Published in final edited form as:
IEEE Trans Med Imaging. 2011 March ; 30(3): 621–631. doi:10.1109/TMI.2010.2089693.

An optimal transportation approach for nuclear structure-based pathology

Wei Wang,
Center for Bioimage Informatics, Biomedical Engineering Department, Carnegie Mellon University, Pittsburgh, PA, 15213 USA

John A. Ozolek,
Department of Pathology, Children's Hospital of Pittsburgh, Pittsburgh, PA, 15201 USA

Dejan Slepčev,
Department of Mathematical Sciences, Carnegie Mellon University, Pittsburgh, PA, 15213 USA

Ann B. Lee,
Departments of Statistics and Machine Learning, Carnegie Mellon University, Pittsburgh, PA, 15213 USA

Cheng Chen, and
Center for Bioimage Informatics, Biomedical Engineering Department, Carnegie Mellon University, Pittsburgh, PA, 15213 USA

Gustavo K. Rohde
Center for Bioimage Informatics, Biomedical Engineering Department, Electrical and Computer Engineering Department, and Computational Biology Program, Carnegie Mellon University, Pittsburgh, PA, 15213 USA. Phone: 412-268-3684. Fax: 412-268-9580

Wei Wang: wwang2@andrew.cmu.edu; John A. Ozolek: ozolja@upmc.edu; Dejan Slepčev: slepcev@math.cmu.edu; Ann B. Lee: annlee@cmu.edu; Cheng Chen: chengchen@cmu.edu; Gustavo K. Rohde: gustavor@cmu.edu

Abstract

Nuclear morphology and structure as visualized from histopathology microscopy images can yield important diagnostic clues in some benign and malignant tissue lesions. Precise quantitative information about nuclear structure and morphology, however, is currently not available for many diagnostic challenges. This is due, in part, to the lack of methods to quantify these differences from image data. We describe a method to characterize and contrast the distribution of nuclear structure in different tissue classes (normal, benign, cancer, etc.). The approach is based on quantifying chromatin morphology in different groups of cells using the optimal transportation (Kantorovich-Wasserstein) metric in combination with the Fisher discriminant analysis and multidimensional scaling techniques. We show that the optimal transportation metric is able to measure relevant biological information as it enables automatic determination of the class (e.g. normal vs. cancer) of a set of nuclei. We show that the classification accuracies obtained using this metric are, on average, as good or better than those obtained utilizing a set of previously described numerical features. We apply our methods to two diagnostic challenges for surgical pathology: one in the liver and one in the thyroid. Results automatically computed using this technique show potentially biologically relevant differences in nuclear structure in liver and thyroid cancers.

Portions of the material in this paper overlap with portions of materials published by the same authors in the IEEE ISBI 2010 conference and in the 26th Souther Biomedical Engineering Conference (SBEC 2010).

Index Terms

Optimal transportation; nuclear structure; pathology; classification

I. Introduction

A. Motivation

Cancer is the second leading cause of death in the United States constituting 23% of all deaths [1]. Basic research has focused on uncovering molecular signatures of tumors and designing new therapies that target specific growth and signaling pathways [2], [3], [4]. Before therapy, however, an accurate diagnosis must be made. Despite advances in radiological imaging, a tissue diagnosis must be obtained using increasingly minimally invasive procedures with the surgical pathologist playing a critical role in this process. Within small tissue samples (needle biopsies and fine needle aspirations), diagnostic information can potentially be lost (microarchitecture, relationships to other structures) and the pathologist then relies heavily on cellular features (cytoplasmic and nuclear) and expensive ancillary techniques (special stains, immunohistochemistry, molecular diagnostics) for a correct diagnosis [5][6].

Surgical pathologists use visual interpretation of nuclear structure to distinguish cancer from normal, benign, and pre-malignant tissue [7]. Many tumors have certain characteristic nuclear appearances or features that clearly aid in narrowing the differential diagnoses (e.g. Langerhans cell histiocytosis, papillary carcinoma of the thyroid). Aberrations in the genetic code and the transcription of different messenger RNAs lie at the heart of transformation from normal to pre-malignant and malignant lesions. These changes occur in the nucleus and are accompanied by the unfolding and repackaging of chromatin that in part or in whole produces changes in nuclear morphology (size, shape, membrane contours, the emergence of a nucleolus, chromatin arrangement, etc.). Figure 1C shows nuclei depicting the complex variation in nuclear structure and chromatin distribution that can occur. Nuclei can be big, small, round, elongated, bent, etc. Cells can have their chromatin distributed uniformly inside the nucleus, along its borders, concentrated into small regions, anisotropically distributed, and with any combination of the above. It has long been known that this information defines phenotypes that are associated with important biological processes, including cancer [8].

We propose a new approach to describe the distribution of nuclear structure in different tissue classes. In contrast to most previous works, in which each nucleus image is reduced to a set of numerical features[9], [10], [11], we utilize a geometric approach, which interprets the data as distribution over carefully constructed mathematical geometries, to quantify the similarity of groups of nuclei (see section III for more detail). Beyond simple automated classification, our approach seeks to provide a visual representation of the nuclear morphometry that characterizes and differentiates normal, premalignant, and cancerous populations of cells. Moreover, instead of seeking to analyze single nuclei, our goal is to describe a method to characterize a distribution of cells of a given tissue, since this distribution may hold important diagnostic clues. In this work we focus on distinguishing lesions within two tissues: one in the liver and one in the thyroid. However, we believe our approach could be used for characterizing nuclear structure of different cancers in different tissues.

B. Previous work on automated digital pathology

In part due to well documented limitations of the human brain and visual system [12], [13], computational approaches have emerged as powerful tools for reproducible and automated

cancer diagnosis based on digital histopathology images. For decades, numerous papers have been published using computational methods to separate diagnostic entities, and some commercial software packages have been developed to screen for cancer cells with varying degrees of success [14]. The overwhelming majority of computational approaches follow a standard feature-based procedure where an image can be represented by a set of numerical features (see [9], [10], [11] for reviews). These methods can be described as a processing pipeline consisting of: image preprocessing (normalization, segmentation), feature extraction, and classification of the state of the tissue (e.g. normal or diseased) (see [14], [15], [16], [17], [18] for a few examples). In addition to these works, and although not directly related to the problem of pathology, we also mention the works of Yang et al [19] and Mangoubi et al [20] on measuring and quantifying chromatin and other nuclear components in time-lapse microscopy images.

These methods have been applied to the diagnosis of several types of cancers including prostate [21], cervix [14], [22], thyroid [23], [24], [25], [26], [27], [28], [29], [30], liver [31], [32], [33], [34], breast [35], and several others. While successful in some cases (see our earlier work [36] where we have applied such an approach to some of the same data used in the results shown below), feature-based methods have some important limitations. First, although classification can be accomplished in some cases, it is at times difficult to obtain useful and relevant biological information from such methods. This is due to the fact that when classifiers are used in multidimensional feature spaces, they rely on combinations (linear or nonlinear) of features each with different units, making physical interpretation notoriously difficult. Secondly, because the operation is usually not reversible, the reduction of each image to a set of features results in compression of information. In this context information from the digital image that may ultimately have diagnostic or biological significance is discarded.

In this paper we describe a geometric approach for classifying and understanding nuclear distributions without first reducing each nucleus to a set of features. Similar techniques have been applied to medical imaging problems at the macroscopic scale where the goal is to build statistical models of different organs (see [37], [38], [39], [40], [41],[42] for a few examples). The main idea in these works is to understand the anatomical variation of organs such as the brain or heart in human populations through analysis of the deformation fields required to warp one anatomy (as depicted in a radiology image) onto another, often with the principal component analysis technique. We explore a similar idea, but with focus on describing nuclear distributions of different tissue classes.

C. Overview of our contribution: a geometric framework for nuclear morphometry using Optimal Transportation

We describe a new technique for nuclear chromatin morphometry and pathology that utilizes the optimal transportation (OT) metric for quantifying the distribution of nuclear morphometry of different tissue classes. Classification of sets of nuclei is achieved with a kernel support vector machine approach, utilizing the distances given by the OT metric, in combination with a majority voting procedure. Distributions of nuclei are characterized and differentiated utilizing the Fisher Discriminant Analysis, in conjunction with the Multidimensional Scaling technique applied to distances computed using OT. Results show that the performance of a classifier using OT distances alone performs at least as well as the same classifier utilizing distances derived from numerical features. In addition, we show that our approach complements traditional feature-based approaches in that combining both OT and numerical-feature derived distances can measurably increase classification accuracy. Finally, we provide results characterizing differences and similarities between the nuclear structure of normal cells and different cancer cells in the liver and thyroid.

