Review of Content Based Retrieval of Malarial Positive Images from Clinical Database

Jasdeep Kaur
Computer Science and Engineering, Guru Nanak Dev Engineering College, Ludhiana, India
Jasdeepkaur09@gmail.com

Kamaljit Kaur
Information Technology, Guru Nanak Dev Engineering College, Ludhiana, India
Kamaldhillon2000@yahoo.com

Abstract—Content-based image retrieval is used for retrieving various medical images from huge database of patients. Query for retrieval of image could be done in form of text or in form of image itself. In this paper four different methods for image retrieval are analyzed. The comparison of all these techniques is done on basis of various parameters e.g. recall, precision, sensitivity and specificity. This paper reviews the method for parasite detection, degree of disease and stage of infection. Major steps involved in CBIR are pre-processing, thresholding, feature extraction, similarity detection and retrieval.

Index Terms—content-Based image retrieval (CBIR), medical Based CBIR, Malaria.

I. CONTENT BASED IMAGE RETRIEVAL

CBIR came up in early 1990’s and it is important part of computer vision and used for processing images. It refers to a system that returns an image matching an image given as query from image database. Each image in the database is defined by a specific feature vector. These feature vectors may consist of color, texture, shape, etc. These vectors are then stored along with the image. The similarity features between the image that we give and image bank are arrived based on the stored vectors [1]. General architecture of CBIR can be represented as in Fig. 1.

There are four principle components of CBIR. These are [2]:
1) Visual feature extraction: mainly color, texture, logical features (identity of objects) and abstract features (significance of the scenes) come under visual features. The semantic gap to measure loss of information while feature representation of image. Also sensory gap measures loss between structure and representation. Along with these visual features we also have features derived from segmentation. The images are annotated with the visual features to make it easy for retrieval purposes.

2) Storage and access methods: Images can be stored in relational databases, inverted files, self made structures or place entire index in main memory for smaller databases. Various feature space reduction techniques, indexing techniques and other methods are used to make retrieval more efficient and quick.

3) Distance measures, similarity calculation: Euclidean vector space model is used to measure distance between image features of image given as query and other images features present with us. The use of probabilistic approach is done to measure that image is relevant or not relevant. Various pattern matching schemes are used for retrieval purposes.
4) Other techniques: Various GUI interaction methods are also used. Relevance feedback is another important technique used for selecting the relevant images and that are not relevant by the researcher.

II. MEDICAL BASED CBIR

These days there are large number of medical images obtained by different methods, so the need of medical image retrieval system is increasing day by day. There are number of systems present for retrieval purposes like webMIRS[3], COBRA[3], medGIFT[3], KMMeD[3], etc. For medical images we require high amount of prior knowledge. Very high level of accuracy is required to make images more reliable. Mainly images are known by features globally (single value assigned to full image) but medical images also require local information (assigned to reach image pixel) without which it will lead to semantic gap- gap between user interpretation and information extracted from visual data. Pixels here are represented by feature vector having minimum feature vectors. Another important feature required here is annotation of features to image [3]. Figure 2 shows typical architecture of the medical system.

Medical CBIR is mainly used for three applications [5]:
1) Primitive queries that are used for retrieval of relevant images and used for further studies.
2) Semantic query used for searching images of desired disease.
3) Browsing: here query maybe semantic or primary and mechanisms such as query refinement and relevance feedback are used.

A. Approach for Image Retrieval in Medical Image [4-5]

(a) Categorization: We have different type of images for example we have ultrasound images, X-ray, MRI, mammograms etc. All these images require different image processing methods. So, we require categorization of images. There are mainly three main classes: modality, region and orientation. Automatic categorization methods are required. These categorize the images on basis of global features derived from image. Many-many mapping from image to categories could be present as one image could be linked to many categories and vice-versa.

(b) Registration: A prototype model is defined by person who has full knowledge in medical field. This prototype is used for measuring registration parameters (RST-parameters). These parameters are used for object identification and retrieval.

(c) Feature extraction: We have global features present for entire image categorization. Also for pixel level or region information for any image we have local features. Processing of queries requires local features or category-free features. Feature vectors are made here. For example features extraction may include extraction of texture, color or shape of objects.

(d) Feature selection: Here the feature set is pre-computed for every image in order to have better performance. This is computed on the basis of a-prior knowledge of category and knowledge about the query. For example, in radiographs to detect bone fracture we have to extract features by edge based method. But if we want knowledge of bone tumors we need texture description of the image.

