International Journal of Psychosocial Rehabilitation, Vol. 24, Issue 7, 2020
ISSN: 1475-7192

Breast Histopathological Whole Slide Image For Retrieval Using Latent Dirichlet Allocation

[1] T.Kavitha, [2] S.Hemalatha, [3] K.Nandhini, [4] K.Chitra

Abstract--- *Whole Slide Image (WSI) has become the major carrier of visual and diagnostic information. Recovery of content-based images among WSIs may help to diagnose an unknown pathological image by finding its similar regions in WSIs with diagnostic information based on deep learning and local morphology nuclei statistical function. This research work implements the breast histopathological image approach, extracting the features such as Gabor and LSFN from the given image and splits the region based on the threshold (detha, beta). Based on threshold value, the matched regions are selected called candidate region. Then, the LDA model is utilized to obtain a histogram for each feature of every region. For this training and testing images are utilized. In this proposed system focused on clustering method such as NN- clustering algorithm that has been widely used for medical image segmentation. Local statistical feature of nuclei (LSFN) is presented to describe morphology and distribution pattern of nuclei is utilized for texture information. The algorithms have been implemented and tested on WSI images. The comparison is made with existing conventional NN-clustering method*

Keywords: *Clustering, Features extraction, Fuzzy C-Means, Histopathological Images, Image Retrieval, Image Segmentation, LDA*.

Introduction :

Digital image processing is a part of digital signal processing. The digital image processing field relates to digital images through a digital computer. Digital image processing has several advantages over analog image processing; it allows for a considerably wider collection of algorithms to be applied to input data and can keep away from issues such as noise build-up and signal deformation during processing. Digital Image Processing involves modifying digital data with the aid of a computer to improve the image qualities. The processing helps maximize the clarity, image sharpness and details of interesting features towards information extraction and further analysis

- T.Kavitha, Assistant Professor, Department of Computer Applications, Kongu Engineering College, Perundurai, Erode, Tamil Nadu.
- S.Hemalatha, Assistant Professor, Department of Computer Applications, Kongu Engineering College, Perundurai, Erode, Tamil Nadu.
- K.Nandhini, Assistant Professor, Department of Computer Applications, Kongu Engineering College, Perundurai, Erode, Tamil Nadu.
- K.Chitra, Assistant Professor, Department of Computer Applications, Kongu Engineering College, Perundurai, Erode, Tamil Nadu.

The results form a new digital image that may be displayed or captured in a pictorial format, or may be further changed by additional computer programs. Digital data is subjected to various image processing operations in order to improve certain data features and to remove noise from the image. Image processing involves changing the image quality to:

- Improve the image's pictorial information for human interpretation
- Making the image more suitable for independent machine perception.

Image processing methods can be grouped into three main functional categories:

Image Restoration

Image Restore compensates for noise, data errors, and geometric distortions that occur during recording, scanning, and playback operations.

- It eliminates the regular line dropouts
- Used to restore periodic line striping
- Perfect for random noise filtering
- Improve geometric distortions

Image Enhancement

Image enhancement processing is an image so that the result is more suitable for a particular application. Remove noise from noisy image, such as sharpening or de-blurring an out - of-focus image, highlighting the edges of the image, improving image contrast or increasing the brightness level of an image.

- Used to enhance contrast
- Intensity, saturation and color transformation
- Edge enhancement and synthetic stereo image production

Image Analysis

The purpose of image analysis is to make a quantitative measurement from an image to produce a description of the image. Image analytics techniques extract the certain characteristics that help to recognize an object. Quantitative analysis of the features of the object allows the description and classification of the image.

- Produce principal component images
- Producing image ratio
- Multi-spectral classification
- Change detection image generated

Image Segmentation

Segmentation partitions an image with similar attributes into distinct regions comprising each pixel. To be relevant and useful for image analysis and interpretation, regions should be closely correlated to objects depicted or interesting features. Significant segmentation is the first step from low-level image processing to turn a gray or color image into one or more other

images into a high-level image description for features, objects, and scenes. Successful image analysis depends on segmentation reliability but accurate image partitioning is typically a very challenging problem.

