Spectral Partitioning with Indefinite Kernels using the Nyström Extension

Serge Belongie[1], Charless Fowlkes[2], Fan Chung[1], and Jitendra Malik[2]

[1] University of California, San Diego, La Jolla, CA 92093, USA
{sjb,fan}@cs.ucsd.edu
[2] University of California, Berkeley, Berkeley, CA 94720, USA
{fowlkes,malik}@cs.berkeley.edu

Abstract. Fowlkes et al. [7] recently introduced an approximation to the Normalized Cut (NCut) grouping algorithm [18] based on random subsampling and the *Nyström extension*. As presented, their method is restricted to the case where W, the weighted adjacency matrix, is positive definite. Although many common measures of image similarity (i.e. kernels) are positive definite, a popular example being Gaussian-weighted distance, there are important cases that are not. In this work, we present a modification to Nyström-NCut that does not require W to be positive definite. The modification only affects the orthogonalization step, and in doing so it necessitates one additional $O(m^3)$ operation, where m is the number of random samples used in the approximation. As such it is of interest to know which kernels are positive definite and which are indefinite. In addressing this issue, we further develop connections between NCut and related methods in the kernel machines literature. We provide a proof that the Gaussian-weighted chi-squared kernel is positive definite, which has thus far only been conjectured. We also explore the performance of the approximation algorithm on a variety of grouping cues including contour, color and texture.

1 Introduction

Among the methods for image segmentation developed in recent years, those based on pairwise grouping arguably show the most promise. By the term "pairwise" we mean that the grouping operation is based on measures of similarity or dissimilarity between pairs of pixels. In contrast, "central" grouping methods proceed by comparing all the pixels to a small number of prototypes or cluster centers; examples include k-means and EM clustering with Gaussian mixture models. Central grouping methods tend to be computationally cheaper, but have difficulty dealing with irregularly-shaped clusters and gradual variation within groups. Moreover, they are sensitive to initialization and require model-selection (i.e. specification of the number of groups). Generally speaking, pairwise grouping methods either eliminate or simplify these problems. Some of the approaches that have been proposed for grouping pairwise data include spectral graph partitioning [18, 19, 13], deterministic annealing [15], and stochastic clustering [8].

The drawback, of course, is that approaches based on pairwise data in principle require measurements between all possible pairs of pixels. Consequently, the number of pairs considered is often restricted by placing a threshold on the number of connections per pixel, e.g. by specifying a cutoff radius. This discourages the use of long-range connections and this can result in over-segmentation of homogeneous regions. A promising solution to this problem for the case of spectral graph theoretic methods was recently proposed by Fowlkes et al. [7]. Their method, based on the Nyström approximation for the integral eigenvalue problem, works by solving a grouping problem on a small set of m randomly sampled pixels and then extending the solution to the complete set of pixels. Using this approach, they produced high-quality segmentations of image sequences in a fraction of the time required to compute the exact solution.

Though not explicitly stated, Fowlkes et al. assume that the function used for computing the simlarity between pairs of pixels is positive definite, i.e. that the weight matrix comprised of all the pairwise similarities is a Gram matrix. While this assumption is generally taken for granted in kernel based methods (e.g. [17]), the same cannot necessarily be said for similarity measures used in the computer vision literature. In the present work, we show that this restriction can be lifted by modifying the orthogonalization step used in [7], which requires positive definiteness. This proposed change necessitates an additional $O(m^3)$ operation; as such it is desirable to know when this alternative is necessary. To this end we discuss the application of the Nyström method to a number of commonly used similarity functions, both positive definite and indefinite (i.e. neither positive definite nor negative definite).

The organization of this paper is as follows. We begin by reviewing in Section 2 the spectral graph theoretic pairwise grouping algorithm used in this work, namely Normalized Cuts (NCut) [18]. Next we review the Nyström extension in Section 3, focusing on its application to NCut. In Section 4 we discuss the issues of definiteness and indefiniteness of commonly used kernels used for measuring pairwise similarity and provide our modification to the method of [7]. Experimental results and discussion are provided in Sections 5. Some properties of the approximation are discussed in Section 6 and finally we conclude in Section 7.

