Published in final edited form as:
Conf Comput Vis Pattern Recognit Workshops. 2010 ; : 79–86. doi:10.1109/CVPRW.2010.5543453.

Two-Point Correlation as a Feature for Histology Images: Feature Space Structure and Correlation Updating

Lee Cooper,
Center for Comprehensive Informatics Emory University Atlanta, GA 30322

Joel Saltz,
Center for Comprehensive Informatics Emory University Atlanta, GA 30322

Raghu Machiraju, and
Department of Biomedical Informatics Ohio State University Columbus, OH 43210

Kun Huang
Department of Biomedical Informatics Ohio State University Columbus, OH 43210

Lee Cooper: Lee.Cooper@Emory.edu; Kun Huang: Kun.Huang@osumc.edu

Abstract

The segmentation of tissues in whole-slide histology images is a necessary step for the morphological analyses of tissues and cellular structures. Previous works have demonstrated the potential of two-point correlation functions (TPCF) as features for tissue segmentation, however the feature space is not yet well understood and computational methods are lacking. This paper illustrates several fundamental aspects of TPCF feature space and contributes a fast algorithm for deterministic feature computation. Despite the high-dimensionality of TPCF feature space, the features corresponding to different tissues are shown to be characterized by low-dimensional manifolds. The relationship between TPCF and the familiar co-occurrence matrix is highlighted, and it is shown that costly cross correlations are not necessary to achieve an accurate segmentation. For computation, the method of correlation updating, based on the linearity of the correlation operator, is proposed and shown to achieve up to a 67X speedup over frequency domain computation methods. Segmentation results are demonstrated on multiple tissues and natural texture images.

1. Introduction

The adoption of digital slide scanning technologies in clinical and research settings is providing Terabytes of high-resolution histology imagery. This data contains a potential wealth of information that can be used to perform or large-scale comparative or correlative analysis of tissue morphologies versus patient outcome or genomic features. A key challenge in the effort to extract this information is the segmentation of tissues in whole-slide images which present themselves as complex arrangements of cellular structures. A popular approach to this problem has been to apply texture based segmentation methods [3, 12, 16]. Conceivably if distinct tissues are represented by different organizations of components such as cell nuclei, cytoplasm, and extracellular matrix then texture measurements can be applied to discover these distinct signatures.

Previous work has shown that popular texutre features such as Haralick features and Gabor filters are insufficient to distinguish the subtle differences in some tissue layers [12]. Instead, a new class of segmentation features, the two-point correlation functions (TPCFs), were proposed and demonstrated effective in difficult scenarios [13]. Despite this advance, problems remain with existing TPCF-based methods. Previous works have all utilized

Monte-Carlo calculation methods without addressing difficult sampling considerations; uncertainty in sampling requirements leads to large sample sizes and increased execution times. For segmentation, due to the high-dimensionality of TPCF features, clustering methods were applied without regard for structure in TPCF feature space. This is significant since the feature space structure informs the method of segmentation or classification in many applications [10, 11].

This paper contributes several results of practical interest on the structure of TPCF feature space, and also a new fast and deterministic method for TPCF feature computation. We show that despite its high dimensionality, TPCF feature space is characterized by remarkably smooth and low-dimensional manifolds. Additionally, we show that costly cross-correlation terms are not necessary to achieve accurate segmentations, and highlight the links between TPCF and the familiar co-occurence matrix. For fast calculation of TPCF features, we propose a deterministic method called *correlation updating*, that uses the linearity of the correlation operator for iterative calculation of features with shared neighborhoods. Experimental results show that correlation updating provides up to a 67X speedup over a simple frequency-domain based implementation, significantly reducing the computational burden for processing giga-pixel digital microscopy images.

The paper is organized as follows: Section 2 provides preliminary and background information on two-point correlation functions. Section 5 describes the use of TPCF as features in image segmentation. Section 3 presents the segmentation and TPCF feature space results on both synthesized image and histological image. Summary and conclusions are provided in Section 6.

2. Two Point Correlation Function

The two-point correlation function originates from the field of statistical geometry where it has been used to study extremes of scale in both materials science [15] and cosmology [18]. Its power in predicting physical phenomena suggests a rich representation of spatial information [7], and lead to its adaptation in computer vision and pattern recognition [6, 12, 13]. This section provides relevant background on the two-point correlation function, a more detailed description can be found in the text by Torquato [15].

2.1. Phase Label Images

The term *phase image* is defined here to describe an image composed of discrete constituents. The phase image I with P phases is a scalar field, partitioned into P exhaustive and disjoint regions $\mathscr{V}_i$. For the purpose of development, assume I is a random entity in sampling space Ω, and that $\omega \in \Omega$ is one realization. For each phase i, an indicator function is defined for $\mathbf{x} = (x, y) \in \mathbb{R}^2$ in I

$$\mathscr{I}^{(i)}(\mathbf{x},\omega)=\begin{cases} 1, & \mathbf{x} \in \mathscr{V}_i(\omega) \\ 0, & \text{else} \end{cases} \quad (1)$$

In practice, the interpretation of phase is application specific. Phases could represent objects in a natural scene, or different cell types in a microscopic image. The flexibility in defining phase implies generalizability of the segmentation framework beyond microscopic tissue images. Phase is discussed further in Section 3.

