Unsupervised Mitosis Detection

Sundar Ram and Andrew Cohen

Department of Electrical and Computer Engineering, Drexel University, Philadelphia, PA.

ABSTRACT

Time lapse microscopy can capture image sequences of live, proliferating neural stem cells. These *in vitro* image sequences show the patterns of development across a clone, or family tree, of cells. When the cells undergo division, or mitosis, they produce two daughter cells, eventually differentiating into neurons or glia. Quantifying the dynamic processes of cell development from a single cell into differentiated neural tissue requires accurate segmentation, tracking and lineaging. A key component in automating these difficult algorithmic tasks is the ability to reliably identify mitotic events. Several supervised techniques have been developed to detect mitoses automatically, however these techniques require large amounts of manually evaluated data for training and they cannot be generalized to new data without retraining. I have implemented an unsupervised mitosis detection algorithm using Spectral Clustering Techniques to identify typical mitotic shape patterns. These are characteristic features of proliferating cells in most phase contrast image sequences so the algorithm is likely to generalize well to new data. In preliminary tests this new approach has resulted in 100% accuracy and has shown a lot of promise and has significantly outperformed all other techniques.

I. INTRODUCTION

Methods for assessing the proliferative activity of the stem cells grown *in vitro* are critical tools to analyze and characterize the behavior of stem cells in an automated fashion. Among these behaviors, mitosis detection, the process wherein a single cell undergoes cell division to form two daughter cells, is one of the major characteristics under study. Analysis based on these, can give detailed information about the development of the stem cells into neurons or glial cells. This data is also useful for the construction of lineage trees which give a detailed picture of the lineage of every single cell in a clone. The number of occurrences of mitotic events is a key factor for cancer screening and assessment and also in regenerative medicine since stem cells proliferate and renew themselves through mitosis.

Currently, there are many cell proliferation assays that are compatible with automated sample handling to measure cell proliferation. But, a majority of these procedures, involve fluorescent, luminescent or colorimetric assays which require destructive methods of cell manipulation, like cell lysis or *in vitro* staining, and thus are unable to continuously monitor cells in culture.

To compensate for the drawbacks of the aforementioned methods, there has been the emergence of a new technique called the Phase Contrast Microscopy, which is a non-destructive imaging modality. Automated time lapse systems employing this imaging modality for monitoring cell populations *in vitro* enable high throughput screening and also facilitate continuous monitoring of

live cells. Moreover, since samples are continuously monitored, sampling at various times is abrogated, resulting in reduced human labor and also eliminating costs of expensive reagents used for staining. With such microscopic systems, automatic mitosis detection can provide quantitative information regarding cell proliferation on a continuous basis. This functionality is also expected to improve automated cell tracking systems [10].

Previous work done in automatic detection of mitotic events can be divided into tracking based and non-tracking based approaches.

Yang *et al.* [1] obtained blob regions along each cell's trajectory produced by a tracking method. Each blob region is then examined to determine if it contains a mitotic event based on several blob properties including area, perimeter, circularity and average intensity. Debeir *et al.* [2] adopted a combination of several mean shift processes to track cells using an ensemble of nested kernels. One of the kernels was designed to model cells in the mitotic state by taking into account their morphological changes. Al-Kofahi *et al.* [3] presented a multiple object matching method that can handle cell divisions in the typical frame by frame segmentation tracking method. Padfield *et al.* [4] investigated cell cycle phases through tracking each nucleus over time. Mitotic events were then identified by linking non mitotic phases using both the Euclidean distance metric and the fast marching method. All these methods are intuitive, however they are highly dependent on the tracking performance, which is more challenging to achieve than mitosis detection performance itself. Moreover, tracking methods are computationally expensive and take up a very long time for processing.

