HEp-2 Cell Image Classification: A Comparative Analysis

Supplementary Material

Praful Agrawal, Mayank Vatsa, Richa Singh

Indraprastha Institute of Information Technology, Delhi, India

Features

Five different feature sets are extracted from the cell images: Boundary, Shape and Size, Statistical, Texture, and Descriptor based features.

- **Boundary Features:** Total 38 boundary features are computed which model various characteristics associated with circumference of HEp-2 cells. Features are briefly summarized below:
  - **Perimeter:** Length of the boundary of cell.
  - **Mean Gradient:** Average of Sobel gradient of boundary pixels.
  - **FFT Spectral Energy Features:** Spectral energy features have been used by researchers to differentiate mitotic nuclei from others present in microscopic images of biopsy slides for breast cancer [1]. To compute these features Fourier domain representation of the radii vector is calculated. Radii vector is a vector which comprises of distances between the centroid of cell and boundary pixels. Radii vector is computed as, $\mathbf{Radii} = \sqrt{(\mathbf{x} - \bar{x})^2 + \mathbf{y} - \bar{y})^2}$ where $\mathbf{x}$ and $\mathbf{y}$ are the vectors containing x and y coordinates of boundary pixels respectively. $(\bar{x}, \bar{y})$ represents the centroid location of the cell. The FFT of Radii vector is computed as,

$$FFT\_Radii = \sum_{n=0}^{N-1} Radii(n) e^{-j2\pi nk/N},\ k = 0, 1, ..., N-1 \tag{1}$$

where $N = 1024$. From above representation following 35 spectral energies are computed

$$FFT_{xxx} = \sum_{n=a}^{b} |FFT\_Radii(n)|^2 \tag{2}$$

where xxx=low, med, and high with values of [a, b] being [0, 340], [341, 683], [684, 1023] respectively. Other spectral energies are computed as,

$$FFTh = \sum_{n=32\cdot h}^{32\cdot(h+1)-1} |FFT\_Radii(n)|^2, \text{ where } h = 0, 1, ..., 31 \tag{3}$$

- **Bending Energy:** Bending energy of a shape is a measure of energy required to deform a metal rod into that shape. It is modeled using the curvature of object boundary, cell boundary in this case. Curvature of cell boundary (θ_{curve}) is quantified by a vector of angles formed between the two subsequent radii. $\theta_{curve} = arctan\frac{y(n+1)-y(n)}{x(n+1)-x(n)}$, where (x,y) are the boundary points and (x(N+1),y(N+1))=(x(1),y(1)). Bending Energy is defined as, $BE = \sum_{n=1}^{N}(\theta_{curve}(n+1) - \theta_{curve}(n))$ where $\theta_{curve}(N+1) = \theta_{curve}(1)$.

– **Statistical Features:** Four statistical measures computed from the intensity values of pixels within the cell image are: Mean, Standard Deviation, Skewness and Kurtosis.

– **Texture Features:** Three types of texture features computed from the HEp-2 cell images are discussed below:

- **Laws Texture Features:** Laws [2] proposed three basic kernels [1 2 1], [−1 0 −1], [−1 2 −1] to represent texture information. Combinations of the three basic kernels further produce five kernels [1 4 6 4 1], [−1 0 2 0 −1], [1 −4 6 −4 1], [−1 −2 0 2 1], [−1 2 0 −2 1] that represent edges, ripples, waves, lines, and spots in a square region. Texture filters are generated from these kernels by multiplication of two same size kernels, as a result 9 filters of size 3×3 and 25 filters of size 5×5 can be formed. These 34 filters are convolved with the cell images to obtain the texture images. From these texture images three statistical measures - mean, absolute mean, and standard deviation are computed as feature values. In total, 102 features are derived from each cell.
- **SGLD Features:** Statistical texture features, also known as Spatial Gray-Level Dependence (SGLD) features, use statistical measures to model the pattern of intensity values in a region [3]. The features are computed using the Gray Level Co-occurrence Matrix (GLCM) which quantifies the number of occurrences of gray level i, in a spatial relation with gray level j. The image is discretized to eight gray levels and GLCM matrices are obtained for four orientations (0°, 45°, 90°, and 135°). Total 65 features are computed, 13 from each of the four GLCM matrices and additional 13 features are computed as mean values of these features.
- **Run Length Texture Features:** Another widely used measure of texture includes features based on the Gray Level Run Length (GLRL) matrix [4]. In total, 44 features are computed from GLRL matrices in four different directions. GLRL matrix $G(x, y|\theta)$ represents the run length matrix in a given direction θ. GLRL matrix for four directions (0°, 45°, 90°, and 135°) are computed.