2.2. n-Point Correlation Functions

Given the set of indicators $\mathscr{I}^{(i)}(\mathbf{x},\omega)$, the n-point correlation function $S_n^{(i)}$ is defined as the probability of finding n points $\mathbf{x}_1, \mathbf{x}_2,\ldots, \mathbf{x}_n$ in phase i

$$S_n^{(i)}(\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_n) \equiv E\left\{\mathscr{I}^{(i)}(\mathbf{x}_{(1)}\mathscr{I}^{(i)}(\mathbf{x}_2)\cdots\mathscr{I}^{(i)}(\mathbf{x}_n)\right\}$$
$$=\Pr\left\{\mathscr{I}^{(i)}(\mathbf{x}_1)=1,\mathscr{I}^{(i)}(\mathbf{x}_2)=1,\cdots,\mathscr{I}^{(i)}(\mathbf{x}_n)=1\right\}.$$

Of particular interest is the *two-point correlation function* (TPCF)

$$S_2^{(i)}(\mathbf{x}) \equiv \mathrm{E}\left\{\mathscr{I}^{(i)}(\mathbf{x}_1)\mathscr{I}^{(i)}(\mathbf{x}_2)\right\}. \quad (2)$$

If I is statistically homogeneous, S_2 is invariant under translation and depends only on $\mathbf{x}_{1,2} = \mathbf{x}_1 - \mathbf{x}_2$. If I is also statistically isotropic then S_2 is rotation invariant and depends only on distance $r = |\mathbf{x}_{12}|$. In this case the TPCF is denoted $S_2(r)$, and can be visualized as a variation of the familiar Buffon's Needle problem (Figure 1).

The assumptions of an isotropic and homogeneous random field are used for illustration. In practice, images are typically anisotropic and the TPCF is measured as a sample average under the isotropic assumption. This produces statistics that are insensitive to orientation, a property that is desirable in many classification and segmentation applications.

The relation in Equation 2 defines the two point *autocorrelation* function, between a phase i and itself. Similarly, the two point *cross-correlation* may be defined between phases i and j

$$S_2^{(i,j)}(\mathbf{x}) \equiv \mathrm{E}\{\mathscr{I}^{(i)}(\mathbf{x}_1)\mathscr{I}^{(j)}(\mathbf{x}_2)\}. \quad (3)$$

The methods presented in this paper exclusively use the auto correlation functions for image segmentation. Previous works have included cross correlation information for segmentation, but as demonstrated in Section 5 this is not always necessary to achieve reasonable segmentations in difficult settings.

2.3. Relationship to Co-Occurrence Matrix

The TPCF represents the probability that phases are separated by a given distance, and is closely related to a popular method for texture image analysis, the *co-occurrence matrix*. Perhaps most widely known for use in calculating the Haralick features [4], the co-occurrence matrix $C_{\mathbf{x}}$ represents the frequencies that image values i, j are separated by $\mathbf{x}$ in the intensity image G

$$C_{\mathbf{x}}(i,j)=\sum_m\sum_n\begin{cases}1, & G(m,n)=i,\\ & G(m+x,n+y)=j\\ 0, & \text{else.}\end{cases} \quad (4)$$

The diagonal frequencies of $C_{\mathbf{X}}$ are related to the sample TPCF of G through a normalization by the total comparisons in Equation 4

$$S_2^{(i)}(\mathbf{x})=\frac{C_{\mathbf{x}}(i,i)}{(N-x)(M-y)}, \quad (5)$$

where N and M are the horizontal and vertical image dimensions. Despite this relationship the use TPCFs for image segmentation is fundamentally different from cooccurrence based

approaches. The primary differences are the range of calculation, and the use of raw TPCF values as features rather than as an intermediate representation.

2.4. Sample TPCF Calculation

Given an $M \times N$ digital phase image $IS_2^{(i)}$ is calculated from the indicator autocorrelation

$$R^{(i)}(\Delta x, \Delta y) = \sum_m \sum_n \mathscr{I}^{(i)}(m,n)\mathscr{I}^{(i)}(m+\Delta x,\ n+\Delta y), \quad (6)$$

where $\Delta x, \Delta y \in \mathbb{Z}$. The correlation of indicators effectively counts the number of pixels of phase i separated by (Δx, Δy), e.g. (0, 0) represents full-overlap and R(0, 0) the pixel count of phase i in I. The values of R are normalized by the overlapping area at each lag to calculate probabilities

$$\widehat{R}^{(i)} = R^{(i)}./(1_{M\times N} * 1_{M\times N}), \quad (7)$$

where $1_{m\times N}$ is an $M \times N$ matrix of ones, ./ is element-wise division, and * is convolution.

The normalized elements of $\hat{R}$ represent the anisotropic but homogeneous TPCF $S_2^{(i)}(\mathrm{x})$. A process of *circumferential sampling* is used to calculate the isotropic $S_2^{(i)}(\mathrm{r})$ from $S_2^{(i)}(\mathrm{x})$. Samples taken at distance r from $\hat{R}^{(i)}$ (0,0) are averaged over angle

$$S_2^{(i)}(r) = \frac{\Delta\theta}{\pi} \sum_{k=0}^{\frac{\pi}{\Delta\theta}-1} \widehat{R}^{(i)}(r\cos(k\Delta\theta), r\sin(k\Delta\theta)), \quad (8)$$

where $\Delta\theta$ is the *angular interval*. This sampling procedure is depicted in Figure 1. Samples off the discrete grid of $\hat{R}^{(i)}$ can be inferred using bilinear interpolation. Due to the symmetry of $\hat{R}^{(i)}$, the sampling angles can be restricted to [0,π).