2 Review of Normalized Cuts

Let the symmetric matrix $W \in \mathbb{R}^{N \times N}$ denote the weighted adjacency matrix for a graph $G = (V, E)$ with nodes V and edges E. We will refer to the function used to compute W_{ij} as a *kernel*; examples of kernels and their properties are discussed in Section 5. Let A and B represent a bipartition of V, i.e. $A \cup B = V$ and $A \cap B = \emptyset$. Let $\mathrm{cut}(A, B)$ denote the sum of the weights between A and B: $\mathrm{cut}(A, B) = \sum_{i \in A, j \in B} W_{ij}$. The degree of the ith node is defined as $d_i = \sum_j W_{ij}$ and the volume of a set as the sum of the degrees within that set: $\mathrm{vol}(A) = \sum_{i \in A} d_i$ and $\mathrm{vol}(B) = \sum_{i \in B} d_i$. The Normalized Cut between sets A

and B is then given as follows:

$$\mathrm{NCut}(A,B) = \mathrm{cut}(A,B)\left(\frac{1}{\mathrm{vol}(A)} + \frac{1}{\mathrm{vol}(B)}\right) = \frac{2 \cdot \mathrm{cut}(A,B)}{\mathrm{vol}(A) \| \mathrm{vol}(B)}$$

where $\|$ denotes the harmonic mean.

We wish to find A and B such that $\mathrm{NCut}(A,B)$ is minimized. Appealing to spectral graph theory [4], Shi and Malik [18] showed that an approximate solution may be obtained by thresholding the eigenvector corresponding to the second smallest eigenvalue of the normalized Laplacian $\mathcal{L}$, which is defined as

$$\mathcal{L} = D^{-1/2}(D - W)D^{-1/2} = I - D^{-1/2}WD^{-1/2}$$

where D is the diagonal matrix with entries $D_{ii} = d_i$. The matrix $\mathcal{L}$ is positive semidefinite, even when W is indefinite. Its eigenvalues lie on the interval $[0,2]$ so the eigenvalues of $D^{-1/2}WD^{-1/2}$ are confined to lie inside $[-1,1]$ (see [4]). Finally, extensions to multiple groups are possible via recursive bipartitioning or through the use of multiple eigenvectors.

3 Review of the Nyström Approximation to NCut

Since N is quite large for typical images (e.g. 256^2), finding the eigenvectors of $\mathcal{L}$ is computationally intensive. One approach to dealing with this difficulty is to connect only to those pixels that are nearby in the image. This makes $\mathcal{L}$ sparse and permits the use of an efficient eigensolver (e.g. Lanczos). However, this discourages the use of long-range connections and the approximation properties are not easily understood. The Nyström approximation provides an alternative approach based on random sampling.

The application of the Nyström approximation to NCut proceeds as follows. First, choose m samples at random from the full set of N pixels. For simplicity in notation, reorder the samples so that these m come first and the remaining $n = N - m$ samples come next. Now partition the weight matrix W as

$$W = \begin{bmatrix} A & B \\ B^T & C \end{bmatrix} \tag{1}$$

with $A \in \mathbb{R}^{m \times m}$, $B \in \mathbb{R}^{m \times n}$, $C \in \mathbb{R}^{n \times n}$, and $N = m + n$, with $m \ll n$. Here A represents the subblock of weights amongst the random samples, B contains the weights from the random samples to the rest of the samples, and C contains the weights between all of the remaining samples. Assuming $m \ll n$, C is huge. The Nyström extension implicitly approximates C using $B^T A^{-1} B$. The quality of the approximation of the full weight matrix

$$\hat{W} = \begin{bmatrix} A & B \\ B^T & B^T A^{-1} B \end{bmatrix} \tag{2}$$

can be quantified as the norm of the Schur complement $\|C - B^T A^{-1} B\|$. The size of this norm is governed by the extent to which C is spanned by the rows of

B. Thus, rather than set the majority of entries in W to zero to produce a sparse approximation, the Nyström method provides (implicitly) an approximation to the entire weight matrix based on a subset of rows/columns.

Fowlkes et al. [7] show that $\hat{W}$ can be diagonalized in an efficient manner. Let $A^{1/2}$ denote the symmetric positive definite square root of A, define $S = A + A^{-1/2}BB^TA^{-1/2}$ and diagonalize it as $S = U\Lambda U^T$. If the matrix V is defined as

$$V = \begin{bmatrix} A \\ B^T \end{bmatrix} A^{-1/2}U\Lambda^{-1/2} \tag{3}$$

then one can show that $\hat{W}$ is diagonalized by V and Λ, i.e. $\hat{W} = V\Lambda V^T$ and $V^TV = I$. We assume that pseudoinverses are used in place of inverses as necessary when there is redundancy in the random samples.

