Hierarchical Image Classification in the Bioscience Literature

Daehyun Kim, PhD, Hong Yu, PhD
Department of Health Science, University of Wisconsin-Milwaukee, Milwaukee, WI

Abstract

Our previous work has shown that images appearing in bioscience articles can be classified into five types: Gel-Image, Image-of-Thing, Graph, Model, and Mix. For this paper, we explored and analyzed features strongly associated with each image type and developed a hierarchical image classification approach for classifying an image into one of the five types. First, we applied texture features to separate images into two groups: 1) a texture group comprising Gel Image, Image-of-Thing, and Mix, and 2) a non-texture group comprising Graph and Model. We then applied entropy, skewness, and uniformity for the first group, and edge difference, uniformity, and smoothness for the second group to classify images into specific types. Our results show that hierarchical image classification accurately divided images into the two groups during the initial classification and that the overall accuracy of the image classification was higher than that of our previous approach. In particular, the recall of hierarchical image classification was greatly improved due to the high accuracy of the initial classification.

Introduction

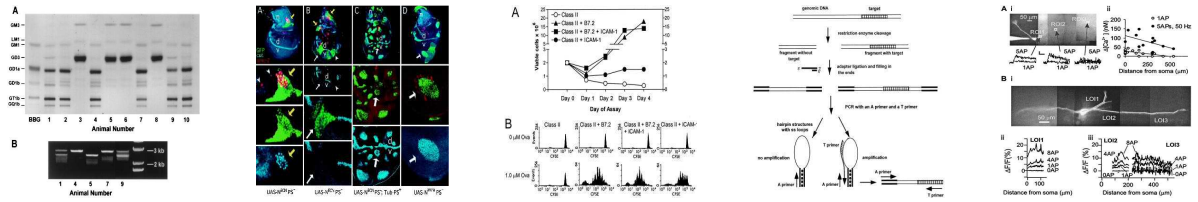
Images are abundant in bioscience articles and they generally provide important information to physicians and biologists. For example, Rafkind et al.[1] found an average of 5.2 images per biological article in the Proceedings of *National Academy of Science* (PNAS) and also found that 43% of the articles in the medical journal *The Lancet* contained biomedical images. Findings such as these suggest the need for a good image classification system for efficiently managing the variety of images used in the biomedical literature.

Previous work has relied on text associated with images for retrieval. For example, Yu and Lee developed natural language processing approaches to make a connection between abstract sentences and biological images in the same article[2]. Hearst et al. developed the BioText search engine, which enables biologists to browse article images by searching figure captions as well as titles and abstracts[3]. Similarly, GoldMiner[4], BioMed Search[6], and Yale Image Finder (YIF)[7] used image captions or text appearing in an image as part of their image search.

However, the significant limitation of text-based image search engines is that they ignore image content. For instance, when a physician searches for "lung cancer" images, he/she might prefer to find microscopic "lung cancer" images that can be used as an aid in diagnosis rather than graphs or charts showing "lung cancer statistics." This suggests the usefulness of image information for an image search engine.

Supporting this hypothesis, studies have found that image classification and retrieval systems benefit from combining image features with text features. For example, Shatkay et al. used image features derived directly from the image data for biomedical document categorization[8]. The SLIF (Subcellular Location Image Finder) system detects fluorescence microscopy images to be incorporated into text information retrieval[9-11]. Furthermore, we have devised a biomedical image classification system implementing both image and text features that has obtained better performance than when relying on text features alone[1].

For this paper, we hypothesized that each image type has specific attributes that can be hierarchically organized for the image classification task and that we could improve the image classification system if



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

Figure 1. Five image types

we could find "signature" features for each image type. In order to find signature image features, we explored a variety of image features and manually evaluated them. We then selected the top-ranked image features and used them for hierarchical image classification.

In the following, we first provide the definitions of the five image types we apply in this study before describing our methods, results, and conclusions.

Image taxonomy

We followed the previous image taxonomy[1] to classify images as one of five types: Gel-Image, Image-of-Thing, Graph, Model, and Mix. Gel-Image presents images such as Northern (for DNA), Southern (for RNA), and Western (for protein). Image-of-Thing includes images of cell, cell components, tissues, organs, or species. Graph includes charts, plots and other graphs drawn either by authors or a computer. Model depicts biological processes, experimental models, protein sequences, or higher protein structures. Mix refers to an image that incorporates two or more of the preceding types of images. Figure 1 shows examples of the five image types.