3. TPCF features for Image Segmentation

The TPCF segmentation workflow consists of four stages: phase labeling, TPCF feature calculation, dimensionality reduction, and feature clustering. The process begins with the identification of phases from a color or grayscale image to generate a phase label image. Feature vectors containing the TPCFs of each phase are then calculated for local regions in a sliding window, throughout the phase image. The TPCF feature vectors typically conform to low-dimensional manifolds, and so the dimensionality of the feature vectors is reduced prior to clustering in feature space. The clustered feature labels are then mapped back to the image domain and refined if necessary to eliminate edge effects and aberrations. Each of these stages is described in further detail below.

3.1. Phase Labeling

Given a color or intensity image, the phase labeling process assigns a label $i \in \{1, 2, \ldots, P\}$ to each pixel. The notion of phase is borrowed from the materials science community where it represents the different constituents in a composite material. In the imaging context phase is a flexible concept that provides a general approach to treating images as mixtures of constituents. These constituents can be identified by either low-level information such as intensity or color, or high-level information such as shape or size. Any number of mode-identifying segmentations such as mean shift [1] or K-means can be used to label constituents. A simple quantization may be effective if the color/intensity is relatively uniform. For high level information phase is certainly application specific since it likely

represents meaningful units e.g. distinct cell types in tissue. A more complex labeling approach that incorporates domain classifications is needed in this case.

3.2. TPCF Feature Vector Extraction

Define $\Phi(x, y)$ as the $w \times w$ sliding region-of-interest with upper left corner $I(x, y)$. The anisotropic sample TPCF is computed inside $\Phi(x, y)$ for $r = 0$ to $w/2$ and for each phase $i \in \{1, 2,\ldots, P\}$ to form the $P(w/2 + 1)$-dimensional feature vector

$$v_{x,y}=[S_2^1(0), S_2^1(1), \ldots, S_2^1(\frac{w}{2}), S_2^2(0), S_2^2(1), \ldots, S_2^2(\frac{w}{2}), \ldots, S_2^P(0), S_2^P(1), \ldots, S_2^P(\frac{w}{2})]^{\mathrm{T}}.$$

Feature vectors are computed at each position of the the sliding ROI $\Phi(x, y) \in \{0, 1,\ldots N - w\}\times\{0, 1,\ldots M - w\}$.

3.3. Dimensionality Reduction and Clustering

Although the feature vectors $v_{x,y}$ reside in $P(w/2 + 1)$ space, their energy is typically concentrated in relatively few modes. Prior to segmentation the dimension of the feature vectors is reduced by projecting $v_{x,y}$ onto the first D primary two-point functions obtained through principal component analysis.

To achieve a segmentation of the image the reduced dimension feature vectors are clustered in the feature space and the clustering result is mapped back to the image space to form a segmentation. The TPCF feature vectors tend to be either restricted to a smooth low-dimensional manifold or distributed among a mixture of low-dimensional linear structures, so we choose to use the unsupervised lossy coding method of [10].

4. Correlation Updating

Previous works using TPCF as a feature for segmentation relied on the use of Monte Carlo methods to estimate $S_2^{(i)}$. Using a simulation similar to Figure 2.4, a needle is repeatedly cast onto the phase image ROI and the endpoints recorded. This process raises the issue of sampling, which certainly depends on image characteristics. In contrast, the deterministic approach described above is exhaustive, effectively integrating information from all possible comparisons over the region. The deterministic method has been employed in the materials science community, where it is implemented using frequency-domain methods [8]. The application of image segmentation if fundamentally different, however, in that TPCFs are calculated in a small sliding window with significant overlap between adjacent regions. This section describes the existing frequency domain method, its weaknesses in segmentation, and proposes a new faster method for deterministic exhaustive calculation that exploits region overlap to reduce computation.

4.1. Frequency Domain Method

The most computationally demanding portion of the TPCF calculations are the correlations of Equation 6, that may be computed efficiently using the Fast Fourier Transform (FFT).

The binary mask $\mathscr{I}_{x,y}^{(i)}$ is padded to the size $2w - 1$

$$\mathscr{P}_{x,y}^{(i)} \equiv \begin{bmatrix} \mathscr{I}_{x,y}^{(i)} & \mathbf{0}_{w\times w-1} \\ \mathbf{0}_{w-1\times w} & \mathbf{0}_{w-1\times w-1} \end{bmatrix} \quad (9)$$

and transformed forward to the discrete frequency domain to obtain the spectrum $\mathscr{F}[k,l]$. The power spectrum is calculated by taking the magnitude of the complex elements and the inverse transformation is computed to obtain the autocorrelation R

$$R^{(i)}=\frac{1}{\sqrt{(2w-1)}}\sum_{l=0}^{2w-1}\sum_{k=0}^{2w-1}\mathscr{F}_{x,y}^{(i)}[k,l]e^{2\pi j\frac{mk+nl}{2w-1}}. \quad (10)$$

The dimension $2w-1$ is critical for the performance of the FFT calculations [2]. The padding of Equation 9 may be manipulated to achieve favorable sizes by adding zeros to achieve the next most favorable size.

4.2. Sparse Sampling

The FFT calculates all $(2w-1)^2$ elements of the autocorrelation R, however only a small set of these are required for the circumferential sampling procedure. This is apparent in Figure 2, where only 10% elements of $R^{(i)}$ are used to interpolate $S_2^{(i)}(\mathrm{r})$. Although algorithms exist for computing subsets of FFT outputs [5, 9], the available implementations of ordinary full-output FFT are optimized to the extent that only a relatively large transform will benefit [2].

4.3. Updating

In addition to the sampling sparsity, the shared content between neighboring ROIs also points to significant amounts of wasted computation. For example, although $\Phi_{x,y}$, $\Phi_{x+1,y}$ differ by only two w-length columns of pixels, a straight-forward FFT method calculates correlations from scratch for each.

