Computerized Medical Imaging and Graphics 40 (2015) 70–87

Contents lists available at ScienceDirect

Computerized Medical Imaging and Graphics

journal homepage: www.elsevier.com/locate/compmedimag

Fourier-ring descriptor to characterize rare circulating cells from images generated using immunofluorescence microscopy

Tegan Emerson[a,*], Michael Kirby[a,1], Kelly Bethel[b], Anand Kolatkar[c], Madelyn Luttgen[c], Stephen O'Hara[d], Paul Newton[e,2], Peter Kuhn[f]

[a] Department of Mathematics, Colorado State University, 841 Oval Drive, Fort Collins, CO 80523, United States
[b] The Scripps Clinic, Department of Pathology, 10666 N Torrey Pines Road, La Jolla, CA 92037, United States
[c] The Scripps Research Institute, The Kuhn Lab, 10550 N Torrey Pines Road, La Jolla, CA 92037, United States
[d] DigitalGlobe, Image Mining Group, 1601 Dry Creek Drive, Longmont, CO 80503, United States
[e] Department of Aerospace and Mechanical Engineering, University of Southern California Viterbi School of Engineering, Los Angeles, CA 90089, United States
[f] The Kuhn Lab, Department of Biological Sciences, University of Southern California, Los Angeles, CA 90089, United States

ARTICLE INFO

Article history:
Received 23 January 2014
Received in revised form 22 May 2014
Accepted 6 October 2014

ABSTRACT

We address the problem of subclassification of rare circulating cells using data driven feature selection from images of candidate circulating tumor cells from patients diagnosed with breast, prostate, or lung cancer. We determine a set of low level features which can differentiate among candidate cell types. We have implemented an image representation based on concentric Fourier rings (FRDs) which allow us to exploit size variations and morphological differences among cells while being rotationally invariant. We discuss potential clinical use in the context of treatment monitoring for cancer patients with metastatic disease.

1. Introduction

Counts of circulating tumor cells (CTCs) have been correlated with outcomes in patients with tumors of epithelial origin including breast cancer, colorectal cancer, non-small cell lung cancer, and prostate cancer [1–5]. For these reasons, research has grown over the last decade. Clinically useful CTC counts have almost exclusively been generated using CellSearch™ [6], which is an FDA approved enumeration method that relies on sample enrichment prior to analysis. Recent research has shown, however, that enrichment prior to analysis results in lower counts of CTCs and the potential loss of entire classes of disease-derived rare cells [7]. CellSearch™ uses an enrichment method based on epithelial biomarkers. However, it is stated in [8] that "invasive tumor cells tend to loose their epithelial antigens by the epithelial to mesenchymal transition process." This finding further suggests that methods like CellSearch™ that enrich based on epithelial biomarkers are unlikely to capture the full population of cells related to tumor progression. Moreover, in [8] they also emphasize that it is known that there can be epithelial cells in blood that do not have their origins in a tumor. These two deficiencies diminish the CTC count accuracy and hence lessens the impact of these methods as diagnostic tools. On the other hand, newer methods that do not enrich prior to analysis have shown the potential to identify a larger and maybe more complete set of candidate cells but require significant computational analysis as well as human verification [7]. In particular, the class of data generated without enrichment provides for the ability to optimize computational methods because of the large amount of data available, over twelve million cells per CTC test. Therefore, there is motivation to improve the level of automation of detection of CTCs on data generated without enrichment. Increasing the automation of detection of CTCs in this setting is the aim of the research presented here.

Over the last 20 years there has been a steady increase in the number of research papers published surrounding morphological cell image analysis [9]. It has been asserted that the five most significant roles of morphological cell analysis in medical imaging are malignant cell identification and cancer detection, following morphological changes during a cell cycle, cell classification, changes in morphology due to treatments, and morphometrical studies [9]. These roles are all strongly related to identifying a set of features that allow us to categorize candidate cells into these rules.

* Corresponding author. Tel.: +1 970 491 5284.
E-mail addresses: emerson@math.colostate.edu (T. Emerson), kirby@math.colostate.edu (M. Kirby), Bethel.Kelly@scrippshealth.org (K. Bethel), anandk@scripps.edu (A. Kolatkar), mluttgen@scripps.edu (M. Luttgen), svohara@ieee.org (S. O'Hara), newton@usc.edu (P. Newton), peter.kuhn@usc.edu (P. Kuhn).
[1] Tel.: +1 970 491 6850.
[2] Tel.: +1 213 740 7782.

http://dx.doi.org/10.1016/j.compmedimag.2014.10.003
0895-6111/

Cellular features are generally related to the geometry of the cell and properties of the interior of the cell, and carry a high-level of interpretability. A review of shape representations and description techniques for general images can be found in [10]. Geometric features of cells commonly include area, radii, perimeter, circularity, eccentricity, and irregularity [9,11]. Internal features of a cell include the intensity, texture, and regularity of the nucleus. The combination of these geometrical and internal features have been employed in many pathological tasks, especially related to human epithelial cells [12,13]. These features can be extracted from images and also visually inspected by a pathologist. However, due to the subjectivity involved in visual inspection, there has been an increase in the efforts to quantify these features using computers and to employ additional computer methods to support pathologists. A layout of one such image-guided decision support function is described in [14].

The set of features described above work well in many pathologist tasks, however interpretable features can have significant drawbacks. A major drawback is the reliance on segmentation. Geometric features are only as accurate as the ability to identify the comprehensive boundary of the cell. There have been several methods proposed for improving the automated segmentation of cell events in blood and tissue [15–17]. A second drawback can be the limitation of these features to differentiate cells. For example, there have not been successful classification methods for differentiating between CTCs in patients of different clinical status using these interpretable features alone. Additional discussion of technical and mathematical challenges of automated screening of epithelial cells can be found in [18]. Features that are generated without concern for interpretability and are mathematical in nature sit in the realm of bioinformatics and are referred to as low-level features. Low-level feature methods have been employed in various cell classification tasks and perform with varying degrees of success. A comparison of different feature types together with varying classifiers is discussed in [19].

In this paper, we have identified a set of low-level features that contain strongly differentiating structural information to separate CTC events of interest from white blood cells as well as identify distinct sub populations. Fourier-ring descriptors (FRDs) extract features from cells by combining techniques from bioinformatics and computer vision. Additionally, FRDs contain interpretable information as well as indirectly containing many geometric features.

In the following sections we will first present the data acquisition process together with the composition of the data. Then, we will describe methods that we have considered, and highlight some of their challenges. Next, we will present our image representation technique and our classification structure for classifying a cell.

2. Material and methods

2.1. Materials

Our analysis has been performed on images generated using the high definition circulating tumor cell (HD-CTC) assay developed by the Scripps Physical Sciences in Oncology Center (Scripps PSOC) as described in [7]. First, we will briefly describe how the data is acquired. Next, we will explain the current two-phase semi-automated algorithm employed at the Scripps PSOC to detect cells of interest. Last, we will define the types of cell populations included in the data set analyzed herein.

Data acquisition. The assay developed at the Scripps PSOC produces images using automated fluorescence microscopy. Data evaluated in this paper was generated using three immunofluorescent stains: *Alexa*555 and pan cytokeratin to highlight epithelial cells, *Alexa*647 conjugated to an anti-CD45 antibody for WBC detection, and *DAPI*(385*nm*) to stain for a nucleus. What separates this assay from many others is the lack of sample enrichment prior to imaging. Images are taken of the entire slide at 10× resolution by an inverted microscope. A set of images corresponding to a slide contains approximately three million nucleated cells, of which less than 0.01 percent are cells of high interest. Cells of high interest are currently detected using a two part semi-automated algorithm.

Phase I: Automated Detection. Phase one of the current algorithm is computer automated and utilizes the medical imaging software ImageJ [20]. Cell centers are determined by detecting on the nuclear image channel and computing a center of mass described in [7]. A set of candidate cells of interest are output, based primarily on the intensity of the cytokeratin and CD-45 channels for any measured cell compared to all cells on the slide. Candidate cells are then passed to a technician or pathologist to manually classify.

Phase II of detection employed at Scripps PSOC: Manual Classification. The second phase of the algorithm consists of the manual classification of candidate events of interest into one of six groups. The four groups of cells of interest are defined as follows:

(1) **CTC-Candidate.** Cells that appear to have a high likelihood of being a CTC. Characterized by bright cytokeratin stain, an intact nucleus, and no CD-45 signal. Cells must be morphologically distinct from surrounding WBCs. This morphological difference typically manifests as a larger nucleus than the neighboring WBCs. These events will then be evaluated by a pathologist to confirm or reject this classification.
(2) **CTC-small.** There exists a population of cells that has appropriate levels of cytokeratin expression to be considered a CTC but has an insufficient nuclear size relative to its surrounding white blood cells (WBCs). This population is considered to be a marginal population of CTCs.
(3) **CTC-dim.** This population accounts for cells which have insufficient levels of cytokeratin expression to be considered CTCs but do have a nuclei that are significantly larger than neighboring cells. This population is considered to be a marginal population of CTCs.
(4) **CTC-Ap.** A last marginal population of CTCs comprised of cells that appear to be apoptotic by identification of nuclear fragmentation or cytoplasmic blebbing.

The two remaining populations of cells are WBCs and imaging noise. This naming scheme is consistent with the cell populations presented in [21]. It is the goal of ongoing research to automate this phase of classification using the methods and experimental design presented here.

Composition of the data set. The analysis herein was performed on a data set of cells which were hand selected for work done in [22]. This data set consists of one thousand cells: five hundred cells of interest and five hundred white blood cells. The five hundred cells of interest contains two hundred CTC-Candidate and one hundred each of CTC-Ap, CTC-dim, and CTC-Small. Cells in the data set were taken from 39 patients with diagnosed lung, breast, and prostate cancers (25 Lung, 7 breast, 7 prostate). This data set is by no means comprehensive, however, it does provide a proof-of-concept to advance this avenue of research. Furthermore, we note that these samples have not been analyzed by any other method, including enrichment based methods, because in order to do so the ability to use these samples in future Scripps PSOC assay studies could be compromised. A detailed comparison of the Scripps PSOC assay to CellSearch^{TM} can be found in [7].

2.2. Methods

We accomplish this classification by identifying a set of low-level features that can structurally differentiate between cell populations of interest. We will present the challenges of the discussed classification tasks and present our solution.

In the current assay used by the Scripps PSOC a single slide is imaged at 10X resolution. Although a technician can manually classify a cell at 10X resolution, the low resolution fails to capture textural variation within each cell channel. This lack of texture, in turn, makes the implementation of gradient-based feature extraction methods impractical. It is impractical to increase the resolution of the images since an n-fold increase in resolution increases the number of images generated from each channel by a factor of n^2. Additionally, low resolution results in s single cell comprising a small area within an image. The small size of the object makes the extraction of patch-based features unnecessary. Alternatively, treating the image of a single cell as a sole patch requires accurate segmentation of the cell. However, because of the frequency of cell overlap and noise from staining in the cytokeratin channel and CD-45 channel it is challenging to determine comprehensive cell boundaries. Furthermore, the cells themselves carry associated challenges. First, a representation technique which is interpretable and captures characteristics of the cells (including size of the nucleus, levels of cytokeratin expression, circularity, and uniformity of the nucleus, for example), is ideal. Next, a cell has no "correct" orientation which makes it imperative to have an image representation which is rotationally invariant.

Concentric circles have been proposed to handle this issue of rotational invariance in other tasks. Circles have been employed to detect objects in multiple rotations of complex colored images [23]. Additionally, in [24] we see the use of concentric rings combined with wavelet transforms for pattern matching. In the method described in [24], a single representative value for each ring is computed, then a wavelet transform of the series of representative values is computed. More recently we see another method involving the sampling of image values along concentric rings, which was developed with pathological tasks in mind [25]. Although the method being applied to these pathological tasks has performed very well, it requires an exhaustive codebook search method for pattern matching [26]. Thus, we have chosen to combine the desirable components of each method and employ concentric rings together with a transformation while not overly restricting the quantity of information taken from each ring.

The Fourier-ring descriptor (FRD) is based on the Fourier transformations of concentric rings about the nucleus of a cell. FRDs treat the image of a cell as a single patch, have limited reliance on segmentation, are rotationally invariant, and the features can be visualized and carry a high level of interpretability. Our solution does not require comprehensive cell boundaries in all channels, but rather only the computation of the center of mass of the cell nucleus. Due to the nuclear channel being the cleanest of the monochrome images, this greatly reduces the reliance on whole cell segmentation. Also, the amplitude spectrum of a Fourier transform is rotationally invariant, and thus the amplitude spectrum of the Fourier transform of a ring is also rotationally invariant.[4] Furthermore, by using concentric rings about the nucleus of the cell we are able to obtain size and morphological information for a cell based on the presence, or lack there of, of features on rings of particular radii. Details of the generation of FRD follow in the Experimental section.

