National Taiwan University

DEPARTMENT OF BIOMEDICAL ENGINEERING

基礎生醫影像處理技術 DBME 5018

TERM PROJECT REPORT

Unsupervised Segmentation of Pathology Images based on Haralick Textures

Student：	余承洋	R13528027
	陳楷霖	B11508002
Advising Professor：	鄒奇軒	教授

OUTLINE

I. Abstract

II. Introduction

III. Materials and Methods

IV. Results

V. Discussion

VI. Conclusion

VII. References

ABSTRACT

Biopsy is a pathology analysis which provides a hint for doctors to perform accurate diagnosis. However, it would take a couple of days to finish an examination, including steps such as sampling, staining and manual inspection. As computational image analysis emerges, there are already several attempts which aim to relieve manual labor by using traditional image analysis, yet a lot of attempts using traditional methods (Thresholding, erosion, dilation …etc.) had proven itself insufficient based on the natures of biopsy images having ill-defined regions and dramatic changes in intra-regional areas (complicated textures).

In this project, we aim to adopt a method using haralick textures to segment biopsy images into areas representing benign and malignant regions. By utilizing this method, we achieved image segmentation via textures, which provided a robust capability of segmenting biopsy images. This method also provided a set of feature vectors which were originally used in image segmentation that can be further used to train a model that can perform auto-segmentation. In conclusion, image segmentation can be performed on biopsy images via haralick texture given its nature of having ill-defined regions and dramatic changes in intra-regional areas.

INTRODUCTION

Traditional methods such as thresholding, edge-detection, region-based methods(Region growing, watershed algorithm) and morphological operations(erosion, dilation) are the most intuitive and simple attempts to perform image segmentation on biopsy images. Nevertheless, these “First Order Statistics” image analysis are computed from histograms of the gray values in images, which do not reflect objects or patterns in the image, only the distribution of gray-levels. Löfstedt *et al.* [1] concluded that this inability makes first order statistics a blunt tool for quantifying changes in images, or any change in the spatial distribution of gray values.

Originally, Haralick *et al.* [6] proposed using a gray-level co-occurrence matrix (GLCM) as a method of quantifying the spatial relation of neighboring pixels in an image. This method of segmenting images using its textures was applied to complicated cases such as skin texture analysis, fabric defect detection and forest type classification …etc. In recent years there has been a rapid increase in the application of Haralick features in medical image analysis on a variety of medical image types such as ultrasounds, MRI images, and X-ray Mammography, and also other cancer related image segmentation.

Additionally, Löfstedt *et al.* [1] developed a method that can create a modified set of Haralick texture features that are asymptotically invariant to the image quantization, while preserving most of the interpretations of the original features. By reinterpreting the GLCM as a discretized probability density function, it is possible to construct a modified set of Haralick texture features that are asymptotically invariant to the image quantization.

This project aims to utilize this modified set of Haralick texture features that are asymptotically invariant to image quantization to segment biopsy images that separates benign and malignant areas in microscopy images. Such feature vectors can also be used in constructing models that are capable of conducting unsupervised segmentation based on extracted features.

MATERIALS AND METHODS

Dataset Introduction

The Warwick-QU dataset [7] is an open datasource consisting of microscopic images designed for benchmarking image analysis algorithms in biomedical applications. It consists of high-resolution histological images, primarily focusing on challenging tasks such as segmentation, classification, and detection of regions of interest within tissue samples. The dataset provides annotated ground truth for accurate evaluation and is widely used in research involving histopathological image analysis.

To enhance the segmentation process, Haralick texture features can capture essential textural patterns from the images. Löfstedt *et al.* [1] offered a robust approach of using a modified set of Haralick texture features that are asymptotically invariant to image quantization to segment biopsy images to delineating different regions based on their unique structural and statistical characteristics.

