UTILIZING IMAGE-BASED FEATURES IN BIOMEDICAL DOCUMENT CLASSIFICATION

Kaidi Ma[1], Hogyeong Jeong[1], Rohith MV[2], Gowri Somanath[2], Ryan Tarpine[3], Kyle Schutter[4], Dorothea Blostein[5], Sorin Istrail[4], Chandra Kambhamettu[2] and Hagit Shatkay[1,5,6]

[1]Computational Biomedicine and [2]Video/Image Modeling and Synthesis Lab, CS Dept., University of Delaware, Newark, DE
[3]Google, Cambridge, MA
[4]Center for Computational Molecular Biology, CS Dept., Brown University, Providence, RI
[5]School of Computing, Queen's University Kingston, Ontario, CA
[6]Center for Bioinformatics & Computational Biology, University of Delaware, Newark, DE

ABSTRACT

Images form a rich information source, which remains under-utilized in biomedical document classification. We present here work that uses both image- and text-based features in order to identify articles of interest, in this case, pertaining to cis-regulatory modules in the context of gene-networks. Extending on our new idea, which we have recently introduced, of using *OCR-based features* to identify DNA contents in images, we combine image and text based classifiers to categorize documents as relevant or irrelevant to cis-regulatory modules. Using a set of hundreds of articles, marked by experts as relevant or irrelevant to cis-regulatory modules, we train/test image and text based classifiers, as well as classifiers integrating both. Our results indicate that the latter show the best performance with *Recall, F-measure* and *Utility* measures all above 0.9, demonstrating the significance of incorporating image data, and specifically OCR-based features, into the document categorization process. Moreover, the use of character distribution properties to represent images is directly relevant to other biomedical images containing text (*e.g.* RNA, proteins). Diagrams and other images containing text are also prevalent outside the biomedical domain, hence the work stands to be applicable and beneficial in other application areas.

Index Terms—image-based features, OCR, document classification, document-representation, bioinformatics

1. INTRODUCTION

A fundamental task in biomedical research is the identification of documents relevant to a specific study area. Given the sheer number of biomedical documents published annually, automating this task is becoming ever more important. The vast majority of methods used for identifying relevant documents rely on text categorization, so far with limited success [13,24,27]. Notably, while images provide significant cues for deciding relevance [25], image-information has gone largely untapped by automatic document classifiers.

A few relatively recent efforts started to examine the value of using images within articles for several biomedical tasks [8,14,25]. Murphy's group [6,16,20] was among the first who used images within articles to study protein subcellular localization, employing standard image-based features such as gray-level histograms and edge-direction statistics [9]. While Murphy's work focuses on protein-subcellular localization, Rafkind *et al.* [21] explored more generally the retrieval of biomedical images and text. In the context of identifying regulatory regions from the literature, Aerts *et al.* [2] examined the extraction of complete DNA sequences directly from *text* (not images) to aid in the cis-regulatory annotation process. In contrast, we focus on the classification task of identifying articles pertaining to cis-regulatory elements, rather than DNA annotation.

Other work seeks to take advantage of text that is associated with the images. Regev *et al.* [22] explored using text from figure *captions* as well as text referencing the image from within the article. Xu *et al.* [30] and Rodriguez *et al.* [23] proposed to use complete words extracted from images to help retrieval of documents. A similar approach was proposed by Gunjan *et al.* [10] in the broader context of classifying images based on words appearing in them. While words within images may provide an additional source for indexing, correctly identifying *complete* words through optical character recognition (OCR) is prone to errors, because individual characters in images are often mis-recognized, introducing noise into the word extraction process.

The classification task addressed here originates from the CYRENE project [7], which aims to obtain highly reliable information about cis-regulatory genomics and gene regulatory networks. We develop document classifiers to identify cis-regulatory related articles, by integrating image and text features, and show significant improvement in performance. Our approach is novel and different as we use OCR to extract individual characters – as opposed to complete words – from images [26]. Properties of the character-distribution within the image are used for representing and classifying images. This approach is more robust in the face of OCR noise, because misidentifying some characters does not have much impact on the whole character distribution obtained.