Techniques of segmentation are either contextual or un contextual. The latter, based on some global attribute, e.g. grey level or color, does not take into account the spatial relationships between features in an image and group pixels together. These relationships are further exploited by contextual techniques, for example grouping together pixels with similar gray levels and close spatial locations.

Image Morphology

Morphology is an image processing technique based upon object shape and shape. Morphological methods apply a structuring function to an image of input, generating an image of output of the same size. The value of each pixel in the image in place is based on a comparison with its neighbors of the corresponding pixel in the image data. By selecting the neighbor's size and shape, you can create a morphological operation in the input image that is responsive to specific forms. On grayscale images, where the source image is planar (single channel), the morphological operations can be described first. Then the description can be extended to pictures in full color.

Morphological operations are a process of image that involves erosion, dilation, opening and closing. Combinations of these operations are often used to analyze morphological images. The mathematical morphology describes many important operators. Dilation, flooding, opening and closing are these. Morphological operations add the structuring elements to an image input, generating an image output of the same dimension. The origin is at its core, regardless of the size of the structuring feature.

2 Existing Problem Definition

WSI Images are sound waves, with frequencies higher than the human hearing upper audible limit. Ultrasound in its physical properties is no different from ' normal ' (audible) sound, only in that it is not heard by humans. This limit varies from person to person, and is about 20 kilohertz (20,000 Hertz) in healthy, young adults. Ultrasound devices operate from 20 kHz to several gigahertz with frequencies.

Ultrasound is used for detecting movements and calculating distances in many different fields, using WSI tools. Ultrasound imaging or sonography is often used in medical applications. Ultrasound is used to detect invisible defects in the non-destructive testing of the items and structures. Ultrasound is used industrially for washing, mixing, and speeding up chemical processes. Animals including bats and porpoises are using ultrasound to identify predators and obstacles

In clinical diagnosis and medical intervention, medical imaging is the method and procedure of producing visual representations of a body's interior. Medical imaging helps to expose internal structures that are concealed by the skin and bones, and to diagnose and treat disease. Medical imaging also creates a database of normal anatomy and physiology so that abnormalities can be detected. While imaging of removed organs and tissues can be done for medical reasons, such procedures are usually considered part of the pathology rather than medical imaging.

3 Proposed Solution

The proposed system must apply the segmentation free of any selection of parameters. Hence an efficient algorithm is required to minimize image segmentation difficulties.

A new method with the name Fuzzy Local Information C-Means (FLICM) Clustering Algorithm is proposed to avoid the disadvantages in the existing system. A new factor in FCM objective function is required in order to overcome the above mentioned disadvantages. The new factor should have some special features of the output of image segmentation.

In FLICM, to replace the parameter used in EnFCM and FCM S and its variants, and the parameter used in FGFCM and its variants, a novel fuzzy factor is set. The new fuzzy local neighborhood factor can determine the spatial and gray level relationship automatically and is completely free of any selected parameter.

4 AN ANTICIPATION OF THE CONCLUSION

Different from supervision-based methods, content-based image retrieval (CBIR) can search in an unlabeled database and return the images similar to the query image. If the images in the database have diagnostic information, CBIR can aid in decision making. Therefore, CBIR has been widely researched in the field of medical imaging and digital pathology. In practical application, a valuable CBIR algorithm for pathology should be an unsupervised, accurate, and fast framework that can serve for a large-scale database.

Image segmentation is the process where an image is partitioned into multiple segments. The segmentation goal is to simplify or modify the image representation into a more meaningful image. It is the process of assigning each pixel of an image to a label, so that pixels with the same label share certain visual characteristics. The result of segmentation of the image is a set of segments which collectively cover the whole image.