Texture Analysis

The process of calculating Haralick texture features for each pixel region is a multi-step approach that involves careful analysis of image texture using statistical measures derived from the Gray-Level Co-Occurrence Matrix (GLCM). The initial step involves preparing the image data. The image can either be converted to grayscale, which simplifies the computational process by reducing the dimensionality of the data, or it can be retained in its original RGB format if color information is essential for the analysis. In the grayscale conversion, each pixel's intensity is represented on a single scale, typically ranging from 0 to 255, allowing for easier computation of texture features.

Next, sliding windows of varying sizes are applied across the image to capture the local texture information. These windows can have dimensions such as 3×3, 5×5, or 7×7 pixels, depending on the level of detail required. Smaller windows capture finer textures, while larger windows provide more contextual information by encompassing a broader area. For each window, the surrounding pixel intensities are extracted, forming a local neighborhood that serves as the basis for calculating texture features.

The core of this process involves computing the gray-level co-occurrence matrix (GLCM), which is a statistical representation of how frequently pairs of pixel intensities occur at a specified spatial relationship within the window. The matrix is defined as:

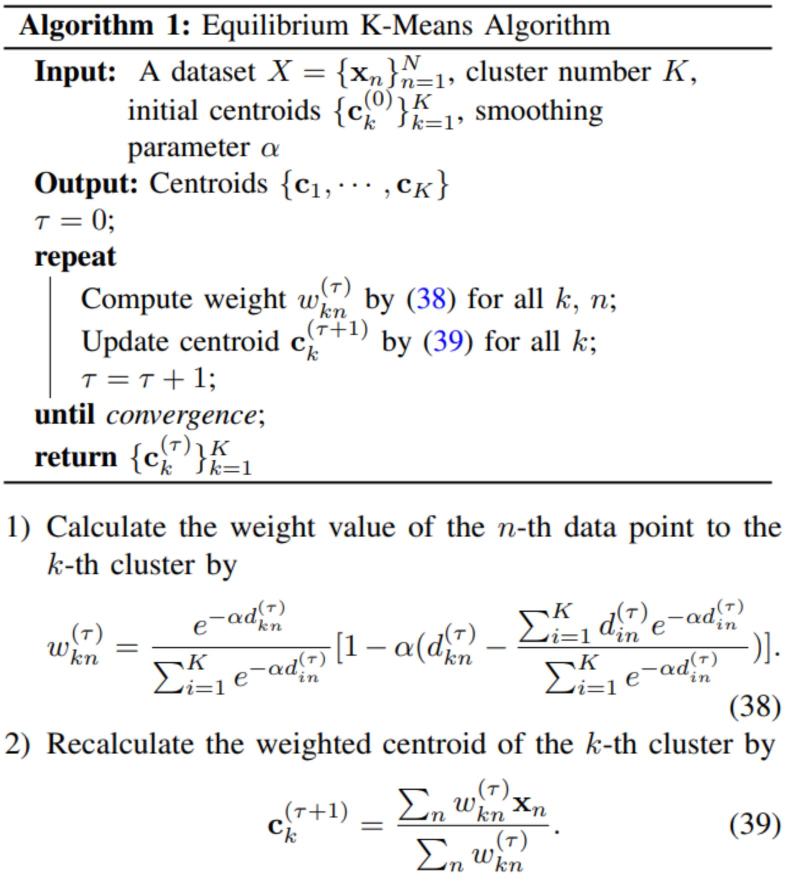
$$\mathrm{GLCM}(i,j) = \sum_{(x,y)} \begin{cases} 1 & \text{if } I(x,y) = i \text{ and } I(x+\Delta x, y+\Delta y) = j, \\ 0 & \text{otherwise,} \end{cases} \tag{1}$$

From the GLCM we can derive 13 Haralick texture features, each quantifying a specific aspect of texture, such as contrast, correlation, energy, and homogeneity. The derived features are then stored as feature maps, with each pixel represented by a set of texture values. These feature maps can include results from multiple window sizes, providing a detailed representation of the image's texture at varying scales. This multiscale approach enhances the ability to capture both fine and coarse texture details, making it useful for applications such as image segmentation, classification, and pattern recognition.