The observations of sparsity and shared content may be simultaneously addressed using the linearity of correlation. Rather than computing $R^{(i)}$ from scratch for each ROI, the portions of neighboring ROIs, say $\Phi_{x,y}$ and $\Phi_{x+1,y}$, that are not shared may be used to update $R^{(i)}$ from $\Phi_{x,y}$ to $\Phi_{x+1,y}$ instead. Furthermore, if this updating is performed directly in the image domain then the updates can be restricted to those locations used in sampling.

Given two horizontally adjacent $w\times w$ ROIs $\Phi_{x,y}$, $\Phi_{x+1,y}$ with corresponding indicators

$$\begin{aligned}\mathscr{I}_{x,y}^{(i)}&=[c_x,c_{x+1},\ldots,c_{x+w-1}]\\ \mathscr{I}_{x+1,y}^{(i)}&=[c_{x+1},c_{x+2},\ldots,c_{x+w}],\end{aligned} \quad (11)$$

where c are w-length pixel columns. The autocorrelation of $\mathscr{I}_{x,y}^{(i)}$ is denoted $R_{x,y}^{(i)}$. Given that $I_{x,y}^{(i)}$, $I_{x+1,y}^{(i)}$ are distinguished only by c_x, c_{x+w}, the autocorrelation $R_{x+1,y}^{(i)}$ can be calculated from $R_{x,y}^{(i)}$ by adding the contribution of c_{x+w} and removing the contribution of c_x.

Define the correlation sums between the columns and their respective regions

$$\begin{aligned}a^{-}_{\Delta x,\Delta y}&\equiv\sum_m\mathscr{I}_{x,y}^{(i)}(\Delta x,m)c_x(m+\Delta y)\\ a^{+}_{\Delta x,\Delta y}&\equiv\sum_m\mathscr{I}_{x+1,y}^{(i)}(\Delta x,m)c_x+w(m+\Delta y).\end{aligned} \quad (12)$$

The *update matrices* containing these correlation sums represent the contributions of c_x to $R_{x,y}^{(i)}$ and c_{x+w+1} to $R_{x+1,y}^{(i)}$

$$A^- \equiv \begin{bmatrix} a^-_{w-1,w-1} & \cdots & a^-_{0,w-1} & a^-_{1,1-w} & \cdots & a^-_{w-1,1-w} \\ a^-_{w-1,w-2} & \cdots & a^-_{0,w-2} & a^-_{1,2-w} & \cdots & a^-_{w-1,2-w} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ a^-_{w-1,0} & \cdots & a^-_{0,0} & a^-_{1,0} & \cdots & a^-_{w-1,0} \\ a^-_{w-1,-1} & \cdots & a^-_{0,-1} & a^-_{1,1} & \cdots & a^-_{w-1,1} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ a^-_{w12;1,1-w} & \cdots & a^-_{0,1-w} & a^-_{1,w-1} & \cdots & a^-_{w-1,w-1} \end{bmatrix}$$

$$A^+ \equiv \begin{bmatrix} a^+_{0,1-w} & \cdots & a^+_{w-1,1-w} & a^+_{w-2,w-1} & \cdots & a^+_{0,w-1} \\ a^+_{0,2-w} & \cdots & a^+_{w-1,2-w} & a^+_{w-1,w-2} & \cdots & a^+_{0,w-2} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ a^+_{0,0} & \cdots & a^+_{w-1,0} & a^+_{w-2,0} & \cdots & a^+_{0,-0} \\ a^+_{0,1} & \cdots & a^+_{w-1,1} & a^+_{w-2,-1} & \cdots & a^+_{0,-1} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ a^+_{0,w-1} & \cdots & a^+_{w-1,w-1} & a^+_{w-2,1-w} & \cdots & a^+_{0,1-w} \end{bmatrix}$$

The relationship between the autocorrelations for adjacent regions is then

$$R^{(i)}_{x+1,y} = R^{(i)}_{x,y} - A^- + A^+. \quad (13)$$

This updating procedure clearly applies to vertically adjacent ROIs as well.

Since only a subset of R is sampled for S_2, the corresponding elements of A^+, A^- may be calculated individually. Each element will then require only $2w$ multiply-add operations from one ROI to the next. Numerical accuracy is not compromised within the updating procedure. Since R represents integer counts of pairwise-separations there is no error accumulation through repeated rounding operations. The updating procedure also provides flexibility in choosing w since performance is not subject to FFT size penalties.

5. Experiments and Results

Segmentation experiments were performed using both natural texture and histology images. Natural texture images from the Brodatz collection were used to compare TPCF features with raw co-occurrence and traditional Haralick features. Microscopic images of human follicular lymphoma and mouse placenta were used to demonstrate the ability of TPCF features to resolve tissue boundaries. Mouse placenta images were used to demonstrate the time performance of the correlation updating method as well.

5.1. Natural Textures

Three images were selected from the Brodatz collection and arranged as in Figure 3(a). This grayscale arrangement was quantized to two bits to produce a four phase image. Three sets of features were calculated: raw co-occurrence, Haralick, and TPCF. Each feature set was reduced to $D = 10$ using PCA and clustered using K-means with $K = 3$. Feature were calculated in a sliding window $w = 32$. TPCF features were calculated at distances $r = 0,1, \ldots, 16$ to generate 68-dimensional features. Raw Co-occurrence features are the unwrapping of $C_{\mathbf{x}}$ into a 16-dimensional vector, with $C_{\mathbf{x}}$ computed at $x = (0, d), (\lceil \sqrt{2}d \rceil, \lceil \sqrt{2}d \rceil), (d, 0), (-\lceil \sqrt{2}d \rceil, \lceil \sqrt{2}d \rceil)$ for $d = 1,2,\ldots,16$, and then averaged to form 256-dimensional features. The Haralick features of contrast, correlation, energy, and

homogeneity were calculated from the co-occurrence matrices averaged over angle to form 64-dimensional features.

