35th Annual International Conference of the IEEE EMBS
Osaka, Japan, 3 - 7 July, 2013

Robust Classification of DNA Damage Patterns in Single Cell Gel Electrophoresis

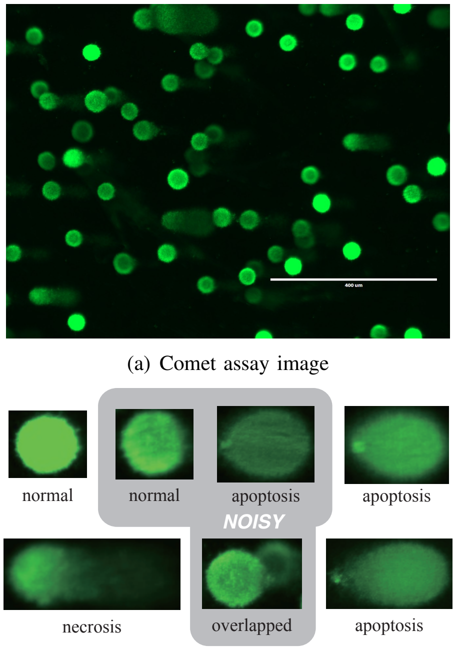
Taehoon Lee, Sungmin Lee, Woo Young Sim, Yu Mi Jung, Sunmi Han,
Chanil Chung, Jay Junkeun Chang, Hyeyoung Min, and Sungroh Yoon

Abstract— **Single cell gel electrophoresis, also known as comet assay, has been widely used for assessing the effect of genotoxicity and detecting DNA damage of individual eukaryotic cells. There exist established imaging techniques for comet-assay analysis, but these platforms have limitations such as required user interventions, low throughput, and weakness to noise caused by incomplete dyeing of fluorescent materials and other experimental errors. To resolve these, we propose a novel procedure for analyzing comet assay images, which considers various DNA damage patterns and classifies them in a robust manner. We tested our approach with twenty golden data sets containing over 300 comets and achieved satisfactory classification accuracy.**

I. Introduction

The *single cell gel electrophoresis* (SCGE) is a method developed for assessing the single cell DNA breakage [1] [2]. As shown in Fig. 1(a), the objects in an image from SCGE, also called *comet assay*, appear as 'comets.' It is a promising technique in analyzing genotoxicity by detecting DNA damage and repair of individual eukaryotic cells. Because of the technique's sensitivity, simplicity, rapidity, and visibility, comet assays have found various applications in biomonitoring, molecular epidemiology and genotoxicology [3].

In a typical comet-assay study, the cells under test are embedded in agarose on a microscope slide and electrophoresed under alkaline conditions. These environments make DNA fragments move away from the cell nucleus. After the genetic materials are stained with a fluorescent dye, cells with DNA damage can be observed with the migration of DNA fragments. These migrated genetic materials form the tail of a comet. The more the DNA strands are damaged, the longer and brighter the tail of the cell extends. Thus, the length of the tail and the percentage of DNA fragments in the tail area are helpful for measuring the degree of DNA breakage. Furthermore, there exist more sophisticated metrics, such as the tail moment and the tail inertia, which can quantify DNA damage more accurately than by simply measuring the tail length [4].



(a) Comet assay image

(b) Types of individual cells

Fig. 1. Comet assay overview.

Test cells can be classified into three types according to their shapes as shown in Fig. 1(b). Normal cells have no DNA damage and form a round shape with bright intensity. Apoptotic cells, which underwent the process of programmed cell death, have smaller nuclei and fanning tails. Necrotic cells, which are dead cells by external factors such as infections or toxins, have spread-out heads and tails. For accuracy diagnoses of DNA conditions, identifying comet types is thus a key issue in studies of DNA damage, aging, and cancer. In this paper, we propose a decision-tree-based classification [5] scheme for categorizing normal and damaged cells. In addition to classification of comets, the proposed method can also provide broadly used parameters for quantitative analysis, such as the tail length, the tail inertia, and the tail moment. Several existing commercial tools can also calculate these parameters and report them, but they have limitations in terms of efficiency and reliability. For example, frequent user interventions of such tools make the analysis low-throughput, and they often cannot handle noisy images well.

This work was supported by the National Research Foundation of Korea funded by the Ministry of Education, Science and Technology (2011-0009963 to S.Y.).

