monotonicity). *For any K,*

$$\max \varepsilon(\Gamma_{\mathbb{V}}^K) \leq \max_{Z^T DZ = I_K} \varepsilon(Z) = \frac{1}{K} \sum_{l=1}^{K} s_l \tag{25}$$

$$\max_{Z^T DZ = I_{K+1}} \varepsilon(Z) \leq \max_{Z^T DZ = I_K} \varepsilon(Z). \tag{26}$$

Next we transform Z back to the space of partition matrices. If f is the mapping that scales X to Z, then f^{-1} is the normalization that brings Z back to X:

$$Z = f(X) = X(X^T DX)^{-\frac{1}{2}} \tag{27}$$

$$X = f^{-1}(Z) = \mathrm{Diag}(\mathrm{diag}^{-\frac{1}{2}}(ZZ^T))\, Z, \tag{28}$$

where diag returns the diagonal of its matrix argument in a column vector. If we take the rows of Z as coordinates of K-dimensional points, what f^{-1} does is to *normalize* their

Proceedings of the Ninth IEEE International Conference on Computer Vision (ICCV'03)
0-7695-1950-4/03 $ 17.00 © 2003 IEEE

lengths so that they lie on the unit hypersphere centered at the origin. With f^{-1}, we transform the continuous optimum Z^*R in the Z-space to the X-space: since $R^TR = I_K$,

$$f^{-1}(Z^*R) = f^{-1}(Z^*)R. \tag{29}$$

This simplification allows the continuous optima to be directly characterized by $f^{-1}(Z^*)$ in the X-space:

$$\{\tilde{X}^*R:\ \tilde{X}^* = f^{-1}(Z^*),\ R^TR = I_K\}. \tag{30}$$

It is now clear that we need K and only K eigenvectors to yield K (not 2^K) partitions. The reason is that group indicators are constrained to be orthogonal. They cannot be chosen freely, as required for a hierarchical partitioning [10]. We also gain more perspective on the first eigenvector. Though $Z_1^* = (1_N^T D 1_N)^{-\frac{1}{2}} \cdot 1_N$ is a trivial multiple of 1_N, $\tilde{X}_1^*$ is not for $K > 1$. The seemingly trivial first eigenvector is as important as any others, since they collectively provide a basis for generating the whole set of optima.

2.2 Finding optimal discrete solutions

The optima of *PNCZ* are in general not feasible to the program *PNCX*. However, we can use them to find a nearby discrete solution. This discrete solution may not be the absolute maximizer of *PNCX*, but it is nearly global-optimal due to the continuity of the objective function. Therefore, our goal here is to find a discrete solution that satisfies the binary constraints of the original program *PNCX*, yet is closest to the continuous optima given in Eqn. (30).

Theorem 1 (Optimal Discretization). *Let $\tilde{X}^* = f^{-1}(Z^*)$. An optimal discrete partition X^* is considered the one satisfying the following program called* POD*:*

$$\textit{minimize} \quad \phi(X,R) = \|X - \tilde{X}^*R\|^2 \tag{31}$$

$$\textit{subject to} \quad X \in \{0,1\}^{N\times K}, \quad X\,1_K = 1_N \tag{32}$$

$$R^TR = I_K, \tag{33}$$

where $\|M\|$ denotes the Frobenius norm of matrix M: $\|M\| = \sqrt{\mathrm{tr}(MM^T)}$. A local optimum (X^*, R^*) of this bilinear program can be solved iteratively.