II. Data and pre-processing

A. Tissue processing and imaging

Tissue blocks were obtained from the archives of the University of Pittsburgh Medical Center (Institutional Review Board approval #PRO09020278). Cases for analysis included five resection specimens with the diagnosis of follicular adenoma of the thyroid (FA) and five cases of follicular carcinoma of the thyroid (FTC). For the other diagnostic challenge, five cases of fetal-type hepatoblastoma (FHB), a tumor of the liver in pediatric patients, and five cases of normal liver were compared. We refer these cases as “diagnostic challenges”, because, within these categories of lesions, the individual diagnostic entities can be difficult to sort from one another by visual methods and usually require additional consultation and testing to determine a diagnosis, particularly on needle biopsy or cytology specimens. The diagnostic challenge of thyroid represents a group of lesions that currently does not use the nuclear features to separate the two entities but rather requires extensive tissue sampling to look for the presence or absence of certain diagnostic microarchitectural features to separate the benign lesion (FA) from the malignant one (FTC). The diagnostic challenge of liver manifests itself more so when a small biopsy of a liver mass is taken and the pathologist must distinguish whether or not the lesion has been sampled (normal liver versus tumor) and then be able to render a diagnosis of FHB. In this case, the distinction between normal liver and FHB is the primary challenge. For each case of the thyroid and liver, nuclei from normal appearing tissues (denoted NL) were also extracted. Tissues were procured at the time of a surgical procedure, and then chosen for our analysis retrospectively over a several year span. All tissues were fixed in 10% neutral buffered formalin and processed by routinely used methods on a conventional tissue processor using a series of graded alcohols and xylenes prior to paraffin-embedding. Tissue sections were cut at 5 micron thickness from the paraffin-embedded block and stained using the Feulgen technique which stains DNA only. This approach has been used in other morphometric studies to specifically isolate nuclei for computational analyses [43], [44] and in our experience allows for more accurate segmentation of the nucleus compared to hematoxylin and eosin, hematoxylin alone, or periodic acid-Schiff stained sections. Counterstaining was not performed to avoid possible interference from the cytoplasm with accurate isolation and segmentation of nuclear membrane boundaries. Only nuclei were stained with a deep magenta hue (see Figure 1A for a sample image).

All images used for analysis in this study were acquired using an Olympus BX51 microscope equipped with a 100X UIS2 objective (Olympus America, Inc., Central Valley, PA) and 2 mega pixel SPOT Insight camera (Diagnostic Instruments Inc., Sterling Heights, MI). Image specifications were 24bit RGB channels and 0.074 microns/pixel, 118 × 89 microns field of view. Slides were chosen by the pathologist (J.A.O.) that contained both lesion (FHB, FA and FTC) and adjacent normal appearing tissue (NL). For each case, between 10 and 20 random fields were imaged to guarantee that at least 200 nuclei were obtained, for both lesion and normal tissue. Nuclei were chosen (using a single mouse click) by the pathologist and engineer (W.W.) for segmentation and analysis that demonstrated a complete and intact nuclear membrane within the focal plane.

B. Segmentation and intensity normalization

Nuclear segmentation consisted of the following three-step procedure. First, a random field graph cut method [45] was utilized to find a near global optimal segmentation, in a computationally efficient manner, which incorporates both region and boundary information. Briefly, the image segmentation problem was formulated as a pixel labeling problem, while the image data was modeled as a Markov Random Field. An energy function can be found to judge the quality of segmentation. This function can be expressed using a

graph structure, and the min-cut of the graph corresponds to an optimal segmentation. Secondly, an efficient level set active contour algorithm [46] is used to refine the initial segmentation (obtained via graph cut) towards the boundary (estimated with the finite difference first derivative) of the nuclei, while constraining a certain smoothness of the final result. The corresponding parameters used were set as $\sigma = 4$, $\varepsilon = 3$, $\lambda = 5$ and $\nu = 1.6$. In the end, the pathologist (J.A.O.) visually inspected all segmented nuclei for quality of segmentation to ensure a circumferential and sharply delineated nuclear membrane. In addition, nuclei were chosen from cells that represented the tissue of interest (tumor or normal) excluding other cells (inflammatory cells, cells containing hemosiderin pigment that obscured the nucleus, biliary epithelial cells (for liver cases), perifollicular cells (for thyroid cases), etc.). In total, using the above criteria, approximately 40% of segmented nuclei were included for analysis. A typical segmentation result is shown in Figure 1B. As our focus is on the analysis of nuclear morphometry, we did not investigate the nuclear segmentation problem extensively and point out that several other approaches already described in the literature could be used [47].

Images containing individual nuclei were converted to grayscale by selecting the green channel from the RGB images, and inverting the intensity values such that a zero (color coded in black) corresponds to the relative minimum amount of chromatin in the nucleus. We note that selecting magenta channel in the CMYK color space yielded very similar results. All nuclei were normalized so that the sum of their intensity values is 1. This was done to guarantee that nonuniformities related to staining and image acquisition, from case to case, were not able to interfere with our method. In total, we extracted 871 normal thyroid nuclei, 489 follicular adenoma and 703 follicular carcinoma nuclei from the thyroid data set. In addition, 461 fetal-type hepatoblastoma and 396 normal liver nuclei from were extracted from the liver data set. A few sample nuclei chosen for the entire data are displayed in Figure 1C.

C. Pre-processing

Nuclei images were pre-processed as in our previous works [48], [49] to eliminate, approximately, variations due to arbitrary rotation, translation, and coordinate inversions of each nucleus. The procedure includes normalization by the center of mass, rotation by major axis reorientation, and coordinate “flips” set up within a least squares minimization problem (see [48], [49] for more details). It is important to note that the objective is to make the distance measurements described below invariant with respect to the uninteresting variations mentioned above. To the best of our knowledge, there is currently no efficient algorithm that can make our metrics invariant under Euclidean transformations.

III. Methods

A. Optimal transportation for comparing nuclear chromatin

We believe the OT metric can capture some of the important information that characterizes the differences in nuclear structure in different cells (see Figure 1C for a few examples, and subsection I-A for their description). More precisely, we utilize the OT metric to quantify how much chromatin, in relative terms, is distributed in which region of the nucleus. There are two benefits of using the OT framework: 1) it provides a distance for comparing two nuclei and 2) provides a shortest (in the OT sense) connection path (geodesic) between them. The distances are used to quantitatively compare two nuclei (and subsequently to classify sets of nuclei). The geodesics (interpolation between nuclei) are used to visualize the data (exemplified in Figure 2 conceptually and in Figure 3 with an actual example). We believe being able to visualize the data in such way (that is, counting, in relative terms, how many nuclei appear similar to each interpolated nucleus displayed on the bottom of Figure

3) is an important addition to the field of pathology that is currently not available through other methods.

Here we describe the optimal transportation metric used for quantifying and classifying nuclear structure. We first do it in a general setting, and then apply it to discrete representations of the images considered. We note that the optimal transportation distance metric has been used in the past for different image analysis problems [50], [51].

Let Ω represent the domain (the unit square $[0, 1]^2$, for example) over which images are defined. Let us consider probability measures I_0 and I_1 on Ω. Recall that probability measures are nonnegative and that the measure of the whole set Ω is 1: $I_0(\Omega) = I_1(\Omega)$. In application to images, the measure of a set is the sum of intensities over all pixels in the set. On the other hand, as customary when discussing optimal transport, we will often refer to the measure of a set as its mass. Let $c: \Omega \times \Omega \rightarrow [0, \infty)$ be the *cost function*. That is $c(x, y)$ is the "cost" of transporting unit mass located at x to the location y. The optimal transportation distance measures the least possible total cost of transporting all of the mass from I_0 to I_1. To make this precise, consider $\Pi(I_0, I_1)$, the set of all *couplings* between I_0 and I_1. That is consider the set of all probability measures on $\Omega \times \Omega$ with the first marginal I_0 and the second marginal I_1. More precisely, if $\mu \in \Pi(I_0, I_1)$ then for any measurable set $A \subset \Omega$ we have $\mu(A\times\Omega) = I_0(A)$ and $\mu(\Omega \times A) = I_1(A)$. Each coupling describes a *transportation plan*, that is $\mu(A_0 \times A_1)$ is telling one how much "mass" originally in the set A_0 is being transported into the set A_1.

We consider optimal transportation with quadratic cost:

$$c(x, y)=|x-y|^2.$$

The optimal transportation distance, also known as the Kantorovich-Wasserstein distance, is then defined by

$$d(I_0, I_1)=\left(\inf_{\mu\in\Pi(I_0,I_1)}\int_{\Omega\times\Omega}|x-y|^2 d\mu\right)^{\frac{1}{2}} \quad (1)$$

It is well known that the above infimum is attained and that the distance defined is indeed a metric (satisfying the positivity, the symmetry, and the triangle inequality requirements), see [52]. For the quadratic cost the space of probability measures is endowed with a structure of a Riemannian manifold [52]. This Riemannian manifold structure is needed to be able to consider paths and in particular the shortest path (i.e. geodesics) connecting any two probability measures, which, in our case, two images of nuclei in the space of images (e.g. in Figures 3,4,5,6,7). Moreover, one can use the geodesic path to interpolate between images I_0 and I_1 in a way consistent with the metric. Namely let μ be the minimizer of (1). For $\alpha \in [0, 1]$ consider the function $\pi_\alpha(x, y) = (1 - \alpha)x + \alpha y$. Then the images on the geodesic are given by $I_\alpha = \pi_{\alpha\#}\mu$, that is $I_\alpha(A) = \mu(\{(x, y): (1 - \alpha)x + \alpha y \in A\})$, with the convenient property that the OT distance between I_0 and I_α is given by $\alpha d(I_0, I_1)$.

