Segmentation Given Partial Grouping Constraints

Stella X. Yu, *Member*, *IEEE Computer Society*, and Jianbo Shi, *Member*, *IEEE Computer Society*

Abstract—We consider data clustering problems where partial grouping is known a priori. We formulate such biased grouping problems as a constrained optimization problem, where structural properties of the data define the goodness of a grouping and partial grouping cues define the feasibility of a grouping. We enforce grouping smoothness and fairness on labeled data points so that sparse partial grouping information can be effectively propagated to the unlabeled data. Considering the normalized cuts criterion in particular, our formulation leads to a constrained eigenvalue problem. By generalizing the Rayleigh-Ritz theorem to projected matrices, we find the global optimum in the relaxed continuous domain by eigendecomposition, from which a near-global optimum to the discrete labeling problem can be obtained effectively. We apply our method to real image segmentation problems, where partial grouping priors can often be derived based on a crude spatial attentional map that binds places with common salient features or focuses on expected object locations. We demonstrate not only that it is possible to integrate both image structures and priors in a single grouping process, but also that objects can be segregated from the background without specific object knowledge.

Index Terms—Grouping, image segmentation, graph partitioning, bias, spatial attention, semisupervised clustering, partially labeled classification.

1 Introduction

A good image segmentation respects not only the structural properties of the image [1] but also the needs of later visual processing such as object recognition [2]. In this paper, we will develop a method that integrates both data-driven and task-driven knowledge for making a global decision on segmentation.

The approach where task-driven knowledge is used to constrain the segmentation at the very beginning contrasts with the sequential processing theory popularized by Marr [3]. According to his theory, visual processing starts with what can be computed directly from an image and ends with the information required to support goals such as navigation or object recognition. Intermediate representations are derived to turn the available information at one level to the required information at the succeeding level. Accordingly, most current image segmentation algorithms adopt a bottom-up approach. They start with an oversegmentation based on low-level cues such as feature similarity and boundary continuity and then build up larger perceptual units (e.g., surface, foreground, and background) by adding high-level knowledge such as statistical properties of regions into the grouping process [4].

Although a sequential system can relieve computational burden from later stages of perceptual processing, such a feed-forward system is vulnerable to mistakes made at each step: the low-level processing alone often produces an unreliable representation, e.g., missing object boundaries of weak contrast caused by lighting and background clutter, which may not be remediable by the later high-level processing.

We demonstrate that it is possible to integrate both bottom-up and top-down information in a *single* grouping process.

We consider the type of task-driven knowledge presented as *partial grouping* information. For example, in Fig. 1, based on intensity distribution and viewers' expectation, for the image with the tiger, a set of bright pixels are likely to be foreground and a set of dark pixels are likely to be background; for the image with the fashion-model, pixels near image boundaries are probably background. Such information provides *bias* to a natural grouping process that is based solely on data themselves.

Our work is concerned with the following issue: What is a simple and principled approach for incorporating these often *sparse* partial grouping cues directly into low-level image segmentation?

A straightforward approach that we adopt in this work is to formulate the problem as a constrained optimization problem, where the goodness of a segmentation is based on low-level data coherence and the feasibility of a segmentation is based on partial grouping constraints. For the normalized cuts criterion under the spectral graph-theoretic framework [5], we show that this straightforward formulation leads to a constrained eigenvalue problem. By generalizing the standard Rayleigh-Ritz theorem, we can compute a near-global optimum efficiently.

We then show through a simple point set example that segmentation performance breaks down especially when partial grouping cues are sparse. This observation leads to a new formulation with smoothed constraints. In the spectral graph framework, the smoothing operator is readily derived from the existing pairwise relationships between grouping elements. We present numerous image segmentation examples to demonstrate the efficacy of the new formulation.

- *S.X. Yu is with the Department of Computer Science, University of California at Berkeley, 549 Soda Hall, Berkeley, CA 94720-1776. E-mail: stellayu@cs.berkeley.edu.*
- *J. Shi is with the Department of Computer and Information Science, University of Pennsylvania. GRASP Laboratory, Levine Hall, 3330 Walnut Street, Philadelphia, PA 19104-6389. E-mail: jshi@cis.upenn.edu.*

Manuscript received 28 Apr. 2002; revised 18 Mar. 2003; accepted 16 June 2003. Recommended for acceptance by M.A.T. Figueiredo, E.R. Hancock, M. Pelillo, and J. Zerubia.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number 118730.

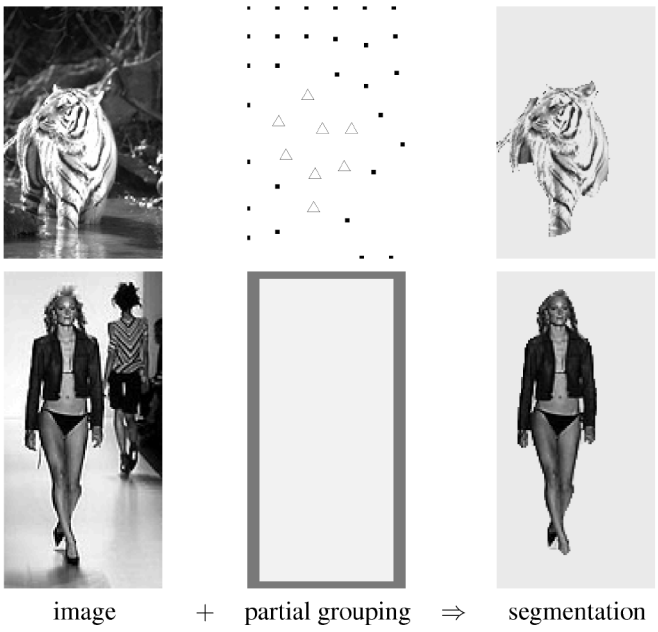

Fig. 1. Segmentation given partial grouping constraints. We desire an algorithm that outputs an object segmentation through integrating partial grouping cues with the data coherence itself. In the middle column, white pixels are unlabeled, whereas marked or gray pixels are a priori known to be in the same group. These cues are derived from feature-driven or location-driven attentional maps. That is, the regions of interest here are defined based on pixel intensities or prior expectation of object locations.

Finally, we summarize the paper after a discussion on its connections to related data clustering methods.

2 Basic Formulation

Given an image of N pixels, the goal of segmentation is to assign one of K prescribed labels to each pixel. Let $\mathbf{V} = [N]$ denote the set of all pixels, where $[n]$ denotes the set of integers between 1 and n: $[n] = \{1, 2, \ldots, n\}$. To segment an image is to decompose $\mathbf{V}$ into K disjoint sets, i.e., $\mathbf{V} = \cup_{l=1}^{K} \mathbf{V}_l$ and $\mathbf{V}_k \cap \mathbf{V}_l = \emptyset$, $k \neq l$. We denote this K-way partitioning by $\Gamma_{\mathbf{V}}^{K} = \{\mathbf{V}_1, \ldots, \mathbf{V}_K\}$.

Let $\varepsilon(\Gamma_{\mathbf{V}}^{K}; f)$ be an objective function that measures the goodness of grouping for some image data f, e.g., $f(i)$ is the intensity value at pixel i, $i \in \mathbf{V}$. In Markov random field (MRF) approaches for image segmentation [6], the objective function is the posterior probability of the segmentation $\Gamma_{\mathbf{V}}^{K}$ given the observation f:

$$\varepsilon_{MRF}(\Gamma_{\mathbf{V}}^{K}; f) = \Pr(\Gamma_{\mathbf{V}}^{K}|f) \propto \Pr(f|\Gamma_{\mathbf{V}}^{K}) \cdot \Pr(\Gamma_{\mathbf{V}}^{K}). \quad (1)$$

The first term $\Pr(f|\Gamma_{\mathbf{V}}^{K})$ describes data fidelity, which measures how well a generative model explains the observed image data and the second term $\Pr(\Gamma_{\mathbf{V}}^{K})$ describes model complexity, which favors the segmentation to have some regularity such as piecewise constancy. In discriminative approaches for segmentation [5], the objective function is some clustering measure which increases with within-group feature similarity and decreases with between-group feature similarity.

Consider partial grouping information represented by n pixel sets: $\mathbb{U}_t$, $t \in [n]$, each containing pixels known to belong together. The labels on these pixels are not known, and they are not required to be different across the n groups. For a unique representation of $\mathbb{U}_t$s, we assume there is no common pixel between any two sets: $\mathbb{U}_s \cap \mathbb{U}_t = \emptyset$, $s \neq t$. In other words, if there is a common pixel, then the two sets should be merged into one.

