Learning Segmentation by Random Walks

Marina Meilă and Jianbo Shi
Carnegie Mellon University
{mmp,jshi}@cs.cmu.edu

Abstract

We present a new view of image segmentation by pairwise similarities. We interpret the similarities as edge flows in a Markov random walk and study the eigenvalues and eigenvectors of the walk's transition matrix. This interpretation shows that spectral methods for clustering and segmentation have a probabilistic foundation. In particular, we prove that the Normalized Cut method arises naturally from our framework. Finally, the framework provides a principled method for learning the similarity function as a combination of features.

1 Introduction

Among the most successful methods in image segmentation combine a global optimality segmentation criterion with local similarity features[3]. Similarity between two pixels i, j is defined as a positive function S_{ij} depending on the local image properties of the pixels(e.g. color, texture, edge flow). Local features are not only computationally convenient, they are also supported by neurological evidence about the human perception of shapes.

The global segmentation criterion is formulated either as energy functions[7, 4] or as weighted graph cut [10, 13]. In both cases, optimizing the chosen criterion turns out to be computationally extremely difficult. Recently[10, 11] connected the graph cuts problems with a set of techniques called *spectral methods* that segment using the eigenvectors and eigenvalues of (certain transformations of) the similarity matrix S_{ij}. As demonstrated in [10, 9, 13], these methods are capable of delivering impressive image segmentation results using simple low-level image features. Moreover, computational efficiency is achieved using sparse and multiscale[9] matrix techniques, which amounts to parallel local computations.

In spite of their practical successes, spectral methods are still incompletely understood. So is the significance of the similarity matrix itself, or, more precisely, the way to combine the various types of lower-level image features into one single matrix S. Is it also of interest to introduce high level knowledge, perhaps through examples, into the definition of the similarity connections S_{ij}.

In this paper, we interpret the local connections as describing a random walk. With this interpretation we achieve:
• a better understanding of spectral methods. We give a simple probabilistic interpretation to the *normalized cut* (NCut) segmentation and show strong connections to other spectral methods.
• the similarity matrix can be learned in a principled way. Given image/segmentation pairs, we optimize the similarity measure as a combination of features
• the framework inspires us to introduce a new feature

Figure 1 depicts the relationship between the main concepts in the paper. Starting from the similarities S_{ij}, we can define a weighted graph G with the set of pixels I as nodes, and S_{ij} as the graph edge weights. Image segmentation can be formulated as a graph partitioning problem which seeks to find small cuts, such as NCut, in the graph. Computationally, the discrete

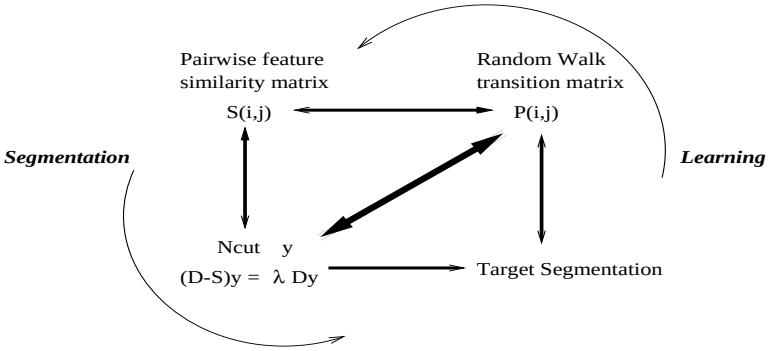

Figure 1: The relationships between the similarity matrix S_{ij}, a Markov random walk, spectral segmentation such as Ncut, and image segmentation. The equivalance between key properties of the Markov random walk and the Ncut critirion offers a principled way of learning the feature similarity matrix S_{ij} in a probabilistic framework.

optimization of NCut is achieved by computing the generalized eigenvectors of $(D-S)\mathbf{y} = \lambda D\mathbf{y}$ (with D the diagonal of S) in the real valued space. This variant of spectral segmentation, called here the NCut algorithm, works well in practice.