In our application, each nuclear structure is represented in a gray level digital image (of size 192 × 192 pixels). Each image I containing one single nucleus can be represented as

$$I=\sum_{i=1}^{M} v_i \delta_{x_i} \quad (2)$$

where δ_{x_i} is a Dirac delta function at pixel location x_i, M is the number of pixels in image I, and v_i are the pixel intensity values. To accelerate the computation, we use a point mass approximation to model the chromatin distribution of each nucleus. In specific, we use Lloyd's weighted K-means algorithm [53] to adjust the position and weights of a set of $N < M$ particle masses to approximate the total intensity distribution of each nuclei. In all of the computations in this paper, $N \leq 800$. The number of particles N was chosen so that there is a good balance between accuracy and speed. Ideally we would like to set these to be the number of pixels in each nuclear image. However, a linear programming based implementation of the distance on a 192 × 192-size image would be impractical for this application. Therefore, the number of particles we chose was so that the average computational time between pairs of nuclei was roughly one minute. When more computational power is available, one could use more particles. The weighted K-means algorithm merges points if two clusters fall within the same pixel coordinate, and this is the reason that N is not fixed to one single number for all images. The problem has now been reduced to finding the OT distance between $I_0=\sum_{i=1}^{N_p} p_i \delta_{x_i}$ and $I_1=\sum_{j=1}^{N_q} q_i \delta_{y_j}$ with N_p and N_q the number of delta-masses chosen for representing images I_0 and I_1. The minimization problem in (1) then reduces to finding an $N_p \times N_q$ matrix $f = [f_{i,j}]$ with $f_{i,j} \geq 0$ which minimizes

$$\min_f \sum_{i=1}^{N_p} \sum_{j=1}^{N_q} c(x_i, y_j) f_{i,j}$$

subject to the constraints for all $j = 1, \ldots, N_q$, $\sum_{i=1}^{N_p} f_{i,j}=q_j$ and for all $i = 1, \ldots, N_p$, $\sum_{j=1}^{N_q} f_{i,j}=p_i$. We utilize Matlab's implementation of a variation of Mehrotras dual interior point method [51] to solve the linear program. The geodesic interpolation between I_0, and I_1 can be approximated by $\sum_{i=1}^{N_p} \sum_{j=1}^{N_q} f_{i,j} \delta_{((1-\alpha)x_i-\alpha y_j)}$ which we denote by I_α for $\alpha \in [0, 1]$.

B. Supervised classification

1) Kernel based support vector machines—From our previous experience with thyroid histopathology data [36], we have found that the support vector machine (SVM) method, when combined with a simple voting strategy, performed best when compared with other classification methods for determining the class of a given set of nuclei [36]. We describe the SVM that utilizes numerical features first, and then show how it can be adapted to utilize only pairwise (OT) distances only.

Given a training data set of images I_i, we can compute a set of n features (stored in vector format $X_i \in \mathbf{R}^n$) describing the morphological properties of the nucleus depicted in image I_i. See [36] for a complete description of the numerical features used in this work. In a two class problem, given the feature-label pairs (X_i, Y_i), $i = 1, \ldots, N$, where $Y_i \in \{-1, +1\}$, the support vector machine seeks to find linear hyperplanes (determined by parameters w and b) that best separate the data set in an "enlarged" space. It can be formalized as the solution of the following optimization problem [54]:

$$\underset{w,b,\xi}{argmin}\left\{\frac{1}{2}w^T w+C\sum_{i=1}^{N}\xi_i\right\}$$
$$subject\ to: Y_i(w^T\varphi(X_i)+b)\geq 1-\xi_i \quad (3)$$
$$\xi_i \geq 0$$

Because the data set is not always linearly separable, the ξ_i represents the distance of each error point i to its correct plane, and C is a penalty constant for the error term. φ is a fixed nonlinear mapping function (known as basis function) that extends training vectors X_i into a higher dimensional space $\varphi(X)$: $\mathbf{R}^n \mapsto \mathbf{R}^m$. This problem is usually solved in its dual representation [54], where the data always occur in pairs, with the aid of the so-called kernel function [55] $K(X_i, X_j) = \varphi(X_i)^T \varphi(X_j)$. In our work we utilize the radial basis function (RBF) kernel $K(X_i, X_j) = exp(-\gamma||X_i - X_j||^2)$, $\gamma \geq 0$ whenever numerical features are used. In order to utilize the OT distances described above, the kernel is modified as $K'(X_i, X_j) = exp(-\gamma' d_{OT}(I_i, I_j)^2)$, $\gamma' \geq 0$. We note that such a technique (replacing a kernel based on Euclidean distances for other distances) has been used previously [56], [57]. While we are unable to show mathematically that this replacement satisfies Mercer's condition (positive semi-definite K'), we have studied the issue empirically and in all tests the matrix $K'(X_i, X_j) = exp(-\gamma' d_{OT}(I_i, I_j)^2)$ was a positive definite matrix, thus allowing for the replacement in the kernel-SVM procedure. It is worth noting that even if the new kernel as modified above does not satisfy the positive semi-definite criterion, it can still be used in the SVM framework. In such case, however, the hyperplane found by the SVM procedure may not be optimal [57]. Finally, for multiple classes problems, we use "one-versus-all" strategy [58] to reduce the single multiclass problem into multiple binary problems. We then use a max-wins voting strategy to combine these binary results for classifying the test instance.

2) Cross validation—Cross validation is performed to select the optimal (C, γ) parameters, as well as to test the average classification accuracy of the system. We use a "leave-one-out" strategy to separate the data into training and testing sets, where data from one case is used for testing and the remaining cases are used for training the classifier.

In order to train a classifier, we used k-fold cross validation to further separate the training set into two parts and searched for optimal parameters (C, γ) that had the best accuracy in this k-fold cross validation. We set $k = 10$, and performed an exhaustive search for the two parameters. After the optimal parameters are selected, we use them to build the classifiers and evaluate their performance on the testing data.

C. Characterizing distributions of nuclei

The geodesics that connect the nuclear structures in the entire data set can be used to characterize and contrast the differences between different tissue classes. The idea is to interpret each nuclear structure as a point in the OT manifold and seek geodesics onto which the projections of nuclear exemplars from different tissue classes most differ according to some quantitative criterion (see Figure 2). The criterion we use is the one described by Fisher [59] and used in the Fisher Linear Discriminant Analysis (LDA) method. However, because explicit "coordinates" for each nuclear structure are not available (only pairwise distances) an Euclidean embedding for the data must be computed before Fisher LDA can be performed. Exemplar nuclei are chosen based on the output of this procedure and used to compute the geodesic over which the nuclear structure of different tissue classes can be approximated.

1) MDS for obtaining Euclidean embedding—Given a set of such multidimensional points (morphological exemplars), and their pairwise distances computed using the OT framework discussed above, multidimensional scaling (MDS) can be used to find a low dimensional "Euclidean" embedding of the data. Let $D_{m,n} = d^2(I_m, I_n)$, with $d(I_m, I_n)$ given by equation (1). The goal in MDS is to find a set of coordinates v_k, $k = 1, \cdots, N$ in an Euclidean space that preserves the OT distances computed [60] (more precisely their inner product). This task can be achieved by choosing L positive eigenvalues and corresponding eigenvectors of the matrix $G = -0.5(\mathrm{Id} - uu^T)D(\mathrm{Id} - uu^T)$, with $u^T = 1/\sqrt{N}(1, \cdots, 1)$, and Id representing the identity matrix. Let $\lambda_1, \cdots, \lambda_N$ represent the eigenvalues of G, arranged in decreasing order of magnitude, and with corresponding eigenvectors $g_1, \cdots, g_N$. The i^{th} component of vector v_k is given by $\sqrt{\lambda_i} g_i^k$. For a given L, the pairwise distance can be reconstructed by its Euclidean embedding $\tilde{D}_{m,n} = \|v_m - v_n\|$. As in [61] L is selected such that the residual variance $1 - R^2(\tilde{D}, D) \leq 0.1$, where $R(\tilde{D}, D)$ the correlation coefficient between these matrices.

2) Fisher LDA for discrimination—Once each nucleus in a given data set has been connected to an Euclidean coordinate through the MDS technique, we utilize Fisher's Linear Discriminant Analysis (LDA) technique [59] to compute the direction in this multidimensional Euclidean space (here denoted h) onto which the data from two classes, if projected, would differ most according to the metric $J(h) = \frac{h^T S_B h}{h^T S_W h}$ where S_B represents the "between classes scatter matrix", S_W represents the "within classes scatter matrix".