Implementation on Unsupervised Segmentation

Unsupervised segmentation based on extracted features involves a series of steps to partition an image into meaningful regions without relying on labeled data. The process begins by incorporating both spatial and pixel-level features into the feature map. Spatial features capture the positional relationships between pixels, while pixel features represent the individual characteristics of each pixel, such as texture, intensity, or color. This combined feature representation ensures that both local and global information is utilized during segmentation.

Next, dimensionality reduction techniques are applied to the feature map to simplify the data while retaining its most significant patterns. Once the dimensionality-reduced features are prepared, segmentation is performed using the K-means clustering algorithm. K-means is favored for its simplicity and effectiveness in grouping similar data points. Additionally, we utilize Equilibrium K-means, a variant of the traditional K-means algorithm designed to enhance cluster stability and convergence. The algorithm of equilibrium K-means is as below, where the smoothing parameter alpha should be assigned manually.



Algorithm 1: Equilibrium K-Means Algorithm

```
Input: A dataset $X = \{\mathbf{x}_n\}_{n=1}^{N}$, cluster number $K$,
       initial centroids $\{\mathbf{c}_k^{(0)}\}_{k=1}^{K}$, smoothing
       parameter $\alpha$
Output: Centroids $\{\mathbf{c}_1, \cdots, \mathbf{c}_K\}$
$\tau = 0$;
repeat
    Compute weight $w_{kn}^{(\tau)}$ by (38) for all $k$, $n$;
    Update centroid $\mathbf{c}_k^{(\tau+1)}$ by (39) for all $k$;
    $\tau = \tau + 1$;
until convergence;
return $\{\mathbf{c}_k^{(\tau)}\}_{k=1}^{K}$
```

1) Calculate the weight value of the n-th data point to the k-th cluster by

$$w_{kn}^{(\tau)} = \frac{e^{-\alpha d_{kn}^{(\tau)}}}{\sum_{i=1}^{K} e^{-\alpha d_{in}^{(\tau)}}}[1 - \alpha(d_{kn}^{(\tau)} - \frac{\sum_{i=1}^{K} d_{in}^{(\tau)} e^{-\alpha d_{in}^{(\tau)}}}{\sum_{i=1}^{K} e^{-\alpha d_{in}^{(\tau)}}})]. \tag{38}$$

2) Recalculate the weighted centroid of the k-th cluster by

$$\mathbf{c}_k^{(\tau+1)} = \frac{\sum_n w_{kn}^{(\tau)} \mathbf{x}_n}{\sum_n w_{kn}^{(\tau)}}. \tag{39}$$

K-means Clustering with Automated Determination of Cluster Count

Due to the diversity of pathological slices, these slices often contain more than just clearly malignant and benign regions. Therefore, we aim to develop a method that can automatically determine the number of clusters.

The approach we adopt involves calculating inertia, which is defined as the total sum of distances between each point and its corresponding cluster center. In this algorithm, a higher cluster count inevitably results in a lower inertia. Consequently, our objective is to identify the point where the second derivative of the inertia decreasing curve indicates that increasing the cluster count no longer significantly reduces inertia. The formula for inertia is as follows:

$$\text{Inertia} = \sum_{i=1}^{K} \sum_{x \in C_i} \|x - \mu_i\|^2 \tag{2}$$

RESULTS

Clustering Results Using Haralick Texture Features and RGB Pixel Values

The figures below present the results of K-means clustering applied to Haralick texture features, extracted using a 9×9 window, derived from both grayscale single-channel images and RGB three-channel images. Clustering was performed with and without the inclusion of pixel values and spatial information in the feature set.

This comparison highlights the differences in segmentation outcomes when color and spatial data are incorporated versus when they are excluded. In the segmented images, the black regions indicate malignant areas, while the white regions correspond to benign areas.