[4] Additional information about the rotational invariance can be found in Section 4

3. Experimental

There are two components of our experiment that will be discussed in this section. First, we will discuss our image representation, the Fourier-ring descriptor, in detail. Next, we will discuss our classification structure and how we implement the structure together with our image representation technique.

3.1. Fourier-ring descriptor

The Fourier-ring descriptor (FRD) is an image representation technique based on the Fourier transform of concentric rings constructed about the center of the nucleus of the cell. We first determine a cell center by computing a center-of-mass for each nucleus that is detected in the DAPI image channel using the medical image processing software ImageJ [20]. Once we have identified the center of a cell, we begin to construct concentric rings that are centered on the cell center.

We use sixteen concentric rings to generate our image representation. The number of rings was determined by experimentation as discussed in [22]. The steps to generate our image representation are as follows:

(i) We sample a number of evenly-spaced points along a ring of a given radius. An example of the location of these points on a ring of radius six pixels is shown in Fig. 1(c).
(ii) Image values at each of the points shown in Fig. 1(c) are interpolated. This set of interpolated image values determines a periodic curve as shown in Fig. 1(d).
(iii) Given the periodic curve of image values, we then perform a Discrete Fourier Transform (DFT) of the image values and keep the magnitude of the Fourier transform. An example of the amplitude spectrum of a single ring is shown in Fig. 1(e).
(iv) The process of sampling points along a ring and computing the DFT of the interpolated image values at the sampled points is repeated for the sixteen different concentric rings in a single monochrome image channel. Examples of the amplitude spectrums for the sixteen different rings in a channel can be seen in Figs. 2–4.
(v) The amplitude spectrums of all rings from a single channel are then concatenated into a single vector according to increasing radius. This process is then repeated in the remaining image channels and the concatenated amplitude spectrums from within a channel are concatenated with the other channels in the following order: Cytokeratin, CD-45, DAPI.

The number of points sampled along a ring is scaled linearly in accordance with the increase in circumference. Eight points were selected for the ring of radius one based on the maximum number of distinct pixels a ring of radius one could encounter. Image values at the sampled locations are determined using cubic interpolation which causes highly associated values across neighboring rings, but this redundancy is minimized by enforcing sparse feature selection in the classifier. Given that we start with eight sampled points on the inner most ring and scale linearly over the sixteen rings, we obtain a total of 1088 features from a single image channel, and a total of 3264 features over all three channels. Each feature of the FRD corresponds to the magnitude of a single frequency along a specific ring.

3.2. Classification structure

The data analyzed for this paper contains five populations of cells. There are white blood cells, CTC-candidates, CTC-Aps, CTC-dims, and CTC-small. We refer to CTC-candidates, CTC-Aps, CTC-dims, and CTC-smalls as populations of interest. As previously

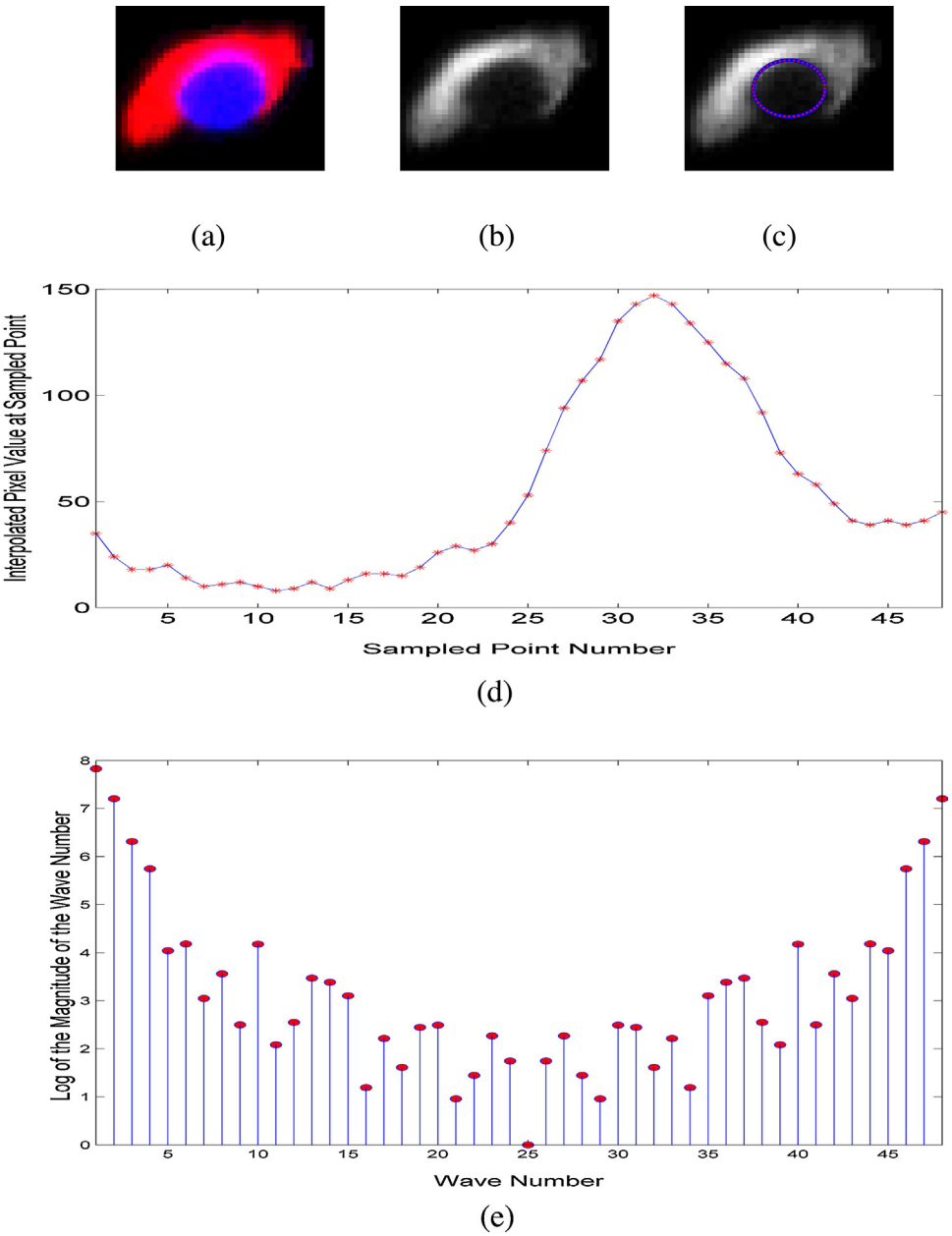

Fig. 1. (a) A composite image of a CTC-Candidate. (b) The monochromatic cytokeratin channel image of the CTC-Candidate in (a). (c) The monochromatic cytokeratin channel image of a circulating tumor cell with the ring of radius six pixels, oriented around the center of the nucleus of the cell, is overlaid in blue. The locations of the sampled points are evenly spaced along the indicated ring and are shown in red. There are 48 points sampled along this ring. (d) A plot of the interpolated image values at the sampling points along the ring of radius six pixels shown (c). Each red point shows an interpolated image value. The image values can be between 0 and 255 as they are taken from 8-bit JPEGs. Lastly, (e) the amplitude spectrum of the Fourier transform of the curve shown in (d). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

described, we consider CTC-Aps, CTC-dims, and CTC-smalls to be marginal populations of CTC-candidates. Initially we want to be able to separate the populations of interest from white blood cells. Separation of white blood cells from populations of interest can first be whittled-down using empirical property values obtained by the Scripps PSOC. Once the population of cells we are looking at has been reduced we can implement a decision tree classification structure. In this decision tree we first want to say with confidence whether or not a given cell is interesting. Once it has been identified as interesting we then ask whether or not the cell is a CTC-Candidate, our most important population of interest. We then proceed to subdivide the populations according to the schematic shown in Fig. 5.

We have chosen to use a decision function determined by an l_1 regularized, l_2 loss function linear support vector machine classifier as implemented by LIBLINEAR [27]. A discussion of the decision function can be found in Section 4. By implementing an l_1 regularized support vector machine we force sparsity of features required for classification. It is desirable for us to encourage sparsity both to limit the number of features we must interpret as well as to account for the potential of over-sampling in the generation of the FRD. By using a support vector machine classifier, we have a way of communicating our confidence in the classification of an event. Cells which are closer to a separating hyperplane may be misclassified while cells far away from the hyperplane may be extremal cases within their class.

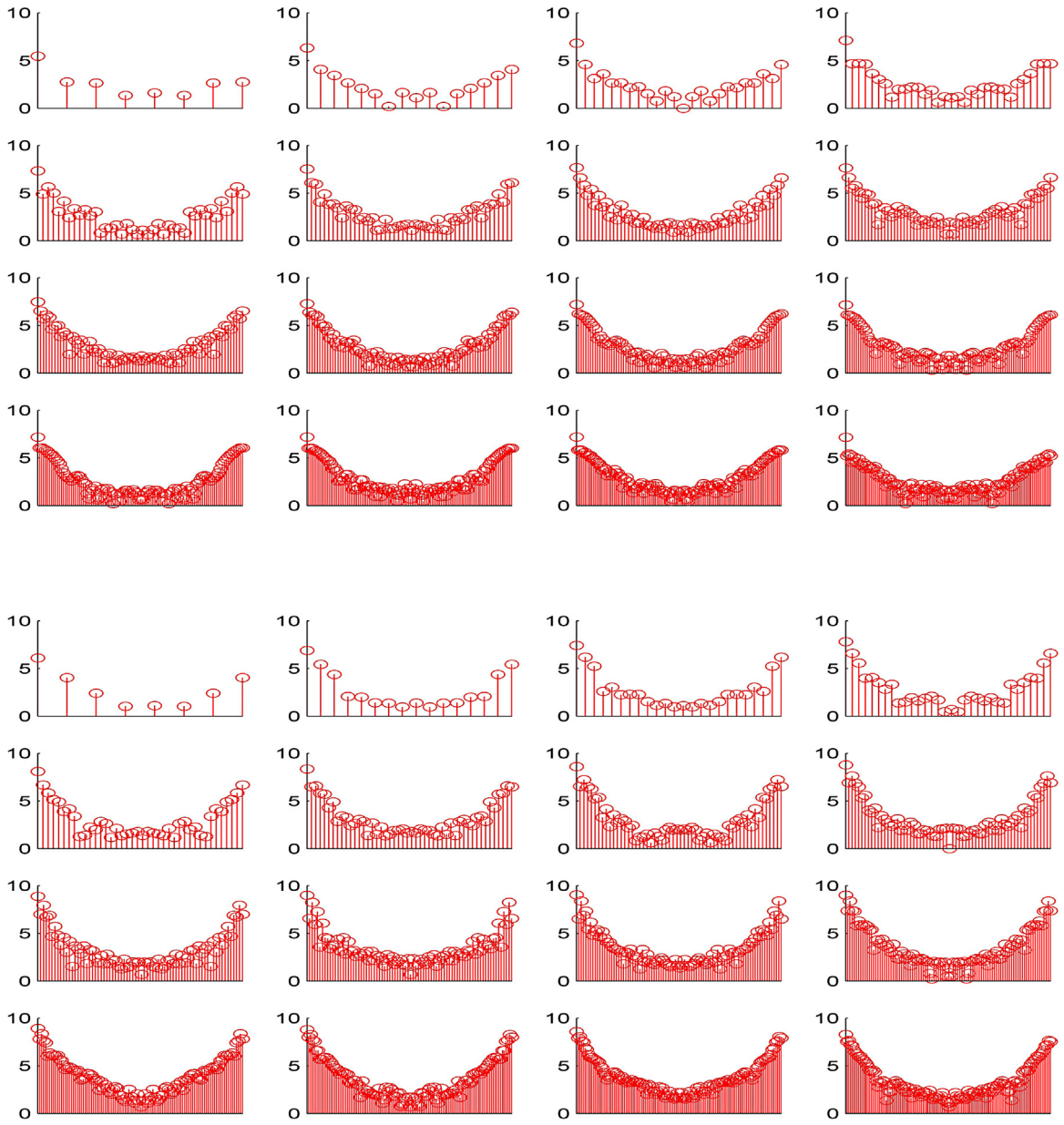

Fig. 2. The top figure shows all 16 rings in the cytokeratin channel for the closest-to-average CTC-Candidate in our data set while the bottom figure shows all 16 rings for the farthest-from-average CTC-Candidate in our data set. The *y*-axis is the log of the amplitude and the *x*-axis corresponds to the wavenumber. A image value curve with m points will generate an amplitude spectrum for m frequencies and we recall for the ring of radius r there will be $8r$ image values sampled. The radius increases left to right and top to bottom.