3) Computing projections in OT space—After finding the discriminating direction h by Fisher LDA, we can compute the projections of all the data on this most discriminant direction, and select the points with smallest and largest projections. The geodesic path linking these two extreme points (denoted here as I_α) can be computed as described in section III-A. The projection of the nucleus I_j over geodesic, interpolated by I_α, can be formalized as $\underset{\alpha}{argmin}\{d_{OT}(I_\alpha, I_j)\}$. This is computed by sampling the path at 11 points ($\Delta\alpha = 0.1$), computing the distance of each nucleus to be projected to all points in the path, and choosing the smallest distance.

IV. Results

Here we describe results obtained in analyzing nuclear structure in two different diagnostic challenges, one in the liver and the other of thyroid cancers. The data set is described in Section II. We begin by demonstrating a sample computation of geodesic path between two sample nuclei. We then show that the distances computed using the OT framework can be used to achieve similar accuracy to the traditional feature approach to this problem described in detail in [36]. Finally, we demonstrate how the OT framework described above can be useful to extract meaningful quantitative information depicting the differences (in a distribution sense) that allow the data to be automatically classified.

A. Computation of OT distances and geodesics

An example geodesic is shown in Figure 3. The larger images I_0, I_1 on top of are the real images chosen for this computation. The red dots placed on them are the final locations for the point mass approximation. The bottom strip shows the actual geodesic. The end images in this strip are the approximated versions of the images shown on top.

B. Classification accuracy comparisons

As a first step, it is beneficial to understand whether the OT metric can capture the morphological information necessary for distinguishing different classes. In a previous work [36] we have described a system that utilizes a combination of 125 numerical features (including shape parameters, Haralick features, and multi-resolution-type features) and an SVM classifier together with a simple majority voting strategy to classify sets of nuclei. For each diagnosis challenge, a classifier is trained (based on labeled data) to determine whether a single nucleus pertains to a normal or a lesion-type class. The class of a group of nuclei of unknown origin can be determined by classifying each individual nucleus from that group and selecting the class to which the majority of nuclei were assigned. Our previous work [36] shows that this system is capable of classifying some of the same data (the thyroid data) used in this paper with 100 % accuracy.

Critical to this performance is the average classification accuracy for individual nuclei. When using a majority voting procedure the overall accuracy for classifying a group of nuclei will follow, approximately, a binomial/hypergeometric distribution. In a two class problem, for example, if the average classification accuracy for each class is greater than 50%, then perfect classification accuracy on a per human case basis can be achieved by selecting sufficiently many nuclei from that patient. Our previous work [36] contains a few Monte-Carlo computations describing the approximate number of nuclei necessary for perfect classification of each case. After extensive testing and fine tuning, our feature-based classification system consisted of training an SVM classifier with all 125 features individually normalized by their standard deviation. We tested whether feature selection approaches could be used to improve on these classification accuracies, but the improvement was negligible.

The results of classifying individual liver and thyroid nuclei using RBF kernel based SVM methods for both features and OT metric are contained in Table I (liver) and Table II (thyroid) respectively. We note that all classification accuracies reported are averaged for all nuclei belonging to a human patient. We also note that both feature-based and OT-based classifiers are identical in their implementation. Since we are using the kernel SVM method described earlier, the only difference is in the actual distance (OT vs. feature-based normalized Euclidean distances). For liver cases, we randomly selected 500 nuclei from the entire 5 cases (evenly distributed between human cases and classes, 100 nuclei per case), and for thyroid, we randomly selected 1050 nuclei from 10 cases (105 per case). All results were computed using the leave one out validation strategy described early. We emphasize again that training and testing data never overlapped, and that nuclei pertaining to each human case were classified without using data from the same case for training.

In Table I, each row corresponds to a testing case, and the numbers correspond to the average classification accuracy (over all nuclei for each case) for Normal liver and Hepatoplastoma. The first column indicates the classification accuracy for feature-based approach; the second column indicates the OT metric; and the third column indicates the classification accuracy for combined metrics (see below). Similarly, Table II shows classification accuracies for the 10 thyroid cases, where case 1 to case 5 consist of FA and NL, and case 6 to 10 consist of FTC and NL. In Table II each row indicates a lesion from one testing case, and for each lesion, we separately report the percentage of nuclei classified as NL, FA and FTC either based on feature-based approach (shown in the first, second and third columns) or based on OT metric approach (shown in the forth, fifth and sixth columns). The seventh column indicates the combined accuracy (see below). The average accuracy for feature-based approach is NL 0.8057, FA 0.6172, FTC 0.5461, and the average accuracy for OT metric is NL 0.8057; FA 0.5935; FTC 0.64. These results show that OT-metric based classification performs as well as feature-based classification (slightly better on

average), and it is more robust in the sense that it has more discriminantion power in the most difficult cases to classify (e.g. thyroid cases 9 and 10).

In addition to individual classification accuracies with features and OT, we have also tested whether a linear combination of these two distances could improve upon the results of each individual metric. The combined kernel was chosen to be $K_c(X_i, X_j) = exp\{-\gamma[d_{OT}(I_i, I_j)^2 + v \times ||X_i - X_j||^2]\}$, $\gamma \geq 0$. Using the same cross validation strategy introduced in Section III-B2, we performed a two-level cross validation to select the parameters (C, γ, v) described earlier. The classification accuracies for this combined kernel are reported in the last columns of Table I and II. We can see that the accuracies always increase in all the thyroid cases, as well as for most of the liver cases. Although 15 human cases is not an extensive data set, we can conclude that the OT and feature-based metrics contain complimentary information as far as this data set is concerned. The complimentary information could be used in conjunction to produce a classification method that, on average, performs better than either metric alone.

C. Characterizing distributions of nuclei

We use the automatic method described in section III-C to identify discriminant geodesic projections for liver and thyroid cases, shown in the bottom of Figure 4 and 5 respectively. Results suggest that, according to the available data, the most important information for discriminating between NL and FHB is the amount, in relative terms, of chromatin concentrated towards the border of the nucleus. The histogram shown in Figure 4 suggests that it is uncommon for FHB nuclei to have a chromatin distribution concentrated exclusively at the nuclear periphery. In thyroid cases, since it is a 3-class classification problem, we use Fisher LDA to find direction that best separates normal vs neoplastic (combining FA and FTC). We find, as shown in Figure 5, that the most discriminant information for differentiating populations of normal and neoplastic thyroid nuclei is size. For example, normal thyroid nuclei are relatively smaller than nuclei in the thyroid neoplasms (FA and FTC); the size of FA nuclei are accumulated to a specific size region, while the FTC nuclei are more evenly distributed in terms of size. We also used Fisher LDA to find the most discriminant direction only for FA and FTC, and we find that the most discriminant information is also size (results not shown).

Finally, the OT framework allows a user to interact manually with the data, and explore *a priori* hypotheses relating to nuclear structure in different tissue classes. For example, the geodesic shown in Figure 6 represents the difference in nuclear chromatin distribution from nearly uniform concentration to chromatin accumulated exclusively along peripheral region of the nucleus. From the histogram of the thyroid cases, we can observe that normal thyroid nuclei, in relative terms, are mostly smooth, while the chromatin distributions of neoplastic nuclei (FA and FTC) tend to distribute more evenly in these two patterns. The geodesic shown in Figure 7 shows another chromatin distribution pattern: from smooth texture to chromatin highly accumulated in the center of the nucleus. Its histogram suggests that the chromatin distributions of neoplastic nuclei (FA and FTC) tend to be more centrally located.

A natural question to ask, in particular for the projections computed by selecting interesting nuclei manually, is whether the projections contain statistically meaningful information. This question can be answered by testing whether or not the projections themselves can be used to classify the data. We have also tested this idea by performing a similar leave one out cross validation strategy, where training consists of computing the histogram projection distribution. An unlabeled case is then classified by first projecting the available data along the same geodesic, and then finding the closest match for histograms (in the L_2 sense) obtained from the training step. For liver cases, the geodesic shown in Figure 4 can correctly classify all the NL and FHB cases. Compared with liver cases, thyroid cases are harder to

classify just based on individual geodesics. The geodesic shown in Figure 5 can classify all the NL cases correctly, but misclassifies 3 cancer cases, for a total classification accuracy of 17/20 groups (including also the normal samples). The geodesic shown in Figure 6 can classify all the NL cases correctly, but misclassify 5 cancer cases (total accuracy of 15/20, including normal samples). The geodesic shown in Figure 7 can classify all the NL cases correctly, but misclassify 4 cancer cases (total classification accuracy of 16/20 groups).

V. Discussion and conclusions

We described an approach for automated digital pathology based on nuclear structure that is complementary to existing feature-based strategies, in particular when it comes to visualizing data distributions. The approach is based on quantifying chromatin morphology in different tissues classes (normal, cancer A, cancer B, etc.) using the optimal transportation (Kantorovich-Wasserstein) metric between pairs of nuclei. These distances are utilized within a supervised learning framework to build a classifier capable of determining the tissue class to which a particular set of nuclei belongs. We compare our approach to the standard feature-based classification approach using image data from a total of 15 human cases. Results show that on average, the optimal transportation metric performs as well or better than a popular feature-based implementation. In all 15 human cases the individual nuclei classification accuracies allow 100 % classification accuracy of the data, as long as multiple nuclei are used in a voting procedure [36].