The most straightforward way to incorporate the partial grouping information is to encode it as constraints. With a little abuse of notation, we use $\Gamma_{\mathbf{V}}^{K}(i, l)$ to denote a Boolean function that returns 1 if $i \in \mathbf{V}_l$. Among the segmentations partially determined by $\mathbb{U}_t$s, we seek one that optimizes the goodness of grouping measured by ε:

$$\text{maximize } \varepsilon(\Gamma_{\mathbf{V}}^{K}; f) \quad (2)$$

$$\text{subject to } \Gamma_{\mathbf{V}}^{K}(i, l) = \Gamma_{\mathbf{V}}^{K}(j, l), i, j \in \mathbb{U}_t, l \in [K], t \in [n]. \quad (3)$$

Since partial grouping cues are encoded as hard constraints, they have to be reliable enough to be enforced. Fig. 1 illustrates two basic scenarios where we can derive such cues. The first type is feature-driven, where pixels conforming to a particular generative model are biased together. For example, we probably perceive a white object against a dark background before we realize that it is a tiger in a river. In this case, $\mathbb{U}_1$ contains pixels of the brightest intensities and $\mathbb{U}_2$ the darkest. The second type is solely location-driven, it reflects our expectation as to where an object is going to appear. For example, pictures taken in a fashion show often have fashion models at the center. To segment out the fashion models, we consider pixels at image boundaries as the background group $\mathbb{U}_1$. Such seemingly insignificant information provides long-range binding cues that are often lacking in low-level grouping.

For some particular forms of ε, such as the above mentioned probability criteria using generative models and the minimum cuts criteria in discriminative approaches [7], [8], [9], the constraints in (3) can be trivially incorporated in an algorithm that optimizes the objective. For the former, Markov Chain Monte Carlo (MCMC) is a general solution technique and the constraints can be realized by generating legitimate samples [10]. For the latter, assuming that $\mathbb{U}_1$ and $\mathbb{U}_2$ take distinct labels, we can solve (3) using maximum-flow algorithms, in which two special nodes called source and sink are introduced, with infinite weights between the source and $\mathbb{U}_1$, and between the sink and $\mathbb{U}_2$ [7]. For others such as the normalized cuts criterion [5], it is not clear whether the solution can be obtained using the same technique that was used for the unconstrained problem. We will explore this criterion further.

3 Constrained Normalized Cuts Criterion

A weighted graph is specified by $\mathbb{G} = (\mathbf{V}, \mathbb{E}, W)$, where $\mathbf{V}$ is the set of all nodes, $\mathbb{E}$ is the set of edges connecting nodes, and W is an affinity matrix, with weights characterizing the likelihood that two nodes belong to the same group. We assume that W is nonnegative and symmetric.

In graph-theoretic methods for image segmentation, an image is first transcribed into a weighted graph, where each node represents a pixel and weights on edges connecting two nodes describe the pairwise feature similarity between the pixels. Segmentation then becomes a node partitioning problem. A good segmentation desires a partitioning that

has tight connections within partitions and loose connections across partitions. These two goals can both be achieved in the normalized cuts criterion [5], a brief self-contained account of which is given below.

3.1 Representation

Given weight matrix W, the multiclass normalized cuts criterion tries to maximize the average of all K linkratios [11]:

$$\varepsilon_{NC}(\Gamma_{\mathbf{V}}^{K}) = \frac{1}{K}\sum_{l=1}^{K}\mathrm{linkratio}(\mathbf{V}_l, \mathbf{V}) \tag{4}$$

$$\mathrm{linkratio}(\mathbf{V}_l, \mathbf{V}) = \frac{\sum_{i\in\mathbf{V}_l, j\in\mathbf{V}_l} W(i,j)}{\sum_{i\in\mathbf{V}_l, j\in\mathbf{V}} W(i,j)}. \tag{5}$$

$\mathrm{linkratio}(\mathbf{V}_l, \mathbf{V}_l)$ is the fraction of the total weights within a group to the total weights all the member nodes have. Its complement $\mathrm{linkratio}(\mathbf{V}_l, \mathbf{V}\setminus\mathbf{V}_l)$ is the fraction of the weights between nodes in one group and the rest nodes in the graph. Since these two quantities sum up to one, maximizing the within-group linkratio is equivalent to minimizing the between-group linkratio. Therefore, this criterion favors both tight connections within partitions and loose connections between partitions.

We use an $N \times K$ *partition matrix* X to represent $\Gamma_{\mathbf{V}}^{K}$, where $X = [X_1, \ldots, X_K]$ and $X(i,l) = 1$ if $i \in \mathbf{V}_l$ and 0 otherwise. X_l is a binary indicator for partition $\mathbf{V}_l$. Since a node is only assigned to one partition, there is an exclusion constraint on X: $X\,1_K = 1_N$, where 1_d denotes the $d \times 1$ vector of all 1s.

For $t \in [n]$, partial grouping node set $\mathbb{U}_t$ produces $|\mathbb{U}_t| - 1$ independent constraints, where $|\cdot|$ denotes the size of a set. Each constraint can be represented by an $N \times 1$ vector U_k with only two nonzero elements: $U_k(i) = 1$, $U_k(j) = -1$, $i, j \in \mathbb{U}_t$ for instance. Let $U = [U_1, \ldots, U_{\bar{n}}]$, where $\bar{n} = \sum_{t=1}^{n}(|\mathbb{U}_t| - 1)$. Then, the partial grouping constraints in (3) become: $U^T X = 0$. We assume that U obtained as such has full rank.

Finally, we introduce the degree matrix D, defined to be the total connections each node has: $D = \mathrm{Diag}(W1_N)$, where Diag denotes a diagonal matrix formed from its vector argument. We assume the degree of each node is nonzero, so that D is invertible.

With these symbols and notation, we write the constrained grouping problem in (3) for the normalized cuts criterion as program *PNCX*:

$$\text{maximize} \quad \varepsilon_{NC}(X) = \frac{1}{K}\sum_{l=1}^{K}\frac{X_l^T W X_l}{X_l^T D X_l} \tag{6}$$

$$\text{subject to} \quad X \in \{0,1\}^{N\times K}, \quad X\,1_K = 1_N \tag{7}$$

$$U^T X = 0. \tag{8}$$

3.2 Computational Solution

We introduce a *scaled partition matrix* Z to make (6) more manageable:

$$Z = X(X^T D X)^{-\frac{1}{2}}. \tag{9}$$

Then, $\varepsilon_{NC}(X) = \frac{1}{K}\mathrm{tr}(Z^T W Z)$, where tr denotes the trace of a matrix. Given the definition in (9), Z naturally satisfies $Z^T D Z = I$, where I is an identity matrix. The grouping constraint in (8) is equivalent to:

$$U^T Z = U^T X(X^T D X)^{-\frac{1}{2}} = 0. \tag{10}$$

Ignoring (7) for the time being, we relax *PNCX* into program *PNCZ*:

$$\text{maximize} \quad \varepsilon_{NC}(Z) = \frac{1}{K}\mathrm{tr}(Z^T W Z) \tag{11}$$

$$\text{subject to} \quad Z^T D Z = I \tag{12}$$

$$U^T Z = 0. \tag{13}$$

PNCZ is a constrained eigenvalue problem [12] in the continuous domain and it can be solved by linear algebra.

In principle, we can solve *PNCZ* by applying the standard Rayleigh-Ritz theorem to its unconstrained version. That is, we first find a basis in the feasible solution space defined by $U^T Z = 0$. Let $U^\perp$ denote an orthonormal basis in this space. Any solution that satisfies the partial grouping constraints can be represented by an $(N - \bar{n}) \times K$ coefficient matrix Y using this basis:

$$Z = U^\perp Y, \qquad U^T U^\perp = 0. \tag{14}$$

We thus reduce *PNCZ* to a program in Y:

$$\text{maximize} \quad \varepsilon_{NC}(Y) = \frac{1}{K}\mathrm{tr}(Y^T W^y Y) \tag{15}$$

$$\text{subject to} \quad Y^T D^y Y = I, \tag{16}$$

where $W^y = (U^\perp)^T W U^\perp$ and $D^y = (U^\perp)^T D U^\perp$ are the equivalent weight and degree matrices for Y. This is a standard Rayleigh quotient optimization problem. If (V^y, S^y) is the eigendecomposition of the matrix pair (W^y, D^y), where $S^y = \mathrm{Diag}(s^y)$ with nonincreasingly ordered eigenvalues in s^y, then the global optimum is given by the eigenvectors corresponding to the first K largest eigenvalues and

$$\varepsilon_{NC}([V_1^y, \ldots, V_K^y]) = \frac{1}{K}\sum_{l=1}^{K} s_l^y = \max_{Y^T D^y Y = I} \varepsilon_{NC}(Y). \tag{17}$$

From (14), we recover the global optimum in the original Z-space as $Z^* = U^\perp[V_1^y, \ldots, V_K^y]$.