Given R^*, POD *is reduced to program* PODX *in X:*

$$\textit{minimize} \quad \phi(X) = \|X - \tilde{X}^*R^*\|^2 \tag{34}$$

$$\textit{subject to} \quad X \in \{0,1\}^{N\times K}, \quad X\,1_K = 1_N. \tag{35}$$

Let $\tilde{X} = \tilde{X}^*R^*$. The optimal solution is given by non-maximum suppression (if there are multiple maxima, only one of them, but any one of them, can be chosen so as to honor the exclusion constraint on a partition matrix):

$$X^*(i,l) = \langle l = \arg\max_{k\in[K]} \tilde{X}(i,k)\rangle, \quad i \in \mathbb{V}. \tag{36}$$

Given X^*, POD *is reduced to program* PODR *in R:*

$$\textit{minimize} \quad \phi(R) = \|X^* - \tilde{X}^*R\|^2 \tag{37}$$

$$\textit{subject to} \quad R^TR = I_K, \tag{38}$$

and the solution is given through some singular vectors:

$$R^* = \tilde{U}U^T, \tag{39}$$

$$X^{*T}\tilde{X}^* = U\Omega\tilde{U}^T, \quad \Omega = \mathrm{Diag}(\omega), \tag{40}$$

where $(U, \Omega, \tilde{U})$ is a singular value decomposition (SVD) of $X^{*T}\tilde{X}^*$, with $U^TU = I_K$, $\tilde{U}^T\tilde{U} = I_K$ and $\omega_1 \geq \ldots \geq \omega_K$.

Proof. First we note that: $\phi(X,R) = \|X\|^2 + \|\tilde{X}^*\|^2 - \mathrm{tr}(XR^T\tilde{X}^{*T} + X^T\tilde{X}^*R) = 2N - 2\,\mathrm{tr}(XR^T\tilde{X}^{*T})$. Thus minimizing $\phi(X,R)$ is equivalent to maximizing $\mathrm{tr}(XR^T\tilde{X}^{*T})$. For *PODX*, given $R = R^*$, as each entry of $\mathrm{diag}(XR^{*T}\tilde{X}^{*T})$ can be optimized independently, Eqn. (36) results. For *PODR*, given $X = X^*$, we construct a *Lagrangian* using a symmetric matrix multiplier Λ:

$$L(R,\Lambda) = \mathrm{tr}(X^*R^T\tilde{X}^{*T}) - \frac{1}{2}\,\mathrm{tr}(\Lambda^T(R^TR - I_K)).$$

The optimum (R^*, Λ^*) must satisfy

$$L_R = \tilde{X}^{*T}X^* - R\Lambda = 0, \quad \text{i.e.} \quad \Lambda^* = R^{*T}\tilde{X}^{*T}X^*. \tag{41}$$

Thus $\Lambda^{*T}\Lambda^* = U\Omega^2U^T$. Since $\Lambda = \Lambda^T$, $\Lambda^* = U\Omega U^T$. From Eqn. (41), we then have: $R^* = \tilde{U}U^T$ and $\phi(R^*) = 2N - 2\,\mathrm{tr}(\Omega)$. The larger $\mathrm{tr}(\Omega)$ is, the closer X^* is to $\tilde{X}^*R^*$. □

Due to the orthonormal invariance of the continuous optima, our method is robust to arbitrary initialization, from either X or R. A good initialization can nevertheless speed up convergence. We find that the heuristic mentioned in [9] is good and fast. It is simply K-means clustering with K nearly orthogonal data points as centers. Computationally, it is equivalent to initialize R^* by choosing K rows of $\tilde{X}^*$ that are as orthogonal to each other as possible. To derive X^* by Eqn. (36) on this non-orthogonal R^* is exactly K-means clustering with the unit-length centers.

Given X^*, we solve *PODR* to find a continuous optimum $\tilde{X}^*R^*$ closest to it. For this continuous optimum, we then solve *PODX* to find its closest discrete solution. Each step reduces the same objective ϕ through coordinate descent. We can only guarantee such iterations end in a local optimum, which may vary with the initial estimation. However, since $\tilde{X}^*R^*$ are all global optima regardless of R^*, whichever $\tilde{X}^*R^*$ the program *POD* converges to, its proximal discrete solution X^* will not be too much off from the optimality.