In addition to automated classification we also describe how optimal transportation-based geodesic paths can be used to summarize differences between the nuclear structure (chromatin distribution) of different tissue classes. The approach involves computing the pairwise distances between all nuclei in the data set and using the MDS technique to find an inner product preserving Euclidean embedding for the data. Fisher LDA is then applied to discover the modes of variation that are most responsible for distinguishing two classes of nuclei. Once the variation, in the form of a optimal transportation geodesic, is computed, a projection of the data can be used to visualize the main differences in chromatin configuration in two or more tissue classes.

We demonstrated that the geometric framework proposed can be used to discover potentially meaningful biological or diagnostic information in liver and thyroid cancers. In many differentiated cells, heterochromatin is associated with the nuclear lamina at the nuclear periphery [62]. However, in cancer cells, this compact peripheral staining is lost in lieu of a more uniform or open chromatin pattern (euchromatin) indicating areas of transcriptional activity. The data suggests that this loss of heterochromatin may be related to the cancerous phenotype itself. Cancer progression often is associated with epigenetic changes including loss of heterochromatin with concomitant increase in transcription of proteins involved in numerous signaling pathways [63], [64]. Our results in Figure 4 show that it is uncommon for FHB nuclei to have a chromatin distribution concentrated exclusively at the nuclear periphery. The compact, dense chromatin (heterochromatin) seen in both the normal thyroid and liver nuclei suggests greater areas of relative transcriptional inactivity than their malignant/neoplastic counterparts. The geometric approach thus provides a new and useful tool to enable visualization of changes in nuclear structure within a group of nuclei from specific pathological lesions. These methods could be employed across any group of pathological lesions providing a visual descriptor of important diagnostic nuclear features that up to this point have not been described. In addition, the information provided by these geometric approaches could be used as a stepping stone for further investigation into the molecular and transcriptional control of both normal and neoplastic nuclei.

Currently the major drawback of the approach we propose is the large computational cost. The codes we used were implemented with Matlab 2008a on a laptop with 2.2GH_z CPU and 2GB memory. It usually takes 60 seconds to compute the distance between two images under OT metric with 800 point masses (see Section III-A). We note however, that recent advances show promise to reduce the computational time of such metrics by an order of magnitude [65], [66]. In addition, we note that other metrics that are less computationally intensive (see [67] for example) could be used within the same framework (in combination with coarsely computed OT distances, for example).

While in all cases available in this dataset both feature-based and OT metrics were able to correctly classify each case to it's gold standard (diagnosis) using a voting procedure, in some cases, the feature-based metric seemed to outperform the OT metric in terms of average nuclear classification accuracy. We analyzed the data visually and detected two possible causes for it. Firstly, for nuclei whose chromatin content seemed fairly uniformly spread throughout the nucleus, our particle-based approximation of that image could be improved by increasing the number of particles. We recomputed the classification accuracy of our liver dataset (the smaller of the datasets) utilizing OT distances computed using N = 1200 particles (as opposed to N = 800). This resulted in an increase of classification accuracy of the OT by 2 and 3 % for cases 2 and 5 of table I, respectively, while the accuracies for the other cases remained the same. This suggests that the overall accuracy of the method could be improved by using more particles to approximate each nucleus. How much so in this application, however, is uncertain. In general, most linear programming based solutions for optimal transportation problems are of order $O(N^3)$ computations (with N being the number of particles chosen) [68]. In our case, this means specifically that if we increase the particle number from N = 800 to N = 1200, run times would be increase by roughly 4 times. Upon visual inspection of the data, we have also noticed that the OT metric seems somewhat sensitive to nuclear size variations. This is evidenced in Figure 5 where Fisher LDA (using the OT metric) detected nuclear size variation as the most discriminating feature in the thyroid datset. We investigated case 2 in this dataset, where the procedure using the OT seems to have misclassified quite a few more nuclei than the feature-based approach. For case 2 NL, for example, we have noticed that all nuclei misclassified by the OT approach that were correctly classified with the feature metric had a nuclear area range of [23.5, 30.1]μ^2. A histogram analysis (not shown) shows that this is more consistent with FA and FTC. For case 2 FA, similar observations can be made, where the OT metric misclassified nuclei whose size was more consistent with the FTC class. We point out again, however, that when all nuclei for all cases, the overall accuracy of the OT metric seems to be as good or better than the overall accuracy obtained using the feature-based metric.

Finally, the methodology we described above is quite general in the sense that it depends only on accurate imaging of the structures of interest. It could also be applied to similar nuclear morphology problems in other benign, preneoplastic, and neoplastic (cancer) lesions. Although several of the methods we described (including feature-based ones) were able to appropriately classify the available testing data in the diagnostic challenges we investigated, we expect this not to be the case in more complicated diagnostic challenges. We also notice that similar work has been applied to medical imaging problems at the macroscopic scale [42] where the goal is to analyze and classify the different structures using other metric space (large deformation diffeomorphic metric mapping). Our future plans including testing and validating our methods on larger data sets (more human cases), as well as more difficult diagnostic challenges involving more types of tissues (classes).

Acknowledgments

We wish to thank the anonymous reviewers for their useful suggestions. This work was partially supported by NIH grant 5R21GM088816-02.

References

1. Kung HC, Hoyert DL, Xu J, Murphy SL. Deaths: final data for 2005. Natl Vital Stat Rep. Apr; 2008 56(10):1–120. [PubMed: 18512336]
2. Ma WW, Adjei AA. Novel agents on the horizon for cancer therapy. CA Cancer J Clin. 2009; 59(2): 111–137. [PubMed: 19278961]
3. Schlotter CM, Vogt U, Allgayer H, Brandt B. Molecular targeted therapies for breast cancer treatment. Breast Cancer Res. 2008; 10(4):211. [PubMed: 18671839]
4. Weiner LM, Dhodapkar MV, Ferrone S. Monoclonal antibodies for cancer immunotherapy. Lancet. 2009; 373(9668):1033–1040. [PubMed: 19304016]
5. Hajdu S, Melamed M. Limitations of aspiration cytology in the diagnosis of primary neoplasms. Acta Cytol. 1984; 28(3):337. [PubMed: 6587711]
6. DeMay, RM. The Art and Science of Cytopathology: Aspiration Cytology. Vol. 2. American Society of Clinical Pathology Press; 1996.
7. Papanicolaou, GN. New cancer diagnosis. Proceedings of the 3rd race etterment conference; Michigan. 1928. p. 528
8. Zink D, Fischer AH, Nickerson JA. Nuclear structure in cancer cells. Nat Rev Cancer. 2004; 4:677–687. [PubMed: 15343274]
9. Bartels PH, Gahm T, Thompson D. Automated microscopy in diagnostic histopathology: From image processing to automated reasoning. I J Imaging Systems and Technology. 1998; 8(2):214–223.
10. Demir, C.; Yener, B. Tech Rep TR-05-09. Rensselaer Polytechnic Institute; 2005. Automated cancer diagnosis based on histopathological images: a systematic survey.
11. Rodenacker K, Bengtsson E. A feature set for cytometry on digitized microscopy images. Anal Cell Pathol. 2003; 25:1–36. [PubMed: 12590175]
12. Alvarez GA, Cavanagh P. The capacity of visual short-term memory is set both by visual information load and by number of objects. Psychol Sci. Feb; 2004 15(2):106–111. [PubMed: 14738517]
13. Hartshorne JK. Visual working memory capacity and proactive interference. PLoS ONE. 2008; 3(7):e2716. [PubMed: 18648493]
14. Bengtsson E. Fifty years of attempts to automate screening for cervical cancer. Med Imaging Tech. 1999; 17:203–210.
15. Gil J, Wu H-S. Application of image analysis to anatomic pathology: realities and promises. Cancer Investigation. 2003; 21(6):950–959. [PubMed: 14735698]
16. Yang L, Chen W, Meer P, Salaru G, Goodell L, Berstis V, Foran D. Virtual microscopy and grid-enabled decision support for large scale analysis of imaged pathology specimens. IEEE Trans Inf Technol Biomed. Apr.2009
17. Kong J, Sertel O, Shimada H, BOyer KL, Saltz JH, Gurcan MN. Computer-aided evaluation of neuroblastoma on whole slide histology images: classifying grade of neuroblastic differentiation. Pattern Recognition. 2009; 42:1080–1092.
18. Padfield, D.; Chen, B.; Roysam, H.; Cline, C.; Lin, G.; Seel, M. Cancer tissue classification using nuclear feature measurements from dapi stained images. Proceedings of 1st Workshop on Microscopic Image Analysis with Applications in Biology; 2006. p. 86-92.
19. Yang S, Köhler D, Teller K, Cremer T, Le Baccon P, Heard E, Eils R, Rohr K. Non-rigid registration of 3d multi-channel microscopy images of cell nuclei. Med Image Comput Comput Assist Interv. 2006; 9(Pt 1):907–14. [PubMed: 17354977]
20. Mangoubi R, Desai M, Lowry N, Sammak P. Performance evaluation of multiresolution texture analysis of stem cell chromatin. ISBI. 2008:380–383.