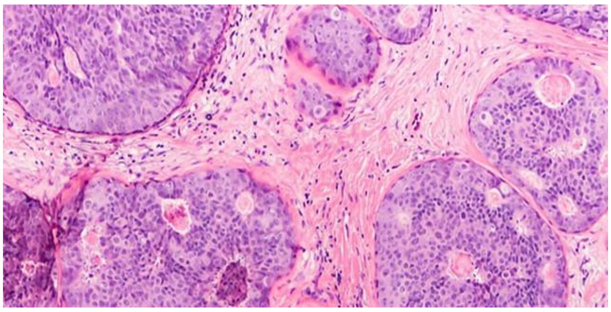


Fig.1. Original image

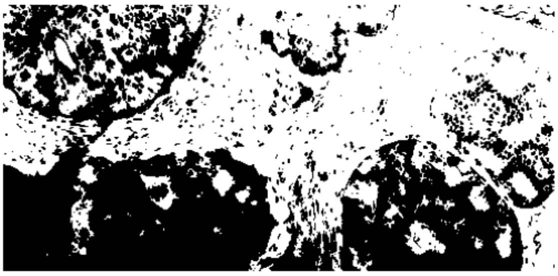


Fig.2. 9*9 graly-level haralick features combined with spatial and pixel features

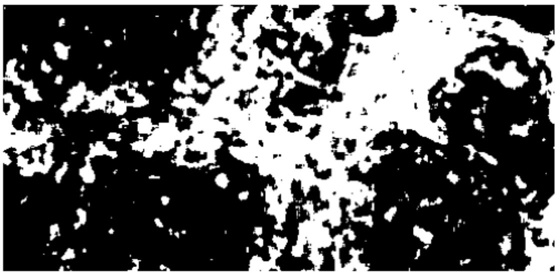


Fig.3. 9*9 gray-level haralick features only

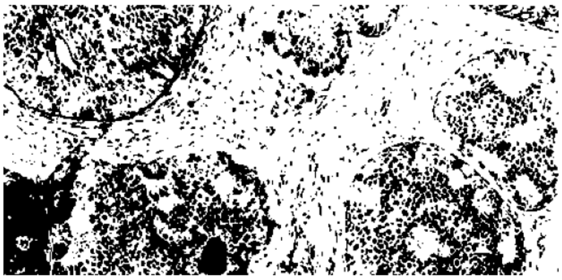


Fig.4. 9*9 RGB haralick features combined with spatial and pixel features

Fig.5. 9*9 RGB haralick features only

Clustering Results Using Different Window Sizes for Extracting Haralick Texture Features

The figure below presents the results of K-means clustering applied to Haralick texture features extracted from RGB three-channel images using various window sizes. Haralick features were computed with window sizes of 3×3, 5×5, 7×7, 9×9, and 11×11 to evaluate the effect of spatial context on texture representation.

This comparison highlights the differences in segmentation outcomes based on the window size used for feature extraction. In the segmented images, the black regions indicate malignant areas, while the white regions correspond to benign areas.