The similarity matrix can also be used to define a Markov random walk with transition matrix P_{ij} through normalization. One of the interesting properties of the Markov random walk is its mixing rate or conductance: the rate at which walks propogates over the entire space. The conductance depends on sets of states where the random walk tends to be "trapped" with high probability. In the image, these sets can be seen as segmented regions. In turns out, the conductance can be measured by the NCut criterion, and the low conductivity sets are exactly the results of NCut. This is described in detail in section 2.

A probabilistic interpretation of NCut as a Markov random walk sheds new lights on why and how spectral methods work in segmentation. In particular, it offers a principled way of learning the weights in S_{ij}. A segmented image can provide a "target" transition matrix to which a learning algorithm matches in KL divergence the "learned" transition probabilities. The latter are output by a model as a function of a set of features measured from the training image. This is described in section 3.

To test our theory, we show an experiment on segmenting objects with smooth and rounded shape. In section 4 we show that by training with synthetic images, one can learn to segment real images.

2 Markov random walks, spectral clustering and normalized cut

This section describes our view of spectral segmentation in the framework of Markov random walks. This view provides a new and better motivation for several spectral segmentation and clustering methods. It is also at the core of the learning algorithm presented in the next section. For the sake of brevity, here we outline only the relationship to the NCut algorithm and criterion; the rest will be treated in a longer version of this paper[1]. Low conductivity cuts have been studied before within spectral graph theory; however, all the consequences pertaining to segmentation presented in this paper (Propositions 1,2, relationship to NCut, the learning model) are new.

Here we assume that the similarity function S_{ij} is given, and concern ourselves with using it to partition the image. The rest of the paper will be devoted to the opposite task, i.e. *learning* a good similarity function from segmented images. First we present "ideal" cases, that demonstrate why spectral methods are expected to work. Then we show that the NCut criterion and algorithm fall out naturally from our representation and that the ideal cases are solved exactly by the NCut algorithm.

[1] `http://www.cs.cmu.edu/ mmp/Papers/segment-long.ps`

Obtaining a Markov chain from a similarity matrix By analogy with a graph's adjacency matrix, we call $d_i = \sum_{j \in I} S_{ij}$ the *degree* of node i. For a subset of nodes $A \in I$ the volume of A is $\mathrm{vol}\, A = \sum_{i \in A} d_i$. Let D be the diagonal matrix consisting of the node degrees. By "normalizing" the similarity matrix S one obtains the stochastic matrix $P = D^{-1}S$ whose row sums are all 1. As it is known from the theory of Markov random walks, P_{ij} represents the probability of moving from node i to j in one step, given that we are in i. The eigenvalues of P are $\lambda_0 = 1 \geq \lambda_1 \geq \ldots \lambda_{n-1} \geq -1$; $x^{0 \ldots n-1}$ are the eigenvectors; λ_0 is called the *first* eigenvalue.

Proposition 0. Disconnected *GIf the graph G has k connected components, P will have k eigenvalues equal to 1 and all the other eigenvalues < 1. (Call this P type 0). The first k eigenvectors are the indicator functions of the respective connected components.*

This fact represents the fundamental idea of spectral segmentation. The number of unit λs tells us the number of segments. Then, because $x^{0 \ldots k-1}$ are indicator functions, we can simply project the pixels on the space spanned by these vectors. All pixels in the same segment will project to the same point in R^k (or close by, if there is noise). K-means (with k known) or some other simple clustering algorithm can then separate the segments. We call this segmentation method the NCut algorithm. Experiments [10] show that NCut works well on many graphs that are not disconnected. The following results motivate this behavior.

Proposition 1. Same connections *Assume that I admits a segmentation into k segments so that all pixels in a segment correspond to equal rows in P (call this P type 1). Let $R = [P_{SS'}]$ where S, S' are segments and $P_{SS'} = Pr[S \to S'|S]$. Let $\mu_0 = 1 \geq \mu_1 \geq \ldots \mu_{k-1}$ and $y^0, \ldots y^{k-1}$ be the eigenvalues/vectors of R. Then*
(i) The first k eigenvalues of P are $\mu_0, \ldots \mu_{k-1}$; the other eigenvalues are 0.
(ii) If x is an eigenvector of P corresponding to a non-zero eigenvalue, then $x_i = x_j$ if pixels i and j belong to the same segment.
(iii) If x^l is an eigenvector of P corresponding to a non-zero eigenvalue μ^l, then $x^l_i = y^l_S$ for all pixels i belonging to segment S.