Proceedings of the Ninth IEEE International Conference on Computer Vision (ICCV'03)
0-7695-1950-4/03 $ 17.00 © 2003 IEEE

2.3 Algorithm

Given weight matrix W and number of classes K:

1. Compute the degree matrix $D = \mathrm{Diag}(W1_N)$.
2. Find the optimal eigensolution Z^* by:

   $D^{-\frac{1}{2}}WD^{-\frac{1}{2}}\ \bar{V}_{[K]} = \bar{V}_{[K]}\,\mathrm{Diag}(s_{[K]}),\ \bar{V}_{[K]}^T\bar{V}_{[K]} = I_K$
   $Z^* = D^{-\frac{1}{2}}\bar{V}_{[K]}$.

3. Normalize Z^* by: $\tilde{X}^* = \mathrm{Diag}(\mathrm{diag}^{-\frac{1}{2}}(Z^*Z^{*T}))Z^*$.
4. Initialize X^* by computing R^* as:

   $R_1^* = [\tilde{X}^*(i,1), \ldots \tilde{X}^*(i,K)]^T$, random $i \in [N]$
   $c = 0_{N\times 1}$
   For $k = 2, \ldots, K$, do:
   $c = c + \mathrm{abs}(\tilde{X}^* R_{k-1}^*)$
   $R_k^* = [\tilde{X}^*(i,1), \ldots \tilde{X}^*(i,K)]^T$, $i = \arg\min c$

5. Initialize convergence monitoring parameter $\bar{\phi}^* = 0$.
6. Find the optimal discrete solution X^* by:

   $\tilde{X} = \tilde{X}^* R^*$
   $X^*(i,l) = \langle l = \arg\max_{k\in[K]} \tilde{X}(i,k)\rangle$, $i \in \mathbb{V}$, $l \in [K]$.

7. Find the optimal orthonormal matrix R^* by:

   $X^{*T}\tilde{X}^* = U\Omega\tilde{U}^T, \quad \Omega = \mathrm{Diag}(\omega)$
   $\bar{\phi} = \mathrm{tr}(\Omega)$
   If $|\bar{\phi} - \bar{\phi}^*| <$ machine precision, then stop and output X^*
   $\bar{\phi}^* = \bar{\phi}$
   $R^* = \tilde{U}U^T$

8. Go to Step 6.

In Step 2, we use $\bar{V}_{[K]}$ as a shorthand for $[\bar{V}_1, \ldots, \bar{V}_K]$, and likewise for $\bar{S}_{[K]}$. In Step 4, $B = \mathrm{abs}(A)$ denotes the absolute values of the elements of A. In Step 3, since $\tilde{X}^* = \mathrm{Diag}(\mathrm{diag}^{-\frac{1}{2}}(Z^*Z^{*T}))Z^*$ scales the lengths of each row to 1, we can skip scaling $\bar{V}$ in order to get V, i.e. $Z^* = [\bar{V}_1, \ldots, \bar{V}_K]$ leads to the same $\tilde{X}^*$.

Step 2 solves the first K leading eigenvectors of an $N \times N$ usually sparse matrix. It is nevertheless the most time consuming. With sampling, our implementation for image segmentation has a time complexity of $O(N^{\frac{3}{2}}K)$. Step 4 has $NK(K-1)$ multiplications in choosing K centers. Step 6 involves NK^2 multiplications to compute $\tilde{X}^*R^*$. Step 7 involves an SVD of a $K \times K$ matrix and K^3 multiplications for computing R^*. Since X^* is binary, $X^{*T}\tilde{X}^*$ can be done efficiently with all additions. Taken together, the time complexity of the algorithm is $O(N^{\frac{3}{2}}K + NK^2)$.

3 Experiments

We first illustrate our method on point set clustering based on proximity. The affinity between two points is defined to be the Gaussian function of their distance. We derive a bipartition using the procedure shown in Fig. 2.