To apply this approximation to NCut, it is necessary to compute the row sums of $\hat{W}$. This is possible without explicitly evaluating the $B^TA^{-1}B$ block since

$$\begin{aligned} \hat{\boldsymbol{d}} = \hat{W}\mathbf{1} &= \begin{bmatrix} A\mathbf{1}_m + B\mathbf{1}_n \\ B^T\mathbf{1}_m + B^TA^{-1}B\mathbf{1}_n \end{bmatrix} \\ &= \begin{bmatrix} \boldsymbol{a}_r + \boldsymbol{b}_r \\ \boldsymbol{b}_c + B^TA^{-1}\boldsymbol{b}_r \end{bmatrix} \end{aligned} \tag{4}$$

where $\boldsymbol{a}_r, \boldsymbol{b}_r \in \mathbb{R}^m$ denote the row sums of A and B, respectively, and $\boldsymbol{b}_c \in \mathbb{R}^n$ denotes the column sum of B.

With $\hat{\boldsymbol{d}}$ in hand, the blocks of $\hat{D}^{-1/2}\hat{W}\hat{D}^{-1/2}$ that are needed to approximate its leading eigenvectors are given as

$$A_{ij} \leftarrow \frac{A_{ij}}{\sqrt{\hat{\boldsymbol{d}}_i\hat{\boldsymbol{d}}_j}}, \qquad i, j = 1, \ldots, m$$

and

$$B_{ij} \leftarrow \frac{B_{ij}}{\sqrt{\hat{\boldsymbol{d}}_i\hat{\boldsymbol{d}}_{j+m}}}, \qquad i = 1, \ldots, m, j = 1, \ldots, n$$

to which we can apply equation (3) as before.

4 Nyström-NCut for indefinite kernels

In diagonalizing $\hat{D}^{-1/2}\hat{W}\hat{D}^{-1/2}$, Fowlkes et al. assume that A is positive semidefinite in order to compute $A^{1/2}$, the positive semidefinite square root of A. Positive definite kernels correspond to a dot products in a "feature space" that is protentially of much higher dimensionality than the input space. This geometric intuition is the essence of the *kernel trick* [16], which serves as the basis for kernel-based methods such as support vector machines (SVM) and kernel principal components analysis (KPCA). As such, in the kernel-machines literature, the term "kernel" is often used synonymously with "positive definite kernel."

The same assumption cannot be made in general for similarity functions used in grouping in the computer vision literature. To be sure, Gaussian-weighted Mahalanobis distance between feature vectors, one of the most common similarity measure used in grouping, is positive definite, as are several other popular choices. However, there are a number of similarity measures one can use that only satisfy the fairly weak requirements that W is symmetric and W_{ij} is "big" if pixels i and j are similar and "small" if they are not. It is therefore important that both cases be properly addressed.

Equation (3) can be thought of as a "one-shot" combined Nyström eigenvector approximation and orthogonalization operation. By keeping these two steps separate, we will show that the positive definiteness requirement can be circumvented.[1] Starting from the approximation of W in Equation (2), let $A = U \Lambda U^T$ denote the diagonalization of A. We may then write $\hat{W}$ as

$$\hat{W} = \begin{bmatrix} U \\ B^T U \Lambda^{-1} \end{bmatrix} \Lambda \left[U^T \; \Lambda^{-1} U^T B \right]$$

where the block $B^T U \Lambda^{-1}$ represents the Nyström extension. As Williams and Seeger [20] noted, this is equivalent to the expression for the projection of a test point onto the feature-space eigenvectors in Kernel PCA. Although this extension appears to give us an approximate diagonalization, the extended eigenvectors are not orthogonal.

We carry out the orthogonalization step as follows. Let $\bar{U}^T = [U^T \quad \Lambda^{-1} U^T B]$ and define $Z = \bar{U} \Lambda^{1/2}$ so that $\hat{W} = Z Z^T$. Let $F \Sigma F^T$ denote the diagonalization of $Z^T Z$. Then the matrix $V = Z F \Sigma^{-1/2}$ contains the leading orthonormalized eigenvectors of $\hat{W}$, i.e. $\hat{W} = V \Sigma V^T$ with $V^T V = I$. As before, a pseudoinverse can be used in place of a regular inverse when A has linearly dependent columns.