Proposition 2. Linear combination and multiplication *Assume we have two stochastic matrices P', P'' that admit the same partition and have types 1 and 0 respectively. The nonzero eigenvalues of P' are $\lambda_{0 \ldots k-1}$ and the corresponding eigenvectors are $x^{0 \ldots k-1}$. Then*
(i) the convex combination $P = \alpha P' + (1-\alpha)P''$ is a stochastic matrix and $\alpha\lambda_{0 \ldots k-1} + 1 - \alpha$ and $x^{0 \ldots k-1}$ are eigenvalues and eigenvectors of P.
(ii) the products $P'P''$ and $P''P'$ are type 1 matrices with first k eigenvalues/vectors equal to $\lambda_{0 \ldots k-1}$, $x^{0 \ldots k-1}$.

Intuitively, proposition 1 says that spectral segmentation will group together pixels that have the same neighbors. Proposition 2 says that spectral segmentation is successful even when the two criteria, disconnection and same neighbors, are combined linearly or by multiplication. These results motivate the practical usage of spectral clustering methods in general and of the NCut algorithm in particular.

The normalized cut criterion of [10] is a graph theoretical criterion of segmenting an image into two by minimizing the the following expression over all subsets A of I

$$NCut(A, \bar{A}) = \frac{\sum_{i \in A, j \in \bar{A}} S_{ij}}{\sum_{i \in A, j \in I} S_{ij}} + \frac{\sum_{i \in A, j \in \bar{A}} S_{ij}}{\sum_{i \in \bar{A}, j \in I} S_{ij}} = \sum_{i \in A, j \in \bar{A}} S_{ij} \left(\frac{1}{\mathrm{vol}\, A} + \frac{1}{\mathrm{vol}\, \bar{A}} \right) \quad (1)$$

NCut measures the weight of the cut normalized by the volumes of the two segments. In [10] it is shown that if the second eigenvector of P is piecewise constant, like in Propositions 0,1,2 then $NCut(A, \bar{A}) = 1 - \lambda_1$. It is also easy to see that $NCut(A, \bar{A}) = P_{A\bar{A}} + P_{\bar{A}A}$. Thus, NCut has a natural probabilistic interpretation in the framework of random walks, and the quality of the cut is indicated by λ_1.

In [10] the practical solution for NCut (which is NP-hard) is given by the generalized eigenvalue/vector problem $(D - S)x = \lambda x$. We show (in the full paper) that this problem has identical solutions with $Px = (1 - \lambda)x$. Hence, the NCut algorithm described here is essentially identical to the original NCut algorithm of [10]. Random walks provide a simple and natural

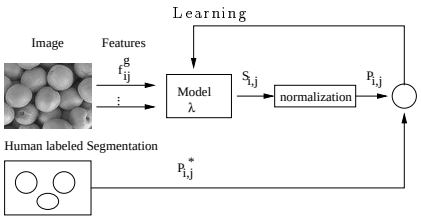

Figure 2: The general framework for learning image segmentation.

alternative interpretation for both the NCut algorithm and criterion.

We found that the NCut algorithm is strongly related to another graph theoretical problem: low conductivity sets [1] as well as to other spectral clustering methods for documents or the web [2]. We discuss these in the light of Markov random walks in the full paper.

3 The framework for learning image segmentation

The previous section has stressed the connection between NCut as a criterion for image segmentation and searching for low conductivity sets in a random walk. Here we will exploit this connection to develop a framework for supervised learning of image segmentation. Our goal is to obtain an algorithm that starts with a training set of segmented images and with a set of features and learns a function of the features that produces correct segmentations. The idea is sketched in figure 2.