Using this application, the problem of segmenting the noise image is eliminated. It reduces the overhead in segmentation calculations. The user interface assists in the effective analysis of the images. The application is well tested, and satisfaction of end users is found to be greater. While the spectral consistency and the WSI image quality of sharpened images appear to

be compatible in nature, i.e. improvements in one quality frequently weaken the other, it seems best to strive for as good spatial quality as possible for classification purposes, since the spectral quality remains above some acceptable minimum.

5 PROPOSED WORK

The following are the concepts which are implemented in this experimental system.

- Local Statistic Feature of Nuclei [*LSFN Extraction]*
- Gabor Feature
- LDA allocation with image retrieval
- Evaluation Image Retrieval

5.1 LOCAL STATISTIC FEATURE OF NUCLEI [*LSFN Extraction]*

In this modules LSFN and Gabor feature are utilized for nucleus description and texture information, respectively: According to previous works, the diagnosis of histopathological images mostly depends on the morphology and distribution of nucle i, and a new feature by considering the neighbors of each nucleus. On this basis, we present LSFN as a novel nucleus descriptor. Meanwhile, the Gabor feature has proved capable in retrieving histopathological images that are abundant in texture. In this module find that the combination of LSFN and Gabor feature can achieve the best performance with limited computational complexity.

The first step to extract our local statistic feature of nuclei (LSFN) is nucleus segmentation and visual of intermediate results. Then, each nucleus is regarded as a feature point, and attributes of the nucleus itself and its circular neighborhood are calculated and along with their significance in histopathology. By concatenating all attributes, to obtain a 27-dimensional feature vector $\boldsymbol{f} = [fi]i=1,...,27$ as the LSFN of the feature point, which is located in the centroid of the nucleus.

The nucleus-based feature is to prove effective in depicting histopathological images. Nonetheless, to find that the feature could be further improved to LSFN by changing some parts

- A large part of the information about variation of a cell lies in the shape and size of its nucleus, so enforce the corresponding description by adding area equivalent diameter f.
- To limit the size of the neighborhood in a reasonable range.
- The density of nuclei is more precise than the absolute number in revealing distribution of the neighborhood, so we change the nucleus distribution item $f9$ from the quantity of nuclei to current one.
- The staining of one nucleus is uniform, which means the standard deviation of pixel grayscales in one nucleus is always close to zero, thus this item is removed.

5.2 GABOR FEATURE

In these modules, Gabor feature implement a global feature for a whole image. In this module, WSI is divided into nonoverlapping square blocks and the Gabor feature is extracted from each block to represent the local texture of WSI. Such block Gabor feature has proved effective for histopathological image retrieval. The entire procedure of block Gabor feature extraction and the histopathological image is divided into blocks of b × b pixels.

For the block a set of Gabor wavelet filters under four scales and eight orientations is used to obtain 32 grayscale response images. The mean value μi and standard deviation of the pixels in the i^{th} response image are concatenated into a 64-dimensional Gabor feature vector **g** which is located in the center of this block, balance the weight of each component, features are normalized by dimension before they are utilized.

5.3 LDA ALLOCATION WITH IMAGE RETRIEVAL

LDA is a generative hierarchical model with four levels of data: word, topic, document, and corpus; meanwhile, each level corresponds to a random variable In the graph, word *w* is the basic discrete data unit, topic *z* is the latent level, the document is a sequence of words, and the corpus consists of multiple documents. The *m*th document is generated by repeatedly sampling topic zmn word *w*mn from multinomial distribution with parameter θm and ϕk, which are sampled from Dirichlet distribution with hyperparameters α and β. Given the word distributions of the documents in the corpus, the inference of LDA is to estimate the topic distribution of any existing or new document. This is usually accomplished through Gibbs sampling In terms of WSI, a feature point corresponds to a word, a region corresponds to a document, and all the regions constitute the corpus. A low-level feature such as LFSN is only a vector that measures the quantitative properties of one nucleus, e.g., area, radius, etc. A word can describe the state of the nucleus such as big, small, round, or flat. The dictionary of BoW consists of the *k*-means clustering centers of training features. A topic may represent user comprehension on the region such as benign, malignant, or specific caner type. Correspondingly the topic histogram of a document can reflect the probability distribution of the region among different histopathology types.