Other mitosis detection algorithms that do not involve cell tracking have also been proposed recently. Li *et al.* [5] applied a fast cascade learning framework adopting AdaBoost classifiers to volumetric Haar-like features extracted from spatio-temporal patches covering the whole image region. This approach requires a large amount of training data. Debeir *et al.* [6] proposed a method to detect mitotic cell regions based on brightness change and link regions in consecutive frames into a cell division candidate linkage. Each candidate was then validated based on its length (no. of frames). This approach is too simple to effectively distinguish between actual mitotic events from other cells. Liu *et al.* [7] developed a method in which Hidden Conditional Random Fields [8] are trained to examine each candidate after mitotic candidate patch sequences are constructed using 3D seeded region growing. This approach adopts explicit candidate detection as well as model based validation but does not detect birth events. Moreover, this method is computationally expensive due to the preconditioning step [9] which was originally devised to segment non mitotic cell regions from background rather than mitotic cell regions. Most recently, Huh *et al.* [10] came up with a similar method involving Event Detection Conditional Random Fields that not only detect mitotic occurrences but also provides time at which mitosis is completed and two daughter cells are born. However all the above methods are supervised learning algorithms that require a lot of training data and also do not have a high probability of generalizing well to new data.

In this paper, we propose an unsupervised mitosis detection algorithm. This algorithm involves turning an affinity matrix of pairwise image similarities into a normalized Markov transition process. The affinity matrix is generated using the Normalized Compression Distance as the notion of Information Distance with the compression obtained using wavelet compression techniques.

The reason behind adopting this approach was because the classification of mitotic patterns essentially depends on whether the cells in the image are ”similar” to those acquired from mitotic or non-mitotic patterns. Conventional methods involving segmentation to acquire cell information for extracting features is just one way to measure similarity. We wanted to find a general and optimal approach for measuring the similarity between two cell images. This fact pushes a great meaning into this approach and makes it more intuitive. The proposed method for 24 images turns out to be 100% accurate.

The remainder of this report is organized as follows. In section II, our approach is discussed in detail which includes the spectral clustering technique and wavelet compression. The next section shows the results obtained using 2D and 3D wavelet transforms. After that, conclusion is laid out in section IV followed by references in section V.

II. IMPLEMENTATION

Given a sequence of phase contrast microscopic images, our goal is to differentiate between mitotic and non-mitotic stem cells. The inputs would be 24 candidate images of different cells cut out from the raw stem cell images. All the images have been resized to the same dimension of 20x20 pixels as shown in Fig. 1.

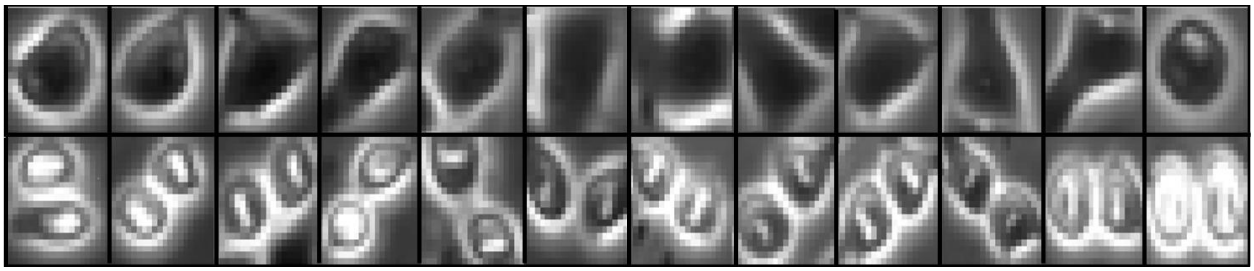

Fig. 1. Montage of the 24 input images. The first row consists of 12 non-mitotic cells and the second row has 12 mitotic cells.

A. Compression Distance

Once I have the outputs, now I have to form an affinity matrix to determine the pairwise similarities of the 24 images. The Kolmogorov complexity of x is the length of the shortest binary program with no input that outputs x; it is denoted as $K(x) = K(x|\lambda)$, where λ denotes the empty input [14].