The ability to characterize images through a character-based distribution provides a straightforward and general way to identify images that depict DNA in which the letters *A, C, G, T* are over-represented (as well as other forms of text-patterns, *e.g.* RNA, protein sequences, and other) within documents. The work presented here builds on the idea that such images can be identified, and shows how DNA-rich images form the basis for improved categorization of documents discussing cis-regulatory regions. In Sections 2 and 3 we describe the representation of documents and the classifiers used; we examine both image and text based classifiers as well as the combination of the two. Our experiments

and results, presented in Section 4, demonstrate the utility of the image-based features for identifying relevant documents.

2. DOCUMENT REPRESENTATION

As a starting point, the CYRENE team has identified 271 articles that contain high-quality validated information about cis-regulatory modules, from prominent journals relevant to the field, such as Molecular and Cellular Biology. Of these, 264 had a full-text PDF file available from which we could obtain images. These *264* articles form the *positive set of relevant articles* for training and testing our classifiers. The CYRENE team also initially surveyed 78 articles that proved irrelevant, which were kept as a *negative set.* To overcome the scarcity of irrelevant articles in the overall dataset, additional 143 negative examples from the Journal of Molecular Cellular Biology were selected, by scanning through the same volumes in which the relevant articles were found. A total of 220 negative articles with an accessible PDF file form the negative set for training and testing. Choosing the negative set this way, helps ensure that the general discourse and style of writing remains consistent across the relevant and the irrelevant articles. That is, there is no shift in time and overall areas of current interest between the relevant and the irrelevant corpus. Such a shift may over-simplify the learning task of separating between relevant and irrelevant documents [5].

Thus, the final dataset for training and testing classifiers through a 5-fold cross-validation scheme comprises 264 relevant articles in the positive set and 220 irrelevant articles in the negative set. Further details on the CYRENE Project can be found elsewhere [7,12]. We next describe the representation methods used for articles and for images.

2.1. Image-Based Representation of Articles

Figures in biomedical publications often consist of multiple sub-figures or panels (*e.g.* Fig.1 has 4 panels,) where each panel is an individual image [16,31]. Thus, before using figures to represent documents, we separate figures into individual panels using a tool we have specifically developed for this purpose based on the Xerox Rossinante utility [29].

As cis-regulatory modules are regions on the DNA, image panels showing DNA segments are typically over-represented in relevant articles discussing such modules. Therefore we hypothesize that automatically identifying images displaying DNA fragments, and finding articles that have an over-abundance of such images, is likely to be useful in identifying the relevant articles. Our experiments indeed support this hypothesis. We refer to an image panel that shows DNA regions as *DNA-rich image panel*. In our preliminary experiments [26], we have trained a decision-tree classifier to identify DNA-rich panels with an average precision of about .93 and average recall of about .90. The image-classifier relies on a novel OCR-based representation of the image panels, as described in Sec. 2.2 below.

Given an article d, we create an image-based representation for it by tagging each image panel within it as *DNA-rich* or *non-DNA-rich*. We count the number of *DNA-rich* panels in the article, denoted A_d, and the number of panels that are *non-DNA-rich*, denoted N_d. An article d is then represented as a 2-dimension vector $<A_d/(N_d+A_d),\ N_d/(N_d+A_d)>$; that is, the article is represented based on the relative frequency of its *DNA-rich* panels and its relative frequency of *non-DNA-rich* panels. Using relative frequency, rather than absolute counts, better generalizes to documents of different lengths and varying numbers of images.

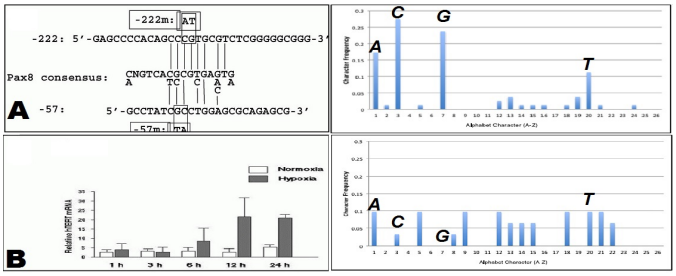

Fig. 1. Two images containing characters are shown on the left. Image A [19] with DNA sequences, and image B [17] with other text. The character histogram for image A (top right) shows four distinct peaks, at *A, C, G* and *T*, while the one for image B (bottom right) does not.