(e) Indexing: Indexing is defined as search of any element from the larger database by representing the data in reduced form. Image is divided into different regions based on feature vector and invariant moment is assigned to them resulting into structures called blob. The image represented in form of blob is then adjusted along with the parameters assigned in registration earlier.

(f) Identification: This involves linking of prior knowledge to blobs by medical experts to build prototype blob trees. It makes up labels for various images from their properties like size and texture of shape.

(g) Retrieval: The query is built from possible category of image, blob structures or from set of local features. Blob structures are compared with selected features in order to make search. Distance measures are used for image retrieval purposes. The images those have minimum
distance are retrieved. Similarity is measured by one to one match and many to many.

Next step is to reduce semantic gap. This can be mainly removed by following methods [6].

1) Object-ontology: Mainly each region of image is defined by specific descriptors like color, position, size and shape. So this is qualitative definition of high level concepts. Ontology is defined as representation of knowledge.

2) Machine learning: It is formal tool used for supervised (to predict output for specific input) and unsupervised learning (to organize input).

3) Relevance feedback: The retrieved results are labeled as relevant or irrelevant images by the user. In next phase i.e. learning phase query parameters are adjusted and again the results are retrieved until user is satisfied.

4) Semantic template: semantic classes are here represented by representative features or templates.

5) WWW image retrieval: This retrieval is done by use of textual and visual features of image on web.

System evaluation is done by various parameters like specificity, sensitivity, accuracy, precision, recall.

III. CONTRIBUTION OF CBIR SYSTEM

The medical information system is made to fulfill the goals that it provides proper quality and efficiency of results required by the patients at right time. To fulfill such goals we need large amount of information more than patient’s record. For the clinical decisions we require to find out similar images of disease for various purposes.

This can be used for research, as patterns that may lead to new knowledge can be more easily found; and diagnostics, as doctors can search for visual similar cases to analyze treatment to be applied.

The medical institutions can obtain several benefits from implementing automated image retrieval system as it can provide by quick diagnosis than manual diagnosis. It also serves the advantage to reduce cost as there would be no need of human resource and increase competitive advantage. Finally, the service provided to the patient is greatly improved.

Teaching efficiency of medical students and postgraduate students can be improved as well by using visual features. For retrieval of images with similar diagnosis we require visual features this also helps in retrieval of images with different diagnosis but different similar visuality. It could help students and lecturers to browse various images and study them deeply.

Therefore CBIR systems have their great contribution to society in form of better research, better diagnosis and much more.

IV. MALARIA DISEASE

Malaria is caused by plasmodium parasite.

PLASMODIUM: THE MALARIA PARASITE

Plasmodium has 4 species that can cause human infection. These are

- Plasmodium Falciparum
- Plasmodium Vivax
- Plasmodium Ovale, and
- Plasmodium Malariae.
V. RULES FOR PARASITE IDENTIFICATION

We can differentiate between different parasites on basis of few rules. These rules are mainly size of parasite, shape, color and number of schuffner’s dots. Size is mainly described in terms of enlarged or not. The shape of each parasite is different, it maybe round, oval or elongated or amoeboid. Color is normal, dark or pale.

These features are mainly used to differentiate between different types of parasites. Table 1.1 explains about the major rules to differentiate between parasites.

<table>
<thead>
<tr>
<th></th>
<th>PF</th>
<th>PV</th>
<th>PO</th>
<th>PM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size</td>
<td>Small</td>
<td>Enlarged</td>
<td>Enlarged</td>
<td>Small</td>
</tr>
<tr>
<td>Shape</td>
<td>Round Crescent Gametocyte</td>
<td>Round or Oval</td>
<td>Round, Oval Amoeboid</td>
<td>Round</td>
</tr>
<tr>
<td>Dots</td>
<td>Large red Spots</td>
<td>Small red dots</td>
<td>Small red dots</td>
<td>Few tiny dots</td>
</tr>
<tr>
<td>Color</td>
<td>Darkened or normal</td>
<td>Normal or Pale</td>
<td>Normal</td>
<td>Normal</td>
</tr>
</tbody>
</table>

VI. RELATED WORK

Khan, M.I. et al in their paper [11] proposed method to automate diagnosis of malarial images. Here the major problem related to classification and detection of type of parasite present in RBCs was solved. Also, positive predictive value (PPV) and sensitivity of the system was calculated. Plus the time required for retrieval was calculated.