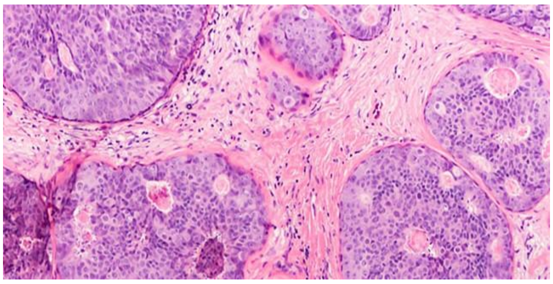


Fig.6. Original image

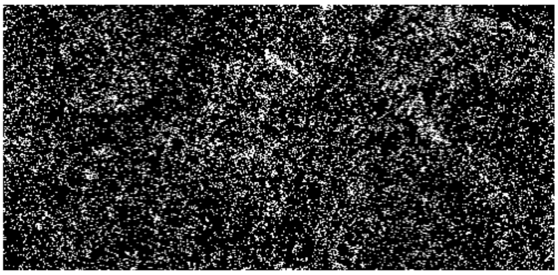


Fig.7. 3×3 sliding window

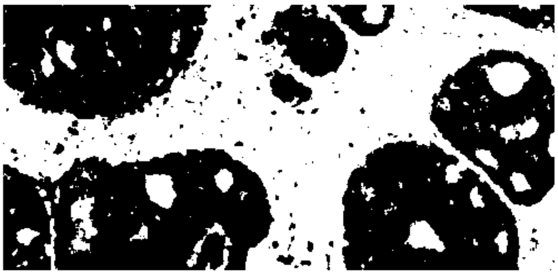


Fig.8. 5×5 sliding window

Fig.9. 7×7 sliding window

Fig.10. 9×9 sliding window

Fig.11. 11×11 sliding window

K-means Clustering with Automated Determination of Cluster Count

After calculating the total distance of each point to its cluster center for different cluster counts, we obtain a line plot of inertia versus cluster count(**Fig. 12**). We define the optimal cluster count as the point where the second derivative of the decreasing curve reverses, indicating that adding more clusters does not significantly reduce inertia.

The figure below shows a biopsy cross-section of oral cancer(**Fig. 13**), which contains multiple components. With a cluster count of 4 determined by this algorithm(**Fig. 14**), the regions of cancer cells are segmented and highlighted in red on the segmentation map.

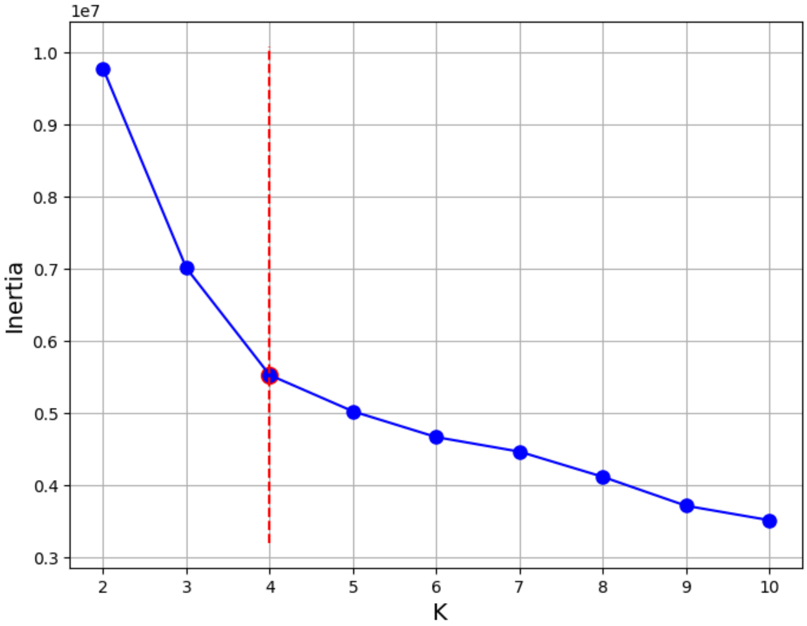
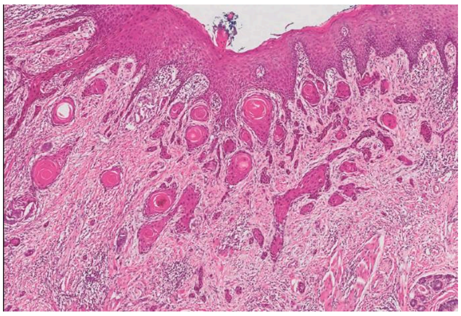


Fig.12. Inertia versus Cluster Count



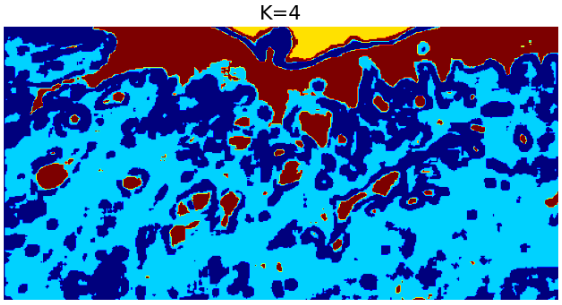


Fig.13. Biopsy cross section of oral cancer

Fig.14. Cluster Count K=4

DISCUSSION

Comparative Analysis of Haralick-Based Clustering: RGB vs. Grayscale and With/Without Spatial and Pixel Features

The clustering results demonstrate that Haralick texture features from RGB images outperform those from grayscale images. The inclusion of color information enhances feature discrimination, enabling more accurate separation of malignant and benign regions by capturing tissue texture heterogeneity absent in grayscale.