T. Lee and S. Lee are with the Department of Electrical and Computer Engineering, Seoul National University, Seoul 151-744, Korea.

W. Sim, Y. Jung, S. Han, C. Chung, and J. Chang are with the Research Division, NanoEnTek Inc., Guro-gu, Seoul 152-740, Korea.

H. Min is with the College of Pharmacy, Chung-Ang University, Seoul 156-756, Korea.

S. Yoon is with the Department of Electrical and Computer Engineering and with Bioinformatics Institute, Seoul National University, Seoul 151-744, Korea.

978-1-4577-0216-7/13/$26.00 ©2013 IEEE

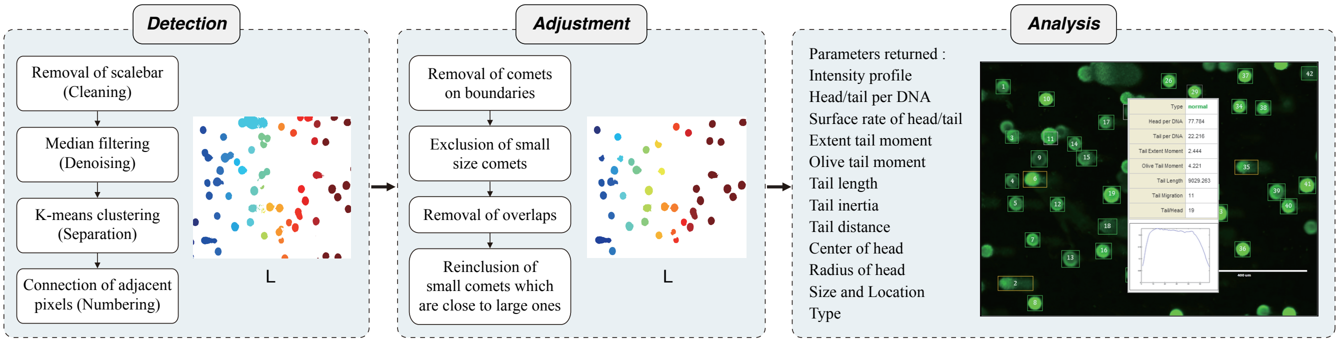

Detection
Removal of scalebar (Cleaning)
Median filtering (Denoising)
K-means clustering (Separation)
Connection of adjacent pixels (Numbering)
L
Adjustment
Removal of comets on boundaries
Exclusion of small size comets
Removal of overlaps
Reinclusion of small comets which are close to large ones
L
Analysis
Parameters returned :
Intensity profile
Head/tail per DNA
Surface rate of head/tail
Extent tail moment
Olive tail moment
Tail length
Tail inertia
Tail distance
Center of head
Radius of head
Size and Location
Type

Fig. 2. Overview of the proposed methodology.

II. Related Work

To identify comets in the input image and classify them, we utilize two techniques developed in the image processing domain. First, image segmentation refers to the process of extracting objects from a digital image [6], and is suitable for the proposed procedure, which requires detection of cells in comet assay images. Most image segmentation algorithms are divided into two types depending on whether side information is needed (spatially guided) or not (spatially blind) [7]. As we can assume that the dominant colors of all comets in one input image are identical and significantly different from the background colors, spatially blind approaches are more appropriate than spatially guided methods are. We use K-means clustering [8] in this paper. K-means clustering is one of the partitioning methods grouping all the pixels of an image into K groups iteratively, and has been widely used due to its simplicity and reliability in performance.

To enhance robustness, we also utilize the edge detection [9] technique, which enables the proposed method to detect overlapped comets and to eliminate occluded comets. In image processing, an edge refers to the points at which the color value changes abruptly and is related to determining the shapes and sizes of objects. Each comet can spread over excessively (not because of DNA fragments but because of gel artifacts) and overlap with other comets. We thus find edges among overlapped comets and extract foreground comets by using edge detection techniques, more specifically the Canny edge detector [10].

III. Methods

Fig. 2 shows the whole procedure of the proposed method. It works in three phases: *detection, adjustment,* and *analysis*. In the first detection phase, we utilize existing image segmentation techniques to separate comets from the background. For automated analysis, the proposed method requires virtually no parameters and bounding boxes around comets unlike existing software do. Next, the Canny edge detector is exploited in the adjustment phase. We can extract foreground comets from overlapped objects by using edge information. To detect apoptotic cells properly, we also redefine the memberships of tiny objects in this phase. After that, the properties of detected comets are calculated and reported with figures and tables. We describe more details of each phase below.