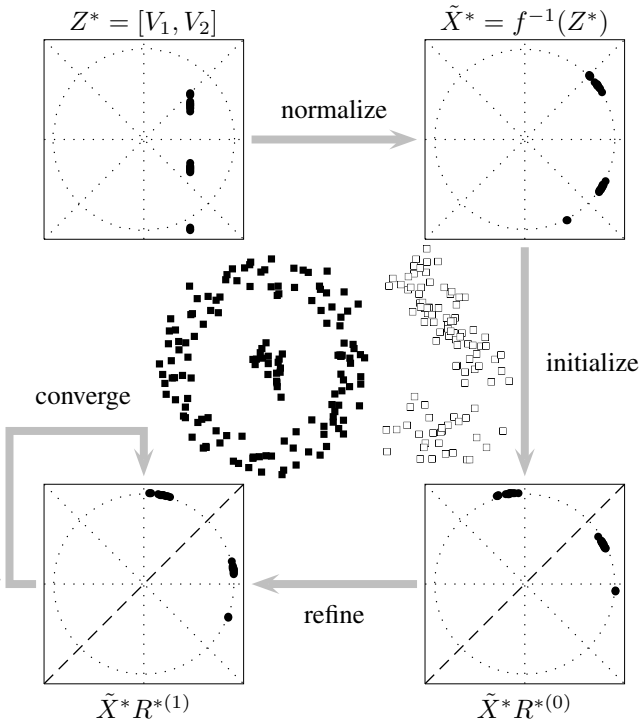

Figure 2. Progression of our algorithm. Each plot shows an $N \times 2$ matrix, with each row taken as (x, y) coordinates of a point in the plane. Though $N = 245$, many of them are mapped to the same planar points, resulting in three visible clusters. 1)Normalize: starting with the eigenvectors Z^*, we first map it back to the X-space by normalizing their lengths so that all of them lie on the unit circle. 2)Initialize: two points with almost orthogonal phases are selected to form $R^{*(0)}$. $\tilde{X}^*R^{*(0)}$ is the projection of all the points to the two chosen directions. An initial clustering $X^{*(0)}$ is obtained by non-maximum suppression: points are divided according to the dashed line $x = y$: points below the line assigned to $(1, 0)$ hence $\mathbb{V}_1$, those above the line assigned to $(0, 1)$ hence $\mathbb{V}_2$. 3)Refine: we find the closest continuous optimal to $X^{*(0)}$ by adjusting the rotation matrix $R^{*(1)}$. Non-maximum suppression produces its closest discrete solution $X^{*(1)}$, which is exactly the same as $X^{*(0)}$. The algorithm converges and stops. The final clustering is shown in the center, with $\varepsilon(X^*) = 0.9997 < \varepsilon(\tilde{X}^*) = 0.9998$.

Images are first convolved with oriented filter pairs to extract the magnitude of edge responses OE [6]. Pixel affinity W is inversely correlated with the maximum magnitude of edges crossing the line connecting two pixels. $W(i, j)$ is

Proceedings of the Ninth IEEE International Conference on Computer Vision (ICCV'03)
0-7695-1950-4/03 $ 17.00 © 2003 IEEE

low if i, j are on the two sides of a strong edge. This measure is meaningful only for nearby pixels. We hence set $W(i,j) = 0$ beyond a city-block distance.

Fig. 3 shows leading eigenvectors for an image. Refining partitions with increasing K can be achieved through a sequential initialization: we use X^* of $\Gamma_{\mathbb{V}}^K$ as a starting segmentation for $\Gamma_{\mathbb{V}}^{K+1}$, with its largest region broken into 2 groups. This produces a pseudo-hierarchical segmentation in Fig. 3: when K increases, regions tend to be successively divided (e.g. $K4$, $K5$), yet the enclosing boundaries are subject to fine adjustment (e.g. $K5$, $K6$).