The energy of the TPCF feature set is concentrated in relatively few modes, 88% in the first three (compare to 79% and 73% for co-occurence and Haralick respectively). Segmentation accuracies are comparable at 94.1%, 97.3%, and 96.6% for Haralick, co-occurrence, and TPCF respectively, with all errors occuring within $w/4$ of the texture boundaries. In this case the presence of cross-correlation information within Haralick and co-occurrence features does not offer significant benefit in terms of segmentation accuracy.

A three-dimensional visualization of the TPCF features using PCA and is presented in Figure 3(c) and (d). The features conform to a smooth manifold-like structure, with different regions representing different textures.

5.2. Follicular Lymphoma Segmentation

The segmentation of follicles in H&E stained lymphoma images presents a challenging tissue segmentation problem [14]. To test the performance of TPCF segmentation a small 5X resolution region of follicles was obtained from the Virtual Slidebox, hosted at the University of Iowa Pathology Department (see Figure 4). A gaussian mixture model was used to cluster the H&E stained region into four classes corresponding roughly to nuclei, cytoplasm, background, and extracellular matrix. Using an ROI size $w = 16$, length $r = 0,1, \ldots, 8$, and angular interval $\Delta\theta = \pi/8$ produced 36-dimensional feature vectors that were then reduced to ten-dimensional space. The reduced features were clustered using the lossy coding method with $\mathscr{E} = 0.01$ [10]. The resulting segmentations are shown in Figure 4. The follicle areas are clearly distinguished.

A visualization of the tissue groups identified by segmentation is shown in Figure 5. The visualization was obtained by projecting the TPCF features into three-dimensional space. The groups remain well-separated in this low dimensional space. Overall they appear as a mixture linear structure with each component corresponding to a separate group. The yellow, dark blue, and red group features representing follicles, perifollicular space, and blood vessels all lie near a two-dimensional surface. The light blue group features representing the edge group vary in a direction normal to this surface.

5.3. Placenta Layer Segmentation

TPCF-based segmentation was applied to a sequence of mouse placenta images in a study similar to [12], where the aim is to characterize the role of the RB gene in mouse development by analyzing variation in tissue morphology [17]. The placenta contains several tissue layers including labyrinth, spongiotrophoblast, trophoblast, and glycogen. The aim of the example segmentation application here is to distinguish the labyrinth layer from the spongiotrophoblast layer as they are the least distinctive pair of adjacent layers. This experiment contrasts with previous works in only autocorrelation TPCFs are utilized in segmentation, no cross correlation information between phases is included.

One 1000 × 1000 pixel area was selected from 25 placenta images to contain approximately half labyrinth layer and half other tissue layers. A Gaussian mixture model maximum-likelihood classifier was applied to these areas to classify the pixels into red blood cell, cytoplasm, nuclei, extracellular matrix and background. These classifications serve as five-phase images from which TPCF feature vectors are calculated. The parameters ROI size $w = 32$, length $r = 0, 1,\ldots, 16$, and angular interval $\Delta\theta = \pi/8$ produced 68-dimensional feature vectors that were then reduced to ten-dimensional space prior to clustering using KNN with one tile as training data. The resulting segmentations are shown in Figure 6. A reasonable segmentation is achieved despite the absence of cross correlation information. The top-left

corner indicates that the autocorrelation functions were even able to distinguish an error in the manual segmentation where giant cells from the spongiotrophoblast were included in the labyrinth.

5.4. Correlation Updating Performance

Experiments were performed to examine the performance gain of correlation updating over a direct FFT-based implementation. The above implementations were tested on a single node from the BALE system at the Ohio Supercomputer Center, equipped with an AMD Opteron 2218 CPU and 8GB DDR2 DRAM. The data used to compare these implementations consists of ten 1000 × 1000 five-component images of H&E stained mouse placenta. TPCF features were calculated for the ten test images using the parameters of Table 5.4, chosen to reflect typical application values. In the power of two cases the $2w - 1$ DFT was padded to $2w$. The transforms for the non power of two cases were not padded to a power of two. This choice is justified since a this padding is detrimental in the $w = 130$ case, and is only marginally beneficial for $w = 34$.

The average per-image execution times and speedup are presented in Table 5.4. There is a strong penalty for non power of two cases for the direct-FFT method. This penalty is absent for the correlation updating implementation. Overall there is a significant speedup for correlation updating, ranging from 7.9–67X. The larger speedup factors correspond to the non-power-of-two sizes due to the large penalty on FFT performance.

6. Conclusion and Discussion

This paper presents results on the structure of TPCF feature space and a new fast and deterministic algorithm for feature computation. The examples provided demonstrate that in spite of their high dimension, TPCF features can form to smooth manifolds and mixture linear structures in low-dimensional space. For histological images these low-dimensional structures represent distinct tissue regions that can be identified using clustering methods such as lossy coding clustering [10] that are well-suited to these arrangements. Future work on feature space structure will explore the space for a larger number of cases, and pursue applications including tissue recognition and synthesis in this space of texture manifolds.