For simplicity, assume the training set consists of one image only and its correct segmentation. From the latter it is easy to obtain "ideal" or *target* transition probabilities

$$P^*_{ij} \;=\; \begin{cases} 0, & j \notin A \\ \frac{1}{|A|}, & j \in A. \end{cases} \qquad \text{for } i \text{ in segment } A \text{ with } |A| \text{ elements} \tag{2}$$

We also have a predefined set of features f^q, $q = 1, \ldots Q$ which measure similarity between two pixels according to different criteria and their values for I. Examples of such features are Euclidean distance between pixels, difference in color, presence of an edge crossing the connecting line between the two pixels , etc. A feature associates to a pair i, j of pixels a value f_{ij}.

The *model* is the part of the framework that is subject to learning. It takes the features f^q_{ij} as inputs and outputs the global similarity measure S_{ij}. For the present experiments we use the simple model

$$S_{ij} \;=\; e^{\sum_q \lambda_q f^q_{ij}} \tag{3}$$

This model has the advantage that S_{ij} is always positive. Intuitively, it represents a set of independent "experts", the factors $e^{\lambda_q f^q}$ voting on the probability of a transition $i \to j$. The λ_q values account for both the features' relative importance and scaling. Being an additive model, (3) has limited representation power. In particular, it cannot model contextual changes or dependencies between features. The goal of learning is to find an optimal S_{ij} of the given form as a function of the features.

In our framework, based on the fact that a segmentation is equivalent to a random walk, optimality is defined as the minimization of the conditional Kullback-Leibler (KL) divergence[2] between the target probabilities P^*_{ij} and the transition probabilities P_{ij} obtained by normalizing S_{ij}. Because P^* is fixed, the above minimization is equivalent to maximizing the *crossentropy* between the two (conditional) distributions, i.e. $\max J$, where

$$J \;=\; \sum_{i \in I} \frac{1}{|I|} \sum_{j \in I} P^*_{ij} \log P_{ij} \tag{4}$$

[2]The KL divergence between distributions P and P' is given by $\sum_x P(x) \log \frac{P(x)}{P'(x)}$.

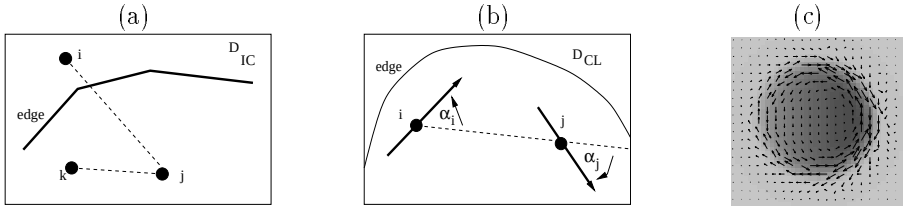

Figure 3: Features for segmenting objects with smooth rounded shape. (a) The edge strength provides a cue of region boundary. It biases against random walks in a direction orthogonal to an edge. (b) Edge orientation provides a cue for the object's shape. The induced edge flow is used to bias the random walk along the edge, and transitions between co-circular edge flows are encouraged. (c) Edge flow for the bump in figure 4. Note that the flow reverses directions on the two sides of an edge.

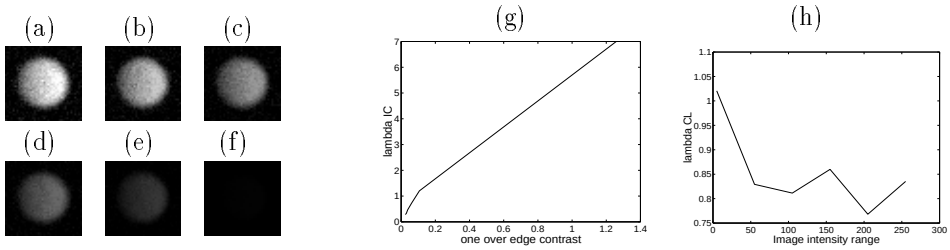

Figure 4: "Bump" images (a)-(f) are used for training. From (a) to (f), the contrast is gradually reduced. (g) shows the relation between the image edge contrast and the learned value of λ_{IC}, demonstrating automatic adaptation to the dynamic range of the IC. (h) shows the dependence on image contrast of λ_{CL}. The importance of the co-linear/co-circular(CL) feature remains relatively constant until the image contrast becomes very low. At low image contrast, CL becomes more important.