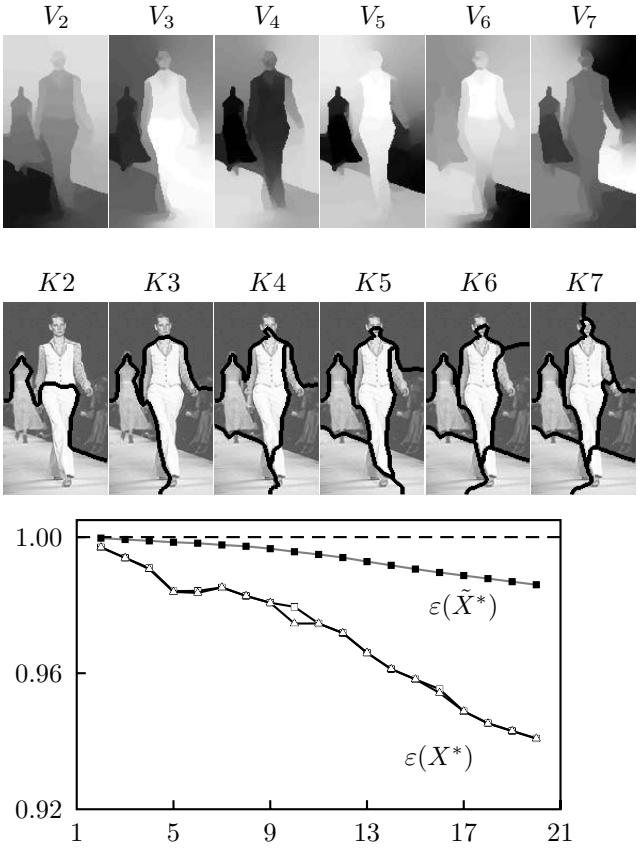

Figure 3. Image segmentation (row 2) based on eigenvectors (row 1). Image size: 120×97, i.e. $N = 11640$. It takes 36 seconds to compute 20 leading eigenvectors in MATLAB on a PC with 1GHz CPU and 1GB memory. The discretization process takes 0.1 up to 1.1 seconds. The above plot is ε over K.

Fig. 3 also shows the discrepancy in the objective value between continuous and discrete optima. The upper bound $\varepsilon(\tilde{X}^*)$ monotonically decreases with larger K, while $\varepsilon(X^*)$ gradually decreases by and large, but not monotonically. Examining these values with the corresponding segmentations, we find that ε itself is not very indicative for selecting the best K.

Fig. 4 shows examples of discrete optima we get after running our method with many different initializations. There are usually only a few discrete optima and they have comparable objective values.

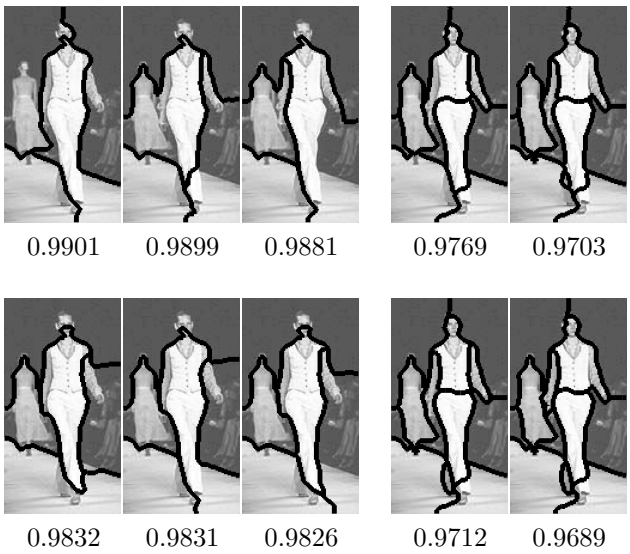

Figure 4. Multiple near-global discrete optima. Row 1: $K = 4, 10$. Row: $K = 5, 11$. Numbers below the images are the ε values for the corresponding segmentations.