Thus the approximate eigenvectors are produced in two steps: first we use the Nyström extension to produce $\bar{U}$ and Λ and then we orthogonalize $\bar{U}$ to produce V and Σ. Although this "two-step" approach is applicable in general, the additional $O(m^3)$ step it requires for the orthogonalization takes extra time and leads to an increased loss of significant figures. Therefore it is expedient to know when the one-shot method can be applied, i.e. when a given kernel is positive definite.

5 Experiments and Discussion

In this section we discuss a number of kernels, both positive definite and indefinite, and show examples of their use.

[1] Since the normalized Laplacian is positive semidefinite even when W is not, it is tempting to try to apply Nyström to $\mathcal{L}$ instead of $D^{-1/2} W D^{-1/2}$. Unfortunately, the Nyström method finds the leading eigenvectors, and the eigenvectors of $\mathcal{L}$ we need are the trailing ones.

Gaussian weighted distance. Perhaps the most commonly used measure of similarity between pixels is Gaussian weighted Mahalanobis distance between feature vectors $\mathbf{x}_i$ and $\mathbf{x}_j$:

$$W_{ij} = e^{-\frac{1}{2}(\mathbf{x}_i - \mathbf{x}_j)^T \Sigma^{-1} (\mathbf{x}_i - \mathbf{x}_j)}$$

This kernel is positive definite and therefore admits the use of the one-shot Nyström method. Fowlkes et al. [7] used this kernel on feature vectors containing position, color, and optical flow. Most of the works cited in the introduction use this kernel (among others); as such, additional experimental results will not be provided here.

Histogram comparison using the χ^2 test. The χ^2 test is a simple and effective means of comparing two histograms. It has been shown to be a very robust measure for color and texture discrimination [15]. Given two normalized histograms $h_i(k)$ and $h_j(k)$ define

$$\chi^2_{ij} = \frac{1}{2} \sum_{k=1}^{K} \frac{(h_i(k) - h_j(k))^2}{h_i(k) + h_j(k)}$$

where it is understood that any term in the sum for which $h_i(k) = 0$ and $h_j(k) = 0$ is replaced by zero.

We can then define the similarity between a pair of histograms as $W_{ij} = e^{-\chi^2_{ij}/\alpha}$. This kernel is widely conjectured to be positive definite (see e.g. [3]) but to our knowledge no proof of this has been published. The appendix A contains our proof that Gaussian-weighted χ^2 is positive definite.

An example of Nyström-NCut on a color image of a tiger is shown in Figure 1. In this example, we computed a local color histogram inside a 5×5 box around each pixel using the color quantization scheme of [14]. Finally, since the weight matrix is positive definite, we used the one-shot Nyström method. We note that the same technique can be applied to texture using the "textons" of Malik et al. [11], i.e. vector-quantized filter responses.

Intervening contour. To integrate contour information into a pairwise region based grouping framework, it is convenient to construct a kernel that indicates points are dissimilar if they lie on oposite sides of an intervening contour [10]. We consider a distance between each pair of pixels that takes into account all possible paths across the image. Each path between a pair is assigned a distance equal to the maximum contour energy encountered along the path. The distance r_{ij} between the pair is then taken to be the minimum energy over all paths and the similarity between two pixels is $e^{-r^2_{ij}/\alpha}$. It is not known whether this kernel is positive definite.

This cue captures the Gestalt notion of closure. If two points are separated by a closed contour then they will have low similarity while if there is a path