4. Theory/calculations

In this section we will address two concepts. First, we have stated that the FRD method is rotationally invariant. We will provide a graphical proof of the rotational invariance of a single FRD which can then be extended to show the invariance of the entire descriptor. Next, we will discuss the type of the decision function that we have implemented at each step in our decision tree classification structure. A description of the optimization problem to be solved and the parameters and variables involved are also included.

4.1. Rotational invariance of FRDs

Here we present graphical proof of the rotational invariance of the amplitude spectrum of the Fourier transform in the context of our application. A rigorous proof of the rotational invariance of the Fourier transform can be found in [28].

Fig. 6 provides us with two images: the first of a cell represented in the cytokeratin channel in its original orientation on the slide image, the second shows the same cell in the cytokeratin channel rotated 60 degrees counterclockwise. For this didactic example, we employed a numerical rotation function that utilizes bicubic interpolation. Next, in Fig. 7(a), we see the curves of interpolated image values along the ring of radius 8 pixels for both the rotated and unrotated versions of the cell. We see that the two image value curves are simply shifted versions of one another, minus slight numerical error. Furthermore, in Fig. 7(b) we see the amplitude spectrums of the two image value curves. From Fig. 7(b), we see that the two amplitude spectrums are nearly identical minus the effect of numerical artifacts present from the rotation. The combination of Figs. 6 and 7 demonstrates that a single FRD generated for a cell is

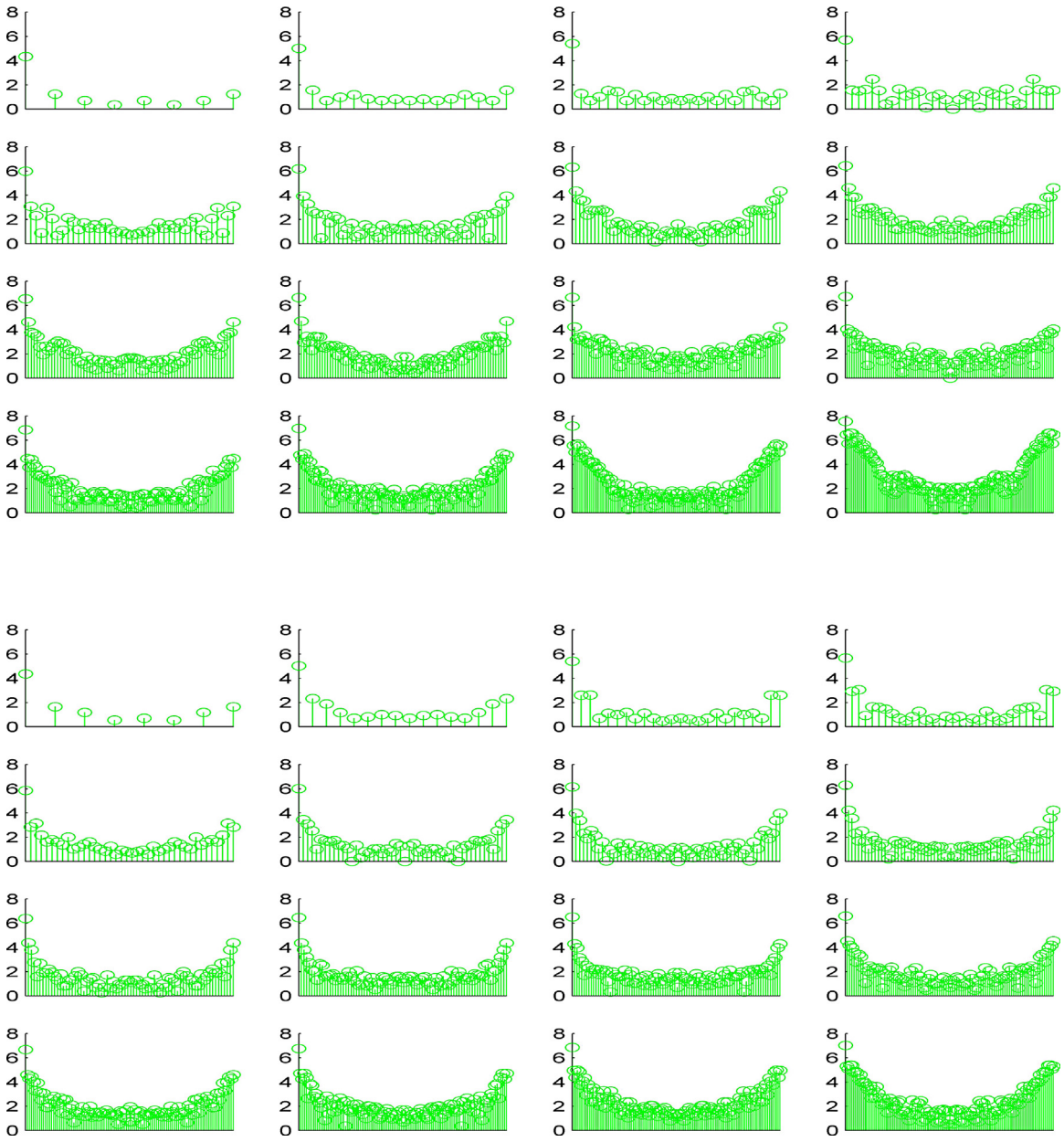

Fig. 3. The top figure shows all 16 rings in the CD-45 channel for the closest-to-average CTC-Candidate in our data set while the bottom figure shows all 16 rings for the farthest-from-average CTC-Candidate in our data set. The *y*-axis is the log of the amplitude and the *x*-axis corresponds to the wavenumber. A image value curve with m points will generate an amplitude spectrum for m frequencies and we recall for the ring of radius r there will be $8r$ image values sampled. The radius increases left to right and top to bottom.

invariant to the orientation of the cell on the slide. The full FRD representation of a cell amounts to stacking the amplitude spectrums for each of the rings inside a channel and then concatenating the channels. Thus, for two orientations of the cell we have shown that the amplitude spectrum for each ring is rotationally invariant and consequently the stacking of those amplitude spectrums will result in the same channel representation for a cell no matter the rotation.

4.2. Discussion of decision function

A linear support vector machine (SVM) aims to define a decision function based on a hyperplane that separates data into two half spaces. Data is classified by which side of the separating hyperplane a datum falls on. We define the separating hyperplane by the equation $y(x) = w^T x + b$. In a classification task involving two classes of data with the standard ± 1 labeling, the classification of a novel data point $\hat{x}$ is determined by

$$D(\hat{x}) = \begin{cases} \hat{x} \in C^+ & \text{if} \quad w^T\hat{x} + b \geq 0 \\ \hat{x} \in C^- & \text{if} \quad w^T\hat{x} + b \leq 0 \end{cases}$$

where C^+ corresponds to the class of positive samples and C^- corresponds to the class of negative samples.

In all of the classification tasks reported here we have implemented the linear, l_1 regularized, l_2 loss function SVM as implemented by LIBLINEAR [27]. The l_1 regularized, l_2 loss function linear SVM is defined by minimizing

$$\min_w \|w\|_1 + C\sum_{i=1}^{n} \max(0, 1 - y_i w^T x_i)^2. \tag{1}$$

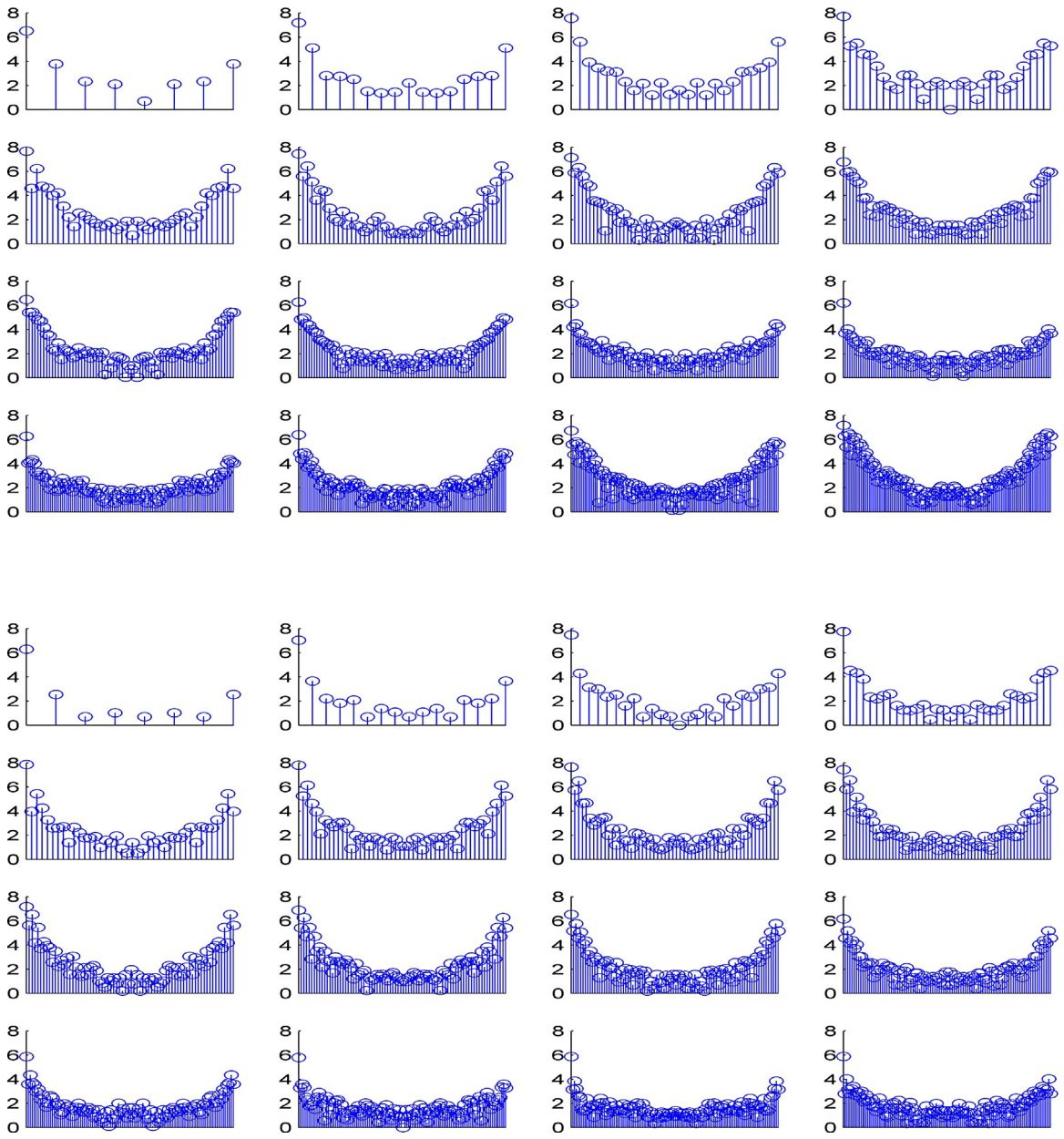

Fig. 4. The top figure shows all 16 rings in the DAPI channel for the closest-to-average CTC-Candidate in our data set while the bottom figure shows all 16 rings for the farthest-from-average CTC-Candidate in our data set. The y-axis is the log of the amplitude and the x-axis corresponds to the wavenumber. A image value curve with m points will generate an amplitude spectrum for m frequencies and we recall for the ring of radius r there will be $8r$ image values sampled. The radius increases left to right and top to bottom.

In the cost function shown in (1) there are several terms to be defined. First, w is defined to be the normal vector to the linear separating surface. Next, x_i is a data point with corresponding label y_i. When trying to separate two groups of data, the standard convention is to label one class as +1 and the other with −1. Lastly, in (1) we see the variable C which is a parameter that determines the contribution of the loss function to the cost function. All accuracies presented here in have been generated using $C = 1/2$. We refer to this formulation as an "l_2 loss function SVM" based on the second component of the cost function referred to as the loss function. A loss function penalizes for generating a hyperplane that misclassifies events. In particular, an l_2 loss function penalizes more for points that are strongly misclassified and less for those that hover near the hyperplane. This formulation of an SVM is referred to as l_1 regularized because of computing the 1-norm on the vector w inside the cost function. Using the 1-norm of w encourages sparsity in the normal vector. This, in turn, encourages sparsity of features used in a classification task due to the definition of the decision function.

5. Results

Our results can be divided into two categories: quantitative and qualitative. In our quantitative results section we will discuss the accuracy of the different decision functions for each branch of our previously discussed decision tree. Next, in the qualitative results section we will provide reconstructions of cells using features selected by the decision functions in 95% of trials for a given classification task.