Deterministic calculation of TPCF features avoids the complex sampling issues associated with a Monte Carlo calculation method, and permits exhaustive calculation of TPCF over a region of interest. An FFT-based method offers deterministic and exhaustive calculation but is accompanied by rigid requirements on w. The FFT method also neglects the sparse autocorrelation sampling pattern and the content shared between neighboring ROIs resulting in significant wasted computation. Correlation updating simultaneously addresses these considerations without any compromise of numerical accuracy. Using the linearity of correlation, the autocorrelation calculations can be updated from one ROI to the next, rather than computed from scratch. Furthermore, performing these updates directly in the image domain permits the sampling locations to be selectively updated, and frees the algorithm from the sensitivity to ROI size. The improvements of correlation updating result in a speedup from 8–67x over the direct-FFT method.

Acknowledgments

This work is partially supported by NIH R21 MH083264.

References

1. Comaniciu D, Meer P, Member S. Mean shift: A robust approach toward feature space analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2002; 24:603–619.
2. Frigo M, Johnson S. The design and implementation of fftw3. Proceedings of the IEEE. Feb.2005 93(2):216–231.
3. Funakubo, N. ICPR84. 1984. Region segmentation of biomedical tissue image using color texture feature; p. 30-32.
4. Haralick RM, Shanmugam K, Dinstein I. Textural features for image classification. Systems, Man and Cybernetics, IEEE Transactions on. 1973; 3(6):610–621.
5. Hu Z, Wan H. A novel generic fast fourier transform pruning technique and complexity analysis. IEEE Transactions on Signal Processing. Jan; 2005 53(1):274–282.
6. Janoos, F.; Irfanoglu, MO.; Mosaliganti, K.; Machiraju, R.; Huang, K.; Wenzel, P.; de Bruin, A.; Leone, G. Proceedings of IEEE International Symposium on Biomedical Imaging. 2007. Multiresolution image segmentation using the 2-point correlation functions. In; p. 300-303.
7. Jiao Y, Stillinger FH, Torquato S. Modeling heterogeneous materials via two-point correlation functions: Basic principles. Physical Review E (Statistical, Nonlinear, and Soft Matter Physics). 2007; 76(3):031110.
8. Jiao Y, Stillinger FH, Torquato S. Modeling heterogeneous materials via two-point correlation functions. ii. algorithmic details and applications. Physical Review E (Statistical, Nonlinear, and Soft Matter Physics). 2008; 77(3):031135.
9. Knudsen K, Bruton L. Recursive pruning of the 2d dft with 3d signal processing applications. IEEE Transactions on Signal Processing. Mar; 1993 41(3):1340–1356.
10. Ma Y, Derksen H, Hong W, Wright J. Segmentation of multivariate mixed data via lossy data coding and compression. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2007
11. Ma Y, Yang AY, Derksen H, Fossum R. Estimation of subspace arrangements with applications in modeling and segmenting mixed data. SIAM Rev. 2008; 50(3):413–458.
12. Mosaliganti K, Janoos F, Irfanoglu O, Ridgeway R, Machiraju R, Huang K, Saltz J, Lenoe G, Ostrowski M. Tensor classification of N-point correlation function features for histology tissue segmentation. Medical Image Analysis. 2009; 13(1):156–166. [PubMed: 18762444]
13. Ridgway, R.; Irfanoglu, O.; Machiraju, R.; Huang, K. Proceedings of the Microscopic Image Analysis with Applications in Biology (MIAAB) Workshop in MICCAI. Dec. 2006 Image segmentation with tensor-based classification of n-point correlation functions.
14. Sertel, O.; Kong, J.; Lozanski, G.; Catalyurek, U.; Saltz, JH.; Gurcan, MN. Computerized microscopic image analysis of follicular lymphoma. Vol. 6915. SPIE; 2008. p. 691535
15. Torquato, S. Random Heterogeneous Materials Microsctructure and Macroscopic Properties. Springer-Verlag; New York, NY: 2002.
16. Tosun AB, Kandemir M, Sokmensuer C, Gunduz-Demir C. Object-oriented texture analysis for the unsupervised segmentation of biopsy images for cancer detection. Pattern Recognition. 2009; 42(6):1104–1112.
17. Wenzel PL, Wu L, de Bruin A, Chong JL, Chen WY, Dureska G, Sites E, Pan T, Sharma A, Huang K, Ridgway R, Mosaliganti K, Sharp R, Machiraju R, Saltz J, Yamamoto H, Cross JC, Robinson ML, Leone G. Rb is critical in a mammalian tissue stem cell population. Genes and Development. Jan; 2007 21(1):85–97. [PubMed: 17210791]
18. Zhao D, Jing YP, Borner G. Pairwise velocity dispersion of galaxies at high redshift: Theoretical predictions. The Astrophysical Journal. 2002; 581(2):876–885.

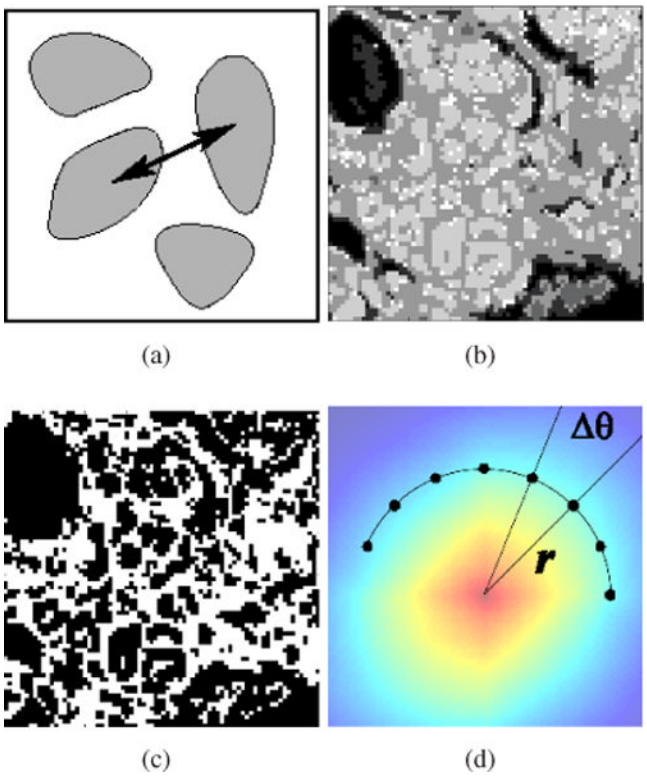


Figure 1.