The database included of 110 images of giemsa stained blood films of 460 x 307 pixels bitmap. The first step was preprocessing of images using 5 x 5 median filters and then disk shape structure with radius 6 pixels was used for filtering. Also SUSAN edge [7] and corner detection enhancement were used to remove noise. The next step involved segmentation by local and global thresholding. Two thresholds were used one to separate erythrocytes from background and another to separate parasites. After thresholding holes were removed using structuring element of disk shaped with radius of 40 % of RBC’s radius. Also remove the objects having radius less than half of RBC’s. Intersection of binary image of RBCs and parasite image was done to remove invalid objects. Global thresholding was done by Otsu’s threshold [8] or by optimal threshold maximizing sum of two class entropies. After this watershed algorithm [9] was used to separate touching cells and Marker controlled watershed algorithm [9] to separate overlapping cells. Granulometry was used for calculation of size and eccentricity of RBCs. Next step was feature extraction that had two sets of features- image characteristics and set of features technicians used to measure infected objects. The feature classification was done by two-stage tree classifier in which at first stage image was weather malaria present or not, and at second stage type of parasite present.

Performance was measured by checking (PPV) and sensitivity.

The results of this method were promising as it detected parasite species and counted number of RBCs with 85.5 % sensitivity and 81% PPV. Also, it gave better execution time of 3.44 sec. In this paper we have learned that firstly image is classified as malarial positive or negative and then parasite type detected. However, this paper had not calculated the other performance evaluation factors that like recall and precision, accuracy and specificity.

Diaz, G.et al in their paper  [12] proposed algorithm for quantification and classification of RBC’s infected with Plasmodium falciparum. Here problem of luminance difference was also solved. Erythrocytes were separated and then only infected one found. Major problem solved here was detection of stage of infection. User intervention was also allowed to reach required results. The performance of system was calculated using sensitivity and specificity.

The database of 450 images of blood cells was used. Firstly the luminance difference in the image was removed by local adaptive low pass filter. Next, for classification of erythrocyte following methods were used.

1) Color space classification and machine learning algorithm were used to label pixel as foreground or background depending on the color present.
2) The pixels in foreground were put into one simplified inclusion tree that have lower value (0) tree and upper value (255) tree. Further reduction to foreground and background was done by considering redundant objects as background. After this binary image clump shape objects were obtained on which pattern match was applied to obtain number of erythrocytes.
3) Classification of obtained erythrocyte was done based on 25 different features and stage of infection found. Classifiers were composed of four learning models three of them for each stage of erythrocyte.

Results show that there were 12,577 erythrocytes, out of which 11,844 healthy cells, 521 ring stage, 109 trypomastigote stage, 83 schizont infection stage. Very good percentage of sensitivity and specificity was achieved for different stages of infection. This method was simple and helped to identify infected erythrocyte and the stage of infection. 95% of clump splitting was possible. But in this paper the detection of type of malaria parasite present was not done and precession and recall of the system was not calculated.

Bugatti, P.H. et al in their paper [13] proposed CBIR technique along with relevance feedback. This paper removed the major drawback of reducing the difference between actual feature representation and what the users expect by using feedback and user profiling (modify distance function). This achieved increase in precision and reduced semantic gap.

Database of 3258 images of CT lung ROI, each of 64x64 pixels and 256 gray levels was used. The dataset contained 6 classes out of which 1 normal and 5 abnormal. Two feature extractors-Zernike moments and Haralick’s texture [13] (256 moments describing shapes) were used. Distance functions like minkowski family (Lp distances) were used to see how similar the feature vectors were. In this paper CBIR, relevance feedback was composed of three steps:

1) Similar images were retrieved to query. Mediator approach was defined to improve similarity.
2) Images labeled as relevant or irrelevant and passed to labeling analyzer.
3) Query adjusted based on relevance and best suited distance function found. Images were retrieved based on this new query.

Last two steps were repeated until user was satisfied. Then query point movement (QPM) [10] was estimated by moving point far from irrelevant images and towards relevant images by use of Rachio’s formula. Information about this relevance feedback cycle was stored in profile. Results of the paper show that \( L_{\text{infty}} \) distance achieved precision gain of 22% than that of Canberra distance in Emhesema class. Reduced semantic gap can increase efficiency. This technique was used for lung images but also could be used for retrieval of malaria parasite image.