Essentially, the Kolmogorov complexity of an image is the length of the ultimate compressed version of the image. A compression distance based on a normal compressor is shown as an admissible distance [11]. A compressor C approximates the information distance E(x,y), based on Kolmogorov complexity, by the compression distance E_c(x,y) defined as

$$E_C(x, y) = C(xy) - \min\{C(x), C(y)\}$$

Here, $C(xy)$ denotes the compressed size of the concatenation of x and y, C(x) denotes the compressed version of x and C(y) denotes the compressed version of y, where x and y are the two

entities between which the similarities must be calculated. However, the normalized version of the admissible distance $E_C(x, y)$, the compressor C based approximation of the normalized information distance is called the Normalized Compression Distance (NCD) is the real world version of the ideal notion of Information Distance [2].

$$NCD(x, y) = \frac{C(xy) - min\{C(x),\ C(y)\}}{max\{C(x), C(y)\}}$$

B. Wavelet based Compression Technique

We have determined the criteria for measuring the similarity or distance between two input images. The Normalized Compression Distance asks for the compressed size of the images. The next step is to compress the images. This is achieved through wavelet based compression techniques. Wavelet transform is just a representation of the image in terms of sets of real coefficients. There are 4 different types of coefficients namely the Approximation, Horizontal, Vertical and Diagonal. Most of the coefficients are nearly zero and so the image can be well approximated with a small number of large wavelet coefficients. This results in high compression ratios. Initially I used the Daubechies-2 family of wavelet transform to compress the 24 images. The number of decomposition levels was set to 3.

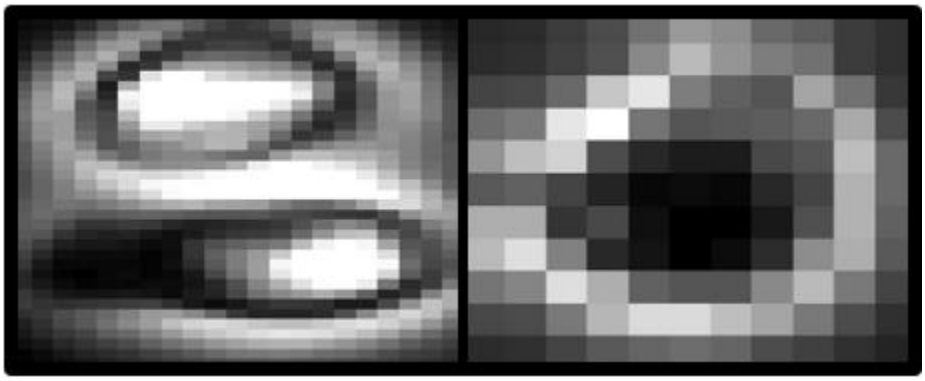

Fig. 2 Sample image of the 2D wavelet transform. On the left is the input image and on the right is the representation of the image using wavelet coefficients.

Before undergoing the wavelet transformation, the candidates were subjected to a preprocessing step involving the standard deviation filter. This was done to emphasize the local variability in the image. It is known that most of the information is stored in the LL band of the wavelet transform which corresponds to the approximation coefficients. Now, the compressed size of the image was calculated by finding the mean of the absolute values of the approximation coefficients. To find the compressed size of the concatenation of x and y, two images were concatenated on the x axis and then made to undergo the same wavelet transform.

Now that we have found the compressed size of all the images, it is convenient to calculate the affinity matrix using the NCD formula. This turns out to be a 24x24 double array.