Two point correlation function. (a) Placing line segments of length r with random orientation on ω, the fraction of times the endpoints both land in phase i represents $S_2^{(i)}(r)$. (b) A phase image contains pixels labeled according to phase. (c) Each phase has an associated indicator. The indicator autocorrelations are used in calculating $\{S_2^{(i)}(r)\}$ (d) Circumferential samples are averaged at radius r from $\hat{R}^{(i)}$ to calculate $S_2^{(i)}(r)$.

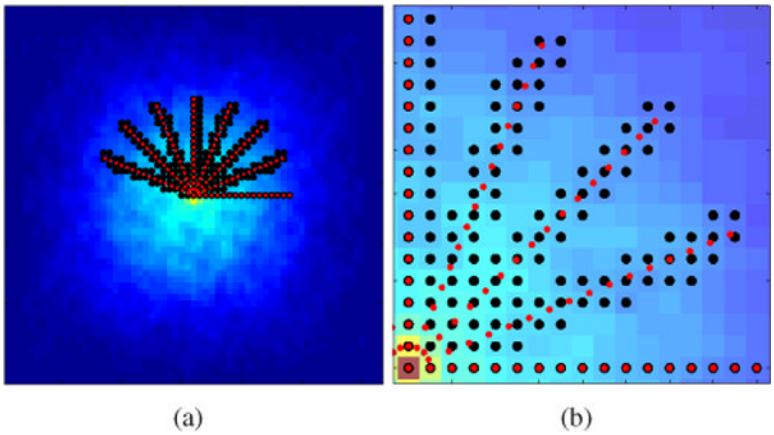


Figure 2.

Sparsity of circumferential sampling. (a) Only a small portion of the autocorrelation matrix is used for TPCF calculation. Here, $w = 32$ and $\Delta\theta = \pi/8$. In this case only 395 of 3969 total grid locations of R are used for interpolation. (b) Zoom of (a). Red indicates interpolation locations, black indicates on-grid autocorrelation locations required for bilinear interpolation.

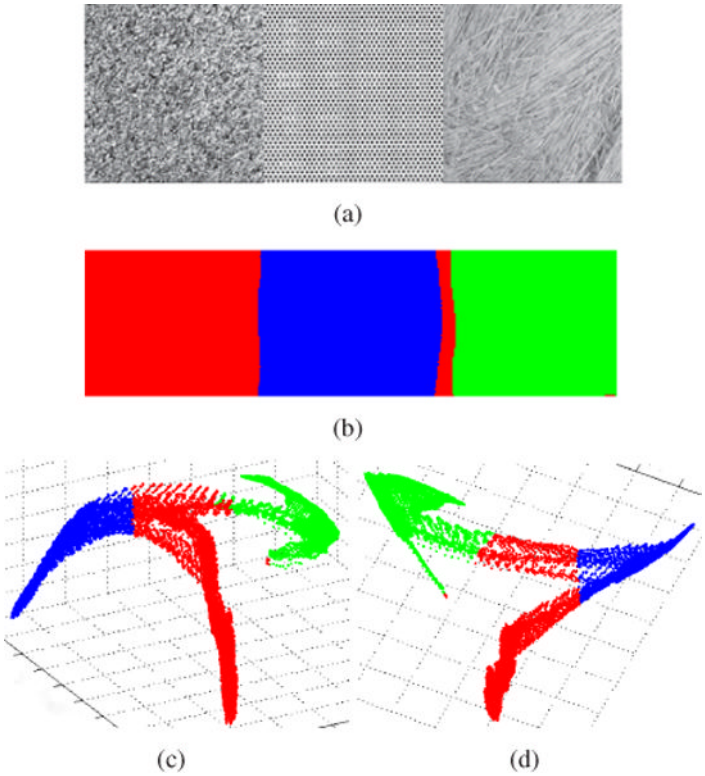


Figure 3.
Natural texture segmentation using TPCF. (a) Brodatz textures grass, holes, straw. (b) Segmentation map for TPCF features. (c), (d) TPCF features form a smooth manifolds in low-dimension.

