Published in final edited form as:

Proc South Biomed Eng Conf. 2010 January 1; 32: 348–352. doi:10.1007/978-3-642-14998-6_89.

Local Histograms for Classifying H&E Stained Tissues

M.L. Massar¹, R. Bhagavatula², M. Fickus¹, and J. Kovačević^{2,3}

¹Department of Mathematics and Statistics, Air Force Institute of Technology, Wright Patterson Air Force Base, USA

²Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, USA

³Department of Biomedical Engineering, Carnegie Mellon University, Pittsburgh, USA

Abstract

We introduce a rigorous mathematical theory for the analysis of local histograms, and consider the appropriateness of their use in the automated classification of textures commonly encountered in images of H&E stained tissues. We first discuss some of the many image features that pathologists indicate they use when classifying tissues, focusing on simple, locally-defined features that essentially involve pixel counting: the number of cells in a region of given size, the size of the nuclei within these cells, and the distribution of color within both. We then introduce a probabilistic, occlusion-based model for textures that exhibit these features, in particular demonstrating how certain tissue-similar textures can be built up from simpler ones. After considering the basic notions and properties of local histogram transforms, we then formally demonstrate that such transforms are natural tools for analyzing the textures produced by our model. In particular, we discuss how local histogram transforms can be used to produce numerical features that, when fed into mainstream classification schemes, mimic the baser aspects of a pathologist's thought process.

Keywords

histology; local histogram; occlusion

I. INTRODUCTION

In this paper, we consider some mathematical theory that arose during the development of an automatic classification scheme for histology, specifically an algorithm that classifies the type and positioning of tissues found in digital microscopy images of hematoxylin and eosin (H&E) stained tissue sections. Here, we focus on the motivation behind the new mathematics itself; more detail on the particular application and classification scheme is given in [1].

The motivating application arose in studies of embryonic stem (ES) cells undertaken by Dr. John A. Ozolek of the Children's Hospital of Pittsburgh and Dr. Carlos Castro of the University of Pittsburgh. Understanding how ES cells differentiate into tissues will yield better insight into early biological development, and could provide advanced research into tissue regeneration and repair, the treatment of genetic and developmental syndromes, and drug testing and discovery [1]. The work here arose from Ozolek and Castro's study of teratomas produced by injecting primate cells into immunocompromised mice; a teratoma is a tumor which is known to contain tissues derived from each of the three primary germ layers of ectoderm, mesoderm and endoderm. Upon removal from the mice, the teratomas are sectioned, H&E stained, and digitally imaged using a microscope. An example of such an image is given in Figure 1.a; here, the purple-pink coloring is characteristic of H&E stain. In normal tissues, different tissue types are arranged in predictable ways. However, in teratomas, the tissues arrange themselves in seemingly chaotic fashions. Nevertheless, using their years of histology experience, Ozolek and Castro are able to look at these images and quickly discern which tissues are present, as well as their locations. In particular, for the image given in Figure 1.a, they have indicated the presence of several tissue types: cartilage, as typified by Figure 1.b, and concentrated in the lower left corner of the overall image; connective tissue, seen in detail in Figure 1.c, and forming a wide oval overall; and bone, detailed in Figure 1.d, and forming much of the center.

Ozolek and Castro have large numbers of such images – many sections of many teratomas – and hope to gain new biological insight by determining the degree to which they contain certain tissues, as well as the spatial relationships between tissues. However, in order to gain this insight, they first need to have the tissues in these images classified according to type. When accomplished by hand, this task, though straightforward, is time-consuming, error-prone and laborious. When analyzing many images, the cost of this manual labor becomes prohibitively high, both in terms of time and money. As such, what is needed is an image processing system which can perform this analysis with minimal user input.

In the following section, we discuss some basic concepts from the theory of image classification that we have borne in mind while designing such a system. These considerations lead to our use of local histogram transforms, whose basic properties are discussed in Section III. We further discuss an occlusion-based mathematical model for the histological images in question, using it to provide a rigorous analysis of the potential use of local histograms as image classification features.

II. CLASSIFICATION

Most classification systems have two main components: a feature extractor and a decision rule. The feature extractor is a collection of transforms which, when applied to a given image, produce a feature vector which is intended to represent the essential properties of that image. The second component of a classification system is a decision rule, namely a function that assigns a label to a given feature vector. For example, for the H&E stained histology images depicted in Figure 1, Ozolek and Castro have indicated that when performing manual tissue classification, they believe their minds are making use of image features such as the color, shape, size and texture of the tissue structures. Based upon these

qualities, and their experience and training, they are able to assign a label, such as “cartilage,” “connective tissue,” or “bone” to a given portion of a histological image. Our goal is to automate this process.

