CLASSIFICATION OF IMAGES IN BIOMEDICAL PUBLICATIONS

Márta ALTRICHTER
Advisors: Gábor HORVÁTH, Bill ANDREOPOULOS

I. Introduction

In the past years the number of publications in biomedical literature was rapidly growing. The number of publications for example in the open accessed database of BioMed Central (BMC) shows a quadratic tendency [1]. As a result the importance of good query systems is increasing too.

Images in publications of biomedical literature are frequent and often provide important details to the reader, while this high semantic content is hard to be indexed, described by automatic methods. The existing approaches to image annotation for reader queries are manual annotations, text based search and content based image retrieval (CBIR) [2].

Manual annotation has obvious disadvantages as the number of images to annotate grow: human time consumption, more annotators can have different views on the same image (an example system is Google Image Labeler). Text based search engines work on querying the title of the image (new techniques try to extend this text mining to the paragraph where the image was referenced too [1]). CBIR systems try to index images based on image features like colors, texture, ... Query-by-image systems use the feature vector of the query image and search for images with similar feature vectors. The presently available CBIR systems mainly work on general image databases, while biomedical databases are harder as images are often very similar while representing totally different content. Hybrid systems try to merge information acquired from images and from the text too.

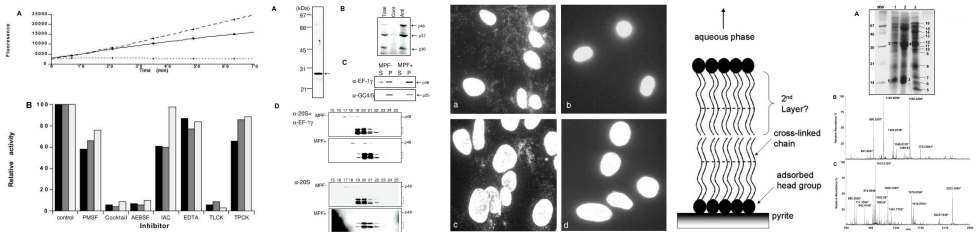
Our main aim is to create a system incorporated to GoPubMed, where the user can search for an image based on text query and by choosing between predetermined image classes. An example query would be: 'fruit fly evolution' search string with the class 'graph'. This query would be expected to return images which are graphs and somehow correlate to 'fruit fly evolution', like a population change graph of a fruit fly community.

In this paper we focused on image classification based on only image features extracted. Moreover we started to propose a way to deal with images containing more panels having different type of images to get closer to the main aim.

II. Image Classes

We downloaded 12598 articles of PubMed publications in years 2000 to 2005. These articles contained 126865 images (of which some are duplicates or thumbnails). Based on our overlook on the images and on the previous work done by Rafkind, Lee, Chang and Yu [3] we decided on the following classes to distinguish between as a first step:

1. Graph: all kind of charts (bar, plot, pie, ...), including hierarchical charts and hand made ones.
2. Gel: images showing chromatographic experiment results.
3. Thing: images, photographs of existing objects like a cell (microscopic), tissues, organs or photo of lab equipment, ... Subclasses: a, Microscopic: as approximately half of the Thing images turned out to be microscopic images it is a feasible way to experiment with the division of Thing to Microscopic and Non-microscopic subclasses. b, Non-Microscopic.
4. Model: any model of a biological process, experiment model, protein sequence or higher protein structure representations, any model of an equipment, drawings ...



(a) Graph (b) Gel (c) Thing (d) Model (e) Gel-Graph

Figure 1: Image Taxonomy

5. Mixed: the black sheep images are the mixed images. These images have more panels on 1 image, and these panels (unlike in Graph type when 2-3 or more graphs on 1 image) differ in types. Most common are Graph-Gel, Graph-Model and Graph-Thing pairing. Images containing 3 or more different types is very rare.

Different from the authors of [3] we plan to have different classes for all of the pairs of Mixed category, as common sense suggests that when a person is looking for a graph of a 'fruit fly' he intends to find those images too which may contain some irrelevant information but having a graph on it. In this case the person wouldn't choose a query separately as Mixed images with the 'fruit fly' text based search as this kind of query could return absolutely irrelevant results for the user like a Gel-Model image too, while the user tries to find Graphs of 'fruit fly'. On Fig 1. you can see some examples for the biomedical image classes.