Image features

We examined five image features – skewnesss, entropy, uniformity, smoothness, and edge difference – and applied them as part of our image classification. The first four image features were extracted from an intensity histogram that was created by quantizing the gray-scale intensity values into a range from 0 to 255 and normalized by the total sum of the histogram before feature extraction to remove the dimension of input image[12]. Skewness is the third moment of the histogram, and it represents the histogram distribution. A histogram in which the tail is heavier on the right has a negative skewness; a histogram in which the tail is heavier on the left has a positive skewness. Entropy is a measure of the variability of the input image, and it is zero for a constant image. Uniformity is a measure of the consistency of the gray level, so it presents a maximum value for an image in which all gray levels are equal. Conversely, smoothness is a measure of gray-level contrast, so it presents a minimum value for an image in which all gray levels are equal.

The only feature we looked at in this paper that was not a histogram feature was edge difference, which is a measure of the resemblance between the input image and its edge image. Generally, since images in the non-texture group are presented by lines, curves, surfaces, etc., their edge images are very similar to original images. On the other hand, images in the texture group are completely different from their edge images. Thus, edge difference actually plays an important role for classifying images into texture and non-texture groups.

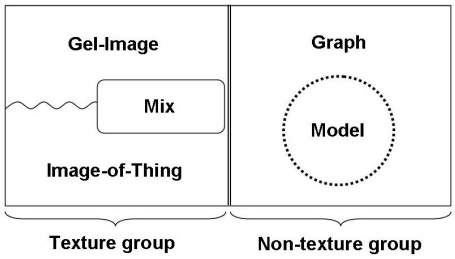


Figure 2. Hierarchical image organization

Image classification

As shown in Figure 1, Gel-Image and Image-of-Thing are presented using a variety of gray-level textures. In contrast, Graph and Model are presented as simple diagrams consisting of lines, curves, surfaces, etc., but their color remains constant. That is, Graph and Model are recognized as having several homogeneous regions, while Gel-Image and Image-of-Thing are noted for their texture. Therefore, we can group Gel-Image and Image-of-Thing as a texture group, and Graph and Model as a non-texture group. Since Mix comprises images belonging to two or more image types, we classify Mix as part of the texture group.

We can hierarchically subclassify images within their respective general classifications based on their image features as shown in Figure 2. In the texture group, the histogram of Image-of-Thing presents a very different distribution than that of Gel-Image and Mix, and Mix has more homogeneous regions than Gel-Image does. In the non-texture group, Graph usually uses a greater variety of diagrams, so image features of Graph have a larger variance than those of Model. If we can accurately divide images into two groups during the first image classification, images in each group will be classified accurately as well.

In the following subsections, we describe how to use image features to divide images into the two groups and to classify an image into one of the five image types.

1) 1st image classification

The first step of image classification is to accurately separate texture images from non-texture images. For this, we used three image features: skewness, entropy, and edge difference.

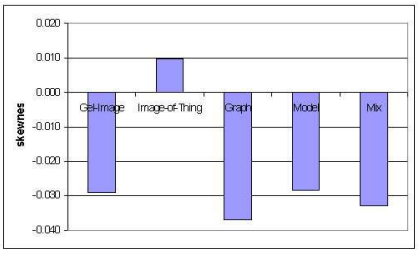
Skewness and entropy were used for extracting texture information. Figure 3 shows average skewness and entropy for the image types, respectively. As the figure shows, the skewness of Image-of-Thing is positive, and the skewness of other image types is negative, while the entropy of the texture group is much higher than that of the non-texture group.

Edge difference is used for extracting non-texture information. Since images in the non-texture group look like their edge images, edge differences for that group are far lower than those of the texture group.

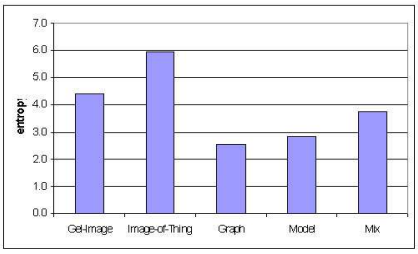
Figure 4 shows the pseudo-code for the first image classification. Since skewness is a unique feature of Image-of-Thing, it has the highest priority. Edge difference is also a unique feature for the non-texture group, so that it is used as a secondary condition. In the last step, images in which entropies are higher than the threshold are classified into the texture group.

2) 2nd image classification in the texture group

Once the texture group is separated from the non-texture group, Gel-Image, Image-of-Thing, and Mix must be distinguished from each other.



(a) Skewness



(b) Entropy

Figure 3. Skewness and entropy

```
If skewness is positive,
    then classify it into texture group
else if edge difference is much lower,
    then classify it into non-texture
    group,
else if entropy is much higher,
    then classify it into texture group,
else classify it into non-texture group
```

Figure 4. Pseudo-code for first image classification

First, the skewness of Image-of-Thing is usually positive and higher than of the skewness of other images. Next, Gel-Image has a higher edge difference and lower uniformity than Mix.

Figure 5 shows the pseudo-code for the second image classification in the texture group. In this step, skewness also has the highest priority due to its uniqueness. Then, Gel-Image is classified using uniformity and edge difference.