Ghosh, S. & Ghosh, A. in their paper [14] proposed an algorithm to make a decision support system that could be used for easy retrieval of malarial positive medical images and also to detect degree of disease present. Efficiency was calculated by use of precession and recall. Processing time was also recorded. The database used here consist of 20 images each of MRI and X-rays, 40 pathological test images both thin and thick smears of malarial positive blood.

Firstly pre-processing was done. Then artifacts were deleted by converting into 4-connected region label and area was calculated. Region with area < 0.3% of total image area was removed and marked as background. Holes were removed if calculated area < area of circle with radius 60% of radius of RBC and replaced as foreground. Next binary image obtained from Otsu’s method was used to classify as thin blood smear or thick blood. Thin smears have large background. Position of chromatin within RBC decided, weather image was malarial positive or not. The edge detection of chromatin was done by sobel mask. Next step included calculation of degree of disease

Degree of disease = \left[ \frac{\text{no of affected RBCs}}{\text{Total no of RBCs}} \right] \times 100 \tag{1}

RBC could be standing, overlapped or partially visible. The centroid, minor axis and major axis of all the regions was calculated. The ratio of major to minor axis was calculated. For free standing cell ratio ~ 1 Cell was partially visible if major or minor axis < diameter of free standing cell and overlapped if major or minor axis > diameter of cell. The query for retrieval of images represented here was in text form or image form. Textual query was retrieved by matching with the attributes stored. But for image as query there were two processes, offline process (database studied and information base prepared) and online process query was formed from image by similarity definition and then compared with the information base to return the image.

This paper resulted in 30 regions, out of which 6 were partially visible and 2 were overlapped cells. Processing of 300x300 images took 1.95 sec with stand deviation of 0.16. But processing of 600x600 image took 5.85 sec with standard deviation of 0.36 leading to three folds increase in processing time. 95% precession and 100% recall was obtained. From this paper we learned that text or image can be used for retrieval purposes and degree of infection could also be calculated. The main problem was that it did not identify the type of parasite. The algorithm proposed here was not used for thick blood smears and gray scale images were not retrieved.

VII. RESEARCH GAP AND PROBLEM FORMULATION

In previous papers, the algorithm automatically adapts to different strains of culture tested by different laboratories. However this does not identify the type of parasite neither automatically nor manually using expert system. Also system does not find the stage of infection.

The above defined system also does not give, a well defined ontology based annotation to each image content, which would make the system highly reliable and accurate. Also this would help to reduce the semantic gap. Various other methods can also be used to remove semantic gap so as to make it more efficient.

Moreover in the given system no time analysis has been done to prove how fast or slow the system works on image retrieval in terms of worst case analysis.
Problem Formulation

With the advancement of technology there is enormous increase in quantity of images generated by hospitals and medical centers. Efficient image searching, browsing and retrieval tools are required to deal with database of patients. For this reason, many image retrieval systems were developed.

Malaria blood sample slides were undergone image processing in order to detect whether the sample is malaria positive or not. In previous papers, the algorithm automatically adapts to different strains of culture tested by different laboratories. However, this does not identify the type of parasite neither automatically nor manually using expert system. The above defined system also does not give, a well defined ontology based annotation to each image content, which would make the system highly reliable and accurate. Moreover in the given system no time analysis has been done to prove how fast or slow the system works on image retrieval.

In proposed work we have detected the type of malaria parasite present. Also using this system user could make query either by text or by image. Query by text could be done because of annotation done using Ontology.

Table II

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Precession</td>
<td>---</td>
<td>---</td>
<td>74%</td>
<td>95%</td>
</tr>
<tr>
<td>Recall</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Specificity</td>
<td>81%</td>
<td>99.7%</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>85.5%</td>
<td>94%</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Time</td>
<td>3.44 sec</td>
<td>---</td>
<td>---</td>
<td>1.95 sec  (for 300x300 image)</td>
</tr>
<tr>
<td>Parasite type detection</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Stage of infection</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Relevance factor</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
</tbody>
</table>

IX. Future Scope

In order to make the research more efficient there is urgent need to develop algorithm for the malaria image content retrieval, storage and annotate images using ontology of malaria ontology (clinical, epidemiological, biological, etc). Performance evaluation must be done on basis of evaluation gap, ground truth, recall and precession. We can use ontology for this purpose as described above in the papers.

REFERENCES