The introduction of Y gets rid of the constraint in (13) and turns program *PNCZ* into an unconstrained eigenvalue problem. However, it requires finding an orthonormal basis for the feasible space first. Given that $\bar{n} \ll N$, this process has a space and time complexity of $O(N^2)$ and $O(N^3)$, respectively, which is prohibitively expensive for a large N. We have to find another way out.

There is such an alternative through the use of matrix projectors. Q is called a *projector* if it is *idempotent*, i.e., $Q^2 = Q$. If Q is a projector onto the space of feasible solutions of *PNCZ*, then QZ is the projection of Z on the feasible space. The key property of QZ is that $QZ = Z$ if and only if Z is feasible. Therefore, we can guarantee the feasibility of a solution by projecting it to the feasible set in the original space without resorting to any reparameterization in a reduced space.

We introduce a few symbols to simplify notation. Let π be a vector of K distinct integers from $[N]$. For any eigenvector matrix V and its corresponding eigenvalue matrix $S = \mathrm{Diag}(s)$, let $V_\pi = [V_{\pi_1}, \ldots, V_{\pi_K}]$ and $S_\pi = \mathrm{Diag}([s_{\pi_1}, \ldots, s_{\pi_K}])$.

Theorem 1 (Generalized Rayleigh-Ritz Theorem). *Let (V, S) be the following eigendecomposition of matrix QPQ:*

$$QPQV = VS \tag{18}$$
$$V^T DV = I, \tag{19}$$

where P is the row-normalized weight matrix and Q is a projector onto the feasible solution space:

$$P = D^{-1}W \tag{20}$$
$$Q = I - D^{-1}U(U^T D^{-1}U)^{-1}U^T. \tag{21}$$

For any local optimum candidate Z^* to program PNCZ, there exists an index vector π and an orthonormal matrix R such that:

$$Z^* = V_\pi R, \quad R^T R = I \tag{22}$$
$$\varepsilon_{NC}(Z^*) = \frac{1}{K}\mathrm{tr}(S_\pi). \tag{23}$$

Assuming that the eigenvectors are ordered according to their eigenvalues, where $s_1 \geq \ldots \geq s_N$, any global optimum of PNCZ can thus be specified by the first K largest eigenvectors and any orthonormal matrix:

$$Z^* = V_{[K]}, \quad R^T R = I \tag{24}$$
$$\varepsilon_{NC}(Z^*) = \frac{1}{K}\mathrm{tr}(S_{[K]}) = \max_{\substack{Z^T DZ = I \\ U^T Z = 0}} \varepsilon_{NC}(Z). \tag{25}$$

Proof. We define a *Lagrangian* for *PNCZ*:

$$L(Z, \Lambda, \Theta) = \frac{1}{2}\mathrm{tr}(Z^T WZ) - \frac{1}{2}\mathrm{tr}(\Lambda^T(Z^T DZ - I)) - \Theta^T U^T Z,$$

where Λ is a $K \times K$ symmetric matrix and Θ is an $\bar{n} \times K$ matrix. An optimal solution $(Z^*, \Lambda^*, \Theta^*)$ must satisfy:

$$L_Z(Z, \Lambda, \Theta) = WZ - DZ\Lambda - U\Theta = 0, \tag{26}$$
$$L_\Lambda(Z, \Lambda, \Theta) = Z^T DZ - I = 0, \tag{27}$$
$$L_\Theta(Z, \Lambda, \Theta) = U^T Z = 0. \tag{28}$$

Multiplying (26) with $U^T D^{-1}$ leads to:

$$\Theta^* = (U^T D^{-1}U)^{-1}U^T D^{-1}WZ^*, \tag{29}$$

where D and $U^T D^{-1}U$ are invertible since both D and U assume full rank. Eliminating Θ in (26) by (29), we obtain

$$QPZ^* = Z^*\Lambda^*. \tag{30}$$

From (28), we also have $QZ^* = Z^*$. Substituting it into the above equation, we obtain $QPQZ^* = Z^*\Lambda^*$. Therefore, there are three necessary conditions for the optimality: Λ^* is symmetric and

$$QPQZ^* = Z^*\Lambda^*, \quad Z^{*T} DZ^* = I. \tag{31}$$

Next, we show that there exists an eigendecomposition (V, S) of QPQ that not only meets these conditions but can also generate all such solutions through orthonormal matrices.

Noting that $QPQZ^* = Z^*\Lambda^*$ is equivalent to:

$$D^{\frac{1}{2}}QD^{-\frac{1}{2}} \cdot D^{\frac{1}{2}}PD^{-\frac{1}{2}} \cdot D^{\frac{1}{2}}QD^{-\frac{1}{2}} \cdot D^{\frac{1}{2}}Z^* = D^{\frac{1}{2}}Z^*\Lambda^*, \tag{32}$$

we rewrite (31) using a transformed variable $\bar{Z}$:

$$\bar{Q}\bar{P}\bar{Q}\bar{Z} = \bar{Z}\Lambda^*, \quad \bar{Z}^T\bar{Z} = I, \tag{33}$$
$$\bar{Z} = D^{\frac{1}{2}}Z^* \tag{34}$$
$$\bar{P} = D^{\frac{1}{2}}PD^{-\frac{1}{2}} = D^{-\frac{1}{2}}WD^{-\frac{1}{2}} \tag{35}$$
$$\bar{Q} = D^{\frac{1}{2}}QD^{-\frac{1}{2}} = I - D^{-\frac{1}{2}}U(U^T D^{-1}U)^{-1}U^T D^{-\frac{1}{2}}. \tag{36}$$

Since both $\bar{P}$ and $\bar{Q}$ are symmetric, $\bar{Q}\bar{P}\bar{Q}$ is symmetric, which means that all its eigenvectors are real and orthogonal. Therefore, if $(\bar{V}, S)$ is an orthonormal eigendecomposition of $\bar{Q}\bar{P}\bar{Q}$, then any K distinct eigenvectors and their eigenvalues, i.e., $(\bar{V}_\pi, S_\pi)$, form a solution to (33).

If $(\bar{Z}, \Lambda^*)$ is a solution that satisfies (33) with $\bar{Z}$ orthonormal and Λ^* symmetric, since $\bar{V}$ is a complete basis in the N-dimensional space, there exists an index vector π and an orthonormal matrix R such that

$$\bar{Z} = \bar{V}_\pi R, \quad R^T R = I \tag{37}$$
$$\Lambda^* = R^T S_\pi R. \tag{38}$$

Multiplying (26) with Z^{*T} and using $\mathrm{tr}(AB) = \mathrm{tr}(BA)$, we derive:

$$K\ \varepsilon_{NC}(Z^*) = \mathrm{tr}(Z^{*T}WZ^*) = \mathrm{tr}(\Lambda^*) = \mathrm{tr}(S_\pi). \tag{39}$$

Therefore, $\{(\bar{V}_\pi, S_\pi) : \pi\}$ produce all possible local optimal values. The global optimal value is thus given by the average of the first K largest eigenvalues. Transforming $\bar{Z}$ back to the Z space based on (34), we have $V = D^{-\frac{1}{2}}\bar{V}$ and (V, S) as an eigendecomposition of QPQ. This completes the proof. □

When there is no constraint, $Q = I$, then $QPQ = P$ can be considered as a transition probability matrix of random walks, and the normalized cuts criterion is equivalent to a maximum conductance problem where subsets of states only occasionally visit each other [13]. When there are constraints, $Q \neq I$, QPQ usually has negative entries and it no longer has a transition probability interpretation. In other words, the solution to constrained grouping can no longer be cast as the equilibrium of a natural diffusion process.

To summarize, the optimal solution to *PNCZ* is not unique. It is a subspace spanned by the first K largest eigenvectors of QPQ by orthonormal matrices:

$$Z^* \in \{V_{[K]}R : QPQV_{[K]} = V_{[K]}S_{[K]}, R^T R = I\}. \tag{40}$$

Unless all K eigenvalues are the same, $V_{[K]}R$ are no longer the eigenvectors of QPQ. Yet, all these solutions have the optimal objective value.