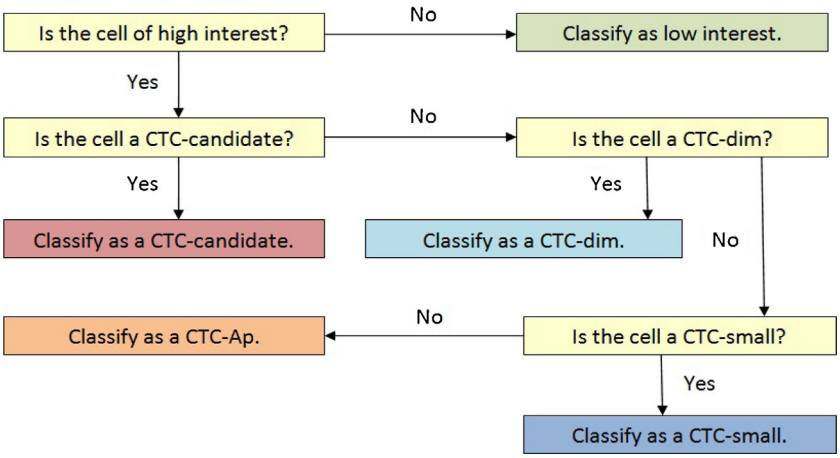

Fig. 5. Schematic of the decision tree classification structure we are building based on strongly differentiating structural features of cells. The answer to each question is determined by a support vector machine classifier.

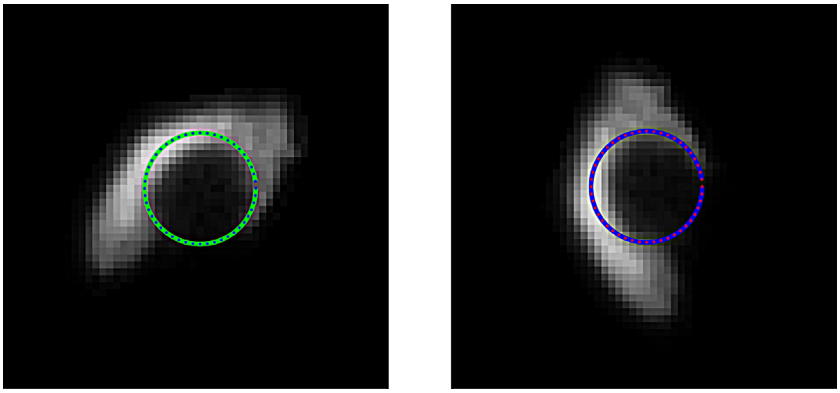

Fig. 6. The result of rotating a cell 60 degrees counterclockwise from the original orientation of the cell. The image on the left the cell in its original orientation on the slide image with the ring of radius 8 pixels overlaid while the right image is the numerically rotated image of the original cell generated using bicubic interpolation, also with the ring of radius 8 pixels overlaid.

5.1. Quantitative results

In accordance with the decision tree structure outlined in the Classification Structure section, we need to determine decision functions that can first differentiate between white blood cells (WBCs) and events of interest (EOI). Next, we want to be able to separate the EOI into CTC-Candidate and all marginal populations. Third, we want to separate CTC-dim from CTC-Ap and CTC-Small within the marginal populations. Last, we want to separate CTC-Small from CTC-Ap. In Table 1 we see that we obtain the accuracy for each decision function in the classification tree to be 99.52± 0.48 %, 92.16± 2.70 %, 89.76± 3.89 %, and 82.48± 5.11 %, respectively. The classification tasks become more challenging and have larger standard deviations the farther down the decision tree we move. Of particular interest is the classification task separating CTC-Ap from CTC-Small which suggests higher levels of overlap between these two populations which concures with the Scipps PSOC visually determined definitions of these two cell populations. Given this result, we decided to also explore all other pairwise classification tasks, the results of which are also shown in Table 1. A discussion of these results and their connection to pathologist experience is included in the discussion section.

Also shown in Table 1 are the number of features that were selected as important for classification by enforcing sparsity in 95% of the trials run. In each trial the data for each class in the classification task is randomly partitioned to have 75% used for training and the remaining 25% for testing, as well as matching the size of the two groups. There does not appear to be a connection between the accuracy of a given classification task and the number of features selected in 95% of the trials. Table 1 also highlights an important fact about the CTC-Ap cells. All of the pairwise classification tasks separating CTC-Aps from another cell of interest have the lowest pairwise classification accuracies. This reflects the biological

Table 1
The first column of the table states the classification task of interest. Each classification task was run 25 times where 75% of the data of each class was randomly selected and used for training, while the remaining 25% was used for testing. In the event that the size of the classes differ, we randomly select samples from the larger class to match the size of the smaller class and then separate into 75/25 partitions. The second column shows the number of features (NOF), out of 3264, that are selected as important for the classification task in 95% of the trials for the given task. The final column gives the overall accuracy on a given classification task.

Classification Task	No. of features	Accuracy
WBC vs Events of Interest	85	99.52 ± 0.48
CTC-Candidate vs. All other	122	92.16 ± 2.70
CTC-Candidate vs. CTC-Ap	63	92.64 ± 3.35
CTC-Candidate vs. CTC-dim	84	93.76 ± 2.79
CTC-Candidate vs. CTC-Small	63	93.20 ± 3.16
CTC-Ap vs. All Other	44	81.20 ± 5.26
CTC-Ap vs. CTC-dim	110	89.50 ± 3.60
CTC-Ap vs. CTC-Small	167	82.48 ± 5.11
CTC-dim vs. All Other	75	86.40 ± 4.32
CTC-dim vs. Other Marginal	55	89.76 ± 3.89
CTC-dim vs. CTC-Small	76	96.00 ± 2.77
CTC-Small vs. All Other	48	87.12 ± 5.42

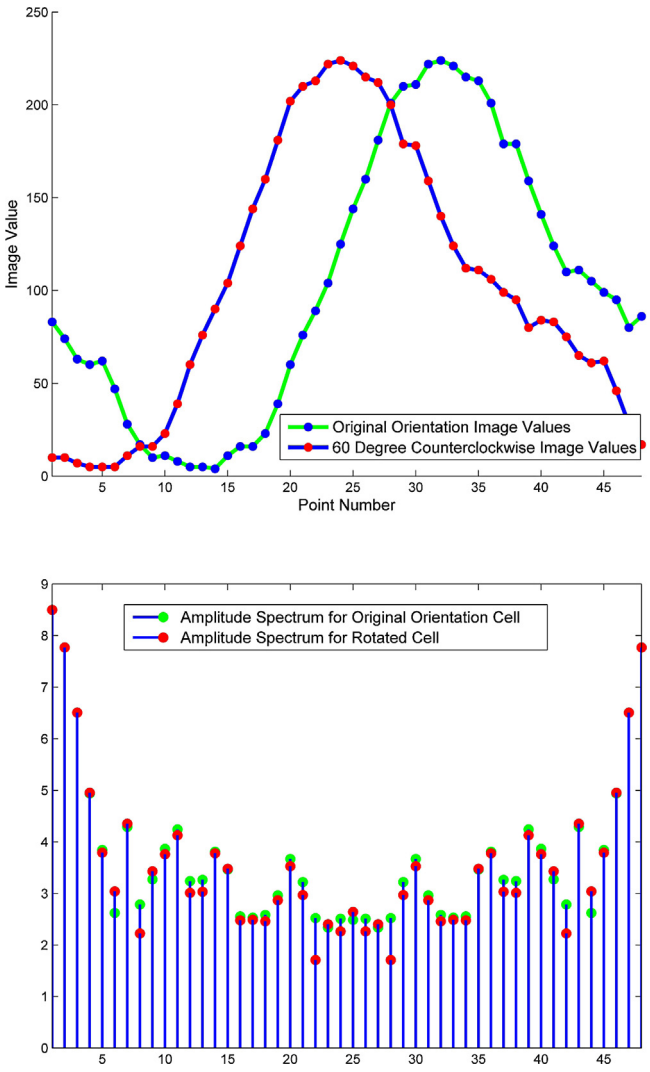

Fig. 7. The rotational invariance of the amplitude spectrum. On the top we see the set of image values interpolated along the ring of radius 8 pixels for both the rotated and unrotated versions of the cell. The bottom figure shows the log of the amplitude spectrums for each of the two sets of image values.

property of this class. CTC-Aps are cells undergoing apoptosis. Apoptosis involves many stages and consequently has the largest variation within the populations of interest as the cells likey represent different events in the apoptotic cascade.

In evaluating the significance of the information provided in Table 1, it is important to acknowledge what the number of reoccurring features means. A large number of reoccurring features means that the classification task is more robust to the training set, while a small set of overlapping features means there is more variation in the classifiers. The features that do reoccur with high frequency, however, provide information as to what are consistently important features. To evaluate what these results reveal, we have broken down the distribution of the reoccurring features across each channel for each channel for all classification tasks in Tables 2–4. In these tables, we have separated a cell into four regions. First, the inner region refers to the area contained within the first four rings, a circle of radius 4 pixels oriented about the cell center. Second, the inner-center of a cell corresponds to the area after the fourth ring up to, and including, the eighth ring. Next, the outer-center of the cell is the area after the eighth ring up to, and including, the twelfth ring. Last, we refer to the area after the twelfth ring up to, and including, the sixteenth as the outer cell.

Table 2 highlights some very interesting observations. First, from the first row we note that the distinguishing information between CTC-Candidates and other cells of interest, in the cytokeratin channel, is contained outside of the the inner-region of the cell. A second observation from this table is that in all of the pairwise classification tasks including CTC-Candidates, the outer-regions (outer-central and outer) contain a higher number of important features for distinguishing CTC-Candidates than the inner-regions. Third, we notice that the pairwise classification tasks involving CTC-Ap have the highest number of cytokeratin features, and that they occur in significant quantity in all of the cell regions. Fourth, we see that no classification task involving CTC-dim as an individual population have a frequently reoccurring cytokeratin feature in the inner-cell. Next, we note that the distribution of reoccurring features are nearly identical for the CTC-Candidate versus CTC-Ap and CTC-Candidate versus CTC-small classification tasks. Additionally, we see that the last four rows of the table contain the smallest number of cytokeratin features reoccurring.

From Table 3 we can draw additional conclusions. Of particular note is the non-trivial nature of this table. This table shows that, despite what we initially thought, the CD-45 channel plays a role in many of the classification tasks involving cells of interest. For example, in the row describing CTC-Ap vs. CTC-Small we see the highest number of CD-45 features. This could suggest one of several hypotheses. First, since CD-45 is a characteristic of white blood cells, does the presence of outer-region CD-45 features suggest frequent overlap with white blood cells for one of the two populations? Second, does the presence of the CD-45 suggest that our method is capturing features that support the knowledge that apoptotic cells non-specifically bind to antibodies as dying cells may have more epitopes exposed? Third, are there existing subpopulations of these cells of interest which express non-trivially in the CD-45 channel?

Finally, we consider the significance of DAPI features, as shown in Table 4. Similar to Table 2, we can immediately note the small

Table 2
This table shows the break down of the location of the reoccurring features in a given classification task within the cytokeratin channel as a fraction of the total number of reoccurring features. Inner refers to occurring within the first four rings. Inner-central refers to occurring within the fifth through eighth rings. Outer-central refers to occurring within the ninth to twelfth rings. Lastly, outer refers to occurring within the thirteenth to sixteenth rings.

Classification Task	Inner	Inner-Central	Outer-Central	Outer
CTC-Candidate vs. All other	1/122	13/122	16/122	22/122
CTC-Candidate vs. CTC-Ap	0/63	4/63	11/63	10/63
CTC-Candidate vs. CTC-dim	0/84	5/84	19/84	12/84
CTC-Candidate vs. CTC-Small	0/63	3/63	12/63	7/63
CTC-Ap vs. All Other	3/44	9/44	8/44	4/44
CTC-Ap vs. CTC-dim	10/110	29/110	25/110	13/110
CTC-Ap vs. CTC-Small	12/167	24/167	16/167	15/167
CTC-dim vs. All Other	0/75	8/75	6/75	5/75
CTC-dim vs. Other Marginal	0/55	5/55	3/55	4/55
CTC-dim vs. CTC-Small	0/76	1/76	8/76	8/76
CTC-Small vs. All Other	2/48	6/48	2/48	7/48

Table 3
This table shows the break down of the location of the reoccurring features in a given classification task within the CD-45 channel as a fraction of the total number of reoccurring features. Inner refers to occurring within the first four rings. Inner-central refers to occurring within the fifth through eighth rings. Outer-central refers to occurring within the ninth to twelfth rings. Lastly, outer refers to occurring within the thirteenth to sixteenth rings.