5.4 EVALUATION IMAGE RETRIEVAL

In comparison with existing methods on labeled dataset, our method is comparable with the state-of-the-art method, which is supervised. LSH dramatically improves the query efficiency with a small cost of precision. In general, our method is accurate and effective for most types of breast histopathology. The proposed approach also shows good potential to construct a powerful classifier. The experiments on a mixed dataset act as a simulation of real-world situation. The output method performs the best when UR > 4. Although the precision of our method is lower than 0.5, the correct results can provide valuable reference to pathologists in diagnosis.

6 Applying Fuzzy C-Means Clustering With New Fuzzy Factor G

During the processing, the values for cluster, weight i.e., fuzziness factor set to value two, and eplison value 10 to the power of -5 are set. Then the fuzzy partition matrix is initialized. Then cluster centers are calculated along with membership matrix with the given fuzzy factor G. The steps are iterated for given number of times to segment the image further. The G Factor is calculated using the Formula

$$G_{ki} = \sum_{\substack{j \in N_i \\ i \neq j}} \frac{1}{d_{ij}+1}(1-u_{kj})^m ||x_j - v_k||^2 \quad (1)$$

Median Filter to filter the noise pixel values

In the form, the noise in the image is filtered by changing the pixel value with median values of surrounding pixels. To apply median filter, for each pixel, the surrounding pixels 3x3 is taken and the gray scale values are summed and median value is found out. The median value is set to the center pixel. This reduces the noise data in segmented image for clear view of output image.

GMM

GMM model, also known as snakes, is a computer vision system designed to delineate an object outline from a potentially noisy 2D image. The snakes model is common in computer vision, and in applications such as object tracking, shape recognition, segmentation, edge detection and stereo matching, snakes are commonly used.

A snake is an energy-minimizing, deformable spline shaped by constraints and image forces pulling it up to object contours and deformation-resistant internal forces. Snakes can be understood as a special case of the general technique of matching a deformable model to an image through minimizing the energy. In two dimensions, the active type model reflects a discrete variant of this approach, using the point distribution model to limit the variety of shapes to a specific domain learned from a training set

The whole problem of finding contours in images is not solved by snakes, since the method requires knowledge of the desired contour shape beforehand. Rather, they rely on other mechanisms such as user interaction, interaction with some process of understanding the image at a higher level, or information from adjacent image data in time or space.

7 CLUSTERING ALGORITHM

EM clustering plays an important role in solving problems in the areas of pattern recognition and fuzzy model identification. A variety of fuzzy methods of clustering have been proposed and most of them are based on distance parameters. The fuzzy c-means (FCM) algorithm is a commonly used algorithm. To measure fuzzy weights it uses reciprocal space. The idea of FCM is using the weights that minimize the total weighted mean-square error:

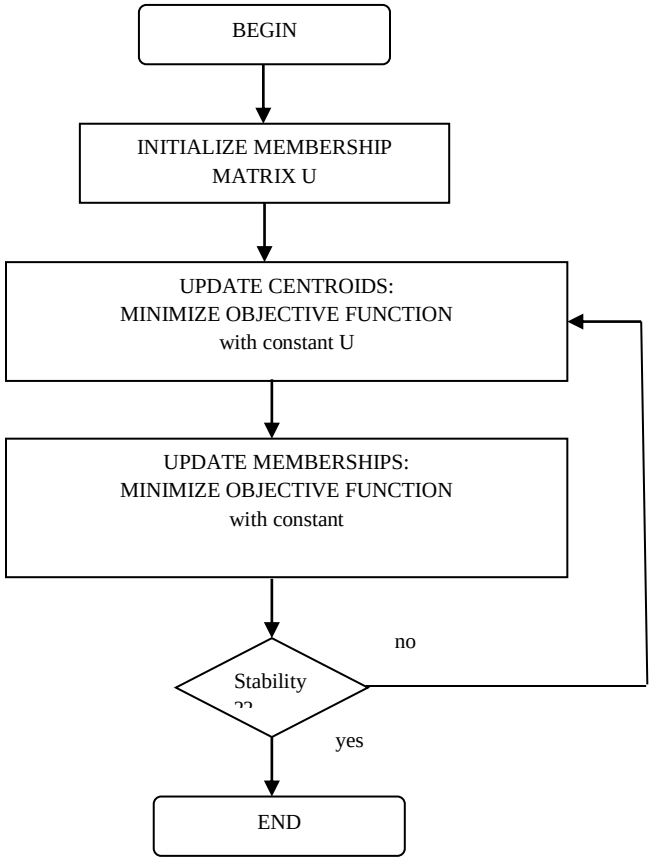
J(wqk, z(k)) = (k=1,K) (k=1,K) (wqk)|| x(q)- z(k)||2 (k=1,K) (wqk) = 1 for each q (2)

wqk = (1/(Dqk)2)1/(p-1) / (k=1,K) (1/(Dqk)2)1/(p-1) , p > 1 (3)

The FCM allows each feature vector to belong to every cluster with a fuzzy truth value (between 0 and 1), which is computed using 2nd Equation. The algorithm assigns a feature vector over all clusters to a cluster, depending on the maximum weight of the feature vector.

Algorithm techniques

Step 1: Set the cluster prototype number c, fuzzification parameter m and stop condition
Step 2: Initialize the fuzzy partition matrix in a random manner.
Step 3:: Set the counter for loops b= 0.
Step 4: Calculate prototypes for the cluster using (1).
Step 5: Use (2) to calculate Membership values.
Step 6: If max{ U(b)-U(b+1)} < π stop, otherwise set b= b+1 and proceed to step 4



7.1 M-FUZZY C-MEAN CLUSTERING ALGORITHM

Given a set of N data points **X =** { xi } **,** a set of kernel functions { K_k }, and the desired number of clusters C, output a membership matrix **U = { uic },** s { w_k } for the kernels.

Procedure MGRCA (Data X, Number C, Kernels { K_k },

1. Initialize membership matrix U(0).
2. repeat calculate normalized memberships
3. Calculate Co-Efficient
4. for (i=1..N;c=1..Ck=1..M) do N
5. Calculate Co-Efficient
6. for (k = 1..M) do N
7. $\beta\kappa \leftarrow \sum \sum$ (u (t)) m αick
8. end for
9. Update Weights for (k = 1..M) do 1/ β_κ
10. $\omega \leftarrow 1/\beta_1 + 1/\beta_2 + \ldots$
11. end for
12. Calculate Distances for (i = 1..N;c = 1..C) do
13. $D \leftarrow \sum_{ic} \alpha_{ick}(\omega_k^{(t)})^2$ $_{k=1}$
14. end for
15. Update Memberships for (i = 1..N;c = 1..C) do U
16. end for
17. until $\| U^{(t)} - U^{(t-1)} \| \in$

The above Algorithm summarizes the MFCM algorithm, which starts by initializing a random membership matrix satisfying nonnegative and unity constraints. Optimal weights are calculated by fixing the memberships, and optimal memberships are updated assuming fixed weights. The process is repeated until the amount of change per iteration in the membership matrix falls below a given threshold.

8 EXPERIMENTAL RESULTS

Requirements

PARAMETER	VALUE
Simulation tool	Matlab R2013
Simulation Type	Segmentation Modle
Image Dataset	100
Compression Rule	Segmetation and Tumor Size
Method	Fuzzy C Mean Clustering
Performance Metrics	Time Analysis

Table 8.1 Environment Setup

Parameter For Evaluation

Table 8.1 shows the experimental results for existing and proposed system. The table contains image datasets, SAR image, Optical Image and size of images details are shows.