– **Size and Shape Features:** These features are widely used in histopathology of breast cancer [1]. These are:

- **Area:** Total number of pixels in the cell.
- **Major and Minor Axis Length:**

$$majlen = 2\sqrt{2} \cdot \sqrt{m_{xx} + m_{yy} + \sqrt{(m_{xx} - m_{yy})^2 + 4m_{xy}^2}}$$
$$minlen = 2\sqrt{2} \cdot \sqrt{m_{xx} + m_{yy} - \sqrt{(m_{xx} - m_{yy})^2 + 4m_{xy}^2}} \quad (4)$$

where, $m_{xx} = \frac{1}{M}\sum_{i=1}^{M}(x_i - \bar{x})^2 + \frac{1}{12}$, $m_{yy} = \frac{1}{M}\sum_{i=1}^{M}(y_i - \bar{y})^2 + \frac{1}{12}$, and $m_{xy} = \frac{1}{M}\sum_{i=1}^{M}(x_i - \bar{x})(y_i - \bar{y}) + \frac{1}{12}$.
($\bar{x}$,$\bar{y}$) represents the centroid of the cell and M is the total number of pixels.

- **Eccentricity:** Deviation in shape of cell from circle. It is computed as, $ecc = \frac{2\sqrt{(\frac{majlen}{2})^2 - (\frac{minlen}{2})^2}}{majlen}$.
- **Orientation:** Angle between the major axis of the ellipse and the x-axis of the original image coordinates denoted by θ_o.

$$\theta_o = \begin{cases} \tan^{-1}\left(\frac{m_{yy} - m_{xx} + \sqrt{(myy - m_{xx})^2 + 4m_{xy}^2}}{2m_{xy}}\right) & \text{if } m_{yy} > m_{xx}, \\ \tan^{-1}\left(\frac{2m_{xy}}{m_{xx} - m_{yy} + \sqrt{(mxx - m_{yy})^2 + 4m_{xy}^2}}\right) & \text{else.} \end{cases} \quad (5)$$

- **Convex Area:** Area of the convex hull of the cell (treated as an object).
- **Convex Deficiency:** Ratio of ($ConvexArea - Area$) to $Area$.
- **Solidity:** Ratio of $Area$ to $ConvexArea$.
- **Extent:** Fraction of cell pixels within the bounding box. It is defined as, $Extent = \frac{Area}{M \times N}$, where M and N are height and width of the bounding box.
- **Aspect Ratio:** Ratio of height to width of the bounding box.
- **Equivalent Diameter:** Diameter of the circle with area same as the area of cell.
- **Sphericity:** Ratio of smallest to largest radii of the cell. Radii of the cell is distance between boundary pixel and centroid of the cell.
- **Compactness:** It models the roundness of its boundary. It is calculated as, $Compactness = \frac{4\pi Area}{Perimeter^2}$.
- **InertiaShape:** It also measures the roundness of the object. It is calculated as, $InertiaShape = \frac{2\pi \sum_n Rad(n)^2}{Area^2}$, where Rad is a vector of size equal to the number of pixels in the cell. Each component of Rad corresponds to distance between a pixel location and centroid of the cell.
- **Deviation in Centre of Mass:** Deviation in centre of mass of the image from centroid due to the intensity distribution of cell is defined as $GCentreMass = \frac{\sqrt{(x_{GCM} - \bar{x})^2 + y_{GCM} - \bar{y})^2}}{mean(Rad)}$, where ($x_{GCM}$, y_{GCM}) and ($\bar{x}$, $\bar{y}$) represent the grayscale centre of mass and centroid of the cell.

– **Descriptor based features:** Normalized histograms of uniform LBP with circle of radius one and HOG in a 3×3 neighbourhood are directly used as descriptor based features. Feature lengths of LBP and HOG histograms are 59 and 162 respectively.

References

1. Boucheron, L.E.: Object- and Spatial-Level Quantitative Analysis of Multispectral Histopathology Images for Detection and Characterization of Cancer. PhD thesis, UCSB (2008)
2. Laws, K.I.: Textured image segmentation. Technical report, USC (1980)
3. Haralick, R.M., Shanmugam, K., Dinstein, I.: Textural features for image classification. IEEE T-SMC (6) (1973) 610–621
4. Tang, X.: Texture information in run-length matrices. IEEE TIP **7**(11) (1998) 1602–1609