Classification Task	Inner	Inner-Central	Outer-Central	Outer
CTC-Candidate vs. All other	4/122	6/122	11/122	16/122
CTC-Candidate vs. CTC-Ap	3/63	4/63	4/63	5/63
CTC-Candidate vs. CTC-dim	4/84	4/84	5/84	6/84
CTC-Candidate vs. CTC-Small	2/63	4/63	4/63	7/63
CTC-Ap vs. All Other	0/44	0/44	5/44	1/44
CTC-Ap vs. CTC-dim	0/110	0/110	5/110	5/110
CTC-Ap vs. CTC-Small	0/167	9/167	16/167	15/167
CTC-dim vs. All Other	0/75	0/75	4/75	9/75
CTC-dim vs. Other Marginal	0/55	0/55	2/55	4/55
CTC-dim vs. CTC-Small	0/76	0/76	5/76	17/76
CTC-Small vs. All Other	1/48	0/48	1/48	8/48

Table 4
This table shows the break down of the location of the reoccurring features in a given classification task within the DAPI channel as a fraction of the total number of reoccurring features. Inner refers to occurring within the first four rings. Inner-central refers to occurring within the fifth through eighth rings. Outer-central refers to occurring within the ninth to twelfth rings. Lastly, outer refers to occurring within the thirteenth to sixteenth rings.

Classification Task	Inner	Inner-Central	Outer-Central	Outer
CTC-Candidate vs. All other	1/122	18/122	7/122	7/122
CTC-Candidate vs. CTC-Ap	1/63	14/63	3/63	4/63
CTC-Candidate vs. CTC-dim	0/84	8/84	6/84	15/84
CTC-Candidate vs. CTC-Small	0/63	17/63	3/63	4/63
CTC-Ap vs. All Other	0/44	8/44	1/44	5/44
CTC-Ap vs. CTC-dim	1/110	3/110	7/110	12/110
CTC-Ap vs. CTC-Small	0/167	17/167	8/167	35/167
CTC-dim vs. All Other	3/75	11/75	18/75	11/75
CTC-dim vs. Other Marginal	2/55	14/55	14/55	7/55
CTC-dim vs. CTC-Small	0/76	13/76	13/76	1/76
CTC-Small vs. All Other	6/48	3/48	9/48	3/48

number of features located in the inner-cell. Next, we note that when separating CTC-Candidates from all other cells of interest, or from either CTC-Ap or CTC-Small in pairwise classification, the bulk of the reoccurring features occur in the inner-central region of the cell. Also, we observe that CTC-dim versus all other marginal and versus CTC-Small show heavily weighted central-regions of the cells. Furthermore, we see that the outer-region of the cell, in the DAPI channel, is the most heavily weighted region in three classification tasks: CTC-Candidate vs. CTC-dim, CTC-Ap vs. CTC-dim, and CTC-Ap vs, CTC-Small.

The above observations reinforce many of the characteristics currently considered in manual classification. CTC-Candidates, for example, are identified, in part, by the intensity of cytokeratin, size, and cytokeratin which encompasses the over sized nucleus. This is reflected in the significance of the inner-central DAPI features together with the heavily weighted outer-region cytokeratin features. Additionally, CTC-dims are characterized by their large nucleus and low cytokeratin expression. Thus, in the the CTC-Candidate versus CTC-dim classification the presence of heavily weighted outer-region DAPI and cytokeratin features is in agreement with manual classification. A CTC-Candidate will have have cytokeratin in the outer-region of the cell but limited DAPI, while a CTC-dim will have limited cytokeratin expression overall but strong DAPI expression in the outer-region of the cell. These reflections of characterizations used in manual classification show that FRDs capture both interpretable, significant and biologically relevant structural information pertaining to different classification tasks.

5.2. Qualitative results

Given the high accuracy of each of our decision functions, and the insights afforded to us by determining reoccurring features, we want to visualize the reoccurring features and observe if these visualizations could be interpreted by pathologists in a meaningful way. For each classification task there were 25 trials run. In each trial, the data was randomly-partitioned for training and testing using the standard 75/25 partitioning. Within each trial, a decision function was built using the LIBLINEAR l_1 regularized, l_2 loss function linear support vector machine. In each trial, a subset of the features are selected as important for the given classification task. After running the trials we can identify the features that are selected in 95% of the trials, the number of which can be seen in Table 1. The design of the FRD allows us to invert the descriptor and visualize individual features or observe a single cell reconstructed using a specific set of features. We have reconstructed specific cells using the set of features selected in 95% of trials for various classification tasks. In this way we can visualize our results and further understand the differentiating structure for an indicated classification task.

We define two different cells of each class. The cell with the prefix “closest” refers to the cell of a given class which had the FRD closest to the average FRD across that class. Similarly, we define the prefix “farthest” to mean the cell of class who’s FRD was the farthest from the average FRD across that class. We determine the closest and farthest from average FRD using the Euclidean distance measure. The following figures show the reconstructions of the closest and farthest from average cells of each class involved in the indicated classification task. In Figs. 8–18 we see reconstructions of cells using features from different classification tasks. For each of these figures, the image directly below a labeled cell image is the reconstruction of the labeled cell for the given feature set. Additionally, all images have been shown on using a color axis bounded between 0 and 255. Thus, though some of the images may appear dim the intensity of the colors in the reconstruction are not artifacts but rather capture important differences.

First, contained in Fig. 8 we see reconstructions of representatives of each cell type using the reoccurring features from the

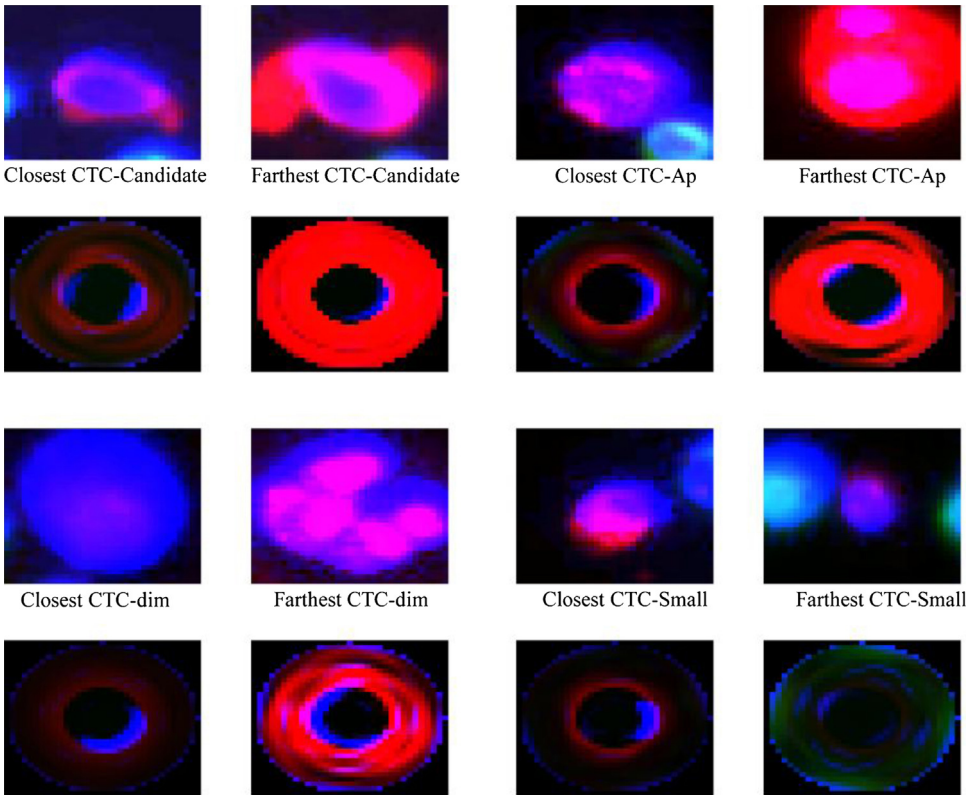

Fig. 8. A visualization of the differential structure used to separate CTC-Candidates from all other cells of interest.

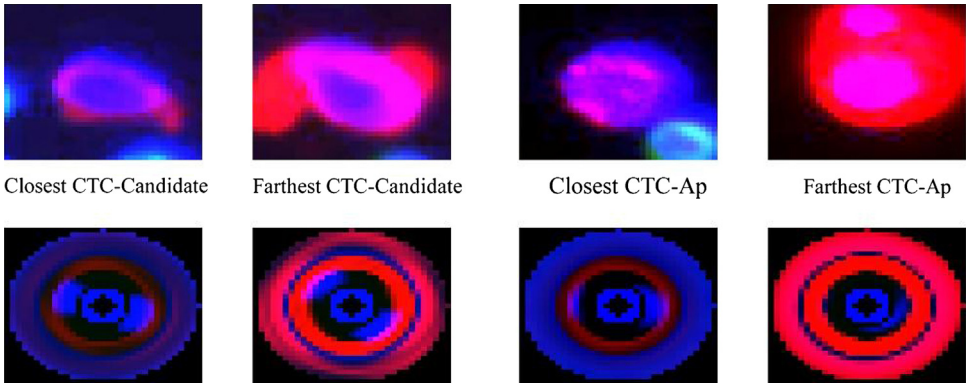

Fig. 9. A visualization of the differential structure used to separate CTC-Candidates from CTC-Aps.

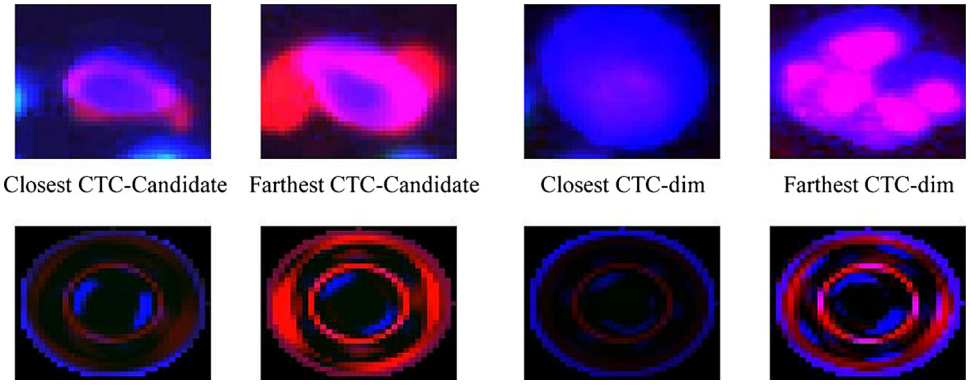

Fig. 10. A visualization of the differential structure used to separate CTC-Candidates from CTC-dims.

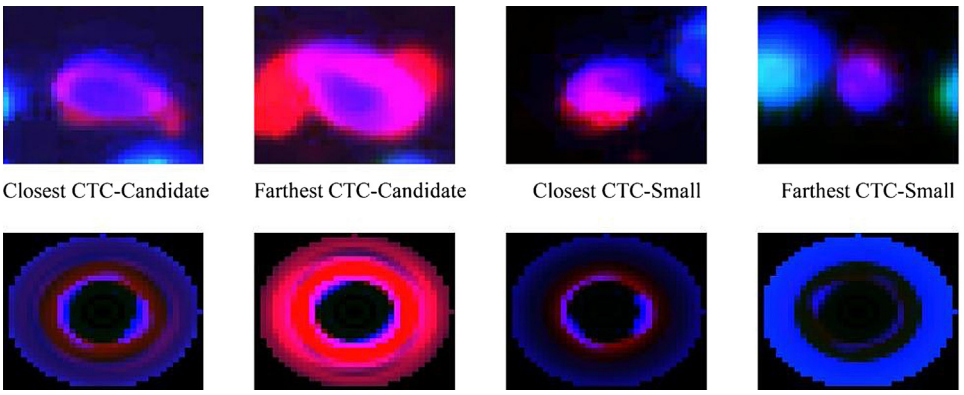

Fig. 11. A visualization of the differential structure used to separate CTC-Candidates from CTC-Smalls.