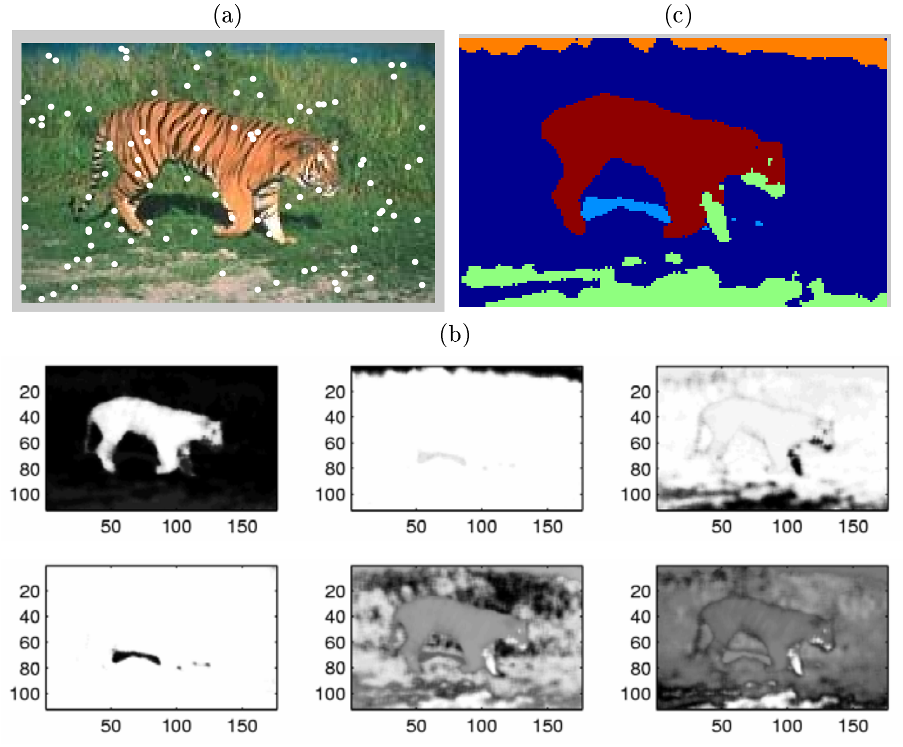

Fig. 1. Segmentation of tiger image based on Gaussian weighted χ^2-distance between local color histograms. The image size is 128 × 192 and the histogram window size is 5 × 5. Color quantization was performed as in [14] with 8 bins. Since the $e^{-\chi^2_{ij}}$ kernel is positive definite, we can use the one-shot method of [7]. (a) Original image shown with $m = 100$ random samples used in approximation. (b) Nyström-NCut eigenvectors 2 through 7, sorted in ascending order by eigenvalue. (c) Segment-label image obtained via k-means clustering on the eigenvectors as described in [7].

connecting two points that doesn't cross an edge then they will have high similarity.

The problem of finding this minimum over all paths has the same structure as the classic shortest path problem and is easily solved by application of Dijkstra's algorithm [5]. Since the problem is sparse it is possible to achieve a running time of $O(m \cdot (N \log N))$ where m is the number of samples and N is the number of pixels. An illustration of this method applied to a sample image is shown in Figure 2.