C. Spectral Clustering

This method is the process of turning an affinity matrix A, in our case the NCD, into a normalized Markov transition process N. In [Meila and Shi, 2001], it is shown that a probability transition matrix N for a Markov chain with k strong clusters will have k high intra-set transition probabilities will have k piecewise constant eigenvectors, and they suggest clustering by finding approximately equal segments in the top k eigenvectors. According to Kamvar, *et al.* [13] the following are the steps to form the Spectral representation and then cluster the given data:

- Form Spectral Representation:
    1. Given data B, form the affinity matrix $A \in R^{nxn} = f(B)$
    2. Define D to be the diagonal matrix with $D_{ii} = \sum_j A_{ij}$
    3. Symmetric Divisive Normalization : $N = D^{-1/2} A D^{-1/2}$ (transition with probability proportional to relative similarity values)
    4. Find $x_1, \ldots, x_k$ be the k largest eigenvectors of N and form the matrix $X = [x_1, \ldots, x_k] \in R^{nxk-1}$
    5. Normalize the rows of X to be unit length
- For Clustering :
    1. Treat each row of X as a point in R^k and cluster into k clusters using K-means
    2. Assign original point x_i to cluster j if and only if row i of X was assigned to cluster j

Here, the affinity matrix formed in step A would be our NCD matrix. After following all the first 5 steps on the NCD matrix, we would have successfully have the spectral representation of our data. The final goal is to cluster this X matrix which has the k constant eigenvectors. I have used K-means as the clustering algorithm. In our case, the data must be clustered into two clusters relating to mitotic and non-mitotic images. So, the value of k is 2. I have explained the complete process in the flow chart shown in Fig. 3.

III. RESULTS

It would be most convenient to show the spectral representation plot of the data. We can easily see if there are two different transition probabilities as one would expect. The graph is shown in Fig. 4. The x and y axes are the largest 2 eigenvector spaces.

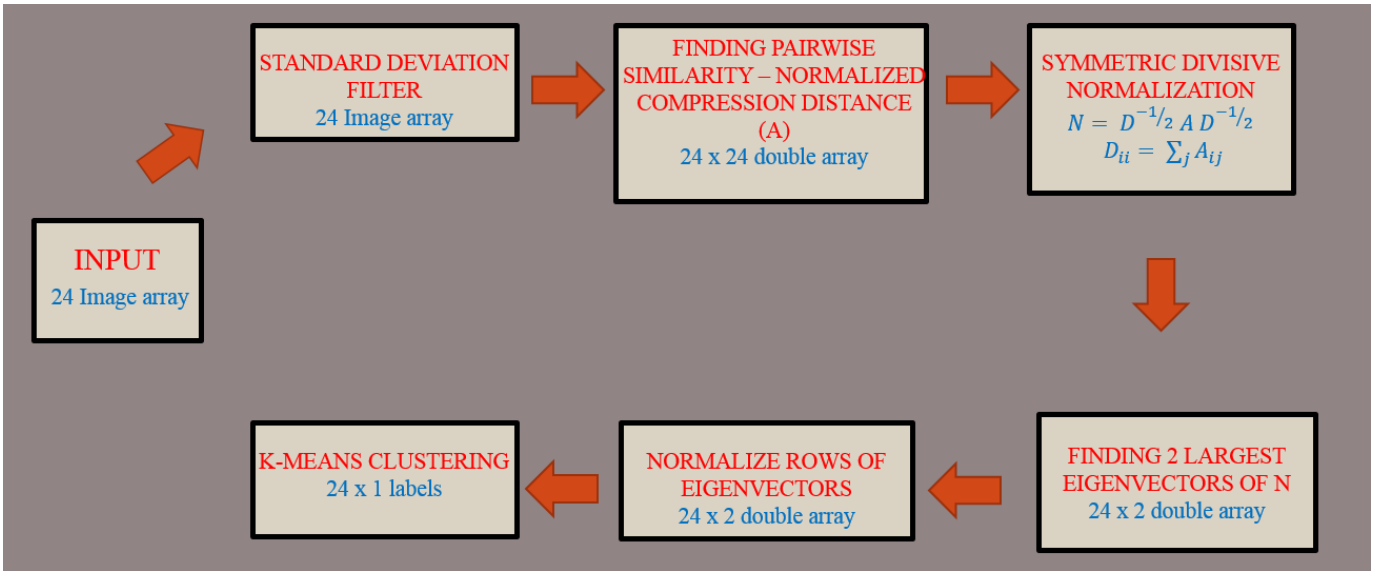

Fig. 3. Flow chart of the complete project.