It is important to note that the average frequency of DNA-rich panels in a *relevant document* is significantly higher than that observed in an *irrelevant* one (12.7% vs. 1.5%, $p<0.001$). This substantial difference in abundance of *DNA-rich* panels between cis-regulatory-relevant and irrelevant papers strongly supports the idea of using the frequency of *DNA-rich* image panels as an informative component.

2.2. OCR-based Representation of Images

To represent images in a way that reveals their DNA content, we use an OCR-based representation of image panels, capturing distributional properties of characters in images. To the best of our knowledge, no other group has used OCR in this way or utilized this idea for image representation. We employ the ABBYY Finereader tool [1] to extract all characters from each image panel, and count the number of times each character (*A-Z, 0-9, Other*) occurs. Each panel is then represented as a 37-dimensional feature vector $<w_1,...,w_{37}>$, where w_i denotes the frequency of the i^{th} character in the panel. An alternative representation can use fewer characters, *e.g.* a 5-dimension vector, maintaining the frequencies of the DNA characters *A, C, G,* and *T* while collapsing all characters into "*Other*". Fig. 1 shows DNA-rich panel (A) with a clearly distinct distribution of characters from that of the non-DNA-rich panel (B). Specifically, four distinct peaks at *A,C,G*, and *T* are associated with the DNA-rich panel, which are not present for the non-DNA-rich panel. Such representation is robust in the face of OCR errors; noise in reading some of the characters has only a small impact on differences between character distributions.

2.3. Text-Based Representation of Articles

The text-based representation follows a standard bag-of-words model that is commonly used in biomedical information retrieval and classification, as we have used in other contexts [3]. Titles and abstracts of all 484 articles are tokenized to build a dictionary of terms consisting of single

words (unigrams) and consecutive words (bigrams). Standard stop-words are removed from the set, as well as rare and frequent terms. The remaining set of terms is further reduced by selecting only terms whose probability to occur in the positive set is statistically significantly different from their probability to occur in the negative set (as estimated by the Z-score test [3]); we call such terms *distinguishing terms*. This process gives rise to a set of 1030 terms. Each document d is then represented as an n-dimensional vector $<d_{t1},...,d_{tn}>$, where $n=1030$ is the number of terms and d_{ti} is an indicator variable whose value is 1 if the term t_i occurs in the document d, and 0 otherwise.

3. CLASSIFIERS

We trained and tested several classifiers including naïve Bayes, decision trees, and random forests, as well as ensemble methods. Bayesian network models were also built, but are not discussed here as results were similar to those from naïve Bayes. All training and testing employed stratified 5-fold cross validation. To ensure stability and statistical significance of the results, we executed five distinct sets of 5-fold cross validation runs, each using a different 5-way split of the data (for a total of 25 training/test runs per classifier). We briefly describe each classification method below.

3.1. Naïve Bayes

The naïve Bayes classifier [15] is based on the assumption that each feature F_i is conditionally independent of every other feature F_j $(j \neq i)$ given the class (in our case, the classes are *relevant* vs. *irrelevant* articles). To determine if a given article d is relevant, we compare the posterior probability $P(d|relevant)$ to the posterior probability $P(d|irrelevant)$ and classify the article as *relevant* if $P(d|relevant)$ is greater, and as *irrelevant* otherwise. The conditional independence assumption allows for a simple calculation of these probabilities using Bayes' rule. Due to its simplicity and speed, this classifier is useful for high-dimensional data, and we therefore use it for text-classification.

3.2. Decision Trees and Random Forests

Decision trees are based on viewing classification as a traversal of a tree-like structure, in which each internal node corresponds to a feature, each branch is labeled by a feature value, and the leaves correspond to classes. For classifying our documents, the leaves correspond to the two classes, *relevant* vs. *irrelevant*. In contrast to the naïve Bayes, decisions trees are more suited for low-dimensional data, and we employ them for classifying the documents under the image-based representation.

We also performed experiments with random forests, which are ensembles of decision trees that use only a subset of the features – selected at random – for each node in a decision tree [4]. The classifier's output is a plurality vote based on the individual decision of each tree. We used a forest size of 2000 decision trees, and a feature subset size of 90. Notably, while the random forest classifier chooses a small subset of features, the selection is done independently at each node in a tree. Thus, features that are more informative are more likely to appear in the tree. Our experiments show that the image-based feature was utilized by every tree in the forest.