A. Detection phase

The proposed method aims to recognize individual (even overlapping) cells and obtain various properties of those cells. To achieve these objectives, image segmentation is applied to detect comet pixels and identify comets by merging contiguous comet pixels. This is why we called this step the *detection phase*. It consists of three steps: preprocessing, clustering, and numbering.

During preprocessing, the scale bar, if any, would be removed from the input image. We can assume that the scale bar is always white and located on the bottom-right. The proposed method thus converts the original image to grayscale and finds white pixels (e.g., all pixels with intensities greater than 200) in the bottom-right region and replaces its colors into those of the background (typically black). Additionally, we detect the comets placed at the boundaries of the image and discard them due to their incomplete shapes.

As another preprocessing, smoothing is applied for denoising. As mentioned previously, there are various experimental errors such as incomplete dyeing of fluorescent materials or errors of artifacts from optical equipments. They appear on a comet assay output image as noise (see Fig. 1) and affect the accuracy of finding nuclei. To alleviate this noise issue, we apply a blurring operation with median filtering or moving-average filtering.

After preprocessing, we perform K-means clustering with $K = 2$ to find comet pixels within an image. All the pixels are considered as 3-dimensional vectors of RGB values and divided into two groups, namely comet pixels (whose intensities are relatively high) and background pixels (relatively low intensities). This clustering process returns a binary matrix in which comet pixels are marked true and background pixels false. Using this matrix, we construct a membership matrix L, in which a positive integer L_{ij} represents the membership of the pixel (i, j) in the binary matrix (and also in the original image); the background pixels have zero values in L. Based on L, we group a set of adjacent pixels with the same membership into a comet. See Fig. 2 for examples of the L matrix.

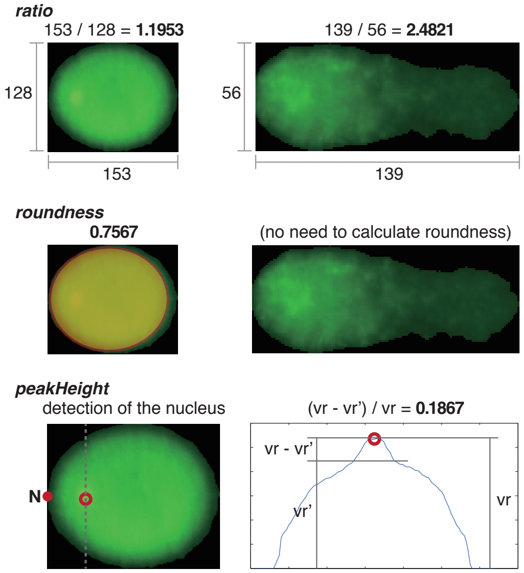


Fig. 3. Three parameters for type classification.

B. Adjustment phase

In this phase, we eliminate the objects on the image boundaries. Using matrix L, we find the objects whose pixels are on the boundaries and set the pixels of these objects to 0. The procedure for removing pixels on the boundary is as follows: out of all the pixels included in a comet, we count the number of pixels located on the first or last row, or the first or last column (top, bottom, left, right), and remove the comet if the number is greater than 3% (empirically) of the image width.

This phase also may be adjusted to properly handle comets representing apoptotic cells. As shown in Fig. 1, the head of an apoptotic comet can be apart from its tail. In this case, it is possible that the head and tail pixels get assigned different comet numbers (different numbers in L). Such assignment should be corrected to detect apoptotic cells properly. To this end, we focus on two groups of pixels within a threshold distance. We test each of these two groups and decide if it is a head or a tail by comparing the number of pixels to the number of pixels in the nucleus of an apoptotic comet (empirically, about 0.07% of the image size). If the group in the left turns out to be a head and the group in the right a tail, then we merge the two groups into one comet.

After this adjustment, we finally perform overlap detection using the Canny edge detector. It outperformed other edge detectors we tested for handling noisy comets.

C. Analysis phase

In this final step, the proposed method characterizes individual cells in terms of well-known metrics such as the tail moments and then classifies them into two groups, normal or abnormal cells. In particular, this classification is crucial for assessing the health status of patients in clinical applications. Algorithm 1 outlines the proposed approach to classify comets.