After we compute $(V_{[K]}, S_{[K]})$ from QPQ, the same procedure for the unconstrained normalized cuts can be applied to find a near global-optimal discrete solution to *PNCX*. The only difference is that now the eigenvectors are from QPQ rather than P. In the discretization procedure, we honor the constraints that were ignored when we relaxed the program *PNCX* into the program *PNCZ*. That is, we find a discrete solution that satisfies the binary and exclusion constraints in (7), yet is closest to the continuous optima given in (40). This is another optimization problem which can be solved efficiently, since the objective function is bilinear in the discrete solution X and the orthonormal transform R. The details can be found in [11].

3.3 Algorithm

To summarize, given data f defined over $\mathbb{V}$, and n partial grouping node sets $\{\mathbb{U}_t : t \in [n]\}$, we use the following constrained normalized cuts algorithm to find an optimal K-way grouping.

1. Compute the affinity matrix W from the data f, e.g.:

$$W(i,j) = e^{-\left(\frac{f(i)-f(j)}{\sqrt{2}\sigma}\right)^2}, i,j \in \mathbb{V}.$$

2. Derive the constraint matrix U from $\{\mathbb{U}_t : t \in [n]\}$:

$$\begin{aligned} &k = 0 \\ &\text{For } t = 1 : n, \\ &\quad \text{For } s = 1 : |\mathbb{U}_t| - 1, \\ &\quad\quad k = k + 1 \\ &\quad\quad U(\mathbb{U}_t(s), k) = 1 \\ &\quad\quad U(\mathbb{U}_t(s+1), k) = 1 \end{aligned}$$

3. Compute the degree matrix $D = \mathrm{Diag}(W1_N)$.
4. Compute $\bar{P}$, $\bar{U}$, and H as:

$$\begin{aligned} \bar{P} &= D^{-\frac{1}{2}} W D^{-\frac{1}{2}} \\ \bar{U} &= D^{-\frac{1}{2}} U \\ H &= (\bar{U}^T \bar{U})^{-1}. \end{aligned}$$

5. Compute the first K eigenvectors of $\bar{Q}\bar{P}\bar{Q}$ by solving:

$$\begin{aligned} &(I - \bar{U}H\bar{U}^T)\bar{P}(I - \bar{U}H\bar{U}^T)\bar{V}_{[K]} = \bar{V}_{[K]} S_{[K]} \\ &\bar{V}_{[K]}^T \bar{V}_{[K]} = I. \end{aligned}$$

6. Compute the first K eigenvectors of QPQ by

$$V_{[K]} = D^{-\frac{1}{2}} \bar{V}_{[K]}.$$

7. Obtain a discrete segmentation X^* closest to $V_{[K]}$ [11].

In Step 5, we avoid directly computing $\bar{Q}\bar{P}\bar{Q}$ since it can become a dense matrix even when U and P are sparse. Specifically, we modify the innermost iteration in an eigensolver. For that, we only need to precompute $\bar{U} = D^{-\frac{1}{2}}U$, which is as sparse as U, and $H = (\bar{U}^T\bar{U})^{-1}$, which is an $\bar{n} \times \bar{n}$ matrix. $\bar{U}$ and H are the only two other matrices apart from those already used for unconstrained cuts. During each iteration of $x := \bar{Q}\bar{P}\bar{Q}x$, we compute:

$$z := \bar{Q}x = x - \bar{U}H\bar{U}^T x \tag{41}$$

$$y := \bar{P}z \tag{42}$$

$$x := \bar{Q}y = y - \bar{U}H\bar{U}^T y. \tag{43}$$

If $\bar{P}$ has an average of k nonzeros per row, then (42) has $O(Nk)$ multiplications. Equations (41) and (43) each requires $O(2N\bar{n} + \bar{n}^2)$ multiplications, which are the only extra computation needed for constrained cuts. Given that $\bar{n} \ll N$ but comparable to k, the increase in time complexity is linear. However, since the solution space is reduced; fewer iterations are needed to converge to the leading eigenvectors. Therefore, the net increase in the computational space and time is negligible if the number of constraints $\bar{n}$ is small. We can further reduce the complexity by sampling the constraints.

We can also avoid the matrix inversion in computing H. To see this, let (A, Σ, B) be the singular value decomposition (SVD) of $\bar{U}$. Since

$$\bar{U} = A_{N\times N}\Sigma_{N\times\bar{n}}B^T_{\bar{n}\times\bar{n}}, \quad A^TA = I, \quad B^TB = I, \tag{44}$$

we have

$$H = (\bar{U}^T\bar{U})^{-1} = (B\Sigma^T\Sigma B^T)^{-1} = B(\Sigma^T\Sigma)^{-1}B^T. \tag{45}$$

Therefore, we can eliminate H altogether since we only need $\bar{U}H\bar{U}^T$ and it becomes:

$$\bar{U}H\bar{U}^T = A\Sigma(\Sigma^T\Sigma)^{-1}\Sigma^TA^T = A_{[\bar{n}]}A^T_{[\bar{n}]}. \tag{46}$$

That is, instead of keeping both $\bar{U}$ and H, we only need to compute the $\bar{n}$ right eigenvectors of $\bar{U}$ and replace $\bar{U}H\bar{U}^T$ with $A_{[\bar{n}]}A^T_{[\bar{n}]}$ in Step 5.

Whether to use $A_{[\bar{n}]}$ or both $\bar{U}$ and H depends on the conditions of the constraint matrix. When the number of constraints is small, H is small and $\bar{U}$ is very sparse, whereas $A_{[\bar{n}]}$ is a full $N \times \bar{n}$ matrix. With the additional cost of computing A from $\bar{U}$, using $A_{[\bar{n}]}$ might not be a good choice. However, when the number of constraints is large, the matrix inversion involved in computing H could be costly and unstable. Later when we smooth the constraints, the columns of $\bar{U}$ can become dense and correlated. In these cases, we can use the SVD of $\bar{U}$ to find a small set of significant constraints, i.e., the first few columns of A, making the computation stable and manageable.

4 Propagating Constraints

The basic formulation works reasonably well if there are enough partial grouping cues. This is not very useful since in reality only a few such cues are given. Sparse cues expose an inherent flaw in the formulation; however, it can be remedied.

4.1 Point Set Example

In Fig. 2, points are naturally organized into four clusters based on proximity. Since the vertical gap is larger than the horizontal gap, an ideal 2-class clustering is obtained by a horizontal cut that divides the four clusters into top and bottom groups. Now, if a few points at the horizontal boundary are grouped together a priori, the horizontal cut violates the partial grouping constraints and the vertical cut becomes optimal. However, when the number of grouping cues is reduced, the formulation in (3) fails to produce the desired vertical cut that divides the four clusters into left and right groups. In particular, the labeled points tend to stand out, while having little impact on the grouping of the rest of the points.

4.2 Why Simple Constraints Are Insufficient

When we preassign points from top and bottom clusters together, we do not just want a group to lose its labeled points to the other group (Fig. 2c), but rather we desire a grouping process that explores their neighboring connections and discovers the left-right division instead.

The formulation in (3), however, does not entail the desire of propagating grouping information on the constrained data points to their neighbors. Often, a slightly perturbed version of the optimal unbiased segmentation becomes the legitimate optimum (Fig. 3). This observation is made from a general optimization point of view and, thus, holds for all choices of ε. The basic formulation in (3), although straightforward, is flawed.

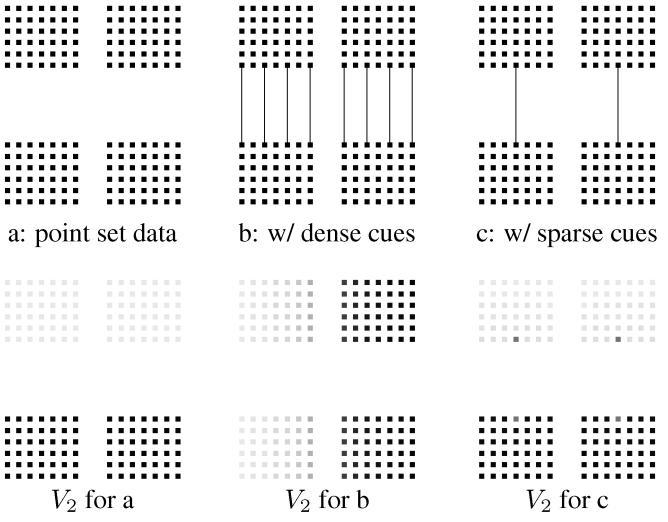

Fig. 2. Three grouping scenarios illustrating the problem of the basic formulation. Row 1: 12×14 dots with a minimum interpoint distance of 1. Pairs of linked points are known to belong together. The weights are computed using a Gaussian function of distance with a standard deviation of 3. Row 2: the continuous optimum V_2 for normalized cuts. For sparse grouping cues, we no longer have the desired vertical cut as the optimal solution.