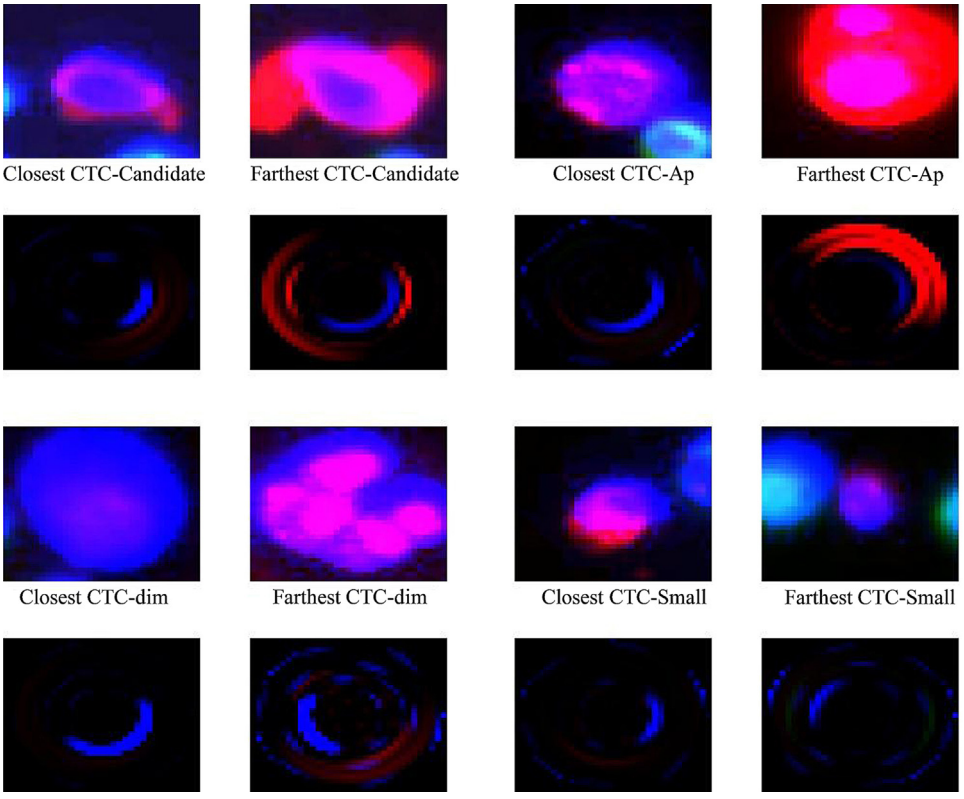

Fig. 12. A visualization of the differential structure used to separate CTC-Aps from all other cells of interest.

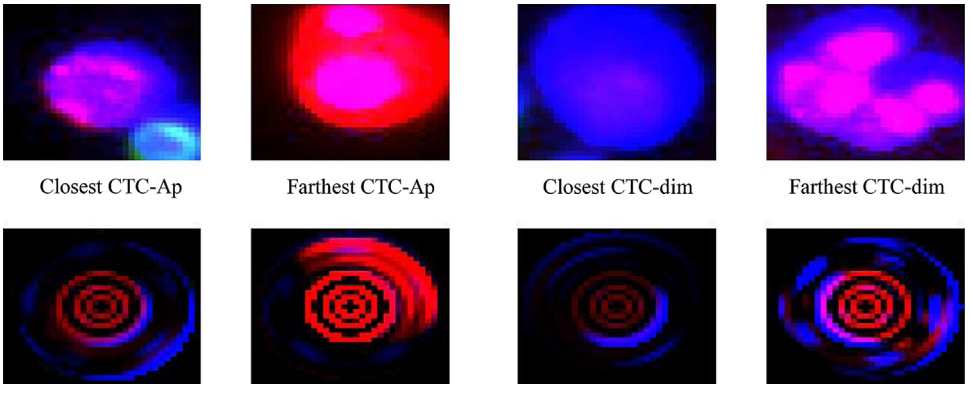

Fig. 13. A visualization of the differential structure used to separate CTC-Aps from CTC-dims.

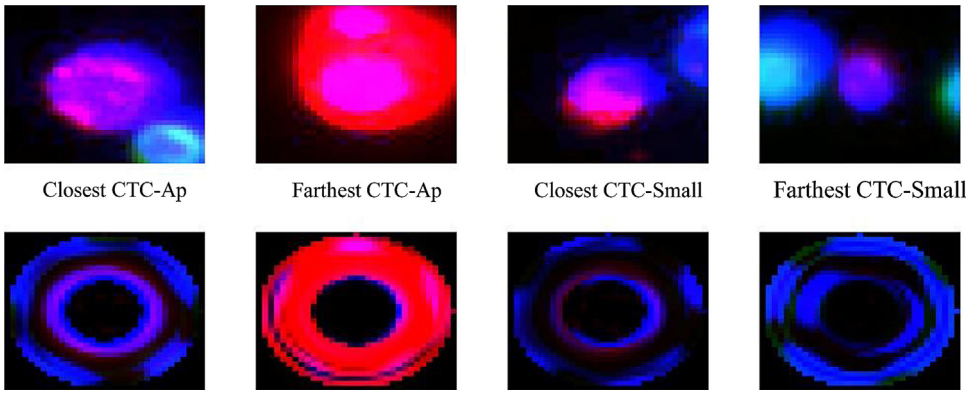

Fig. 14. A visualization of the differential structure used to separate CTC-Aps from CTC-Smalls.

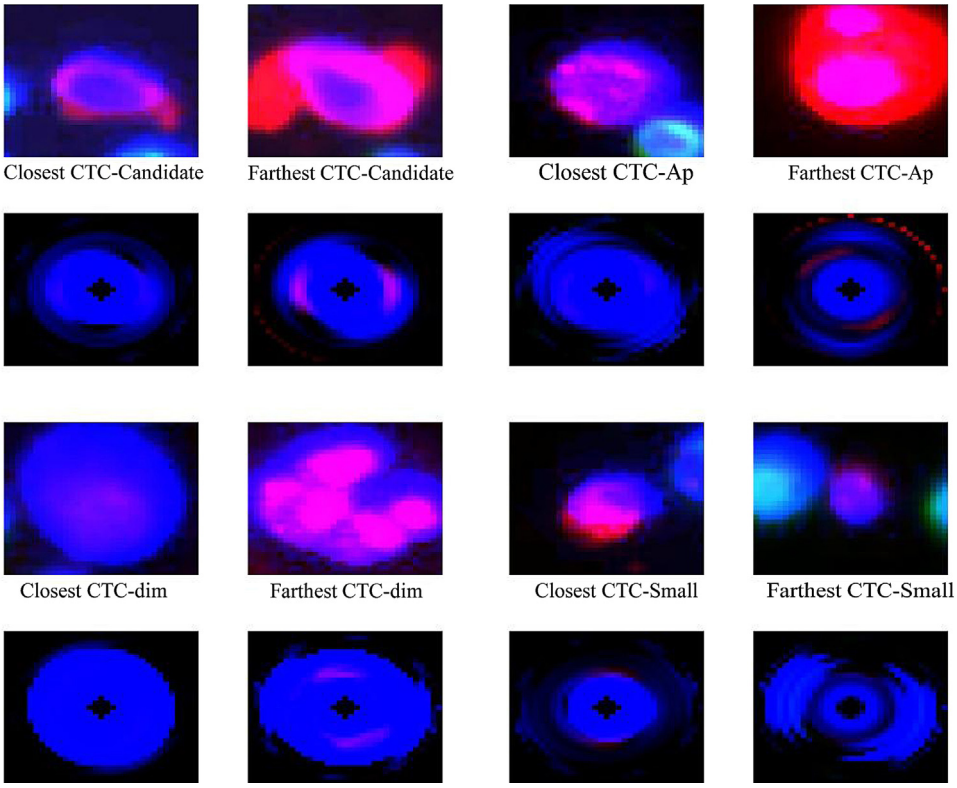

Fig. 15. A visualization of the differential structure used to separate CTC-dims from all other cells of interest.

CTC-Candidate versus all other cells of interest classification task. We see, in this figure, that CTC-Candidates are distinguished from the other cells of interest based on the uniformity and thickness of cytokeratin in the central-cell together with greater amounts of inner-central DAPI present. Within this figure we can also observe the similarities between the reconstructions for each of the farthest-from-average cells. The similarities between the farthest-from-average cells for these cell types illustrate the high-level of variation in the current labeling of data. Furthermore, we observe the lack of features/structure from the inner cell. This lack of inner structure suggests that at the given resolution, the central cell/center of the nucleus contains limited structurally differentiating information. Biologically, this is significant since research has shown that there is differentiating information contained in the nuclei and that at the current resolution this information is not being be captured.

Next, Fig. 9 captures discriminating information between CTC-Candidates and CTC-Aps. Based on the reconstructions shown, we can infer that CTC-Candidates can be separated from CTC-Aps based on a more uniform and intense nucleus, more uniform cytokeratin extending into the outer-center region of the cell, and a less circular shape. The first inference is based on the bright inner cell DAPI. Our second inference is based on the purple hue of the outer regions of the closest-to-average CTC-Candidate, as purple suggests the presence of both DAPI and cytokeratin. Additionally, the purple in the closest-to-average CTC-Candidate appears uniform in its intensity suggesting uniformity in the two contributing channels.

According to Fig. 10 the combination of outer-central cytokeratin expression and outer DAPI expression contribute to the differential structure between CTC-Candidates and CTC-dims. Visually, the category of CTC-dim is more uniform than many of the others, but still contains size variation in the nucleus of the cells. Thus, since CTC-dims typically have larger nuclei than CTC-Candidates, the presence of outer cell DAPI features for discrimination is in agreement with visual classification criterion.

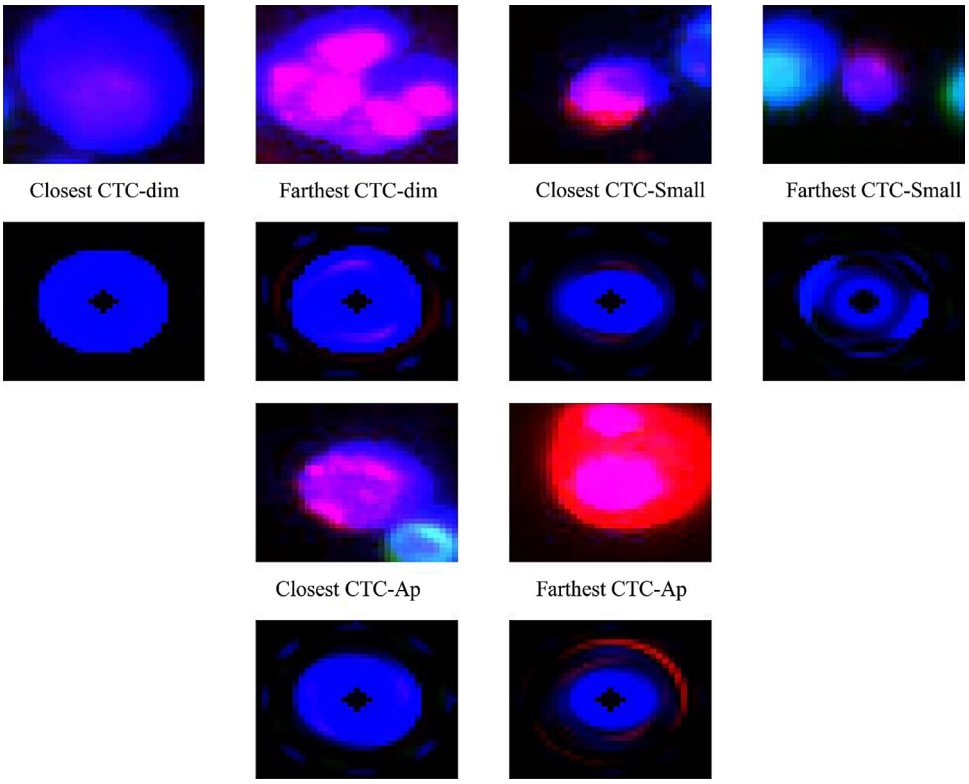

Fig. 16. A visualization of the differential structure used to separate CTC-dims from other marginal cell populations.

Fig. 11 illustrates the differentiating structure between CTC-Candidates and CTC-Smalls. Of note in this figure are the inner-central cytokeratin features and outer DAPI features. CTC-Smalls are biologically classified based on their high cytokeratin expression, but small nucleus and overall cell size. Thus, the presence of DAPI in the outer cell highlights the size variation between these two cell populations, as does the location of the cytokeratin which would not extend beyond the inner-central rings in a cell of small size. We also note, in this figure, the presence of outer DAPI features in the CTC-Small shown can be attributed to the closeness of the neighboring cells in both the closest-to-average and farthest-from-average cells.

The visualizations shown in Fig. 12 make the low performance of the CTC-Ap versus all other cells of interest classification task very understandable. From Table 1 we see that the CTC-Ap one-versus-all classification task has the lowest accuracy of all one-versus-all tasks. Visually, there appears to be much less structurally differentiating information contained in this classification task. Both the closest-to-average and farthest-from-average cells of each cell type are very similar to those of the CTC-Aps. As mentioned in the qualitative results section, CTC-Aps are cells undergoing apoptosis. This figure suggests that the labeling of CTC-Aps, based on visual inspection, likely has captured cells of the other cell populations which may be undergoing apoptosis. Additionally, we note that this classification task had the fewest number of reoccurring features suggesting that the classifier is highly dependent on the set of cells used for training. A consequence of the small number of reoccurring features is a less complete reconstruction.