Image Dataset (n)	SAR Image	Optical Images	Size of Images (MB)
25	Simage_1	Opimage_1	200
50	Simage_2	Opimage_2	250
75	Simage_3	Opimage_3	300
100	Simage_4	Opimage_4	350
125	Simage_5	Opimage_5	400

Table 8.2 **Dataset Collections**

The above table shows the experimental results for existing and proposed system. The table contains image datasets, size of images and EM and FCM number of change detection image details are shows.

Image Dataset (n)	Size of Images (MB)	EM (n)	FCM (n)
25	200	11	14
50	250	23	31
75	300	36	42
100	350	45	51
125	400	58	62

Table 8.3 Number of change Detection Images

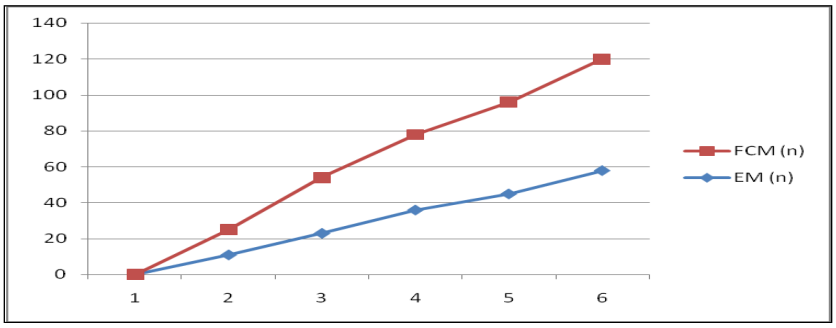


Figure 8.1 Number of change Detection Images-EM-FCM

The above figure shows the experimental results for EM and FCM algorithm system. The figure contains image datasets, size of images and EM and FCM number of change detection image details are shows. Table 8.3 shows the experimental results for EM and FCM algorithm system. The table contains image datasets, similarity values and SAR image change detection duration (ms) details are shows.

IMAGE Datasets	Similarity (p)	EM (ms)	FCM (ms)
Simage_1	0.045	0.00.03	0.00.02
Simage_2	0.066	0.00.42	0.00.34
Simage_3	0.079	0.00.58	0.00.46
Simage_4	0.089	0.00.63	0.00.54
Simage_5	0.097	0.00.78	0.00.65