If we interpret the factor $1/|I|$ as a uniform distribution over states π^0 then the criterion in (4) is equivalent to the KL divergence between two distributions over transitions $KL(P^{*1}_{i\to j}||P_{i\to j})$ where $P^{(*)}_{i\to j} = \pi^0_i P^{(*)}_{ij}$[3].

Maximizing J can be done via gradient ascent in the parameters λ. We obtain

$$\frac{\partial J}{\partial \lambda_q} = \frac{1}{|I|}\left(\sum_{ij} P^*_{ij} f^q_{ij} - \sum_{ij} P_{ij} f^q_{ij}\right) \tag{5}$$

Hence, the gradient of J w.r.t to a parameter λ^q measures the difference between the means of the corresponding feature f^q under the target and the current distribution. The optimal parameters are attained when the two means are equal. One can further note that the optimum of J corresponds to the solution of the following maximum entropy problem:

$$\max_{P_{j|i}} H(j|i) \quad \text{s.t.} \quad < f^q_{ij} >_{\pi^0 P_{j|i}} = < f^q_{ij} >_{\pi^0 P^*_{j|i}} \text{ for } q = 1, \ldots Q \tag{6}$$

Since this is a convex optimization problem with convex constraints, it has a unique optimum (if any). Thus for this simple model the problem of local maxima is avoided. Knowing that the values of λ_q may grow indefinitely during learning, we shall stop the parameters from growing after they reach a certain upper bound.

4 Segmentation with shape and region information

In this section, we exemplify our approach on a set of synthetic and real images and we use features carrying contour and shape information. First we use a set of local filer banks as edge

[3]We choose to minimize this criterion instead of e.g. $KL(P^*_{i\to j}||\pi^\infty P_{ij})$ because (1) π^∞ converges to π^0 if P converges to P^* so the two criteria are asymptotically equivalent and (2) we prefer to weight the contribution of P_{ij} to the KL divergence by the image statistics represented by π^0.

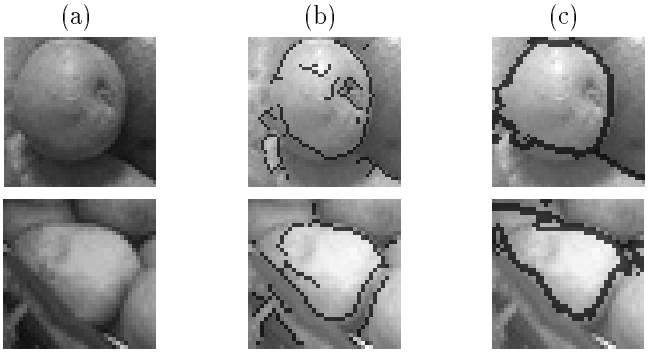

Figure 5: Testing on real images: (a) test images; (b) canny edges computed with the Matlab "edge" function; (c) NCut segmentation computed using the weights learned on the image in 5(c). The system learns to prefer contiguous groups with smooth boundary. The canny edge map indicates that simply looking for edges is likely gives brittle and less meaningful segmentations.

detectors. They capture both edge strength and orientation. From this basic information we construct two features: the *intervening contour* (IC) and the *co-linearity/co-circularity* (CL).

The first feature is based on the assumption that if two pixels are separated by an edge, then they are less likely to belong together(figure 3). In the random walk interpretation, we are less likely to walk in a direction perpendicular to an edge. The intervening contour [6] is computed by [8]

$$f_{ij}^{IC} = \mathrm{MAX}_{k \in l(i,j)} \mathrm{Edge}(k), \tag{7}$$

where $l(i,j)$ is a line connecting pixel i and j, and $\mathrm{Edge}(k)$ is the edge strength at pixel k.

While the IC provides a cue for region boundaries, the edge orientation provides a cue for object shape. Human visual studies suggest that the shape of an object's boundary has a strong influence on how objects are grouped. For example, a convex region is more likely to be perceived as a single object [5][4] Thinking of segmentation as a random walk provides a natural way of exploiting this knowledge. Each discrete edge in the image induces an *edge flow* in its neighborhood. This can be used to bias random walks of non-edge pixels in a direction following the edge orientation. To favor convex regions, we can further bias the random walk by enhancing the transition probabilities between pixels with co-circular edge flow. Thus we define the CL feature as:

$$f_{ij}^{CL} = \frac{2 - cos(2\alpha_i) - cos(2\alpha_j)}{1 - cos(\alpha_l)} + \frac{2 - cos(2\alpha_i + \alpha_j)}{1 - cos(\alpha_o)}, \tag{8}$$

where α_i, α_j are defined as in figure 3(b).