Algorithm 1 Type classification

```
procedure $type$ = GETTYPE($img$)
    $img$ : $rows \times cols$ matrix (grayscale intensity)
    if $ratio < 0.85$ then
        $type \leftarrow$ ‘fail’
    else if $ratio < 1.3$ then
        if $roundness < 0.6$ then
            $type \leftarrow$ ‘fail’
        else
            if $peakHeight > 0.12$ then
                $type \leftarrow$ ‘abnormal’
            else
                $type \leftarrow$ ‘normal’
            end if
        end if
    else if $ratio < 4.5$ then
        $type \leftarrow$ ‘abnormal’
    else
        $type \leftarrow$ ‘fail’
    end if
    return $type$
end procedure
```

First of all, we define three parameters used for characterizing comets: $ratio$, $roundness$, and $peakHeight$ (see Fig. 3). The $ratio$ is calculated as the width of a comet divided by its height. The $roundness$ is defined as the correlation coefficient between the input image and an oval shape, which has the same size as the comet image. For the definition of the $peakHeight$ parameter, refer to Fig. 3. First, we detect the position of the nucleus and assume that a vertical line passes through the nucleus on the image. Then, we measure the intensity values over this line, as depicted in the bottom-right plot in Fig. 3. There, v_r is the maximum intensity over the line, and v_r' is calculated as follows.

Let img be the input image represented in grayscale. We obtain the maximum intensity m in img and find all the pixels whose intensities are greater than $0.9m$. Out of those pixels, we identify the closest pixel to the location marked N and define the location of that pixel (i.e., the nucleus) as the column location c, the estimated location of the nucleus on the x-axis. Next, we employ a temporary vector $v = (v_1, \ldots, v_i, \ldots, v_{rows})^T$ where $v_i = \sum_{j=c-2}^{c+2} img_{i,j}$. Thus, v denotes the intensity distribution on the y-axis in column c and its surrounding columns. Finally, we extract the peak of v and store its location in r. As a result, (r, c) is the coordinate of the estimated location of the nucleus, and the $peakHeight$ is finally defined as

$$peakHeight = \frac{v_r - v_r'}{v_r}, \text{ where } v_r' = \frac{1}{|R|} \sum_{i \in R} v_i. \quad (1)$$

In the equation above, R represents the set of neighborhoods of r. Taken together, $peakHeight$ is defined the relative intensity of the nucleus over the vertical line as depicted in Fig. 3.

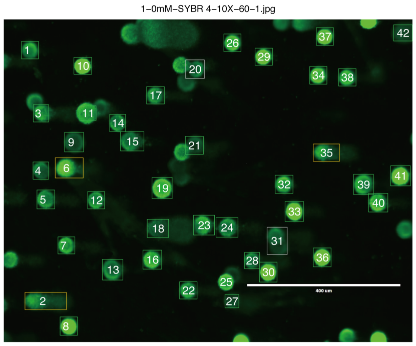


Fig. 4. Classification results of individual cells. Three types are represented by using the colors: green (normal), orange (abnormal), and gray (fail).

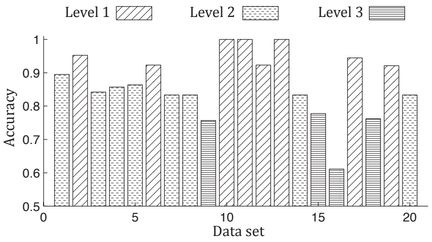
With the definitions above, the decision-tree-based algorithm can be described as in Algorithm 1. The range of $ratio$ is divided into four intervals: (A) $[0.85, 1, 3]$, (B) $[1, 3, 4.5]$, and (C) the others ($[0, 0.85]$ and $[4.5, \infty]$). The comets in case of (B) are classified as ‘abnormal’ cells, because the comets are distributed widely over the x-axis. The comets in group (C) are classified as ‘fail’ comets because their $ratio$ are unrealistic. In addition, we classify a comet in (A) as ‘fail’ if its shape is not a circle, ‘abnormal’ if the nucleus is presented on the image significantly, and ‘normal’ if a nucleus does not exist.