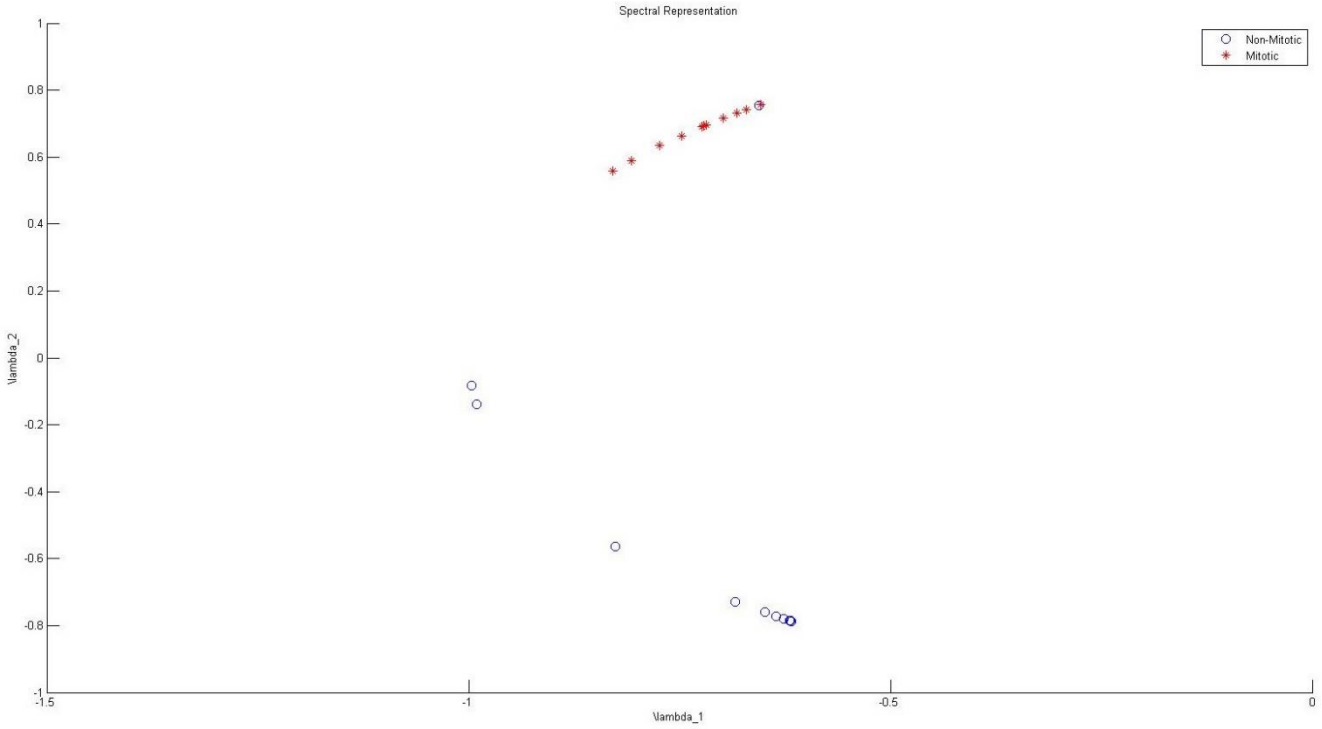

Fig. 4. Spectral representation plot of the 24 images. Blue circles correspond to non-mitotic cells and red stars correspond to mitotic cells.

From the graph, it is evident that all except one image were clustered correctly resulting in an accuracy of 95%.