For training, we have constructed the set of "bump" images with varying image contrast, shown in figure 4a–f. Figure 4g,h shows the learned $\lambda_{IC}, \lambda_{CL}$. To check that this system is indeed able to pick up the relevant features, we introduced a cue called *rand*, which assigns random connections to each pixel pair from 0 to 1. The learned value of λ_{rand} is -0.0002, negligible w.r.t. both λ_{IC} and λ_{CL}.

Most real images are not as simple as these synthetic "bump" image. Can the segmentation knowledge learned in these synthetic images be transfer to the real images? Figure 5 shows segmentation results using the weights trained with the "bump" image in figure 4(c). We see that although both feature that we use are local, i.e computed from small image neighborhoods, and although the NCut algorithm has no built-in notion of contiguity, the segmentations are able to produce large contiguous regions with mostly smooth boundaries. Thus suitably chosen local features are able to achieve meaningful global effects.

[4]The numerous efforts in this area[12] have been mostly focused on grouping together discrete edge elements into smooth curves. The question remains on how to transfer this shape information into image region segmentation.

5 Discussion

The main contribution of our paper is showing that spectral segmentation methods have a probabilistic foundation. In the framework of random walks, we give a new interpretation to the NCut criterion and algorithm and a better understanding of its motivation. The probabilistic framework also allows us to define a principled criterion for supervised learning of image segmentation.

We see supervised learning as feasible in this traditionally unsupervised domain because: (1) We are proposing to learn a combination of fixed features. This is a relatively simple model and we expect it to require proportionally little training data. (2) Since we are using only *local* features, the training set may consist of synthetic images that reproduce the feature statistics of the real images we want to segment. Both arguments are supported by our preliminary experiments where one 60×60 synthetic noisy image was sufficient. Learning is an alternative to the current lack of a principled approach to constructing similarity functions. In domains like medical imaging, cell biology, where the relative importance of features is less clear, learning has a strong potential in automatic segmentation.

References

[1] Fan R. K. Chung. *Spectral Graph Theory*. American Methematical Society, 1997.

[2] P. Drineas, Ravi Kannan, Alan Frieze, Santosh Vempala, and V. Vinay. Clustering in large graphs and matrices. In *Proc. of the 10th ACM-SIAM Symposium on Discrete Algorithms*, 1999.

[3] S. Geman and D. Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *PAMI*, 6:721–741, November 1984.

[4] T. Hofmann and J. Buhmann. Hierarchical pairwise data clustering by mean-field annealing. In *International Conference on Artificial Neural Networks*, 1995.

[5] I. Kovacs. Gestalten of today: early processing of visual contours and surfaces. *Behavioural brain research*, 1996.

[6] T. Leung and J. Malik. Contour continuity in region based image segmentation. In *European Conference on Computer Vision*, 1998.

[7] D. Mumford and J. Shah. Optimal approximations by piecewise smooth functions, and associated variational problems. *Comm. Pure Math.*, pages 577–684, 1989.

[8] P. Perona and W. Freeman. A factorization approach to grouping. In *European Conference on Computer Vision*, 1998.

[9] E. Sharon, A. Brandt, and R. Basri. Fast multiscale image segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2000.

[10] J. Shi and J. Malik. Normalized cuts and image segmentation. *PAMI*, 2000.

[11] Y. Weiss. Segmentation using eigenvectors: a unifying view. In *International Conference on Computer Vision*, 1999.

[12] L. Williams and K. Thornber. A comparison of measure for detecting natural shapes in clutter backgrounds. In *European Conference on Computer Vision*, 1998.

[13] M. Werman Y. Gdalyahu, D. Weinshall. A randomized algorithm for pairwise clustering. In *NIPS*, 1998.