Table 8.4 Number of change Detection Images Similarity -EM-FCM

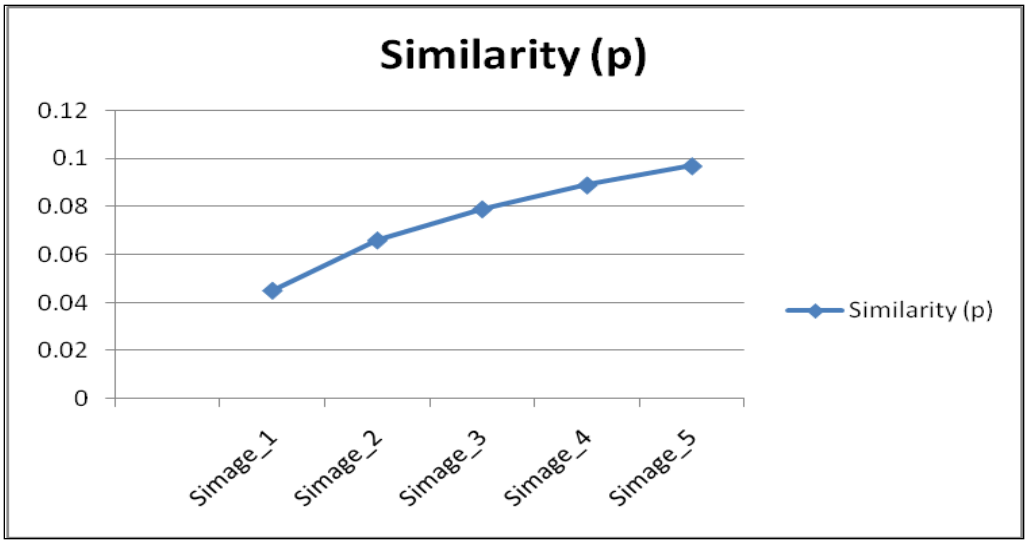


Figure 8.2 Number of change Detection Images Similarity -EM-FCM

Fig 8.2 shows the experimental results for EM and FCM algorithm system. The table contains similarity values and SAR image change detection duration (ms) details are shows.

Table 8.4 shows the experimental results for EM and FCM algorithm system. The table contains image datasets, similarity values and Optical image change detection duration (ms) details are shows.

IMAGE Datasets	Similarity (p)	EM (ms)	FCM (ms)
Opimage_1	0.056	0.00.05	0.00.03
Opimage_2	0.068	0.00.46	0.00.36
Opimage_3	0.082	0.00.62	0.00.55
Opimage_4	0.092	0.00.72	0.00.62
Opimage_5	0.096	0.00.81	0.00.73

Table 8.5 Number of change Detection Images Time -EM-FCM

Fig 8.3 shows the experimental results for EM and FCM algorithm system. The table contains similarity values and Optical image change detection duration (ms) details are shows

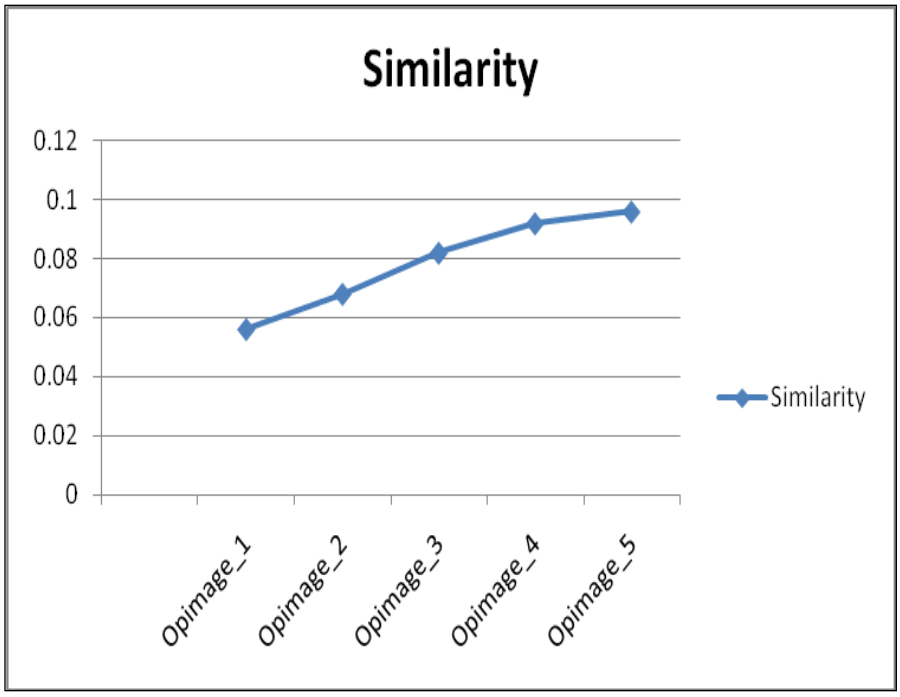


Figure 8.3 Number of change Detection Images Time -EM-FCM

9 RESULTS AND DISCUSSION

- Finding the surface detection in SAR and Optical images.
- The GSM clustering concept is improved by using FCM (Fuzzy C-Means Clustering) concepts for the Segmentation purpose.
- Improving the image segmentation by applying median filter for removing noisy pixels.
- Comparing the input image with the stored image in database that is already founded to be affected by area.
- Intermediate images can be viewed from starting to final segmented image.
- RGB image can be converted into gray scale image.