There are two reasons for such a solution to be undesirable. First, the solution is not smooth. One of the biased data points takes a label that is very different from its nearby points. This is not acceptable especially to those neighbors with which it has high affinity. In other words, we need to explicitly encode *data-driven smoothness* into our discriminative formulation.

The second reason is that such a biased grouping lacks *fairness* with regard to labeled points. Intuitively, if two labeled points, i and j, have similar connections to their neighbors, we desire a fair segmentation so that if i gets grouped with i's friends, j also gets grouped with j's friends. In Fig. 3, the two points in a labeled pair have similar affinity patterns to their nearby points, yet their local segmentations are dissimilar in any solution resulting from the perturbation of the unbiased optimal grouping.

These two conditions, smoothness and fairness of the local segmentations on biased data points, provide a remedy to our basic formulation. Rather than strictly enforcing exactly the same labels on biased data points, we desire an average of their labels to be the same. The average is taken based on the coherence among data points. The more similar a data point is to the biased ones, the heavier the weight is on the label that it takes. Formally, let $g_1 \circ g_2$ be the compound function of g_1 and g_2. Let S_f denote a smoothing function contingent on the data f. We modify the formulation in (3) to be:

$$\begin{aligned} \text{maximize} \quad & \varepsilon(\Gamma^K_{\mathbf{V}}; f), \\ \text{subject to} \quad & S_f \circ \Gamma^k_{\mathbf{V}}(i,l) = S_f \circ \Gamma^k_{\mathbf{V}}(j,l), \\ & i,j \in \mathbb{U}_t, l \in \mathbb{K}, t \in [n]. \end{aligned} \tag{47}$$

Such smoothed constraints on the biased data points can condition a grouping to the extent that many trivial near-optimal unbiased grouping solutions are ruled out from the feasible space.

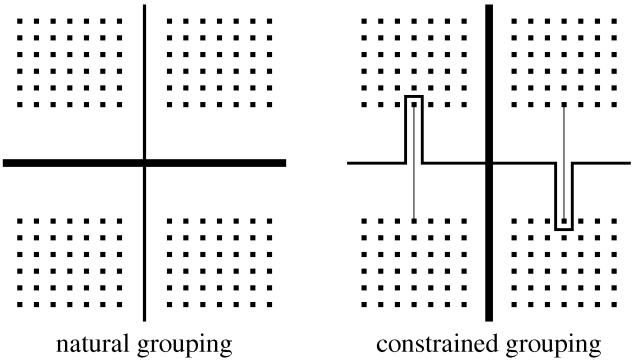

Fig. 3. Undesired grouping from sparse constraints. Left: In the 2-class grouping based on proximity, the horizontal division is optimal while the vertical division is suboptimal. Right: When we add constraints that the points linked together have to belong to the same group, the vertical division becomes the desired partitioning. However, the slightly modified horizontal division is a partitioning that satisfies the constraints, while producing the maximum objective value ε.

Our new formulation is *not* equivalent to the introduction of smoothness priors in a generative approach. There, prior knowledge such as piecewise constancy is usually imposed on the solution independently of the goodness of fit [6], whereas ours is closely coupled with the coherence of the data. Our essential message, in this regard, is that an effective propagation of priors requires an intimate interaction with the data themselves.

4.3 Smooth Constraints for Normalized Cuts

A natural choice of S_f for the normalized cuts criterion is the normalized weight matrix P:

$$S_f \circ \Gamma^k_{\mathbf{V}}(i,l) = \sum_j P_{ij} X(j,l), \quad i \in \mathbf{V}, l \in [K]. \tag{48}$$

This value measures the average density of $\mathbf{V}_l$ from node i's point of view, with nodes of high affinity to it weighted more in the density. This discourages i to take a label different from those of its close neighbors. We may not know in advance what this density is for the optimal partitioning, but the fairness condition requires it to be the same for the labeled pair (i,j): $S_f \circ \Gamma^K_{\mathbf{V}}(i,l) = S_f \circ \Gamma^k_{\mathbf{V}}(j,l)$. The partial grouping constraints in (8) then become:

$$U^T P X = (P^T U)^T X = 0. \tag{49}$$

Since the only change here is that the constraint matrix U becomes $P^T U$, the same solution technique applies. That is, the eigensolution to the program *PNCZ* is given by the eigenvectors of QPQ, where Q is a projector onto the solution space specified by $(P^T U)^T X = 0$ instead of $U^T X = 0$.

In Fig. 4, we show new results with the smoothed constraints. In addition to the basic results in Fig. 2, we also consider two other alternatives that directly utilize partial grouping cues. The simplest case of encoding the labeled pair (i,j) is to modify their weights so that

$$W_{ij} = W_{ji} := 1, \tag{50}$$

where an originally vanishingly small value increases to the maximum affinity. The influence of this change depends on the number of connections the biased nodes have. For example, if node i connects to 10 other nodes, this one more

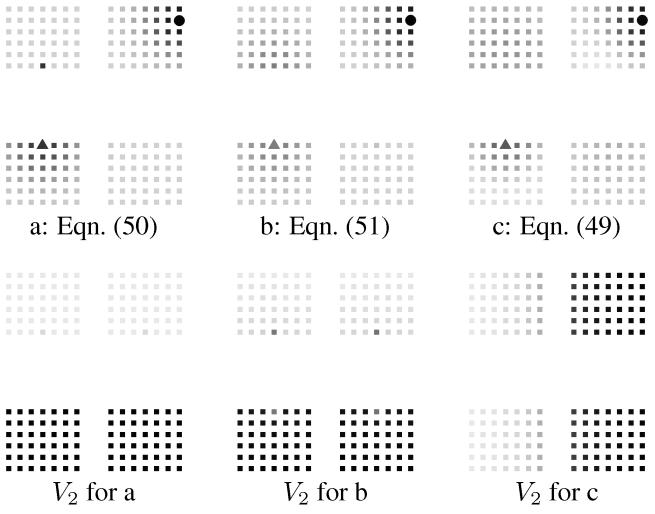

Fig. 4. Propagating partial grouping constraints. Row 1: QPQ values for one labeled point (△) in Fig. 2c and one unlabeled point (•). They are superimposed, with darker gray for larger values. a: Direct modification according to (50) only adds the other labeled point as its neighbor. b: Direct modification according to (51) doubles the neighborhood size for the labeled point. c: Smoothed constraints allow the labeled point to have extensive correlations with all the nodes yet still maintaining fine differentiation toward its own neighbors and those of its labeled peer. The QPQ values on the unlabeled point change little. Row 2: The continuous optimum V_2 for normalized cuts in the three cases. The corresponding discrete 2-class segmentations are omitted as they are obvious from these eigensolutions.

connection would matter little after being normalized by the total connections. Unlike minimum cuts, where a change in one link can change the global optimum completely, normalized cuts are insensitive to perturbation in the weights. Another approach is to let i and j share each other's neighboring connections since i and j are indistinguishable in a desired grouping:

$$W_{ik} = W_{ki} = W_{jk} = W_{kj} := \max(W_{ik}, W_{jk}),\ k \in \mathbf{V}. \quad (51)$$

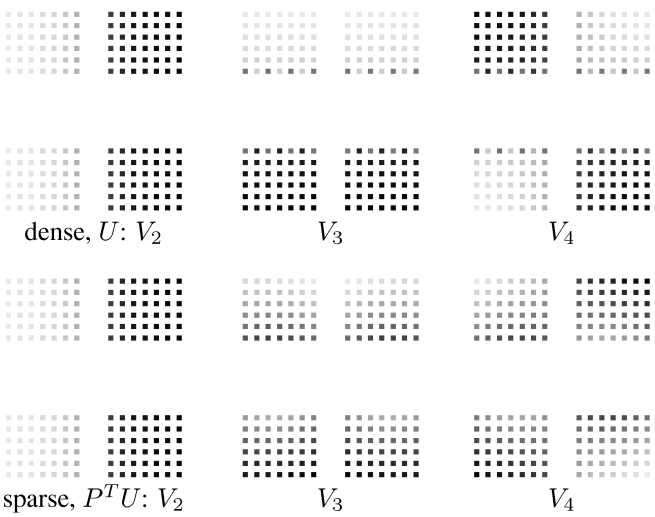

Fig. 5. The importance of smoothing partial grouping constraints. Each row shows three leading eigenvectors. Row 1 are those for the dense grouping case in Fig. 2b, with simple constraints U. Row 2 are those for the sparse grouping case in Fig. 2c, with smoothed constraints P^TU. The first uniform eigenvectors (1_N) are omitted.