3.3. Combination Classifier

We also conducted experiments combining image-based and text-based representations. A simple way to combine image and text data is to directly represent each document as a vector including both features. We apply both naïve Bayes and random forest classifiers to data represented this way, and also use the union of the two as a combination classifier. Another way to integrate classifiers, is to combine results obtained from classifiers that are trained separately, as an ensemble or a union. As the simplest combination, we consider the union of the image-based decision tree and the text-based naïve Bayes: if any one of the individual classifiers identifies a document as *relevant*, the union classifier tags it as *relevant*. In contrast, both classifiers must tag a document as *irrelevant* for the union classifier to deem it *irrelevant*. We also built an ensemble classifier combining the decision tree image-based classifier, naïve Bayes text classifier, and random forest text classifier, using a majority vote among their outputs to assign the label; if two or more of the classifiers tag a document as *relevant*, the document is classified as *relevant*; otherwise as *irrelevant*.

4. EXPERIMENTS AND RESULTS

We conducted two sets of experiments. The first set separately examines the performance of a decision tree classifier (J48) trained on image-based representation, and of a naïve Bayes classifier (NB) trained on text-based representation. The second set examines classifiers combining image and text-based features, as described in Sec. 3.3. We used the Weka software suite [28] to train/test our classifiers. The seven classifiers used are listed in the leftmost column of Table 1.

As mentioned before, we employ five *complete sets* of *stratified 5-fold cross validation* (a total of 25 runs) for training and testing each classifier. Notably, the feature selection step for each of the text classifiers uses only the training part of the data and is repeated as part of each cross-validation run. This ensures that the test set is never used for feature selection and is completely excluded from the representation and the classification process.

4.1. Evaluation Measures

We use the standard measures that are widely employed for classification evaluation: *Precision, Recall, F-measure* (*F*), and *overall accuracy* (*Acc*) [18, 24]. In biomedical research, *Recall* is often more important than precision because while an irrelevant article may impose some extra burden on the researcher, missing published information may compromise the integrity of the CYRENE project. Hence, we include the *utility* measure introduced by TREC Genomics, which biases the evaluation in favor of high recall [11]. We use two versions of this measure: One as originally introduced in TREC, *Utility-20*, which gives 20 times the weight to true positives (where a *positive* instance is a CYRENE-relevant

Table 1. Results obtained from 7 classifiers: 1) Decision tree (J48) on image data (img); 2) Naive Bayes classifier (NB) on img; 3) NB on text data (txt); 4) NB on txt and img; 5) Random forests (RF) on txt and img; 6) Union of classifiers 1 and 3; 7) Union of classifiers 4 and 5; 8) Ensemble of classifiers 1, 3, and RF on txt. Type of data is listed in brackets. The performance measures are averaged over 25 cross-validation runs. The top two performing classifiers in each column are shown in boldface. Standard deviations are shown in parentheses.

CLASSIFIER	RECALL	PRECISION	F-MEASURE	ACCURACY	UTIL-10	UTIL-20
J48 [img]	0.890 (.013)	0.868 (.005)	0.878 (.009)	0.865 (.009)	0.876 (.013)	0.883 (.013)
NB [img]	0.614 (.009)	**0.916 (.005)**	0.735 (.006)	0.760 (.005)	0.608 (.009)	0.611 (.009)
NB [txt]	0.850 (.007)	0.893 (.001)	0.870 (.004)	0.862 (.003)	0.839 (.007)	0.844 (.007)
NB [img txt]	0.868 (.011)	**0.896** (.005)	0.880 (.006)	0.872 (.006)	0.858 (.010)	0.863 (.010)
RF [img txt]	0.946 (.004)	0.870 (.001)	0.906 (.002)	0.893 (.002)	0.932 (.004)	0.939 (.004)
Union (J48 [img] + NB [txt])	**0.976** (.004)	0.855 (.005)	0.911 (.004)	0.896 (.005)	**0.959** (.004)	**0.968** (.004)
Union (NB [img txt] + RF [img txt])	**0.976** (.003)	0.865 (.002)	**0.917** (.002)	**0.903** (.003)	**0.961** (.003)	**0.968** (.003)
Ensemble (J48 [img]+NB [txt] + RF[txt])	0.937 (.006)	0.893 (.003)	**0.914** (.002)	**0.904** (.003)	0.926 (.006)	0.932 (.006)