Despite the lack of differential structure for separating CTC-Aps from all other cells of interest, there is differential structure in the remaining CTC-Ap pairwise classifications. For example, in Fig. 13 we see that inner cytokeratin and non-uniform outer DAPI are involved in separating CTC-Aps from CTC-dims. In fact, this classification task is the only task with apparent cytokeratin in the inner cell. Next, in Fig. 14 we again see the differential structure returning to the outer regions of the cell to differentiate between

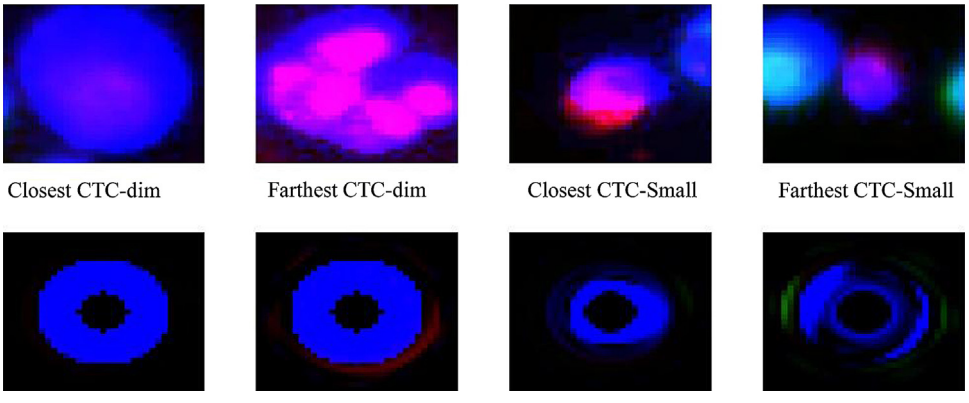

Fig. 17. A visualization of the differential structure used to separate CTC-dims from all CTC-Small.

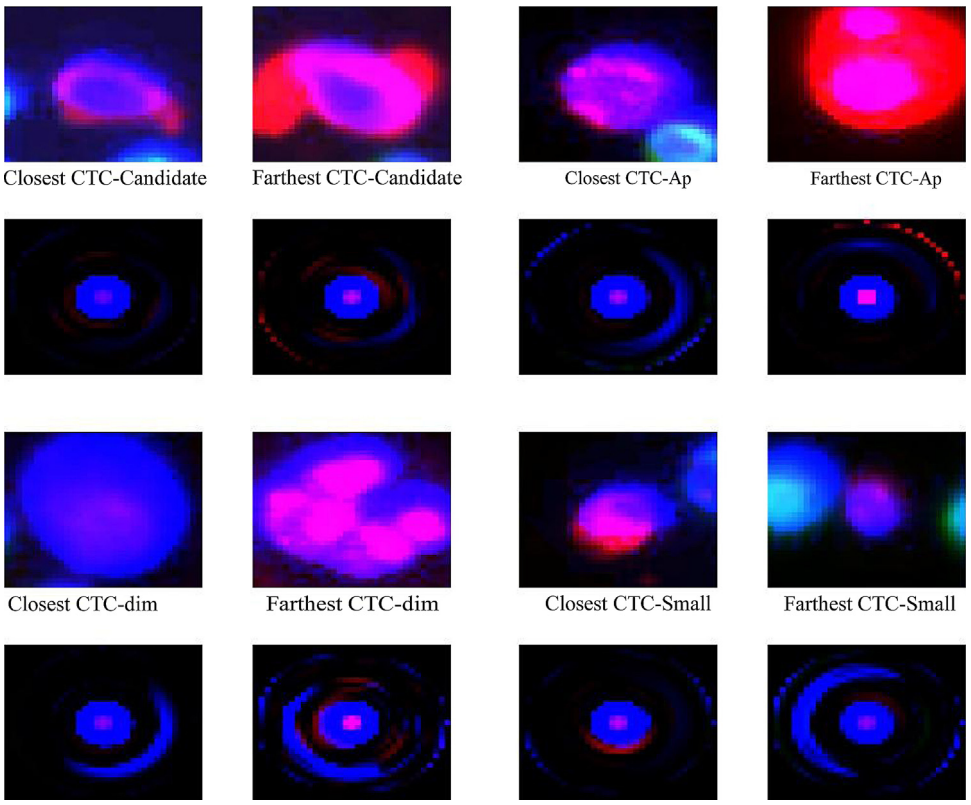

Fig. 18. A visualization of the differential structure used to separate CTC-Smalls from all other cells of interest.

CTC-Aps and CTC-Smalls. Although there is a weaker level of separation between CTC-Aps and CTC-Smalls, according to Table 1, by observing the reconstructions, there are clear differences between the cell populations involved. These differences are especially pronounced between the farthest-from-average cells, although the farthest-from-average CTC-Small closely resembles the closest-to-average CTC-Ap. In many of the classification tasks shown, we see greater similarity between the farthest-from-average cells of different types. Biologically, the patters of similarity in the CTC-Ap versus CTC-Small classification task could suggest a greater overlap of these two populations.

Moving to Fig. 15 we see our first case of a heavily-weighted inner DAPI, and the DAPI channel being used almost exclusively, for separating cell populations. Fig. 15 shows reconstructions based on the classification task of separating CTC-dims from all other cell populations. Observing the reconstructions we see DAPI features, the other channel features are overwhelmed, and note the size, circularity, and uniformity of the DAPI in the CTC-dim reconstructions. The other populations of interest, however, have more asymmetry and smaller areas of intense DAPI expression. Again, this mirrors the characterization used in visual classification in which a CTC-dim is characterized largely by the size of the nucleus and limited to nonexistent cytokeratin.

The remaining classification tasks for distinguishing CTC-dims from other populations are visualized in Figs. 16 and 17. In both the classification task separating CTC-dims from the remaining marginal populations and separating CTC-dims from CTC-Smalls, we again see the majority of the features coming from the DAPI channel. Also, we again see that the reconstructions of CTC-dims show a larger region of uniform DAPI intensity which are more uniform and circular than the reconstructions of the other cells.

Finally, we consider Fig. 18 which contains the reconstructions of all cells using the reoccurring features selected to separate CTC-Smalls from all other cells of interest. This classification task also uses the second fewest number of reoccurring features for reconstruction, suggesting, like CTC-Ap versus all other cells of interest, that the classifier is highly dependent on the training set. We do, however, see that the inner and inner-central DAPI and cytokeratin play a role in differentiating between the cells. Thus, we are again capturing the size information of the CTC-Smalls.

The culmination of these visualizations allow us to describe the features that are consistently evaluated at each branch of our decision tree. First, we classify according to the features shown in Fig. 8. Next, we classify based on the features shown in Fig. 16. Finally, we separate based on the features shown in Fig. 14. This amounts to first looking at the combination of central cell cytokertain and inner cell DAPI, then looking for at the uniformity, size, and circularity of the DAPI features from the inner cell to the outer-central cell, and then finally looking at the combination of outer-central to outer DAPI features and outer-central cytokeratin.

The variation across reconstructions of a single cell show that there is strongly differentiating structure within a cell that changes depending on the classification the reconstruction is based on. By observing the various reconstructions we can highlight tasks where there is more or less differentiating structure. Again, this knowledge allows us to understand the classifications and define hard lines between what we look for in a cell in order to belong to a certain cell population.

While many of our features can be connected to features that are currently used in visual classification, there are several benefits to computer automated classification. First, in addition to being objective an automated classification method, when properly optimized, could reduce valuable labor hours needed to evaluate a single patient sample. Second, due to the construction of the FRD, we are evaluating information extracted from a pixel level which cannot be accomplished by the human eye. Thus, by using an automated method, we have the ability to use all available information in a cell to classify all events. Finally, and most importantly, by using

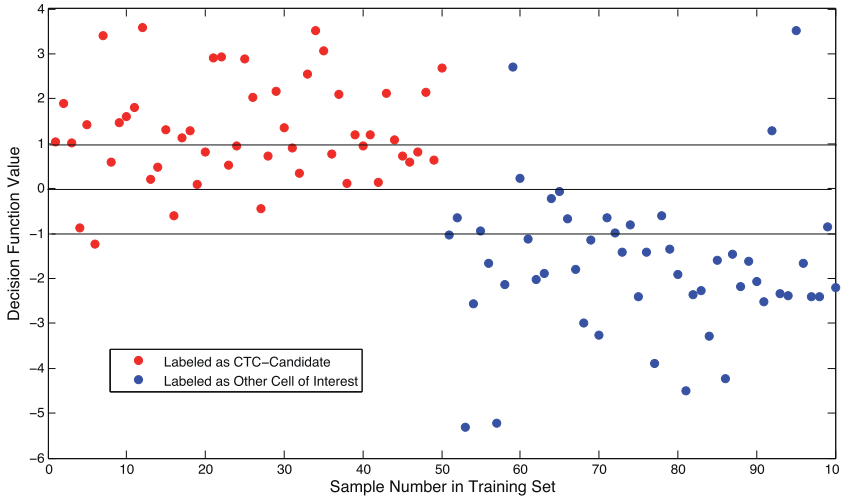

Fig. 19. A plot of the decision function values for the samples in the training set for one trial of the CTC-Candidate versus all other cells of interest classification task.

a numerical approach to classification we are able to express the unbiased confidence of our classification of a single event.

Using a support vector machine classifier, we determine the class of a cell in a training set based on the sign of the decision function value. For data which is linearly separable, the decision values of one class will all be greater than or equal to one, while all of the decision values of the other class will be less than or equal to one. Thus, for data points in the training set, the larger the magnitude of the decision function value, the greater our confidence in the classification of the event. For example, in Fig. 19 we see a plot of the decision function values for all the points in a training set for one trial of the CTC-Candidate versus all other cells of interest classifier. In this example we consider the points that have been visually classified as CTC-Candidates to be the positive (red) class, while the cells which have been visually classified as other cells of interest are the negative (blue) class. Therefore, any point with a decision value greater than zero has been classified by our classifier as a CTC-Candidate, and any with a decision value less than zero has been classified as another type of cell of interest. For any point with a decision value magnitude greater than one, we have confidence in the classification (more confidence the larger the magnitude). However, points with decision values between one and negative one, we are less confident in the classification. In this way we can quantify the confidence of a cell's classification, which cannot easily be routinely accomplished using visual classification. The results here show that perhaps some cells visually classified into certain cell populations may actually fit better within a different cell population.

To conclude, from the images provided, one can readily observe a significant amount of variation within a single cell class. This is important to be able to quantify for several reasons. Of primary concern is ongoing research connecting quantities of circulating tumor cells to cancer status. If these correlations between quantity of circulating tumor cells and cancer status are based on highly subjective classifications, then there will be doubt as to the validity of those claims. However, if using these visualization techniques and classification methods allow us to not only confidently classify, but also capture all cell types of interest, then we will have confidence in any connections that we may find in the future. While the variation we observe is over a small dataset, it allows us to understand the challenges in generating a set of classifiers that can be readily applied to many patient samples across various cancer types. In supervised learning methods, it is very important to select the right set of data to train your classifiers. From these results, we see that we will need to systematically scale our training set to produce a more universal set of classifiers both within a cancer type and across all cancers.

6. Discussion

The current methods for CTC isolation and characterization are both representative of traditional pathology practice of visual evaluation but also lend themselves to sophisticated and rigorous mathematical frameworks as the data sets are all digital by nature. Human inspection, while of high quality, will always be qualitative and limited to certain scales of data sets leading to a self-evident risk of false negatives. The risk of false negatives assosciated with manual classification has been present for years in the context of visually based cell classification. A notable area in which problematic false negative rates have occured is the pap smear. Often, the high false negative rates can be attributed to the fact that manual screening of slides is very labor intensive and requires that technicians, or pathologists, be capable of high levels of concentration for extended periods [29]. For two different cell types of interest in [30] they state the rates of false negative diagnoses in manual rescreening to be 100% and 73.2%. Additional results in [30] show a greater than 25% reduction in the laboratory's false negative rate when using an automated method for rescreening. Alternatively, in [31] they state a 17.6% mannual false negative diagnosis rate and provide results indicating almost a 7% improvement in the false negative rate when asisted by an automated method in primary screening. Although no exact percentage is guaranteed, there are several scenarios in which a statistically significant reduction in false negative rates has been reported after an automated or semi-automated method has been implemented.

The FRD method, when applied to large datasets, could provide more confidence around having more completely interrogated a patient's sample. There are significant advantages in using computational methods in that patterns can be identified more rapidly. For example, no single parameter currently exists that would distinguish CTCs originating from different organs but a large scale computational approach might indeed be able to do so. However, the more detailed and comprehensive method described herein may lend further insight into distinguishing cancer types and stages of disease.

The Fourier-ring descriptor (FRD) methods developed in this paper are the first that we know of that exploit the rotational structure of cells identified in peripheral blood samples from patients with metastatic breast, prostate and lung cancer cells. We have used these FRDs, along with a linear support vector

machine decision tree classifier to exploit the size variations and morphological distinctions among the cell populations to obtain reasonable and quantifiable accuracy benchmarks. While we are not proposing that this automated technique be used in place of visual inspection in making clinical decisions, we do see our methods as offering clinically relevant quantifiable support in their decision processes. Hence, we view this as a first step in developing a useful computer-vision based clinical support tool for the active monitoring of cancer patients based on peripheral blood samples.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant Nos. DMS-1228308 and DMS-0915262. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.compmedimag.2014.10.003.