We run our algorithm on $450+$ real images with the same set of parameters. Fig. 5 is a sample of our results on a set of fashion pictures and Berkeley test dataset. How to choose K remains an open problem.

4 Relations to other works

Fig. 6 shows that K-means on $\tilde{X}^*$ [9] can produce similar results but it may take twice as long to converge. In [9], a perturbation rationale is given for the need to normalize the eigenvectors, while the use of K-means is unjustified. K-means' similar results are a consequence of the continuous optima greatly reducing the chance for mis-clustering. Yet we observe that a good initial estimation is crucial for K-means, whereas our method is robust to a random initialization. This is not surprising because K-means introduces additional unwarranted assumptions, while our principled account has a clear criterion ϕ to optimize, which guarantees the near global optimality of discrete solutions under the orthonormal invariance of continuous optima.

Finally, since various average cuts are variations of the normalized cuts criterion [10], our principled account from obtaining a relaxed solution to final discretization can be extended trivially to those methods.

Proceedings of the Ninth IEEE International Conference on Computer Vision (ICCV'03)
0-7695-1950-4/03 $ 17.00 © 2003 IEEE

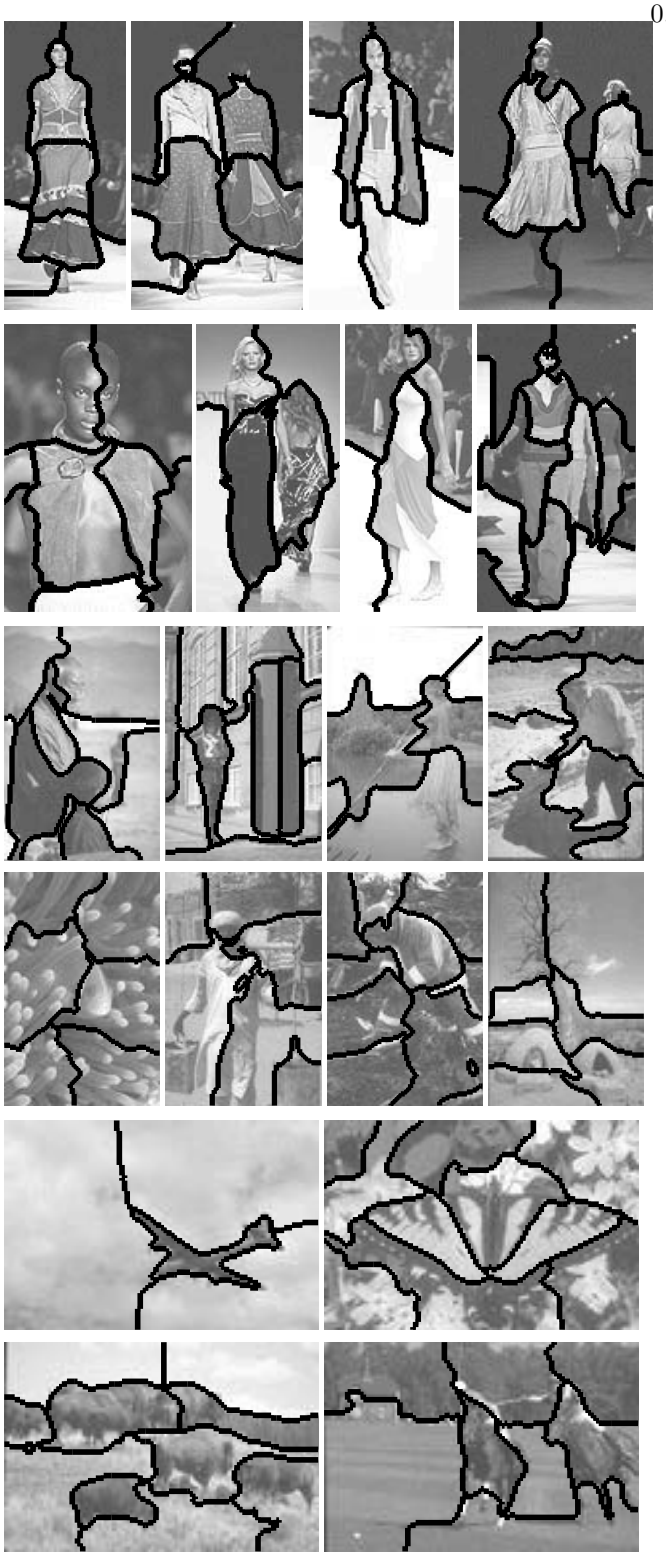

Figure 5. Multiclass segmentation on New York Spring 2002 fashion pictures (fashionshowroom.com) and Berkeley test images [7] (cs.berkeley.edu). The number of classes K is manually chosen.

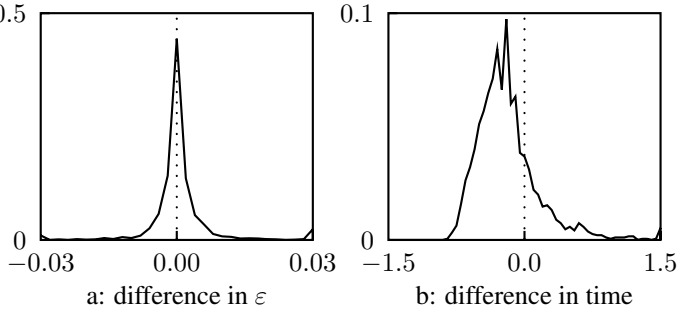