III. Feature extraction and preliminary image classification

We decided on extracting image features to use machine learning to see preliminary image classification. Based on experimenting with different features and using up the suggestions in [3] we decided on to use intensity histogram features and edge based projection features. At extraction of these features, first the RGB images are converted to grayscale by eliminating the hue and saturation information while retaining the luminance.

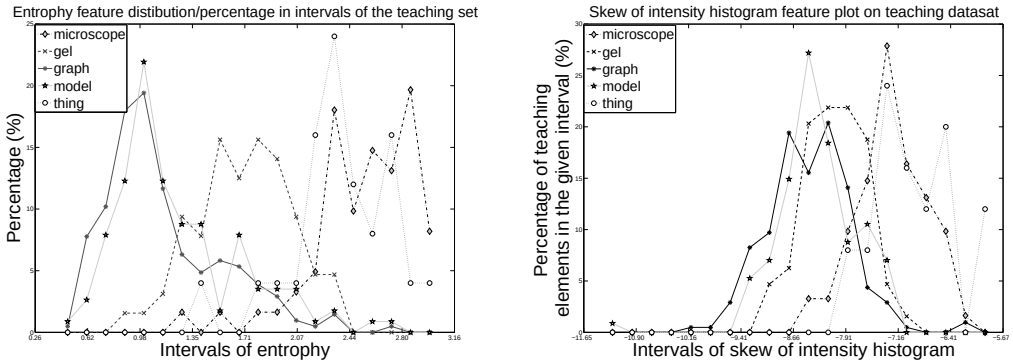
For the extraction of the histogram features the histogram containing the 0 to 256 grayscale value is created and normalized by dividing the sum of all the bins. The mean, variance, entropy, kurtosis and skew of this plot is determined. Fig. 2. shows what percentage of the teaching set is in certain intervals for some of these features.

For the edge based projection features first the grayscale image is filtered by a Sobel operator, than the image is projected onto x-axis (summed vertically) and onto y-axis (summed horizontally). The two plots are again normalized, and the entropy, mean, variance, kurtosis and skew calculated again. Altogether histogram and edge-based features result in 15 features.

To classify the images into the classes mentioned in Section II. we constructed a neural network with 10 neurons in hidden layer and 4 at output layer. Each of the output neurons gives a classification probability of the image being 1 of the 4 classes (Graph, Gel, Model, Thing) based on the feature vector. The desired output vector is a vector having 1 or 0s, e.g a Graph-Gel image has vector $[1, 1, 0, 0]$.

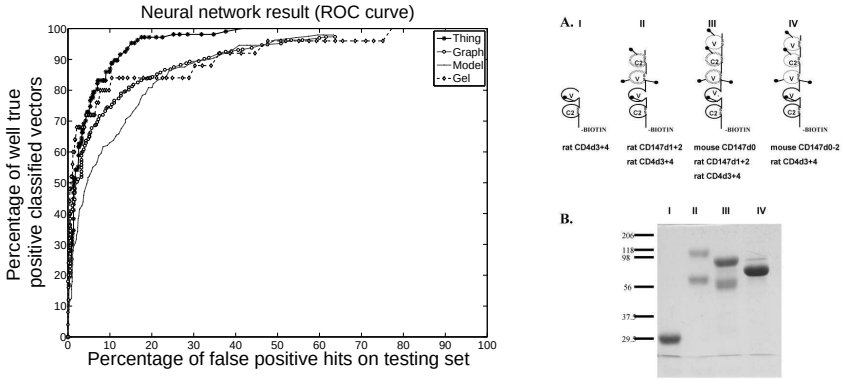
The network was taught by 93*4 teaching vectors, testing set contained 25*4 vectors for early stopping, and the remaining 886 vectors were used for validation on how good the taught network's classification ability is.