21. Singh SS, Kim D, Mohler JL. Java web start based software for automated quantitative nuclear analysis of prostate cancer and benign prostate hyperplasia. Biomedical Engineering Online. 2005; 4:1. [PubMed: 15631635]
22. Abulafia O, Sherer DM. Automated cervical cytology: meta-analysis of the performance of the PAPNET system. Obstet Gynecol Surv. 1999; 54:253–264. [PubMed: 10198930]
23. Burger, G.; Ploem, JS.; Goerttler, K. Clinical Cytometry and Histometry. Academic Press; 1987.
24. Murata S, Mochizuki K, Nakazawa T, Kondo T, Nakamura N, Yamashita H, Urata Y, Ashihara T, Katoh R. Morphological abstraction of thyroid tumor cell nuclei using morphometry with factor analysis. Microsc Res Tech. Aug; 2003 61(5):457–462. [PubMed: 12845572]
25. Karslioglu Y, Celasun B, Gunhan O. Contribution of morphometry in the differential diagnosis of fine-needle thyroid aspirates. Cytometry B Clin Cytom. May; 2005 65(1):22–28. [PubMed: 15779051]
26. Gupta N, Sarkar C, Singh R, Karak AK. Evaluation of diagnostic efficiency of computerized image analysis based quantitative nuclear parameters in papillary and follicular thyroid tumors using paraffin-embedded tissue sections. Pathol Oncol Res. 2001; 7(1):46–55. [PubMed: 11349221]
27. Vasko VV, Gaudart J, Allasia C, Savchenko V, Di Cristofaro J, Saji M, Ringel MD, De Micco C. Thyroid follicular adenomas may display features of follicular carcinoma and follicular variant of papillary carcinoma. Eur J Endocrinol. Dec; 2004 151(6):779–786. [PubMed: 15588246]
28. Frasoldati A, Flora M, Pesenti M, Caroggio A, Valcavil R. Computer-assisted cell morphometry and ploidy analysis in the assessment of thyroid follicular neoplasms. Thyroid. Oct; 2001 11(10): 941–946. [PubMed: 11716041]
29. Murata S, Mochizuki K, Nakazawa T, Kondo T, Nakamura N, Yamashita H, Urata Y, Ashihara T, Katoh R. Detection of underlying characteristics of nuclear chromatin patterns of thyroid tumor cells using texture and factor analyses. Cytometry. Nov; 2002 49(3):91–95. [PubMed: 12442308]
30. Tsai TH, Chang TC, Chiang CP. Nuclear area measurements of parathyroid adenoma, parathyroid hyperplasia and thyroid follicular adenoma. a comparison. Anal Quant Cytol Histol. Feb; 1997 19(1):45–48. [PubMed: 9051185]
31. Albregtsen F, Nielsen B, Danielsen H. Adaptive gray level run length features from class distance matrices. Int Conf on Pattern Recognition. 2000:3746–3749.
32. Baheerathan S, Albregtsen F, Danielsen H. New texture features based on complexity curve. Pattern Recogn. 1999; 32:605–618.
33. Kerenji A, Bozovic Z, Tasic M, Budimlija Z, Klem I, Polzovic A. Fractal dimension of hepatocytes' nuclei in normal liver vs hepatocellular carcinoma (hcc) in human subjects - preliminary results. Arch of Oncology. 2000; 8:47–50.
34. Ikeguchi M, Sato N, Hirooka Y, Kaibara N. Computerized nuclear morphometry of hepatocellular carcinoma and its relation to proliferative activity. J Surg Oncol. Aug; 1998 68(4):225–230. [PubMed: 9721707]
35. Pereira RR, Marques PMA, Honda MO, Kinoshita S, Engelmann R, Muramatsu C, Doi K. Usefulness of texture analysis for computerized classification of breast lesions on mammograms. Journal of Digital Imaging. Sep; 2007 20(3):248–255. [PubMed: 17122993]
36. Wang W, Ozolek J, Rohde G. Detection and classification of thyroid follicular lesions based on nuclear structure from histopathology images. Cytometry Part A. 2010; 77(5):485–494.
37. Dryden, IK.; Mardia, KV. Statistical Shape Analysis. Chichester: Wiley; 1998.
38. Bookstein FL. Size and shape spaces for landmark data in two dimensions. Stat Sci. 1986; 16:181–242.
39. Joshi S, Davis B, Jomier M, Gerig G. Unbiased diffeomorphic atlas construction for computational anatomy. Neuroimage. 2004; 23:S151–S160. [PubMed: 15501084]
40. Rueckert D, Frangi AF, Schnabel JA. Automatic construction of 3-d statistical deformation models of the brain using nonrigid registration. IEEE Trans Med Imag. 2003; 22(8):1014–1025.
41. Grenander U, Miller MI. Computational anatomy: an emerging discipline. Quart Appl Math. 1998; 56(4):617–694.
42. Miller MI, Priebe CE, Qiu A, Fischl B, Kolasny A, Brown T, Park Y, Ratnanather JT, Busa E, Jovicich J, Yu P, Dickerson BC, Buckner RL, Morphometry BIRN. Collaborative computational

anatomy: an mri morphometry study of the human brain via diffeomorphic metric mapping. Hum Brain Mapp. Jul; 2009 30(7):2132–41. [PubMed: 18781592]

43. Salmon I, Kiss R, Franc B, Gasperin P, Heimann R, Pasteels JL, Verhest A. Comparison of morphonuclear features in normal, benign and neoplastic thyroid tissue by digital cell image analysis. Anal Quant Cytol Histol. Feb; 1992 14(1):47–54. [PubMed: 1558615]
44. Salmon I, Gasperin P, Pasteels JL, Heimann R, Kiss R. Relationship between histopathologic typing and morphonuclear assessments of 238 thyroid lesions. digital cell image analysis performed on feulgen-stained nuclei from formalin-fixed, paraffin-embedded materials. Am J Clin Pathol. Jun; 1992 97(6):776–786. [PubMed: 1375804]
45. Boyokov Y, Funka-Lea G. Graph cuts and efficient n-d image segmentation. Intern J Comp Vis. 2006; 70(2):109–131.
46. Li C, Xu C, Gui C, Fox MD. Level set evolution without reinitialization: A new variational formulation. CVPR '05: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2005; 1:430–436.
47. Coelho LP, Shariff A, Murphy RF. Nuclei segmentation in microscope cell images: A hand-segmented dataset and comparison of algorithms. Proc IEEE Int Symp Biomed Imaging. 2009:518–521. [PubMed: 20628545]
48. Rohde GK, Ribeiro AJS, Dahl KN, Murphy RF. Deformation-based nuclear morphometry: capturing nuclear shape variation in hela cells. Cytometry A. Apr; 2008 73(4):341–50. [PubMed: 18163487]
49. Rohde GK, Wang W, Peng T, Mphy RF. Deformation-based nonlinear dimension reduction: applications to nuclear morphometry. Proc IEEE Int Symp Biomed Imaging. 2008:500–503.
50. Haker S, Zhu L, Tennembaum A, Angenent S. Optimal mass transport for registration and warping. Intern J Comp Vis. 2004; 60(3):225–240.
51. Rubner Y, Tomassi C, Guibas LJ. The earth mover's distance as a metric for image retrieval. Intern J Comp Vis. 2000; 40(2):99–121.
52. Villani, C. Graduate Studies in Mathematics. Vol. 58. Providence, RI: American Mathematical Society; 2003. Topics in optimal transportation, ser.
53. Lloyd SP. Least squares quantization in pcm. IEEE Trans Inf Theory. 1982; 28(2):129–137.
54. Bishop, CM. Pattern Recognition and Machine Learning (Information Science and Statistics). Springer; Aug. 2006
55. Aizerman A, Braverman EM, Rozoner LI. Theoretical foundations of the potential function method in pattern recognition learning. Automation and Remote Control. 1964; 25:821–837.
56. Zamolotskikh A, Cunningham P. An assessment of alternative strategies for constructing emd-based kernel functions for use in an svm for image classification. Technical Report UCD-CSI-2007-3. 2007
57. Chapelle O, Haffner P, Vapnik V. Support vector machines for histogram-based image classification. IEEE Trans Neural Networks. 1999; 10(5)
58. Kreß el U. Pairwise classification and support vector machines. Advances in Kernel Methods Support Vector Learning. 1999:255–268.
59. Fisher RA. The use of multiple measurements in taxonomic problems. Annals of Eugenics. 1936; 7:179–188.
60. Cox T, Cox M. Multidimensional scaling. Chapman and Hall. 2001
61. Tenembaum J, de Silva V, Langford JC. A global geometric framework for nonlinear dimensionality reduction. Science. 2000; 290:2319–2323. [PubMed: 11125149]
62. Dundr M, Misteli T. Functional architecture in the cell nucleus. Biochem J. Jun; 2001 356(Pt 2): 297–310. [PubMed: 11368755]
63. Moss TJ, Wallrath LL. Connections between epigenetic gene silencing and human disease. Mutat Res. May; 2007 618(1–2):163–74. [PubMed: 17306846]
64. Dialynas GK, Vitalini MW, Wallrath LL. Linking heterochromatin protein 1 (hp1) to cancer progression. Mutat Res. Dec; 2008 647(1–2):13–20. [PubMed: 18926834]

65. Delzanno GL, Chacón L, Finn JM, Chung Y, Lapenta G. An optimal robust equidistribution method for two-dimensional grid adaptation based on Monge-Kantorovich optimization. J Comput Phys. 2008; 227(23):9841–9864. [Online]. Available: http://dx.doi.org/10.1016/j.jcp.2008.07.020.