Furthermore, using only Haralick texture features without spatial and pixel values yields better segmentation than incorporating these additional features. This indicates that Haralick features alone effectively capture the necessary textural patterns for distinguishing tissue types, while spatial and pixel features may add noise and reduce clustering accuracy.

In summary, combining RGB color information with pure Haralick texture features provides more robust and accurate segmentation of oral cancer tissues.

Comparative Analysis of Sliding Window Sizes

The use of a 3×3 sliding window for feature extraction proved inadequate for effective clustering. This limitation is likely due to the insufficient size of the local region, which fails to capture the necessary texture details required for distinguishing between different tissue types.

As the window size increases, the quality of clustering noticeably improves. Window sizes of 7×7 or larger facilitate the identification of well-defined boundaries and contribute to a reduction in noise within the segmentation outcomes. These larger windows provide a more comprehensive capture of the tissue's regional textural characteristics, leading to more accurate and reliable segmentation of malignant and benign areas.

K-means Clustering with Automated Determination of Cluster Count

The clustering analysis identified four distinct clusters corresponding to key tissue components: air, cancerous tissue, connective tissue, and connective tissue with inflammatory cell infiltration. This result demonstrates that determining the optimal cluster count using the inertia-based method effectively separates tissue types into meaningful categories.

CONCLUSION

By treating the GLCM as a discretized probability density function, we can derive an adjusted set of Haralick texture features that remain asymptotically unaffected by image quantization. Apart from maximum probability and entropy-related features, the adjusted features preserve their original meanings.

Our analysis demonstrates that these invariant features perform better than the original definitions across various classification scenarios. This indicates that the invariant Haralick texture features maintain their consistency and reproducibility across different gray-level quantizations, unlike the traditional definitions.

REFERENCES

[1] Löfstedt, T., Brynolfsson, P., Asklund, T., Nyholm, T., & Garpebring, A. (2019). Gray-level invariant Haralick texture features. PLOS ONE, 14(2), e0212110. https://doi.org/10.1371/journal.pone.0212110

[2] Naira Elazab, Wael Gab Allah, & Elmogy, M. (2024). Computer-aided diagnosis system for grading brain tumor using histopathology images based on color and texture features. BMC Medical Imaging, 24(1). https://doi.org/10.1186/s12880-024-01355-9

[3] Öztürk, Ş., & Akdemir, B. (2018). Application of Feature Extraction and Classification Methods for Histopathological Image using GLCM, LBP, LBGLCM, GLRLM and SFTA. Procedia Computer Science, 132, 40–46. https://doi.org/10.1016/j.procs.2018.05.057

[4] Belsare, A. D., Mushrif, M. M., Pangarkar, M. A., & Meshram, N. (2015, November 1). Classification of breast cancer histopathology images using texture feature analysis. IEEE Xplore. https://doi.org/10.1109/TENCON.2015.7372809

[5] Eizan Miyamoto1 and Thomas Merryman Jr.2, Fast calculation of haralick texture features

[6] Haralick, R. M., Shanmugam, K., & Dinstein, I. (1973). Textural Features for Image Classification. IEEE Transactions on Systems, Man, and Cybernetics, SMC-3(6), 610–621. https://doi.org/10.1109/tsmc.1973.4309314

[7] Sirinukunwattana, K., Snead, D. R. J., & Rajpoot, N. M. (2015). A Stochastic Polygons Model for Glandular Structures in Colon Histology Images. IEEE Transactions on Medical Imaging, 34(11), 2366–2378. https://doi.org/10.1109/tmi.2015.2433900