In automated classification schemes, image features are produced using mathematical formulae. For example, we have investigated an automated classification system [2] that computes Haralick texture features of discrete wavelet transforms of the images in question. Once computed, this feature vector is then fed into a decision rule, typically a neural network or a support vector machine, to produce a label. In this paper, we will not comment further about decision rules, and will rather focus entirely on our choice of histogram-based image features.

There are two reasons why we shall make use of histograms, as opposed to other features. One reason is that histograms are easy to understand intuitively, and this intuition has been the key for us to conjecture and prove rigorous results concerning them. The second, more significant, reason is that histograms are directly related to the image features that Ozolek and Castro have indicated they themselves use when classifying histological images. For example, again consider the histological image given in Figure 1.a – a 2-D histogram of its pixel values is given below it in Figure 1.e. Here, the 2-D histogram is obtained by counting how many pixels have a given value of red (x-axis) and blue (y-axis); we have discarded the green channel of the RGB image, as it contains little distinguishing information in the purple-pink class of H&E stained images. As this histogram is taken over the entire image, it combines information from all tissues. Meanwhile, tissue specific histograms are given in Figures 1.f, 1.g and 1.h. For example, in Figure 1.f, the low-valued, off-diagonal blob corresponds to the dominant, more-blue-than-red purple color of cartilage, while the long tail of the distribution is indicative of the white cell interiors. Meanwhile, the light pink of connective tissue and deeper red of bone can be discerned in the histograms of Figures 1.g and 1.h, respectively. We note that while these histograms themselves may be regarded as image features, we have only been using the locations and heights of their dominant peaks. More importantly, we see that little information can be gleaned from looking at the histogram of the entire image – our goal is to determine which tissues are present at any given location, and as such, our histograms must have some location dependence. As such, it is natural to consider *local histograms*, that is, histograms that are only computed over a small neighborhood of any given pixel.

III. LOCAL HISTOGRAMS AND OCCLUSIONS

Local histograms are a well-studied signal processing tool [3–8]. Here, we define local histograms for images f which are regarded as functions from one finite abelian group G of pixel locations into another finite abelian group P of pixel values. For example, for the 1200 by 1200 image given in Figure 1.a, we have $G = \mathbb{Z}_{1200} \times \mathbb{Z}_{1200}$, where $\mathbb{Z}_N$ denotes the group of integers from 0 to $N-1$, in which arithmetic is performed modulo N . Here, the pixel values themselves lie in $P = \mathbb{Z}_{256} \times \mathbb{Z}_{256}$; we are considering only the 8-bit red and blue channels of the original RGB image. The local histogram of such an image f is defined in terms of a *window*, that is, a nonnegative function w over G that sums to one. To be precise, the *local histogram* of f with respect to w is:

$$(\text{LH}_w f)(g, p) = \sum_{h \in G} w(h) \delta_{f(g+h)}(p).$$

For any fixed pixel value p in P , the corresponding portion of the local histogram may be computed by filtering the function which indicates where f assumes this particular pixel value with the window w . Even with this realization, the computation and storage of a local histogram requires nontrivial resources: for our running example, the local histogram is a four-dimensional array of size $1200 \times 1200 \times 256 \times 256$.

In order to determine the appropriateness of using local histograms as feature transforms for histological images, we consider an occlusion-based model for synthesizing test images. Similar models have previously been considered in [9–13]. To be clear, given an *indicator function* I , which assigns a label from 0 to $K-1$ to each pixel location in G , we define the corresponding *occlusion* of a collection of images $\{f_0, \dots, f_{K-1}\}$ to be the composite image:

$$(\text{occ}_S \{f_k\}_{k=0}^{K-1})(g) := f_{I(g)}(g).$$

An example of an image generated in this fashion is given in Figure 2.d. Here, the number of images is $K=2$, with f_0 given in Figure 2.a, with f_1 given in Figure 2.b, and the indicator function I given in Figure 2.c. Though by no means photorealistic, this synthesized image nevertheless possesses much of the basic color and shape information of cartilage. The reason we use such a simple image model is that it permits a rigorous analysis of the properties of local histograms. Indeed, examining the (global) histograms of the background, foreground and composite images of Figure 2, we note that the histogram of the composite is nearly a convex combination of the histograms of the background and foreground. Indeed, it is possible to show that such a result will always occur, even for local histograms, provided one does not focus on a single means of occlusion, but rather computes an expectation over every possible occlusion. To be precise, consider a probability density function $\mathbf{P}$ defined over the class of all $\{0, \dots, K-1\}$ -valued indicator functions I over G . We say that $\mathbf{P}$ is fair if for all $k = 0, \dots, K-1$, there exists some real scalar λ_k such that:

$$\sum_I \mathbf{P}(I) \delta_k(I(g)) = \lambda_k.$$

When $\mathbf{P}$ is fair, we can prove that, on average, the local histogram of a composite image is indeed a convex combination of the local histograms of each image:

Theorem. *If $\mathbf{P}$ is fair, then*

$$E_I [\text{LH}_w(\text{occ}_I \{f_k\}_{k=0}^{K-1})] = \sum_{k=0}^{K-1} \lambda_k (\text{LH}_w f_k).$$

This begs the question of whether or not fairness is a realistic assumption. Our current research is focused on answering this question. In particular, we are studying methods of producing more complicated occlusion indicator functions from simpler examples, as given in Figure 3. In particular, new indicator functions may be produced by overlaying other known examples of them. More significantly, we can extend this notion of overlay to probability density functions over the set of all indicator functions, and can use this idea to build more complicated fair probabilities from simpler fair ones.

Acknowledgments

Massar and Fickus were supported by AFOSR F1ATA09125G003. Bhagavatula and Kovačević were supported by NIH through award NIH-R03-EB009875 and the PA State Tobacco Settlement, Kamlet-Smith Bioinformatics Grant. The authors would like to thank Dr. John A. Ozolek of the Children's Hospital of Pittsburgh and Dr. Carlos Castro of the University of Pittsburgh. The views expressed in this article are those of the authors and do not reflect the official policy or position of the United States Air Force, Department of Defense, or the U.S. Government.

REFERENCES

1. Bhagavatula R, Fickus M, Ozolek JA, Castro CA, Kovačević J. Automatic identification and delineation of germ layer components in H&E stained images of teratomas derived from human and nonhuman primate embryonic stem cells. To appear in Proc. IEEE Int. Symp. Biomed. Imag. 2010
2. Chebira A, Ozolek JA, Castro CA, Jenkinson WG, Gore M, Bhagavatula R, Khaimovich I, Ormon SE, Navara CS, Sukhwani M, Orwig KE, Ben-Yehudah A, Schatten G, Rhode GK, Kovačević J. Multiresolution identification of germ layer components in teratomas derived from human and nonhuman primate embryonic stem cells. Proc. IEEE Int. Symp. Biomed. Imag. 2008:979–982.
3. Koenderink JJ, van Doorn AJ. The structure of locally orderless images. Int. J. Comput. Vis. 1999; 31:159–168.
4. van Ginneken B, ter Haar Romeny BM. Applications of locally orderless images. J. Vis. Commun. Image Represent. 2000; 11:196–208.
5. Koenderink JJ, van Doorn AJ. Blur and disorder. J. Vis. Commun. Image Represent. 2000; 11:237–244.
6. van de Weijer J, van den Boomgaard R. Local mode filtering. Proc. IEEE Comput. Soc. Conf. Comput. Vis. & Pattern Recognit. 2001; 2:428–433.
7. Hadjidemetriou E, Grossberg MD, Nayar SK. Multiresolution histograms and their use for recognition. IEEE Trans. Pattern Anal. & Mach. Intell. 2004; 26:831–847.
8. Dalal N, Triggs B. Histograms of oriented gradients for human detection. Proc. IEEE Comput. Soc. Conf. Comput. Vis. & Pattern Recognit. 2005; 1:886–893.
9. Lee AB, Mumford D. An occlusion model generating scale-invariant images. Proc. IEEE Workshop Stat. & Comput. Theor. Vis. 1999
10. Lee AB, Mumford D, Huang J. Occlusion models for natural images: A statistical study of a scale-invariant dead leaves model. Int. J. Comput. Vis. 2001; 41:35–59.
11. Mumford D, Gidas B. Stochastic models for generic images. Quart. Appl. Math. 2001; 59:85–111.
12. Ying Z, Castanon D. Partially occluded object recognition using statistical models. Int. J. Comput. Vis. 2002; 49:57–78.
13. Bordenave C, Gousseau Y, Roueff F. The dead leaves model: A general tessellation modeling occlusion. Adv. Appl. Probab. 2006; 38:31–46.

Fig. 1.

Histological images and the two-dimensional histograms of their red and blue channels.

Fig. 2.

Synthesizing cartilage-like textures. A $\{0,1\}$ -valued function (c) indicates where to occlude a background texture (a) by a foreground texture (b). The histogram of the synthesized image is a combination of the histograms of the background and foreground; the relative heights of the peaks of (g) can be used to infer how much background (cell exterior) and foreground (cell interior) is characteristic of a given tissue.

Fig. 3.

Building complicated indicator functions from simple ones: (a) and (b) give two $\{0,1\}$ -valued indicator functions; (c) gives a $\{0,1,2\}$ -valued indicator function obtained by laying (a) over (b); (d) produces a distinct $\{0,1,2\}$ -valued indicator function obtained by laying (b) over (a).