66. Haber E, Rehman T, Tannenbaum A. An efficient numerical method for the solution of the L^2 optimal mass transfer problem. SIAM J Sci Comput. 2010; 32(1):197–211. [PubMed: 21278828]

67. Lafon S, Lee AB. Diffusion maps and coarse-graining: a unified framework for dimensionality reduction, graph partitioning, and data set parameterization. IEEE Trans Pattern Anal Mach Intell. 2006; 28(9):1393–1403. [PubMed: 16929727]

68. Rubner Y, Tomassi C, Guibas LJ. The earth mover's distance as a metric for image retrieval. Intern J Comp Vis. 2000; 40(2):99–121.

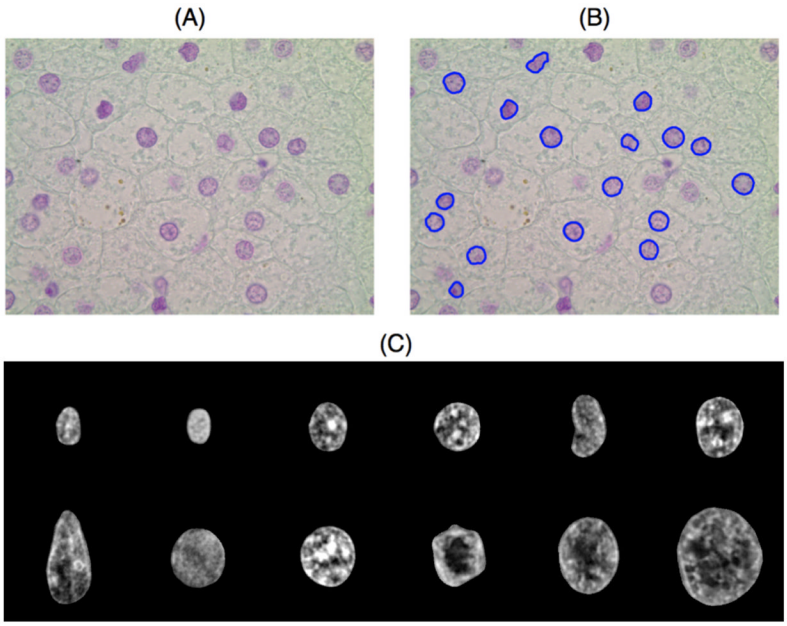

Fig. 1.
Feulgen stained image and segmentation results. A: raw image. B: segmented image. C: individual segmented nuclei after preprocessing. Some sample nuclei show variations in size, shape, membrane contours and etc. Note that each of these images has been contrast stretched for best visualization.

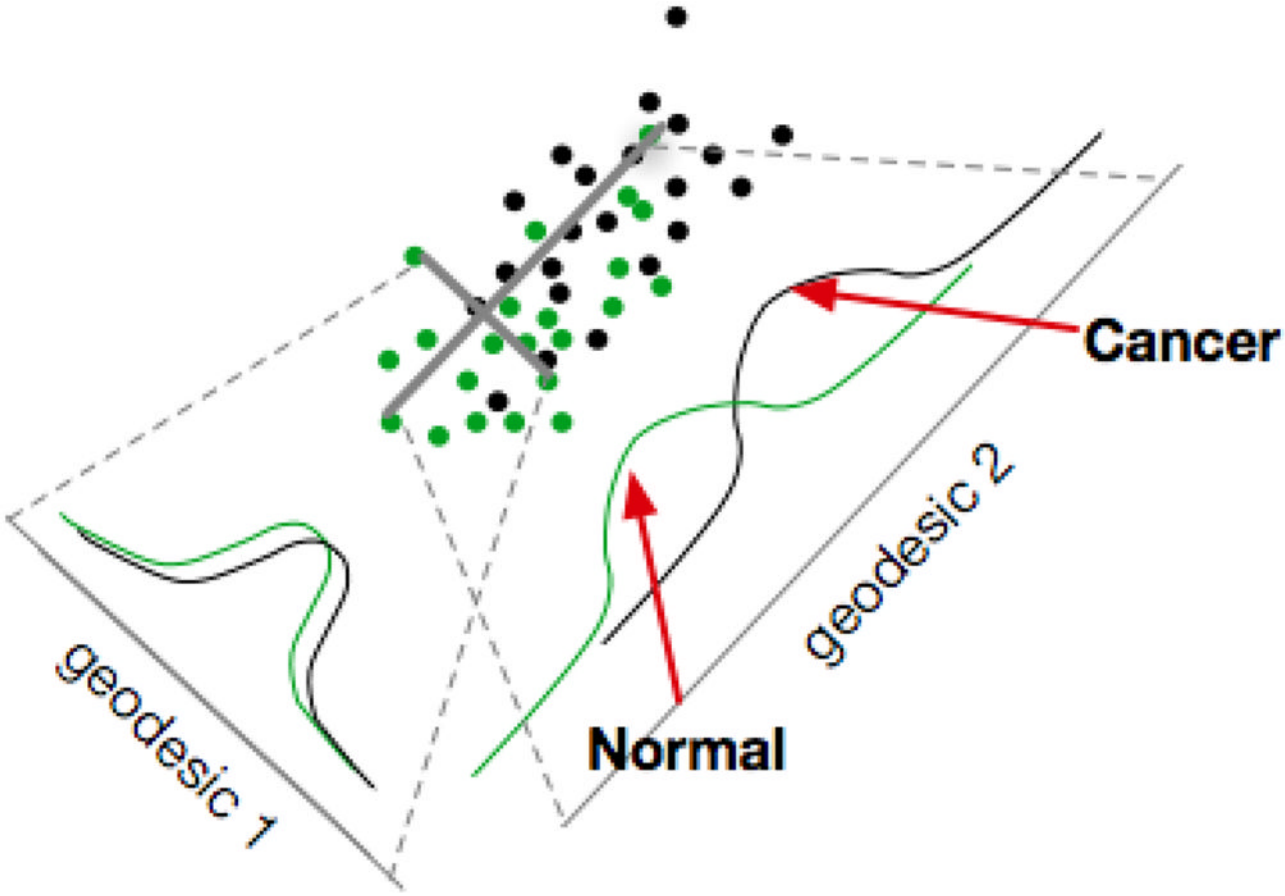

Fig. 2.

Schematic illustration of geometric approach for decoding discriminant information between normal and cancerous nuclei. Each black dot denotes a nucleus from a cancerous tissue, while the green (gray) dots denote the nuclei from a normal tissue. Geodesic paths between any two nuclei can be computed using the approaches described in the text. Projections over these can also be computed utilizing the same geometric metric being utilized. Discriminating paths (see Figure 3 for an actual example) are those over which the projection of the two (or more) populations differ most. In this case geodesic 1 is not as informative as geodesic 2.

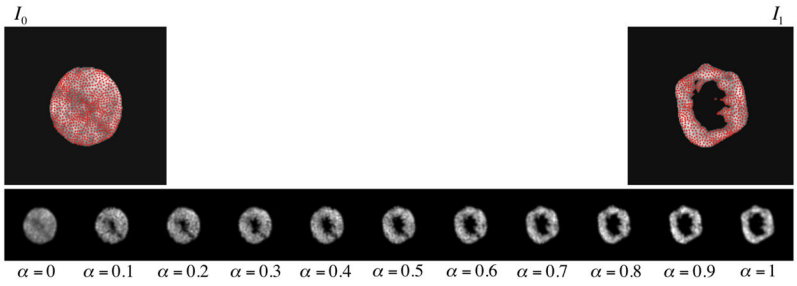

Fig. 3.

A geodesic generated by I_0, and I_1. The larger images on top are the real nuclei images. All the other images are interpolated based on I_0, and I_1. The red dots in I_0 and I_1 are locations of particle masses used to approximate each image (see text for a complete description).