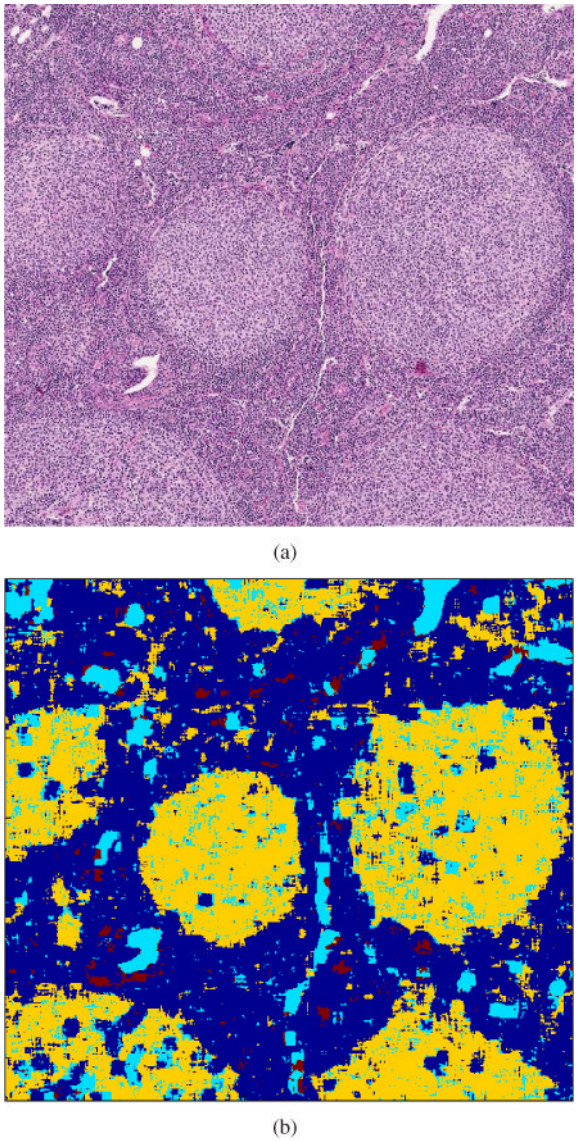


Figure 4.
Segmentation of follicles in lymphoma. (a) Follicle areas appear as large elliptical areas. (b) Follicles segmented using TPCF with lossy coding clustering.

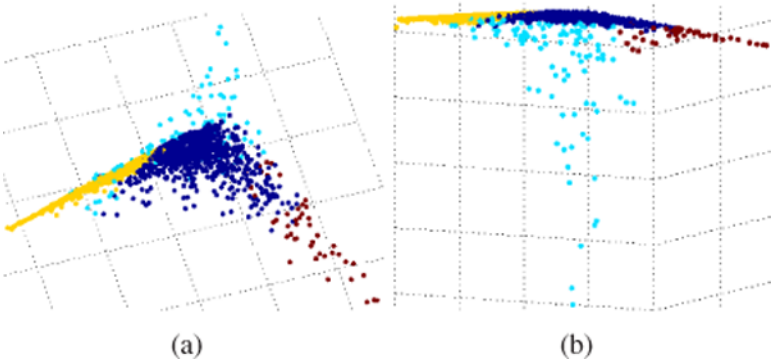
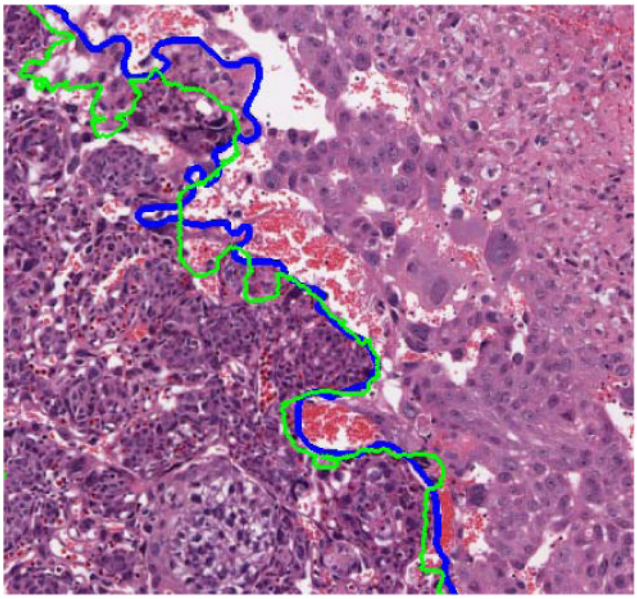
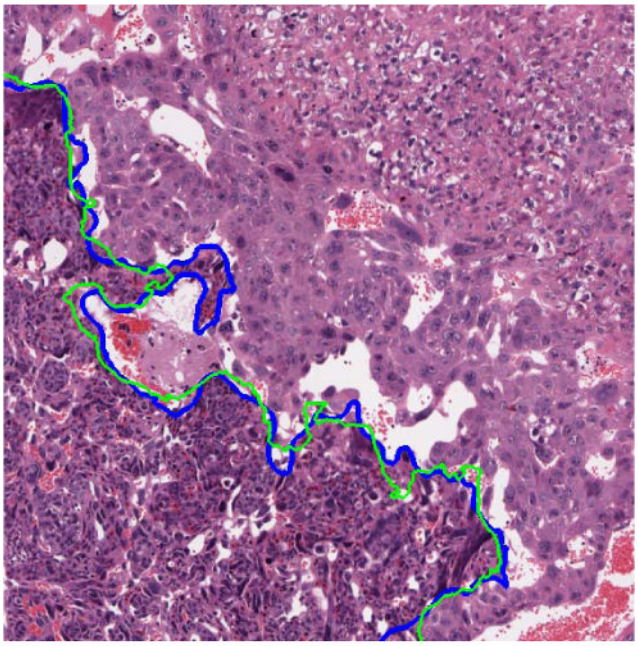


Figure 5.
Visualization of lymphoma TPCF features. The features exhibit a low-dimensional mixture distribution. Colors are coded to correspond with the classes of Figure 4.



(a)



(b)

Figure 6.
Placenta segmentation without cross correlation information. The blue line represents the manual segmentation. The green line indicates TPCF segmentation.

Table 1

Correlation updating test parameters, average execution times, and speedup.

case	small-pow2	small	large-pow2	large
w $\Delta\theta$	32 π/8	34 π/8	128 π/16	130 π/16
FFT(s)	1280	11637	43129	126489
updating(s)	162	178	3474	3557
speedup	7.9X	67.0X	12.4X	35.6X