3) 2nd image classification in the non-texture group

Mix images are often misclassified into the non-texture group due to ambiguity in the image features they contain, which means that these misclassified images must be separated from the non-texture group before image classification. The edge difference of Mix is generally higher than that of Graph and Model, which enables misclassified Mix images to be distinguished from other images in the non-texture group.

On the other hand, there are an insufficient number of

```
If skewness is positive or higher than threshold,
    then classify it into Image-of-Thing
else if uniformity is lower and edge difference is
higher than threshold,
    then classify it into Gel-Image,
else classify it into Mix
```

Figure 5. Pseudo-code for 2nd image classification in the texture group

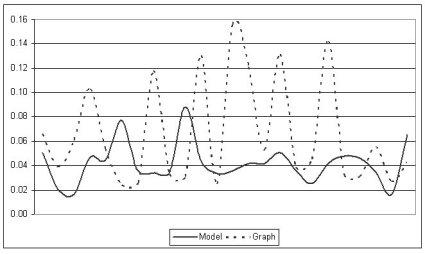


Figure 6. Smoothness of Graph and Model

```
If edge difference is higher,
    then classify it into Mix
else if smoothness and skewness are lower
than threshold,
    then classify it into Model,
else classify it into Graph
```

Figure 7. Pseudo-code for second image classification in the non-texture group

features to discriminate between Graph and Model. Figure 3(a) and Figure 6 show, however, that the skewness is bigger and the smoothness less variable for Model than they are for Graph, so the use of these two characteristics can be used to limit the range of Model. By limiting the range of Model, more instances of Graph can be accurately classified and although less instances of Model can be accurately classified.

Figure 7 shows the pseudo-code for the second classification of the non-texture group. In this step, edge difference has the highest priority for distinguishing Mix from the others. Then, Model is classified using smoothness and skewness.

Experimental results

We randomly selected a subset of 450 images from the image poo l[1]: Gel-Image (64), Image-of-Thing (68), Graph (196), Model (24), and Mix (98). These images were split up such that half could be used for training and the other half for testing. The training set was then used to determine the reference and threshold values, and the testing set was used to evaluate the classification system.

We used recall, precision, and F-score as the evaluation metrics for image classification. Recall is the number of true positives divided by the total number of elements that actually belong to the type. Precision is the number of true positives divided by the total number of elements labeled as belonging to the type. F-score is the harmonic mean of recall and precision, and it represents the accuracy of the classification.

In addition, we tried to change the gray level of images. Although 256 gray levels often provide greater information, they also create ambiguity by implying that Graph and Model have texture information. Thus, we adjusted the gray level of image to remove the ambiguity between Graph and Model.

Table 1 shows the number of correctly classified images in the first image classification according to gray levels, and as it shows, the accuracy is changed according to the gray level. Eventually, we obtained the best accuracy at the gray levels of 128, 64, and 32 gray levels rather than at 256 gray levels.

Table 2 shows the number of correctly classified images in the second image classification according to gray level. The image classification presents the highest accuracy (57.3%) at the 32 gray levels, and this value is about 3.8% higher than the accuracy of 53.5% obtained by Rafkind et al[1].

Table 3 shows the confusion matrix for the image classification at the 32 gray levels, and Table 4 shows recall, precision, and F-score based on Table 3 for evaluating its performance. The number in the round bracket shows the performance of Rafkind et al[1].

The results show that the precision of Gel-Image decreased, while its recall increased. Conversely, the precision of Mix increased, but its recall decreased. Table 3 shows that 20 Mix images were misclassified as Gel-Image. These misclassifications lowered the performance of Gel-Image in terms of precision and

gray level	Texture Group			Non-texture Group		Prec.
	Gel	Thing	Mix	Graph	Model	
256	31	32	33	82	7	82.2%
128	**30**	**32**	**36**	**84**	**8**	**84.4%**
64	**30**	**32**	**36**	**83**	**9**	**84.4%**
32	**29**	**32**	**37**	**83**	**9**	**84.4%**
16	31	32	40	69	7	79.6%

Table 1. Number of correctly classified images for 1st image classification

gray level	Texture Group			Non-texture Group		Prec.
	Gel	Thing	Mix	Graph	Model	
256	23	30	12	52	2	52. 9%
128	26	30	15	53	3	56.4%
64	26	30	14	53	4	56.4%
32	**25**	**30**	**17**	**54**	**3**	**57.3%**
16	21	31	18	49	3	54.2%

Table 2. Number of correctly classified images for 2nd image classification

Actual Types	Predicted Types				
	Gel	Thing	Mix	Graph	Model
Gel	25	2	3	1	1
Thing	3	30	0	0	1
Mix	20	1	17	9	2
Diagram	2	2	17	54	23
Model	2	0	1	6	3