IV. Results and Discussion

To test our approach, we prepared 20 golden data sets, which were generated by a micro comet-assay system (PICASSo, currently under development, NanoEnTek Inc., Korea). In this system, all the cells in one pallet were exposed to a toxic material and captured by a microscope (EVOS, AMG Inc., USA) after being loaded into multi-microchannels. These comet assay images contain 140 normal and 229 abnormal cells in total (on average, 7 and 11.45 cells, respectively). Fig. 4 shows a classification result, in which different comet types are color coded differently (green, orange, and gray; see the caption for Fig. 4).

For evaluation of the classification, independent domain experts marked the labels of individual comets and categorized the golden data sets into three groups according to the difficulty of image processing. Fig. 5 summarizes the performance of the proposed classifier, in which each bar is hashed according to the level of difficulty. The average accuracy is 86.8% overall, and the average accuracies of levels 1, 2, and 3 are 95.8%, 84.9%, and 72.7%, respectively (level 1 is the easiest).

Level 3 samples contain a number of noisy comets. As mentioned previously, incomplete dyeing of fluorescent materials makes only the borders of comets bright, whereas the center areas remain dim. We can frequently misclassify such noisy normal cells as ‘abnormal.’ Also, necrotic cells whose $ratio$s are less than 1.3, are hard to classify, because their heads have round shapes and their tails are too short to be ‘abnormal.’ For these reasons, those necrotic cells are often misclassified as ‘normal.’



	Level 1	Level 2	Level 3
Average of Accuracy (Standard Deviation)	95.8% (3.41)	84.9% (2.06)	72.7% (6.73)
Average of Sensitivity (Standard Deviation)	89.3% (7.05)	77.8% (39.1)	90.2% (9.08)
Average of Specificity (Standard Deviation)	92.1% (16.1)	73.0% (23.0)	58.5% (18.2)

Fig. 5. Performance of the proposed type classifier.

To solve these challenges, we are planning to make the nucleus detection process more robust. The adjustment phase, especially the edge detection step therein, also needs some enhancements to increase the classification accuracy. In addition, we could revise the classification algorithm further so that it can detect the subtypes of abnormal types, such as apoptotic or necrotic cell types.

V. Conclusions

The proposed procedure aims to handle comet assay images and consists of three phases: detection, adjustment, and analysis. Our approach is one of the first attempts to fully automate comet assay analysis. The average classification accuracy achieved was 86.8% for 20 test data sets (over 300 comets) with varying difficulty levels. We hope that the proposed tool may be useful for assessing the health status of patients in clinical applications.

References

[1] D. W. Fairbairn, P. L. Olive, and K. L. O’Neill, “The comet assay: a comprehensive review,” *Mutation Research/Reviews in Genetic Toxicology*, vol. 339, no. 1, pp. 37–59, 1995.

[2] J. H. Hoeijmakers, “Dna damage, aging, and cancer,” *New England Journal of Medicine*, vol. 361, no. 15, pp. 1475–1485, 2009.

[3] A. Collins, “The comet assay for dna damage and repair,” *Molecular Biotechnology*, vol. 26, pp. 249–261, 2004.

[4] B. Hellman, H. Vaghef, and B. Boström, “The concepts of tail moment and tail inertia in the single cell gel electrophoresis assay,” *Mutation Research/DNA Repair*, vol. 336, no. 2, pp. 123–131, 1995.

[5] A. Vlahou, J. O. Schorge, B. W. Gregory, and R. L. Coleman, “Diagnosis of ovarian cancer using decision tree classification of mass spectral data,” *Journal of Biomedicine and Biotechnology*, vol. 2003, no. 5, pp. 308–314, 2003.

[6] K. Fu and J. Mui, “A survey on image segmentation,” *Pattern Recognition*, vol. 13, no. 1, pp. 3–16, 1981.

[7] S. R. Vantaram and E. Saber, “Survey of contemporary trends in color image segmentation,” *Journal of Electronic Imaging*, vol. 21, no. 4, 2012.

[8] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, “An efficient k-means clustering algorithm: Analysis and implementation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 881–892, 2002.

[9] X. Jiang and H. Bunke, “Edge detection in range images based on scan line approximation,” *Computer Vision and Image Understanding*, vol. 73, no. 2, pp. 183–199, 1999.

[10] J. Canny, “A computational approach to edge detection,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. PAMI-8, no. 6, pp. 679–698, 1986.