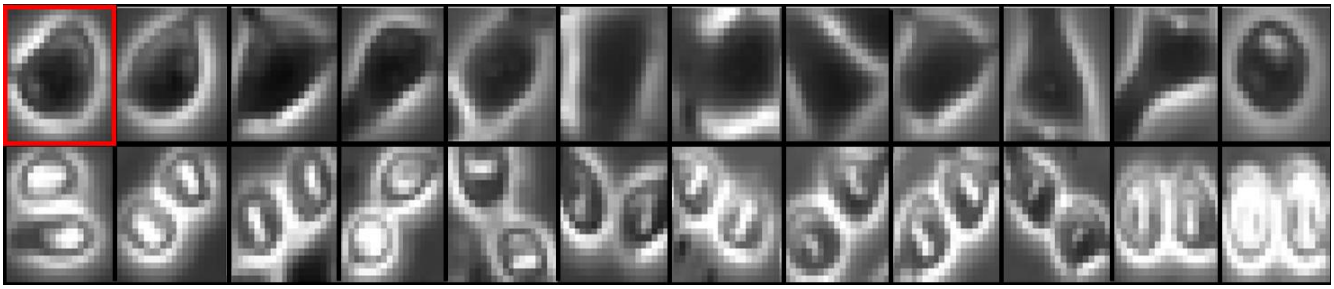

Fig. 5. The red box represents the misclassified cell.

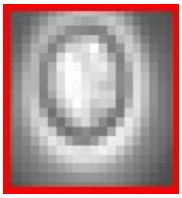

Fig. 6. Another misclassified example.

Fig. 5 shows the misclassified image. It was also seen that if a high intensity non-mitotic cell as shown in Fig. 6 was classified as a mitotic cell. So, I suspected that the algorithm was using the intensity as one of the important features. So, I felt there was a need to get a time series filter which captures the mitotic process. In other words, I wanted the algorithm to capture the difference between a cell which is single in the previous frame but divides into two in the next frame and a cell which remains single in the previous and the next frames. In the next experiment, I used 48 images as shown in Fig. 7.

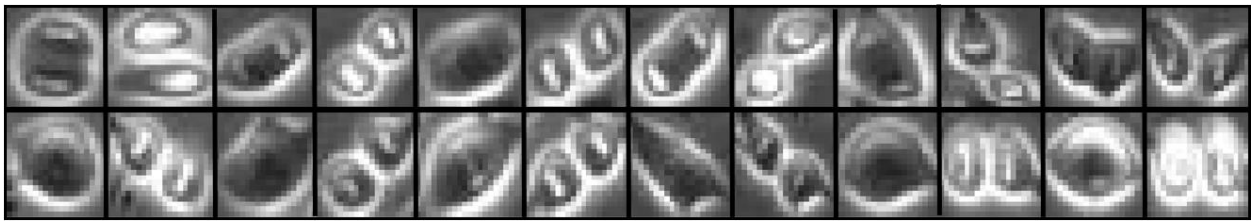

Fig. 7a) 24 images belonging to mitotic cluster.

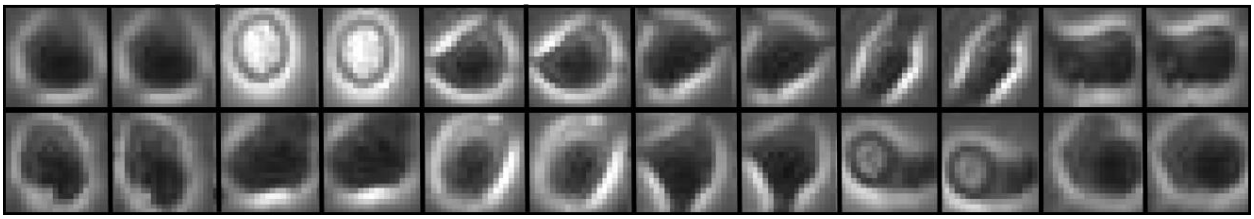

Fig. 7b) 24 images belonging to non-mitotic cluster.