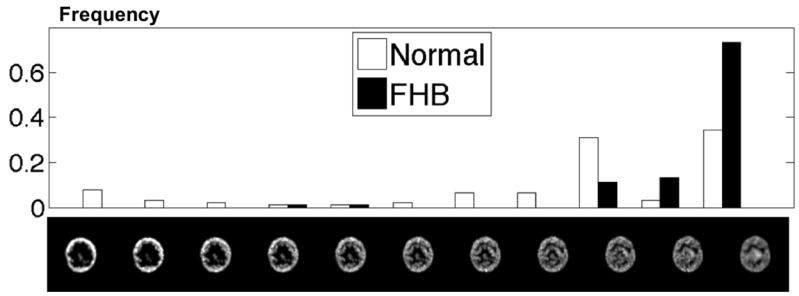

Fig. 4.

Geodesic identified automatically by our method. In the histogram, the height of the bar directly above each nucleus indicates the proportion of nuclei in each data class was most similar (in the OT sense) to the nucleus directly below it. In this plot normal liver and FHB nuclei are compared, with FHB nuclei having their chromatin more evenly spread over the entire nucleus (see text for more details).

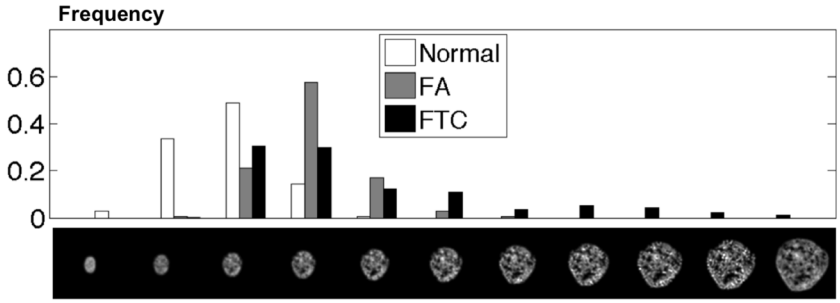

Fig. 5.

Geodesic identified automatically by our method. In the histogram, the height of the bar directly above each nucleus indicates the proportion of nuclei in each data class was most similar (in the OT sense) to the nucleus directly below it. In this plot, normal thyroid, FA and FTC nuclei are compared. We can observe that normal thyroid nuclei are relatively smaller than nuclei in the thyroid neoplasms (FA and FTC); the size of FA nuclei are accumulated to a specific size region, while the FTC nuclei are more evenly distributed in terms of size(see text for more details).

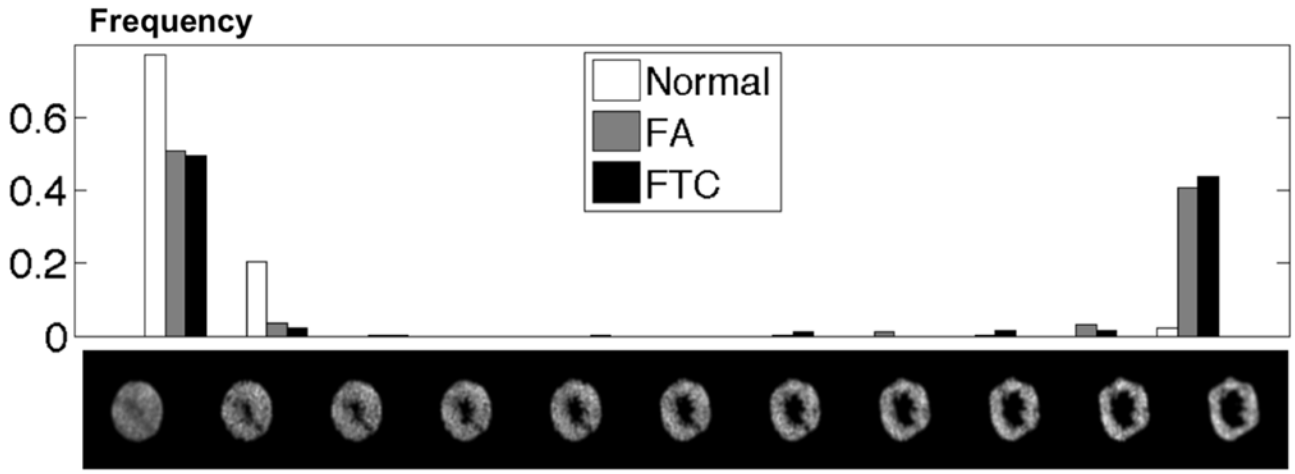

Fig. 6.

Geodesic selected manually to investigate an interesting projection. In the histogram, the height of the bar directly above each nucleus indicates the proportion of nuclei in each data class (normal vs FA vs FTC) was most similar (in the OT sense) to the nucleus directly below it. This geodesic in the bottom mainly shows variation in nuclei texture from chromatin smoothly distributed to chromatin accumulated exclusively along peripheral. We can observe that normal thyroid nuclei, in relative terms, are mostly smooth, while the chromatin distributions of neoplastic nuclei (FA and FTC) tend to distribute more evenly in these two patterns.

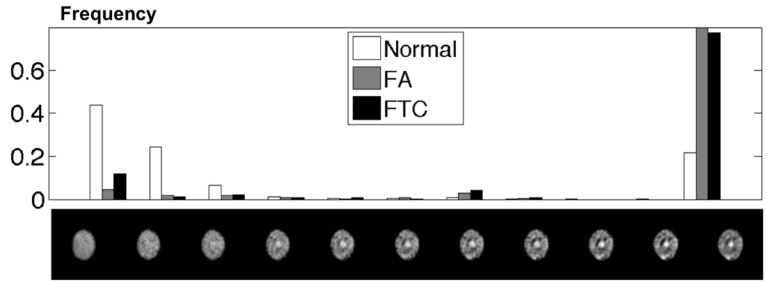

Fig. 7.

Geodesic selected manually to investigate an interesting projection. In the histogram, the height of the bar directly above each nucleus indicates the proportion of nuclei in each data class (normal vs FA vs FTC) was most similar (in the OT sense) to the nucleus directly below it. This geodesic in the bottom mainly shows variation in nuclei texture from smooth texture to chromatin highly accumulated in some locations. We can observe that the chromatin distributions of neoplastic nuclei (FA and FTC) tend to be more centrally located.

TABLE I

Average classification accuracy in liver data

	Feature	OT	Combined
Case 1	89%	86%	93%
Case 2	92%	87%	91%
Case 3	94%	91%	92%
Case 4	80%	88%	89%
Case 5	71%	78%	84%
Average	85.2%	86.0%	89.8%

TABLE II

Average classification accuracy of thyroid data. NL: Normal Thyroid; FA: Follicular Adenoma; FTC: Follicular Carcinoma. Each row indicates a lesion from one testing case, and for each lesion, we separately report the percentage of nuclei classified as NL, FA and FTC either based on feature-based approach (shown in the first, second and third columns) or based on OT metric approach (shown in the forth, fifth and sixth columns). The seventh column indicates the combined accuracy

	Feature NL	Feature FA	Feature FTC	OT NL	OT FA	OT FTC	Combined
Case 1 NL	68.6%	14.3%	17.1%	71.4%	14.3%	14.3%	71.4%
Case 1 FA	5.7%	62.9%	31.4%	5.7%	54.3%	40.0%	65.7%
Case 2 NL	82.9%	8.6%	8.6%	74.3%	8.6%	17.1%	82.9%
Case 2 FA	17.1%	68.6%	14.3%	20.0%	54.3%	25.7%	65.7%
Case 3 NL	82.9%	11.4%	5.7%	80.0%	11.4%	8.6%	88.6%
Case 3 FA	8.6%	62.9%	28.5%	11.4%	60.0%	28.5%	68.6%
Case 4 NL	85.7%	2.9%	11.4%	77.1%	5.7%	17.1%	85.7%
Case 4 FA	7.1%	64.3%	28.5%	17.1%	60.0%	22.9%	74.3%
Case 5 NL	74.3%	5.7%	20.0%	74.3%	2.9%	22.9%	77.1%
Case 5 FA	20.0%	68.6%	11.4%	8.6%	62.9%	28.5%	68.6%
Case 6 NL	85.7%	8.6%	5.7%	85.7%	0.0%	14.3%	88.6%
Case 6 FTC	0.0%	30.0%	70.0%	0.0%	15.7%	84.3%	82.9%
Case 7 NL	82.9%	14.3%	2.9%	80.0%	8.6%	11.4%	85.7%
Case 7 FTC	0.0%	38.6%	61.4%	0.00%	32.9%	67.1%	71.4%
Case 8 NL	74.3%	22.9%	2.9%	74.3%	11.4%	14.3%	80.0%
Case 8 FTC	12.9%	40.0%	47.1%	2.9%	40.0%	57.1%	58.6%
Case 9 NL	77.1%	22.9%	0.0%	80.0%	11.4%	8.6%	82.9%
Case 9 FTC	11.4%	40.0%	48.6%	22.9%	28.5%	48.6%	55.7%
Case 10 NL	91.4%	8.6%	0.0%	91.4%	0.0%	8.6%	94.3%
Case 10 FTC	32.9%	27.1%	40.0%	17.1%	28.5%	54.3%	58.6%

IEEE Trans Med Imaging. Author manuscript; available in PMC 2012 August 13.