We observed a significant increase in the precision by the introduction of the edge based projection features, most probably cause these features quite well characterize the x and y axis line's in Graph images for example. Also we have to note that based on these features Thing images already quite well separate, most probably cause those have very distinct characteristics like dense filled square like



(a) Entropy of intensity histogram (b) Skew of intensity histogram

Figure 2: Some examples of feature distribution on the teaching sets



(a) ROC curve (b) Image with panels

Figure 3: ROC curve of NN results and an example for a difficult paneled Gel-Model image

panels (microscope panels for example), while the worst performance is on models, as model images have big variance from hand drawn images to UML box diagrams, or protein sequences involving lot of characters and connecting lines. The ROC curve represents the TP/FP percentage ratio of the neural network's results on the validation set. To create this ROC curve a threshold from 0 to 1 is applied on each output neuron's result, and if the output number is higher than the threshold the given input sample is classified as one of the class represented by that neuron. Note that in this way one input sample can become member of more classes, which is important in order to handle Mix type images. See Fig. 3(a).

IV. Panel separation

For any chance to handle Mixed Images an algorithm is needed to separate the sub-images, many times called panels on the biomedical image. Unfortunately in general biomedical databases images containing several panels are not always of compact panels (like microscope image panels of Fig. 1(c)) but sometimes a panel does not have a distinct square like boundary as in Fig. 3(b).

Panel separation starts by converting RGB images to grayscale, than finding a background threshold by Otsu's method and converting the greyscale image to binary image. On this binary image the bounding boxes of separated objects are calculated. The shortest distance between each object pairs' boundary boxes is calculated. As not the distance between object centroids are used but the distance measured between boundaries the standard clustering methods (K-means, subtractive clustering, ...) cannot be applied.

(a) Original image (b) It = 0 (c) It = 1 (d) It = 2 (e) It = 4

Figure 4: Iteration process for object clustering to find panels: small objects are moved toward the big grey ones, overlapping objects are brighter

A modified gravitational clustering algorithm of the one suggested in [4] is used as a heuristic way to find out which smaller bounding boxes are together nonetheless (like in most cases the numbers of x and y axis on graphs are separate objects, that need to be clustered in with the big bounding box of the graph to form the actual panel).

Gravitational forces are calculated on each object. There is a normal force on o_1 exerted by an other object o_2 on it: $F_g = G * m_{o_2}/(dist_{o_1-o_2})$, where G is a gravitational constant decaying with each iteration, and m_{o_2} is the mass (object size) of o_2 . There is a repulsive force on o_1 exerted by an other object o_2 if both objects are big: $F_r = -2 * G * m_{o_2}/(dist_{o_1-o_2})$, in this way the big bounding boxes which most probably belong to 2 different panels try to push each other away. And a third repulsive force on o_1 exerted by an other object o_2 is when o_2 is a well filled object, meaning that most of it's bounding box region is not filled with background pixels but useful information. This is to make the filled regions like microscope images push other objects away from them. $F_r2 = -o_{2coverage} * G * m_{o_2}/(dist_{o_1-o_2})$, where $o_{2coverage}$ is a value from 0 to 1 describing how big ratio of the bounding box is non-background pixel. All the forces applied on an objects are summed up making the resultant, than the object is moved in the direction of resultant divided by its own mass m_{o_1} (so heavier objects move slower): $s = F_sum/m_{o_1}$. All movement is calculated for all the objects, than G is decreased and the process is iterated again till there is movement or the maximum iteration limit is reached. One process is illustrated on Fig. 4.

V. Conclusions

We created a system giving preliminary classification of images. We proposed a new way for panel separations on images made up of small sub-figures.

There is still a need to improve the features on which the classification is based, like incorporating color (RGB) information too. An extensive research of the panelization is needed, to determine when the heuristic fails, and the best gravitational constants to give general good results. Future work is to combine image based classification results with the text mining based classification detailed in Strobert's work [1].

References

[1] H. Strobert, "An image retrieving system for biomedical literature using text-mining and ontologies," M.S. thesis, Technological University of Dresden.

[2] Y. Eui and T. S. Huang, "Image retrieval: current techniques, promising directions and open issues," *Journal of Visual Communication and Image Representation*, 10(1):39–62, Mar. 1999.

[3] B. Rafkind, M. Lee, S.-F. Chang, and H. Yu, "Exploring text and image features to classify images in bioscience literature," in *Proceedings of the BioNLP Workshop on Linking Natural Language Processing and Biology HLT-NAACL 06*, New York City, USA, June 73–81 2006.

[4] J. Gomez, D. Dasgupta, and O. Nasraoui, "A new gravitational clustering algorithm," in *Proceedings of the Third SIAM International Conference on Data Mining 2003*, pp. 83–94, 2003.