CONCLUSION

This method removes the difficulties of segmenting the noise picture. It reduces the overhead in segmentation calculations. The user interface helps in the efficient processing of the images. The program is well checked, and satisfaction of end users is found to be greater.

The application works well in windows environment for given tasks. Any node with installed Simulation tool framework can run the application and identify the best location. The underlying framework can be applied to any and all web servers and even to multi-platform systems such as Linux, Solaris and more. The program is also expected to expand the services to include IBM software as data.

The main conclusion is that while the spectral consistency and the WSI image quality of sharpened images appear to be compatible in nature, i.e. improvements in one quality often weaken the other, it seems best to strive for as good spatial quality as possible for classification purposes, provided that the spectral quality remains above some acceptable minimum.

The system removes problems in the existing system. It is user friendly developed. In applying the segmentation algorithm the system is very quick. This software is very specific in reducing the problem of segmentation algorithms.

REFERENCES

[1] X. Zhang, H. Su, L. Yang, and S. Zhang, "Fine-grained histopathological image analysis via robust segmentation and large-scale retrieval," inProc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2015, pp. 5361–5368.

[2] S. Naik, S. Doyle, S. Agner, A. Madabhushi, M. Feldman, and J. Tomaszewski, "Automated gland and nuclei segmentation for grading of prostate and breast cancer histopathology," inProc. 5th IEEE Int. Symp. Biomed. Imag., Nano Macro, May 2008, pp. 284–287.

[3] H. Fatakdawalaet al., "Expectation-maximization-driven geodesic active contour with overlap resolution (EMaGACOR): Application to lymphocyte segmentation on breast cancer histopathology,"IEEE Trans. Biomed. Eng., vol. 57, no. 7, pp. 1676–1689, Jul. 2010.

[4] H. Kong, M. Gurcan, and K. Belkacem-Boussaid, "Partitioning histopathological images: An integrated framework for supervised colortexture segmentation and cell splitting,"IEEE Trans. Med. Imag., vol. 30, no. 9, pp. 1661–1677, Sep. 2011

[5] R. Gutirrez, F. Gmez, L. Roa-Pea, and E. Romero, "A supervised visual model for finding regions of interest in basal cell carcinoma images,"Diagnostic. Pathol., vol. 6, no. 26, Mar. 2011. [Online]. Availableat: http://diagnosticpathology.biomedcentral.com/articles/ 10.1186/1746-1596-6-26

[6] C.-R. Angel, D. Gloria, R. Eduardo, and G. Fabio, "Automatic annotation of histopathological images using a latent topic model based on nonnegative matrix factorization,"J. Pathol. Informat., vol. 2, no. 2, p. 4, 2011.

[7] P. Ghosh, S. Antani, L. Long, and G. Thoma, "Review of medical image retrieval systems and future directions," inProc. 24th Int. Symp. Comput.-Based Med. Syst., Jun. 2011, pp. 1–6.

[8] A. Kumar, J. Kim, W. Cai, M. Fulham, and D. Feng, "Content-based medical image retrieval: A survey of applications to multidimensional and multimodality data,"J. Digital Imag., vol. 26, no. 6, pp. 1025–1039, 2013.

[9] [22] X. Zhang, W. Liu, M. Dundar, S. Badve, and S. Zhang, "Towards largescale histopathological image analysis: Hashing-based image retrieval," IEEE Trans. Med. Imag., vol. 34, no. 2, pp. 496–506, Feb. 2015

[10] J. Caicedo, F. Gonzalez, and E. Romero, “A semantic content-based retrieval method for histopathology images,” in Information Retrieval Technology (Series Lecture Notes in Computer Science 4993), Berlin, Germany: Springer-Verlag, 2008, pp. 51–60