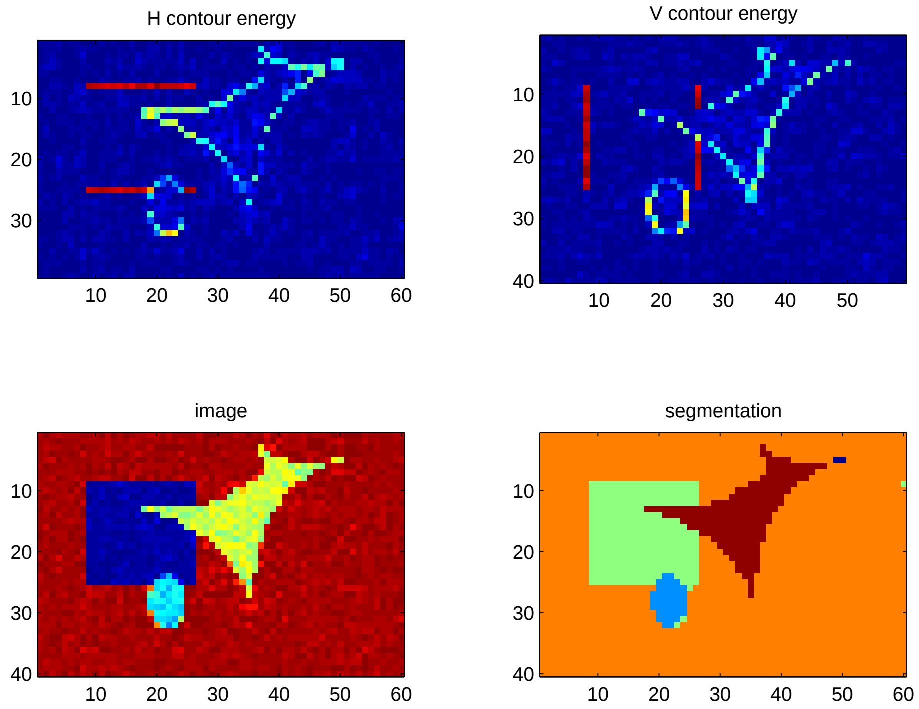

Fig. 2. Segmentation using intervening contour. Original image of synthetic shapes with noise is shown at lower left. At top, the horizontal and vertical boundary energy is shown; this is computed by squaring the x and y components of the smoothed gradient. The connection weight between a pair of pixels is based on the contour energy encountered along all paths between the pair of pixels; see text for details. The segmentation label map, obtained via k-means on the Nyström-NCut eigenvectors, is shown at lower right.

One Minus Squared Distance. A simple choice of kernel for expressing similarity between pixels is the following,

$$W_{ij} = 1 - \frac{r_{ij}^2}{\alpha}$$

where r_{ij}^2 represents the squared distance between feature vectors at i and j. This kernel is in general indefinite[2]; moreover, it takes on negative values.[3] Nevertheless, this kernel makes intuitive sense and, empirically, NCut works well with it. An example using the two-step Nyström method with this kernel on color and proximity is shown in Figure 3.

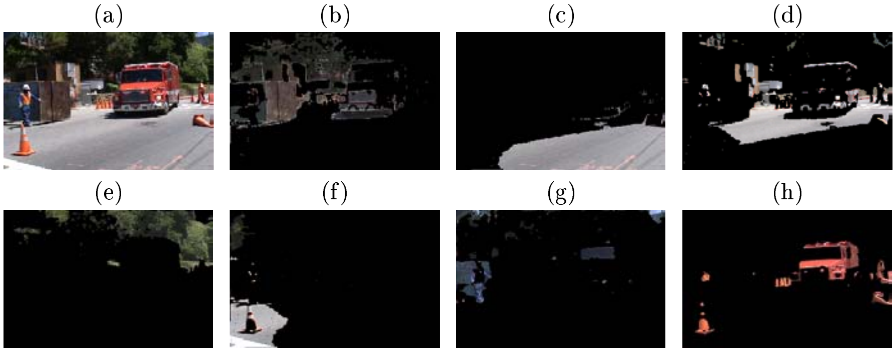

Fig. 3. Segmentation of Firetruck image using 2nd degree polynomial kernel. The feature vector for each pixel contains RGB color values and (x, y) coordinates. The form of the kernel is $W_{ij} = 1 - r_{ij}^2/\alpha$ where r_{ij}^2 represents the Mahalanobis distance between feature vectors at pixels i and j. Since this kernel is indefinite, we applied Nyström-NCut using the proposed two-step method. (a) Original image. (b-h) Segments obtained via k-means clustering on Nyström-NCut eigenvectors as in Figure 1.

6 Properties of the Approximation

Since the Nyström method only requires us to diagonalize an $m \times m$ matrix to find the leading eigenvectors of $\hat{D}^{-1/2}\hat{W}\hat{D}^{-1/2}$, this approach can be very efficient. A key question is how well a given set of samples allows us to approximate these eigenvectors. Fowlkes et al. [7] provide empirical results on a large set of natural images to show that roughly 100 samples do a good job when using color and proximity. In this section we wish to shed some light on the geometric interpretation of the use of $B^T A^{-1} B$ as an approximation to C.

[2] In Multidimensional Scaling (MDS) [6] one applies a "centering operation" to a squared distance matrix to isolate the positive semidefinite component corresponding to the inner products between the embedded coordinates. This centering operation (not repeated here) is more complicated than the one-minus transformation used in this kernel, and though interesting in its own right, is beyond the scope of the current discussion.

[3] In principle this means the degree could be negative, viz. if enough negative entries conspire in a single row of W to dominate the positive entries. In such cases, one could do clipping, however in our experiments we found that this was unnecessary.