Short-circuiting labeled nodes as well as their neighbors produces a similar result as the simple biased grouping in Fig. 2. Their common problem is that only the labeled nodes expand their neighborhoods significantly, which make them distinct from the rest unlabeled data. If we extend (51) to modify the weights *among* the neighbors of labeled points, we can overcome the discontinuity of the segmentation. That's what (49) does, and in a principled way.

The inherent flaw in our basic formulation is also evident in the undesirable results from even dense grouping cues. Though it is unclear for this point set what the best 4-class clustering is with either dense or sparse partial grouping cues, as shown in Fig. 5, the labeled data points never stand out with smoothed constraints. In general, we don't know how many classes there are and whether the partial grouping cues are sufficient. Therefore, partial grouping constraints should always be smoothed with the coherence exhibited in the data in order to produce a meaningful segmentation.

5 Experiments

We calculate pixel affinity using a Gaussian function on the maximum magnitude of intensity edges separating two pixels. $W(i, j)$ is low if i, j are on the opposite sides of a strong edge [14]. Using this simple feature, we will demonstrate how simple extra-image knowledge can improve low-level segmentation and how smoothed partial grouping constraints make a difference.

In Fig. 6, we derive partial groupings based on brightness values, e.g., the foreground is more likely to be lighter and the background is darker. We choose two thresholds to find the pixels at the two intensity extremes and then use morphological operations to further remove pixels appearing in the other set due to noise. As we have already seen in Fig. 2, with simple constraints, biased pixels stand out in segmentation, while with smoothed constraints, they bring

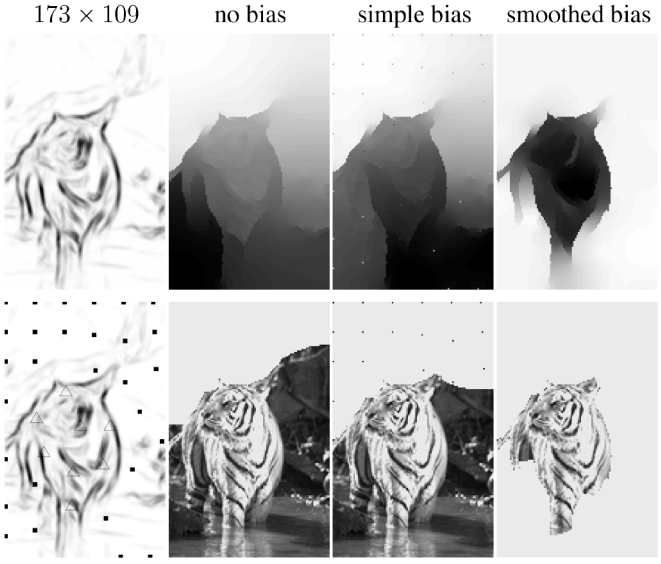

Fig. 6. Segmentation with partial grouping from brightness. Column 1: edge magnitudes and biased nodes (29 pixels marked as ■, 8 pixels marked △) having extreme intensities. Columns 2, 3, and 4: The second eigenvector and foreground images obtained with no constraints, simple constraints U and smoothed constraints P^TU, respectively.

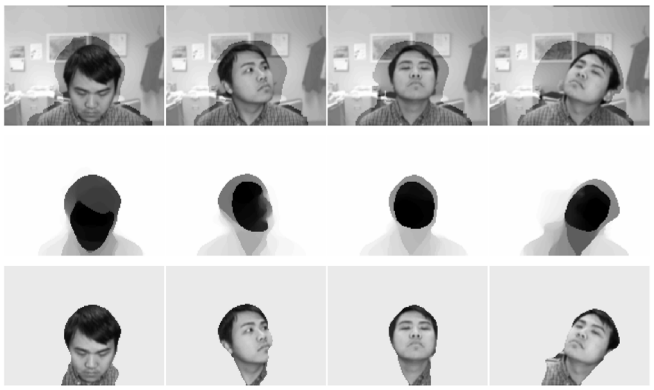

Fig. 7. Segmentation with partial grouping from motion. A sequence of 120×160 images taken every 40 frames from a head tracking system. Row 1: images with peripheries masked out (contrast reduced) according to the difference with neighboring images. The peripheries are pregrouped together. Row 2: the second eigenvectors of constrained normalized cuts. Row 3: foreground images from discrete segmentation.

their neighbors along and change the segmentation completely. This image has rich texture against a relatively simple background. Compared to segmentation using morphological operations on such images, our method can fill the holes caused by thresholding without losing thin structures or distorting region boundaries.

Partial grouping cues can also be derived from motion cues in a video sequence. In Fig. 7, for every image, we compute its difference with two preceding images in a video sequence, threshold and then apply morphological operations to the difference image to create a mask for the foreground. Our constrained segmentation can effectively shrink it to the head in motion.

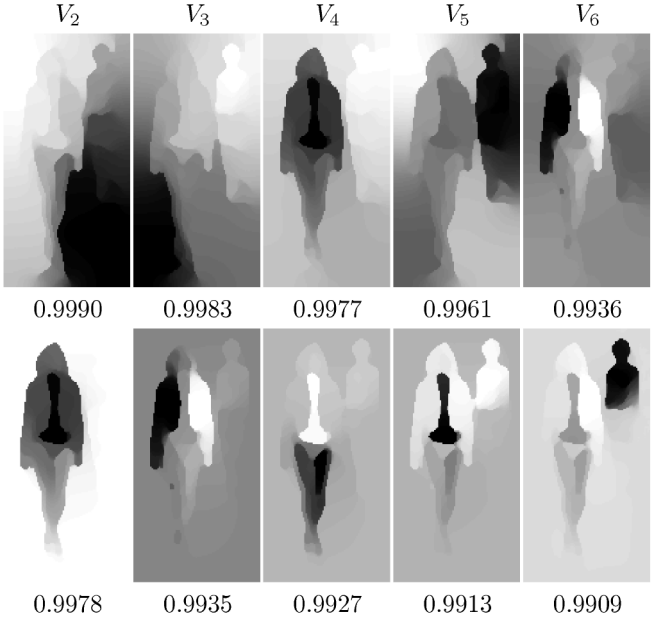

Fig. 8. Segmentation with partial grouping from spatial attention. Image size: 180×90. Rows 1 and 2: leading eigenvectors of unconstrained and constrained normalized cuts, respectively. Uniform V_1s are omitted. Numbers are eigenvalues. It takes 27.2 and 19.7 seconds, respectively, to compute these eigenvectors in MATLAB on a PC with 1 GHz CPU and 1 GB memory.

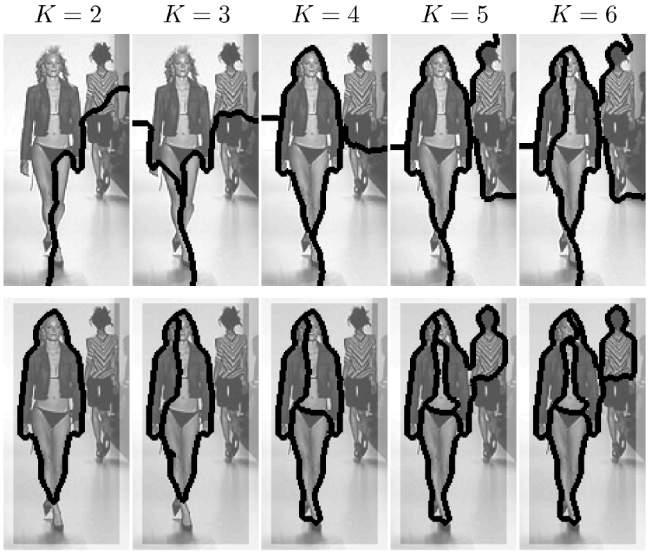

Fig. 9. Multiclass segmentation derived from the eigenvectors shown in Fig. 8. Row 1: unconstrained cuts. Row 2: constrained cuts. The contrast is reduced for biased pixels at the image boundaries.