Figure 6. Performance comparison to K-means clustering on $\tilde{X}^*$. Both are estimated probability distribution of the relative difference between the two methods: $\frac{g-g_{Kmeans}}{g}$, where g is ε in a and running time in b. These statistics are collected over 100 Berkeley test images. Each image is segmented into 2 to 20 classes. Both codes are optimized to take advantage of the unit lengths of all data points, with the same initialization method.

Acknowledgments

This research is funded by (DARPA HumanID) ONR N00014-00-1-0915 and NSF IRI-9817496.

References

[1] C. J. Alpert and A. B. Kahng. Multiway partitioning via geometric embeddings, orderings and dynamic programming. *IEEE Transactions on Computer-aided Design of Integrated Circuits and Systems*, 14(11):1342–58, 1995.

[2] E. R. Barnes. An algorithm for partitioning the nodes of a graph. *SIAM J. ALG. DISC. METH.*, 3(4):541–50, 82.

[3] P. K. Chan, M. D. F. Schlag, and J. Y. Zien. Spectral k-way ratio-cut partitioning and clustering. *IEEE Transactions on Computer-aided Design of Integrated Circuits and Systems*, 13(9):1088–96, 1994.

[4] B. Hendrickson and R. Leland. An improved spectral graph partitioning algorithm for mapping parallel computations. *SIAM J. Sci. Comput.*, 16(2):452–459, 1995.

[5] H. Ishikawa and D. Geiger. Segmentation by grouping junctions. In *CVPR*, 1998.

[6] J. Malik, S. Belongie, T. Leung, and J. Shi. Contour and texture analysis for image segmentation. *IJCV*, 2001.

[7] 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 *ICCV*, 2001.

[8] M. Meila and J. Shi. Learning segmentation with random walk. In *NIPS*, 2001.

[9] A. Y. Ng, M. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *NIPS*, 2002.

[10] J. Shi and J. Malik. Normalized cuts and image segmentation. *PAMI*, 22(8):888–905, 1997.

Proceedings of the Ninth IEEE International Conference on Computer Vision (ICCV'03)
0-7695-1950-4/03 $ 17.00 © 2003 IEEE