The first image of the first row in Fig. 7a) is the image of a cell before mitosis and the next image is of the same cell after mitosis dividing into 2 daughter cells. These two form a pair of mitotic images. Similarly the images in Fig. 7b) form pairs of non-mitotic images. So, essentially, the input to calculate the NCD would be the same 24 entities, but instead of giving a single image, a sequence of two images would be considered as one entity. So, now I had to use 3D wavelet

transform to compress the input images. 3D wavelet transform is when we combine the images of the same cell in the ‘t-1’ frame and ‘t’ frame, where ‘t’ represents the frame at which the cell undergoes mitosis. The concatenation of x and y was performed such that the ‘t-1’ frame of two different cells were concatenated together and the ‘t’ frames of these two cells were concatenated together in the X axis. Also, instead of calculating the mean of the absolute values of the approximation coefficients to determine the size, this time, I used the BZIP2 compressor which gave more accurate compressed size values of the wavelet coefficients. The results came out to be very good as shown in Fig. 8.

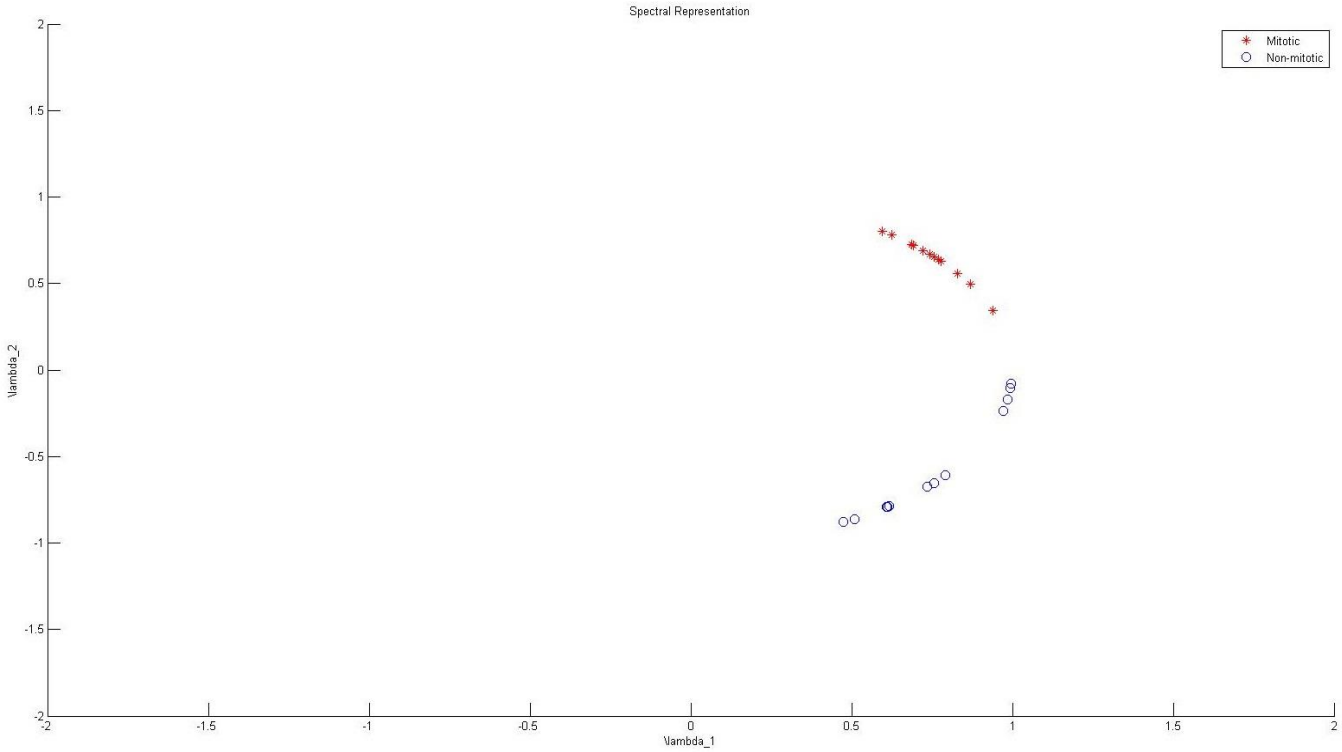

Fig. 8. Spectral representation plot. Blue circles correspond to non-mitotic cells and red stars to mitotic cells.