Partial grouping cues can come not only from low-level cues, but also from high-level expectation. For fashion pictures featuring a fashion model at the center, we choose the background to be: 4-pixel wide at left and right sides, and 7-pixel high at top and bottom sides. Figs. 8 and 9 show the results with and without such background knowledge. Notice that all eigenvectors of QPQ satisfy the constraints and pixels at the four image sides always have similar values in the eigensolutions. Through these constraints, the large uniform background is never broken up in a segmentation, which focuses on the more interesting foreground-background separation or a division within the foreground itself.

Using the same spatial mask and the same set of parameters for computing pixel affinity, we apply our

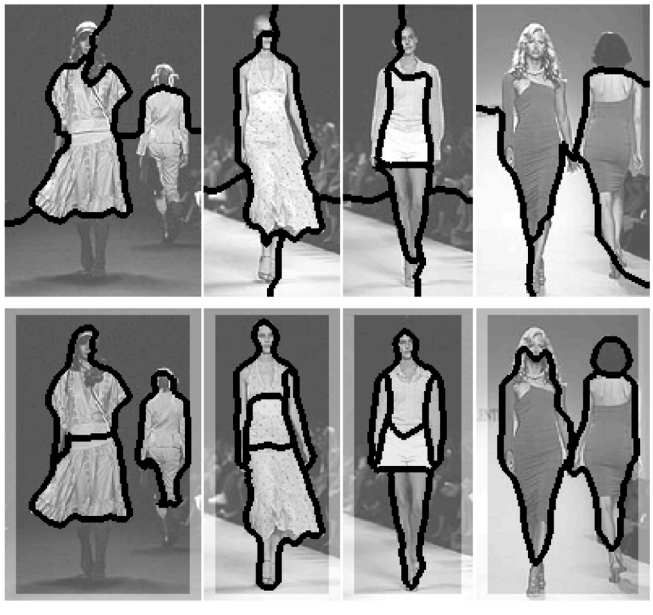

Fig. 10. Segmentation without (Row 1) and with (Row 2) partial grouping at image boundaries, where contrast is reduced. Pictures are from New York Spring 2002 fashion shows.

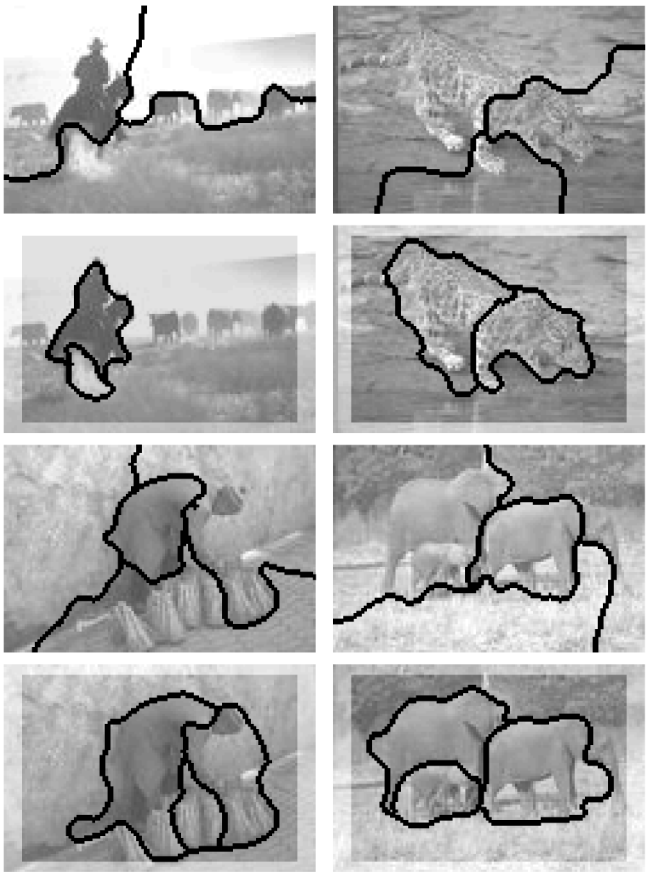

Fig. 11. Segmentation without (Rows 1 and 3) and with (Rows 2 and 4) partial grouping at image boundaries, where contrast is reduced.

constrained normalized cuts to other fashion pictures (http://www.fashionshowroom.com) and Berkeley image data sets [15]. See sample results in Figs. 10 and 11. The number of classes K is chosen manually. When there is an object in the center of the image, such spatial priors always help the segmentation process to pick out the object. If the prior is wrong, for example, when the background spatial mask touches the object of interest, e.g., the tip of shoes in the rightmost fashion picture, the final segmentation also removes the feet from the foreground. The extent of this detrimental effect depends on the connections of the constrained nodes, since partial grouping information is propagated to neighboring nodes that they have large affinity with. Our formulation can neither spot nor correct mistakes in priors.

Technically, (49) can be replaced by an up-to sth order smoothness condition (or a subset of it): $S_f = [P^0, P^1, \ldots, P^s]$. However, higher-order smoothness constraints propagate the partial grouping further at the cost of more computation. In our experiments, we also observe no significant improvement over $S_f = P$ in the eigensolutions.

6 Discussion of Related Work

Our work can be regarded as a small step toward bridging generative approaches and discriminative approaches for grouping. Generative models, including MRF [6] and variational formulations [16], [17], can be naturally cast in a Bayesian framework, where data fidelity and model specificity are treated at equal footing. However, they are sensitive to model mismatches and are usually solved by MCMC methods, which often find local optima with slow convergence.

Discriminative methods, for example graph approaches on image segmentation [5], [18], [19], [20], [21], [22], [23], achieve a global decision based on local pairwise relationships. These algorithms often have efficient computational solutions. These local pairwise comparisons can encode general grouping rules such as proximity and feature similarity. Promising segmentation results on a wide range of complex natural images were reported in [14]. Such pairwise comparisons, however, often have difficulty in deriving reliable long-range grouping information.

Attempts have been made to find MRF solutions by graph partitioning algorithms [7], [8], [24], [25], [26]. In particular, sufficient and necessary conditions on the properties of energy functions that can be solved by *minimum* cuts have been proven in [27], [28]. The work here shows that prior knowledge can be used to guide grouping for discriminative criteria such as normalized cuts [5] and that their global optima in the continuous domain can be solved algebraically with little extra cost.

Our work is also closely linked to the transduction problem, the goal of which is to complete the labeling of a partially labeled data set [29], [30], [31], [32]. If the labeled data set is rich enough to characterize both the structures of the data and the classification task, then using the induced classifier on the labeled set and interpolating it to the unlabeled set shall suffice, which is a supervised learning problem that has many efficient algorithms. However, usually the labeled set is small, so the problem becomes how to integrate the two types of information from both sets to reach a better solution. In [29], the classification problem is formulated in the support vector machine (SVM) framework and labeled data are treated similarly to the rest except that their labels have been instantiated. In [30], information about the labeled data is encoded in the prior distribution of the labeling and the goal is to find a projection of the best SVM discriminator onto the prior space. Through model averaging, partial labeling constraints are softly enforced. In [31], class-dependent data generation models are assumed and the labeled data can be used to estimate the parameters involved in the models. This might be the most effective way to propagate priors. However, these generative models are often too simple to be realistic. In [32], the class-dependent probability models are hidden in the pairwise affinity matrix of all the data points. Again, the labeled set is used to estimate the class-dependent label generation process.

Though our work was initially motivated by the gap between discriminative and generative approaches, we are aware of other works that put similar constraints into clustering algorithms such as K-means [33], [34]. Two types of constraints, *must-link* and *cannot-link*, are considered. Earlier versions of our work [35], [36] also considered cannot-link constraints, that is, two nodes cannot assume the same label. Such constraints are not transitive, which makes them difficult to propagate. In [35], repulsion weights

are used to help enforcing such cues. It also involves approximation in the constraint formulation. For clarity, we choose not to include cannot-link constraints here. Our work is distinct from all these methods in two aspects. Rather than instantiating the labels or the constraints on labeled data points, we use them to regulate the form of a segmentation. We gave an intuitive computational account for the need of constraint propagation and provided a principled way to implement it. Secondly, we can solve near-global optima of our formulation, whereas most other works can only guarantee local optimality.