As we saw in Section 3, the quality of the approximation depends on the extent to which the rows of C are spanned by the rows of B. This is true for W in general. When W is positive semidefinite, we can say more. In particular, we can express the blocks A and B as $A = X^T X$ and $B = X^T Y$ where $X \in \mathbb{R}^{(m+n)\times m}$ and $Y \in \mathbb{R}^{(m+n)\times n}$; in the parlance of the kernel-machines literature [16], the columns of X and Y represent the *empirical feature mapping.* Let $X = QR$, with $Q \in \mathbb{R}^{(m+n)\times m}$, $Q^T Q = I$, and upper-triangular $R \in \mathbb{R}^{m\times m}$ denote the QR decomposition of X. In other words, Q represents an orthonormal basis for the space spanned by the columns of X. Then the matrix $B^T A^{-1} B$ simplifies as follows:

$$\begin{aligned} B^T A^{-1} B &= Y^T X (X^T X)^{-1} X^T Y \\ &= Y^T QR(R^T Q^T QR)^{-1} R^T Q^T Y \\ &= Y^T QRR^{-1} R^{-T} R^T Q^T Y \\ &= Y^T QQ^T Y \\ &= (Q^T Y)^T (Q^T Y) \end{aligned}$$

Recall that the exact values of C are given by $Y^T Y$, i.e. the inner products between the columns of Y. The quantity $(Q^T Y)^T (Q^T Y)$ represents the inner products of the columns of Y after projecting them onto the subspace spanned by X. Thus if Y is spanned well by X then $Y^T QQ^T Y$ will be a good approximation to $Y^T Y$.

7 Conclusion

In this paper we have introduced a modification to the Nyström approximation to Normalized Cuts (Nyström-NCut) that does not require the measure of similarity between pairs of pixels to be a positive definite function. The proposed change involves separating the steps of the Nyström extension and orthogonalization. As this necessitates an additional $O(m^3)$ operation, where m is the number of samples used in the approximation, it is important to know whether a kernel is positive definite in order not to waste computation and sacrifice numerical precision unnecessarily. In light of this, we examined a number of kernels, both positive definite and indefinite, and showed image segmentation results using both versions of Nyström-NCut. In the process we have provided what we believe is the first proof that the Gaussian weighted χ^2 kernel is positive definite. Finally, we provided some geometrical insight into the nature of the approximation for the case of positive definite kernels.

Acknowledgments

We wish to thank Olivier Chapelle and Gianluca Donato for helpful discussions. This research was supported in part by NSF Grant no. DMS 0100472 and Office of Naval Research grant no. N00014-01-1-0890, as part of the MURI program.

A Proof of positive definiteness of $e^{-\chi^2_{ij}}$

We now prove that $e^{-\chi^2_{ij}}$ is positive definite. We begin by considering the χ^2_{ij} term by itself. Noting that $(h_i(k) - h_j(k))^2 = (h_i(k) + h_j(k))^2 - 4h_i(k)h_i(k)$, we can rewrite χ^2_{ij} as

$$\chi^2_{ij} = 1 - 2\sum_{k=1}^{K} \frac{h_i(k)h_j(k)}{h_i(k) + h_j(k)}$$

We wish to show that the matrix Q with entries given by

$$Q_{ij} = 2\sum_{k=1}^{K} \frac{h_i(k)h_j(k)}{h_i(k) + h_j(k)}$$

is positive definite. Consider the quadratic form $c^T Qc$ for an arbitrary finite nonzero vector c:

$$\begin{aligned}
c^T Qc &= \sum_{i,j=1}^{n} c_i c_j Q_{ij} \\
&= 2\sum_{k=1}^{K} \sum_{i,j=1}^{n} c_i c_j \frac{h_i(k)h_j(k)}{h_i(k) + h_j(k)} \\
&= 2\sum_{k=1}^{K} \sum_{i,j=1}^{n} c_i c_j h_i(k) h_j(k) \int_0^1 x^{h_i(k)+h_j(k)-1} dx \\
&= 2\sum_{k=1}^{K} \sum_{i,j=1}^{n} \int_0^1 c_i h_i(k) x^{h_i(k)-\frac{1}{2}} c_j h_j(k) x^{h_j(k)-\frac{1}{2}} dx \\
&= 2\sum_{k=1}^{K} \int_0^1 \left(\sum_{i=1}^{n} c_i h_i(k) x^{h_i(k)-\frac{1}{2}} \right) \left(\sum_{j=1}^{n} c_j h_j(k) x^{h_j(k)-\frac{1}{2}} \right) dx \\
&= 2\sum_{k=1}^{K} \int_0^1 \left(\sum_{i=1}^{n} c_i h_i(k) x^{h_i(k)-\frac{1}{2}} \right)^2 dx \\
&> 0
\end{aligned}$$

Thus Q is positive definite.