IV. CONCLUSIONS

The algorithm gave a 100% accuracy for the 48 images. Moreover, it was able to differentiate between high intensity non-mitotic cells as seen in the 3rd and 4th images of the first row in Fig. 7b and actual mitotic cells. This time it classified them correctly as non-mitotic cells. Thus, the algorithm was made more robust and efficient enough to capture the difference in transformation of cells during a mitotic process. This algorithm is purely an unsupervised algorithm and clearly outperforms previous methods in preliminary tests. Since it is an unsupervised approach, there is no need for training data and labels for mitotic events. This approach, when combined with the tracking based mitosis detection tool would play a big role in the analysis of cell proliferation.

V. REFERENCES

[1] F. Yang, M. A. Mackey, F. Ianzini, G. Gallardo, and M. Sonka, "Cell segmentation, tracking, and mitosis detection using temporal context," in *Proc. Int. Conf. Med. Image Computing Computer Assist. Intervent.*,2005, pp. 302–309.

[2] O. Debeir, P. Van Ham, R. Kiss, and C. Decaestecker, "Tracking of migrating cells under phase-contrast video microscopy with combined mean-shift processes," *IEEE Trans. Med. Imag.*, vol. 24, no. 6, pp. 697–711, Jun. 2005.

[3] O. Al-Kofahi, R. J. Radke, S. K. Goderie, Q. Shen, S. Temple, and B. Roysam, "Automated cell lineage construction: A rapid method to analyze clonal development established with murine neural progenitor cells," *Cell Cycle*, vol. 5, no. 3, pp. 327–335, 2006.

[4] Padfield, J. Rittscher, N. Thomas, and B. Roysam, "Spatio-temporal cell cycle phase analysis using level sets and fast marching methods," *Med. Image Anal.*, vol. 13, no. 1, pp. 143–155, 2009.

[5] K. Li, E. D. Miller, M. Chen, T. Kanade, L. E.Weiss, and P. G. Campbell, "Computer vision tracking of stemness," in *Proc. IEEE Int. Symp. Biomed. Imag.*, 2008, pp. 847–850.

[6] O. Debeir, V. Mégalizzi, N. Warzée, R. Kiss, and C. Decaestecker, "Videomicroscopic extraction of specific information on cell proliferation and migration in virto," *Exp. Cell Res.*, vol. 314, no. 16, pp. 2985–2998, 2008.

[7] A.-A. Liu, K. Li, and T. Kanade, "Mitosis sequence detection using hidden conditional random fields," in *Proc. IEEE Int. Symp. Biomed. Imag.*, 2010.

[8] Quattoni, S.Wang, L. Morency, M. Collins, and T. Darrell, "Hidden conditional random fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 10, pp. 1848–1853, Oct. 2007.

[9] K. Li and T. Kanade, "Nonnegative mixed-norm preconditioning for microscopy image segmentation," in *Proc. Int. Conf. Inf. Process. Med. Imag.*, 2009, pp. 362–373.

[10] S. Huh, Dai Fei Elmer Ker, Mei Chen and Takeo Kanade, "Automated Mitosis Detection of Stem Cell Populations in Phase-Contrast Microscopy Images," in *IEEE Trans. Med. Imag.*, vol. 30, no. 3, Mar. 2011.

[11] Rudi Cilibrasi and Paul Vitanyi, "Clustering by Compression", *IEEE transactions on Information Theory*, Vol. 51, No. 4, April 2005.

[12] Sepandar Kamvar, Dan Klein and Christopher Manning, "Spectral Learning", *International Joint Conference on Artificial Intelligence*, pp.561-566, 2003.

[13] M. Li and P. M. B. Vitányi, *An Introduction to Kolmogorov Complexity and Its Applications*, 2nd ed. New York: Springer-Verlag, 1997