Our experimental results on image segmentation demonstrate that simple grouping bias can approach figure-ground segregation without knowing what the object is. Our spatial priors effectively take advantage of the asymmetry between figure and ground [37]. In other words, since the outcome of a grouping depends on global configurations, figure-ground segregation can be obtained not only by enhancing the saliency of object structures, but also by suppressing background structures, the latter of which is often easier than the former. Our next step is to explore the integration of more complicated priors in order to segment out only objects known a priori.

7 Summary

We developed a method that integrates both bottom-up and top-down information in a *single* grouping process. The former is based on low-level cues presented in the image, whereas the latter is based on partial grouping cues known a priori; the former defines the goodness of a segmentation, whereas the latter defines the feasibility of a segmentation. The two are unified in a constrained optimization problem. We showed that it is essential to propagate sparse partial grouping cues based on the coherence exhibited in the data. In particular, we developed an efficient solution for such constrained normalized cuts and applied the method successfully to segmenting a wide range of real images.

Acknowledgments

The authors would like to thank Shyjan Mahamud, David Tolliver, Jing Xiao, Charless Fowlkes, and the anonymous reviewers for valuable comments. Jing Xiao also kindly provided the video sequence. This research is supported by (DARPA HumanID) ONR N00014-00-1-0915 and NSF IRI-9817496. S.X. Yu has also been supported in part by NSF LIS 9720350 and NSF 9984706.

References

[1] A. Witkin and J.M. Tenenbaum, "On the Role of Structure in Vision," *Human and Machine Vision,* Beck, Hope, and Rosenfeld, eds., pp. 481-543, New York: Academic Press, 1983.
[2] C. Xu, D.L. Pham, and J.L. Prince, "Medical Image Segmentation Using Deformable Models," *Handbook of Medical Imaging: Progress in Medical Image Processing and Analysis,* pp. 129-174, SPIE, 2000.
[3] D. Marr, *Vision.* Freeman, 1982.
[4] S.C. Zhu and A. Yuille, "Region Competition: Unifying Snakes, Region Growing, and Bayes/MDL for Multiband Image Segmentation," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 18, no. 9, pp. 884-900, Sept. 1996.
[5] J. Shi and J. Malik, "Normalized Cuts and Image Segmentation," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 22, no. 8, pp. 888-905, Aug. 2000.
[6] S. Geman and D. Geman, "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 6, no. 6, pp. 721-741, 1984.
[7] H. Ishikawa and D. Geiger, "Segmentation by Grouping Junctions," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 1998.
[8] S. Roy and I.J. Cox, "A Maximum-Flow Formulation of the *n*-Camera Stereo Correspondence Problem," *Proc. Int'l Conf. Computer Vision,* 1998.
[9] Y. Boykov, O. Veksler, and R. Zabih, "Fast Approximate Energy Minimization via Graph Cuts," *Proc. Int'l Conf. Computer Vision,* 1999.
[10] S. C. Zhu, "Embedding Gestalt Laws in Markov Random Fields," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 21, no. 11, Nov. 1999.
[11] S.X. Yu and J. Shi, "Multiclass Spectral Clustering," *Proc. Int'l Conf. Computer Vision,* 2003.
[12] W. Gander, G.H. Golub, and U. von Matt, "A Constrained Eigenvalue Problem," *Linear Algebra and Its Applications,* vol. 114/115, pp. 815-839, 1989.
[13] M. Meila and J. Shi, "Learning Segmentation with Random Walk," *Neural Information Processing Systems,* 2001.
[14] J. Malik, S. Belongie, T. Leung, and J. Shi, "Contour and Texture Analysis for Image Segmentation," *Int'l J. Computer Vision,* 2001.
[15] 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," *Proc. Int'l Conf. Computer Vision,* 2001.
[16] A. Blake and A. Zisserman, *Visual Reconstruction.* Cambridge, Mass. : MIT Press, 1987.
[17] D. Mumford and J. Shah, "Boundary Detection by Minimizing Functionals," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 22-26, 1985.
[18] A. Amir and M. Lindenbaum, "Quantitative Analysis of Grouping Process," *Proc. European Conf. Computer Vision,* pp. 371-384, 1996.
[19] Y. Gdalyahu, D. Weinshall, and M. Werman, "A Randomized Algorithm for Pairwise Clustering," *Neural Information Processing Systems,* pp. 424-430, 1998.
[20] J. Puzicha, T. Hofmann, and J. Buhmann, "Unsupervised Texture Segmentation in a Deterministic Annealing Framework," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 20, no. 8, pp. 803-818, Aug. 1998.
[21] P. Perona and W. Freeman, "A Factorization Approach to Grouping," *Proc. European Conf. Computer Vision,* pp. 655-670, 1998.
[22] E. Sharon, A. Brandt, and R. Basri, "Fast Multiscale Image Segmentation," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 70-77, 2000.
[23] A. Robles-Kelly and E. R. Hancock, "An EM-Like Algorithm for Motion Segmentation via Eigendecomposition," *Proc. British Machine Vision Conf.,* pp. 123-132, 2001.
[24] D.M. Greig, B.T. Porteous, and A.H. Seheult, "Exact Maximum A Posteriori Estimation for Binary Images," *J. Royal Statistics Soc., Series B,* vol. 51, no. 2, pp. 271-279, 1989.
[25] P.A. Ferrari, A. Frigessi, and P. Gonzaga De SA, "Fast Approximate Maximum A Posteriori Restoration of Multicolour Images," *J. Royal Statistics Soc., Series B,* vol. 57, no. 3, pp. 485-500, 1995.
[26] Y. Boykov, O. Veksler, and R. Zabih, "Markov Random Fields with Efficient Approximations," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 1998.
[27] V. Kolmogorov and R. Zabih, "What Energy Functions Can Be Minimized via Graph Cuts?" *Proc. European Conf. Computer Vision,* 2002.
[28] H. Ishikawa, "Exact Optimization for Markov Random Fields with Convex Priors," *IEEE Proc. Trans. Pattern Analysis and Machine Intelligence,* vol. 25, no. 10, pp. 1333-1336, Oct. 2003.
[29] T. Joachims, "Transductive Inference for Text Classification Using Support Vector Machines," *Proc. Int'l Conf. Machine Learning,* 1999.
[30] T. Jaakkola, M. Meila, and T. Jebara, "Maximum Entropy Discrimination," *Neural Information Processing Systems,* vol. 12, 1999.
[31] K. Nigam, A. Kachites McCallum, S. Thrun, and T. Mitchell, "Text Classification from Labeled and Unlabeled Documents Using EM," *Machine Learning,* pp. 1-34, 1999.
[32] M. Szummer and T. Jaakkola, "Partially Labeled Classification with Markov Random Walks," *Neural Information Processing Systems,* vol. 14, 2001.

[33] K. Wagstaff, C. Cardie, S. Rogers, and S. Schroedl, "Clustering with Instance-Level Constraints," *Proc. Int'l Conf. Machine Learning*, 2000.
[34] K. Wagstaff, C. Cardie, S. Rogers, and S. Schroedl, "Constrained K-Means Clustering with Background Knowledge," *Proc. Int'l Conf. Machine Learning*, 2001.
[35] S.X. Yu and J. Shi, "Grouping with Bias," Technical Report CMU-RI-TR-01-22, Robotics Inst., Carnegie Mellon Univ., Pittsburgh, Pa., July 2001.
[36] S.X. Yu and J. Shi, "Grouping with Bias," *Neural Information Processing Systems*, 2001.
[37] A. Amir and M. Lindenbaum, "A Generic Grouping Algorithm and Its Quantitative Analysis," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 20, no. 2, pp. 168-185, Feb. 1998.

Stella X. Yu received the BS degree in information science and technology from Xi'an Jiaotong University, the MS degree in pattern recognition and intelligent control from Tsinghua University, Beijing, P.R. China, and the PhD degree in robotics from Carnegie Mellon University. Her research interests include adaptive signal processing, Markov decision theory, machine learning, and computational vision. She is a member of the IEEE Computer Society.

Jianbo Shi studied computer science and mathematics as an undergraduate at Cornell University where he received the BA degree in 1994. He received the PhD degree in computer science from the University of California at Berkeley in 1998. From 1999 to 2002, he was a research faculty at the Robotics Institute at Carnegie Mellon University. Since 2003, he has been a faculty at the University of Pennsylvania, where his primary research interests include image segmentation, human recognition, and machine learning. He is a member of the IEEE Computer Society.