References

[1] Giuliano M, Giordano A, Jackson S, Hess KR, De Giorgi U, Mego M, et al. Circulating tumor cells as prognostic and predictive markers in metastatic breast cancer patients receiving first-line systemic treatment. Breast Cancer Res 2011;13(3):R67.
[2] Cohen SJ, Punt CJA, Iannotti N, Saidman BH, Sabbath KD, Gabrail NY, et al. Prognostic significance of circulating tumor cells in patients with metastatic colorectal cancer. Ann Oncol 2009;20(7):1223–9.
[3] Nieva J, Wendel M, Luttgen MS, Marrinucci D, Bazhenova L, Kolatkar A, et al. High-definition imaging of circulating tumor cells and associated cellular events in non-small cell lung cancer patients: a longitudinal analysis. Phys Biol 2012;9(1):016004.
[4] Danila DC, Heller G, Gignac GA, Gonzalez-Espinoza R, Anand A, Tanaka E, et al. Circulating tumor cell number and prognosis in progressive castration-resistant prostate cancer. Clin Cancer Res 2007;13(23):7053–8.
[5] Pachmann K, Camara O, Kavallaris A, Krauspe S, Malarski N, Gajda M, et al. Monitoring the response of circulating epithelial tumor cells to adjuvant chemotherapy in breast cancer allows detection of patients at risk of early relapse. J Clin Oncol 2008;26(8):1208–15.
[6] Riethdorf S, Fritsche H, Müller V, Rau T, Schindlbeck C, Rack B, et al. Detection of circulating tumor cells in peripheral blood of patients with metastatic breast cancer: a validation study of the CellSearch system. Clin Cancer Res 2007;13(3):920–8.
[7] Marrinucci D, Bethel K, Kolatkar A, Luttgen MS, Malchiodi M, Baehring F, et al. Fluid biopsy in patients with metastatic prostate, pancreatic and breast cancers. Phys Biol 2012;9(1):016003.
[8] Paterlini-Brechot P, Benali NL. Circulating tumor cells (ctc) detection: clinical imapct and future directions. Cancer Lett 2007;253:180–204.
[9] Chen S, Zhao M, Wu G, Yao C, Zhang J. Recent advances in morphological cell image analysis. Comput Math Methods Med 2012.
[10] Zhang D, Lu G. Review of shape representation and description techniques. Pattern Recogn 2004;37(1):1–19.
[11] Elbischger P, Geerts S, Sander K, Ziervogel-Lukas G, Sinah P. Algorithmic framework for hep-2 fluorescence pattern classification to aid auto-immune diseases diagnosis. IEEE international symposiumon biomedical imaging: fromnano to macro (ISBI'09). 2009. p. 562–5.
[12] Perner P, Perner H, Müller B. Mining knowledge for hep-2 cell image classification. Artif Intell Medicine 2002;26(1):161–73.
[13] Wiliem A, Wong Y, Sanderson C, Hobson P, Chen S, Lovell BC. Classification of human epithelial type 2 cell indirect immunofluoresence images via codebook based descriptors. ArXiv 2013.
[14] Comaniciu D, Meer P, Foran DJ. Image-guided decision support system for pathology. Mach Vis Appl 1999;11(4):213–24.
[15] Huang Y-L, Jao Y-L, Hsieh T-Y, Chung C-W. Adaptive automatic segmentation of hep-2 cells in indirect immunofluorescence images. IEEE international conference on sensor networks, ubiquitous and trustworthy computing (SUTC'08). 2008. p. 418–22.
[16] Wählby C, Lindblad J, Vondrus M, Bengtsson E, Björkesten L. Algorithms for cytoplasm segmentation of fluorescence labelled cells. Anal Cell Pathol 2002;24(2):101–11.
[17] Wählby C, Sintorn I-M, Erlandsson F, Borgefors G, Bengtsson E. Combining intensity, edge and shape information for 2d and 3d segmentation of cell nuclei in tissue sections. J Microsc 2004;215(1):67–76.
[18] Hiemann R, Büttner T, Krieger T, Roggenbuck D, Sack U, Conrad K. Challenges of automated screening and differentiation of non-organ specific autoantibodies on hep-2 cells. Autoimmun Rev 2009;9(1):17–22.
[19] Agrawal P, Vatsa M, Singh R. Hep-2 cell image classification: a comparative analysis. Machine learning in medical imaging. Springer; 2013. p. 195–202.
[20] Abramoff MD, Magalhaes PJ, Ram SJ. Image processing with imageJ. Biophoton Int 2004;11(7):36–42.
[21] Nieva J, Wendel M, Luttgen MS, Marrinucci D, Bazhenova L, Kolatkar A, et al. High-definition imaging of circulating tumor cells and associated cellular events in non-small cell lung cancer patients: a longitudinal analysis. Phys Biol 2012;9(1):016004.
[22] Emerson T. Automated detection of circulating cells using low level features. Colorado State University; 2013. Master's thesis.
[23] Tsai D-M, Tsai Y-H. Rotation-invariant pattern matching with color ring-projection. Pattern Recogn 2002;35(1):131–41. ISSN 0031-3203.
[24] Tsai D-M, Chiang C-H. Rotation-invariant pattern matching using wavelet decomposition. Pattern Recogn Lett 2002;23(1-3):191–201. ISSN 0167-8655.
[25] Hipp J, Cheng J, Hanson JC, Yan W, Taylor P, Hu N, et al. Sivq-aided laser capture microdissection: a tool for high-throughput expression profiling. J Pathol Inform 2011;2(1):19.
[26] Hipp J, Cheng J, Toner M, Tompkins R, Balis U. Spatially invariant vector quantization: a pattern matching algorithm for multiple classes of image subject matter including pathology. J Pathol Inform 2011;2(1):13, http://dx.doi.org/10.4103/2153-3539.77175.
[27] Fan R-E, Chang K-W, Hsieh C-J, Wang X-R, Lin C-J. Liblinear: A library for large linear classification. J Mach Learn Res 2008;9:1871–4.
[28] Kirby M. Geometrical data analysis: an empirical approach to dimensionality reduction and the study of patterns. John Wiley and Sons, Inc.; 2001.
[29] Birdsong G. Automated screening or cervical cytology. Hum Pathol 1996;27:468–81.
[30] Renshaw AA, Lezon KM, Wilbur DC. The human false-negative rate of rescreening pap tests. Cancer Cytopathol 2001;93(2):106–10.
[31] Biscotti CV, Dawson AE, Dziura B, Galup L, Darragh T, Rahemtulla A, et al. Assisted primary screening using the automated thinprep imaging system. Am Clin Pathol 2005;123(2):281–7.

Tegan Emerson is a doctoral candidate in the Department of Mathematics at Colorado State University. She received her B.S. in Mathematics from Oregon State University and her M.S. in Mathematics from Colorado State University. She is interested in applications of geometric data analysis to biomedical challenges. Her research projects have included mathematical modeling of the human immune response, signal reconstruction for magnetic resonance imaging, data dimensionality reduction, and using Fourier-ring descriptors to characterize rare cells.

Michael Kirby received his B.S. in Applied Mathematics from Massachusetts Institute of Technology and his M.S. and Ph.D. in Applied Mathematics from Brown University. Upon completion of his degrees, he was awarded the DARPA Fellowship, Center for Fluid Dynamics, Turbulence and Computation, Brown University. He then was awarded the Alexander von Humboldt Fellowship, Institute for Information Sciences University of Tuebingen, Federal Republic of Germany. He was a visiting research fellow for the Engineering and Physical Sciences Research Council in the United Kingdom. Currently, he is a Professor in both the Department of Mathematics and the Department of Computer Science at Colorado State University.

Kelly Bethel joined Scripps Clinic Medical Group in 1999 as a member of the Scripps Clinic Medical Group and the Department of Pathology at Scripps Green Hospital and Scripps Clinic. She has academic appointments as an Adjunct Professor of Cell Biology at The Scripps Research Institute and Voluntary Assistant Clinical Professor at UCSD School of Medicine. As an active staff hematopathologist, Dr. Bethel supervises the diagnostic hematopathology service at Scripps Clinic/Scripps Green Hospital, serving patients at the Scripps Cancer Center and Bone Marrow and Stem Cell Transplant Program. After serving as the Program Director for a decade, she is currently the assistant program director of the hematopathology fellowship program at Scripps Green Hospital/Scripps Health. She has co-authored multiple research papers on subjects ranging from hairy cell leukemia to extramedullary and choroidal hematopoiesis, as well as circulating tumor cells. Currently the senior clinical investigator for the NIH/NCI funded Scripps Physical Sciences in Oncology Center, she studies circulating tumor cells and their relationship to primary and metastatic tumors in patients. Dr. Bethel works extensively at the translational research interface, having collaborated with the physical sciences research team of Dr. Kuhn for the past seven years and having supervised numerous fellows and trainees in translational research.

Anand Kolatkar received his B.A. in Chemistry and Biology from Augustana College and his Ph.D. from Rice University in Biochemistry. He then took a postdoctoral position at Stanford University in the field of Biochemistry and Crystallography. Trained as a biophysicist using macromolecular X-ray crystallography and diffuse X-ray scattering, Dr. Kolatkar has broadened his expertise to include fluorescence microscopy. He has been directly involved in the design and development of the latest generation of automated fluorescence microscopy hardware and software analysis systems in use at the Kuhn laboratory. He has evaluated a wide variety of different microscopy and imaging equipment and has extensive experience with image analysis and processing which will provide the required background necessary for this proposal. He is currently leading the automated imaging and image analysis program in the translational science laboratory and has developed the algorithms and interfaces for circulating tumor cell identification, relocation, and representation of that imagery to the pathologists.

Madelyn Luttgen received her B.S. in Biomedical Engineering from the University of California, Irvine. She was a research assistant Siemens Healthcare Diagnostics in Los Angeles, CA. She is currently a research assistant in the department of cell biology at The Scripps Research Institute in La Jolla, CA. The laboratory she works for has developed a reliable way to detect and to characterize circulating tumor cells isolated from the blood of cancer patients. She is a principal component in the analysis of cell images and identification of cellular events.

Stephen O'Hara is a Senior Scientist for DigitalGlobe, Inc., where he develops algorithms for largescale analysis of satellite imagery. He was a post-doctoral fellow at the Pattern Analysis Lab, Department of Mathematics, at Colorado State University from 2012 to 2013. He was awarded his Ph.D. from the Computer Science Department at Colorado State University in 2013.

Paul Newton received his B.S. in Applied Math/Physics at Harvard University and his Ph.D. in Applied Mathematics from Brown University. After a post-doctoral fellowship at Stanford University, he was Assistant and Associate Professor of Mathematics and The Center for Complex Systems Research at the University of Illinois Champaign–Urbana. He has held visiting appointments at Caltech, Brown, Hokkaido University, The Kavli Institute for Theoretical Physics at U.C. Santa Barbara, and The Scripps Research Institute. He is currently Professor of Applied Math, Engineering, and Medicine in the Viterbi School of Engineering and the Norris Comprehensive Cancer Center at the University of Southern California. He serves as Managing Editor of The Journal of Nonlinear Science, Advisor on Texts in Applied Mathematics Series, Springer-Verlag, New York, and is on The Center Advisory Committee for The Physical Sciences Oncology Center at The Scripps Research Institute in La Jolla, CA, where he serves as Project Leader, Mathematical Modeling: Physics and Mathematics of Cancer Metastasis.

Peter Kuhn received his Vordiplom in Physics from Julius Maximilians Universität Würzburg, Germany and his M.S. and Ph.D. in Physics from State University of New York at Albany. Professor Kuhn then became a visiting scientist at New England Biolabs, Inc. before becoming a staff scientist at Stanford Synchroton Radiation Laboratory, Stanford University. He was an associate professor at Stanford University in the SSRL and the Medical School. Since 2002 Dr. Kuhn has held the positions of Life Sciences Director, Scripps-PARC Institute for Advanced Biomedical Sciences, Research Fellow, Palo Alto Research Center, and Associate Professor, Dept. of Cell & Molecular Biology, The Scripps Research Institute. His laboratory has developed a reliable way to detect and to characterize circulating tumor cells isolated from the blood of cancer patients. His research is focused on the development of a fluid biopsy for the diagnosis, prognosis and therapy management of ovarian, breast, lung, colon, prostate and other tissue cancers. Additionally, he is the principal investigator of the NIH NCI Physics Oncology Center on the 4-dimensional biopsy, investigating the physics and mathematics of cancer metastasis.