Table 3. Confusion matrix for 2nd image classification at 32 gray levels

	TP	FP	FN	Precision	Recall	F-score
Gel	25	27	7	48.1% (52.9%)	78.1% (69.2%)	0.595 (0.600)
Thing	30	5	4	85.7% (85.7%)	88.2% (88.6%)	0.870 (0.800)
Mix	17	21	32	44.74% (42.9%)	34.7% (75.0%)	0.391 (0.300)
Graph	54	16	44	77.1% (52.7)	55.1% (23.1%)	0.643 (0.661)
Model	3	27	9	10.0% (33.3%)	25.0% (12.0%)	0.143 (0.176)

Table 4. Performances of image classification

Mix in terms of recall. In other words, Gel-Image and Mix are inversely correlated with each other, and system performance is too weak to accurately classify Gel-Image and Mix.

On the other hand, the precision and recall of Graph greatly increased. This is due to 92 of 110 images in the non-texture group being correctly classified during the first image classification and 54 Graph images being stably classified during the second image classification, which also accounts for the better recall of Model. However, as mentioned in the previous section, there is no way to avoid the loss of precision in distinguishing Model since the accuracy of the non-texture group depends on the results of the first image classification.

Conclusion

In this paper, we reported on the use of hierarchical image classification for bioscience literature. We manually evaluated image features and explored the features for hierarchical image classification. Our results show that our methods outperform our previous system, although in future work we will need to integrate text features into our model to optimize system performance.

Future Work

In future work we will continue to identify and explore the use of features that could benefit image classification. The image types Graph and Model, for instance, share many similar components (e.g., points, lines, and surfaces) that make it difficult to discriminate them from one another; the use of additional features might aid in distinguishing these types. Mix incorporates multiple image types, which creates problems for classification that might require segmentation and the extraction of additional features to solve. In addition, we will integrate image features with text features to explore how this might improve image classification.

Acknowledgement

We acknowledge the support of 1R21RR024933-01A1 to Hong Yu. Any opinions, findings, or recommendations are those of the authors and do not necessarily reflect the views of the NIH. We thank Lamont Antieau for editing the article.

Reference

1. Rafkind B, Lee M, Chang S, Yu H. Exploring text and image features to classify images in bioscience literature. proceedings of the BioNLP workshop on Linking Natural Language Processing and Biology HLT-NAACL 06, New York City, USA, pp.73-80, June 2006.
2. H. Yu, M. Lee. Accessing Bioscience Images from Abstract Sentences. Bioinformatics, Vol. 22, No. 14, pp. 547-556, 2006.
3. M. A. Hearst, A. Divoli, H. Guturu, A. Ksikes, P. Nakov, M. A. Wooldridge, J. Ye, BioText Search Engine: beyond abstract search, Bioinformatics, Vol. 23, No.16, pp. 2196-2197, 2007.
4. C. E. Kahn, C. Thao. GoldMiner: A Radiology Image Search Engine. American Journal of Roentgenology, Vol. 188, pp. 1475-1478, 2007.
5. Howarth P, Ruger S, Robust texture features for still-image retrieval. IEE Proceedings-Vision, Image and Signal Processing, pp.868- 874, December 2005.
6. M. A. Hearst, A. Divoli, H. Guturu, A. Ksikes, P. Nakov, M. A. Wooldridge, J. Ye. BioText Search Engine: beyond abstract search. Bioinformatics, Vol. 23, pp. 2196-2197, 2007.
7. S. Xu, J. McCusker, M. Krauthammer. Yale Image Finder (YIF): a new search engine for retrieving biomedical images. Bioinformatics, Vol. 24, No.17, pp. 1968-1970, 2008.
8. H. Shatkay, N. Chen, D. Blostein. Integrating Image Data into Biomedical Text Categorization. Bioinformatics, Vol. 22, No. 14, pp.446-453, 2006.
9. R. F. Murphy, M. Velliste, J. Yao, G. Porreca. Searching Online Journals for Fluorescence Microscope Images depicting Protein Subcellular Location Patterns. IEEE International Symposium on Bio-Informatics and Biomedical Engineering, pp. 119-128, 2001.
10. R. F. Murphy, Z. Kou, J. Hua, M. Joffe, W. Cohen. Extracting and structuring subcellular location information from on-line journal articles: the subcellular location image finder. IASTED International Conference on Knowledge Sharing and Collaborative Engineering (KSCE2004), pp. 109-114, St. Thomas, US Virgin Islands, 2004.
11. Y. Qian, R. F. Murphy, Improved recognition of figures containing fluorescence microscope images in online journal articles using graphical models, Bioinformatics, Vol. 23, No.4, pp. 569-576, 2008.
12. Gonzalez R, Woods R. Digital image processing. Prentice Hall, 2nd edition, Upper Saddle River, New Jersey, 2002.