(Alternatively, one can show the positive definiteness of Q using properties of Hadamard products as follows. We begin by noting that Q can be written as a sum of K matrices of the form

$$\left[\frac{2x_i x_j}{x_i + x_j} \right]$$

where $x_i > 0$. That is to say, Q is a sum of matrices of harmonic means between all pairs of entries in $h_i(k)$ over all k. Using $\circ$ to denote the Hadamard (componentwise) product [2], this matrix can be rewritten as

$$\left[\frac{2x_i x_j}{x_i + x_j}\right] = [2x_i x_j] \circ \left[\frac{1}{x_i + x_j}\right]$$

The first matrix is positive definite since it is simply a constant times the outer product of x with itself. The second matrix is also positive definite since it is a Hilbert matrix [12]. By Schur's theorem [2], the Hadamard product of two positive definite matrices is also positive definite. Finally, since the sum of positive definite matrices is also positive definite [9], this establishes that Q is positive definite.)

Returning now to $e^{-\chi^2_{ij}}$, we note that it can be written as a positive constant times $e^{Q_{ij}}$. Since the exponential of a positive definite function is also positive definite [1], we have established that $e^{-\chi^2_{ij}}$ is positive definite.

References

1. C. Berg, J. P. R. Christensen, and P. Ressel. *Harmonic Analysis on Semigroups.* Springer-Verlag, New York, 1984.
2. R. Bhatia. *Matrix Analysis.* Springer Verlag, 1997.
3. O. Chapelle, P. Haffner, and V. Vapnik. SVMs for histogram based image classification. *IEEE Trans. Neural Networks*, 10(5):1055–1064, September 1999.
4. F. R. K. Chung. *Spectral Graph Theory.* Number 92 in CBMS Regional Conference Series in Mathematics. AMS, 1997.
5. T. H. Cormen, C. E. Leiserson, and R. L. Rivest. *Introduction to Algorithms.* MIT Press, 1991.
6. J. de Leeuw. Multidimensional scaling. UCLA Dept. of Statistics, Preprint no. 274, 2000.
7. C. Fowlkes, S. Belongie, and J. Malik. Efficient spatiotemporal grouping using the Nyström method. In *Proc. IEEE Conf. Comput. Vision and Pattern Recognition*, December 2001.
8. Yoram Gdalyahu, Daphna Weinshall, and Michael Werman. Stochastic image segmentation by typical cuts. In *CVPR*, 1999.
9. R.A. Horn and C.R. Johnson. *Matrix Analysis.* Cambridge Univ Press, 1985.
10. T. Leung and J. Malik. Contour continuity in region-based image segmentation. In H. Burkhardt and B. Neumann, editors, *Proc. Euro. Conf. Computer Vision*, volume 1, pages 544–59, Freiburg, Germany, June 1998. Springer-Verlag.
11. J. Malik, S. Belongie, T. Leung, and J. Shi. Contour and texture analysis for image segmentation. *Int'l. Journal of Computer Vision*, 43(1):7–27, June 2001.
12. R. Mathias. An arithmetic-geometric-harmonic mean inequality involving Hadamard products. *Linear Algebra and its Applications*, 184:71–78, 1993.
13. P. Perona and W. T. Freeman. A factorization approach to grouping. In *Proc. 5th Europ. Conf. Comput. Vision*, 1998.
14. J. Puzicha and S. Belongie. Model-based halftoning for color image segmentation. In *ICPR*, volume 3, pages 629–632, 2000.
15. J. Puzicha, T. Hofmann, and J. Buhmann. Non-parametric similarity measures for unsupervised texture segmentation and image retrieval. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 267–72, San Juan, Puerto Rico, Jun. 1997.

16. B. Schölkopf and A. Smola. *Learning with Kernels.* Cambridge, MA: MIT Press, 2001. in preparation.
17. B. Schölkopf, A. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10:1299–1319, 1998.
18. J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(8):888–905, August 2000.
19. Y. Weiss. Segmentation using eigenvectors: a unifying view. In *Proc. 7th Int'l. Conf. Computer Vision*, pages 975–982, 1999.
20. C. Williams and M. Seeger. Using the Nyström method to speed up kernel machines. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 13: Proceedings of the 2000 Conference*, pages 682–688, 2001.