article). The other, *Utility-10*, assigns only 10 times the weight to true positives. The respective formulae for *Utility-10* and *Utility-20* are:

$$\textit{Utility-10} = \frac{10 \times TP - FP}{10 \times Pos}, \quad \textit{Utility-20} = \frac{20 \times TP - FP}{20 \times Pos},$$

where *TP* is the number of true positives, *FP* is the number of false positives, and *Pos* is the number of articles that are relevant (*i.e. TP + FN*). To evaluate the statistical significance of performance differences among the different classifiers, we use the two-sample paired t-test [32].

4.2. Results

Table 1 summarizes the results obtained from each of the seven classifiers, averaged over five independent runs of stratified 5-fold cross validation. Fig. 2 depicts results obtained from six of these classifiers along the *recall*, *precision*, *F-measure*, and *Utility-20* measures.

The image-based J48 classifier demonstrates significantly higher recall and utility measures than the naïve Bayes text classifier. The differences are also statistically significant ($p<0.05$). The precision of the text-based classifier is higher (and this difference is also highly statistically significant, $p<0.005$). By introducing image features into the naïve Bayes text classifier (NB [txt img] in Table 1), the performance improves according to all measures; the improvement in recall, F-measure, accuracy, and utility are statistically significant (with $p<0.05$ or better).

Both of the union classifiers (J48[img] + NB[txt], as well as NB[txt img] + RF[txt img]) demonstrate the highest recall, along with highest utility values. They also perform among the top in terms of overall accuracy and F-measure, but have relatively low precision (as expected, by the definition of the union). The ensemble classifier that comprises an image-based classifier and two text-based classifiers, retains a relatively high performance across all measures. While it is not the top performer according to any of the measures, it demonstrates a good balance between precision and recall.

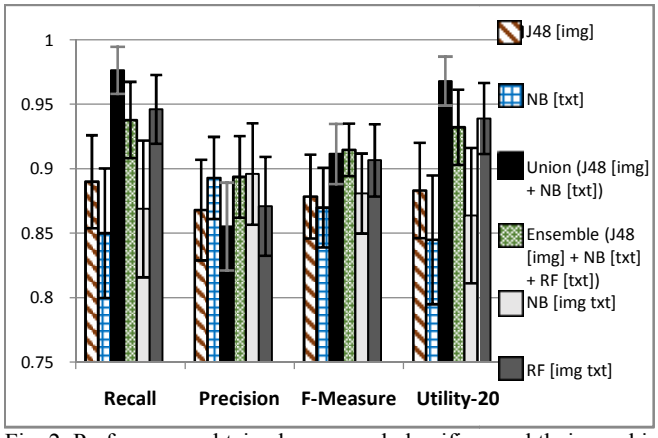

Fig. 2. Performance obtained on several classifiers and their combinations: 1) J48 on image data; 2) NB on text data; 3) Union of classifiers 1 and 2; 4) Ensemble of classifiers and RF on text data; 5) NB on text and image data; 6) RF on text and image data. *Recall, precision, F-measure,* and *utility-20* are shown from left to right. Error bars indicate ± 1 standard deviation.

5. DISCUSSION AND CONCLUSION

Two different ways of incorporating images into document classification were shown in our work: By combining image-based and text-based classifiers, and by using images as a source of additional features that directly augment the document text features within a single multi-modal vector representation. Both methods demonstrate improvement with respect to image or text-only classifiers, along several performance measures. Notably, our results, which show high recall levels along with well over 80% precision are particularly relevant and useful. While most combined classifiers sacrifice precision for high recall, we demonstrate that the vote-based ensemble classifier shows a relatively high level of precision while retaining a high recall.

The work presented here uses the pre-classification of image panels as DNA-rich vs. DNA-poor as the main basis for image-based document categorization. This particular classification scheme for image panels, which is based on genomic sequence contents that can be detected via OCR, is a novel promising direction within the biomedical domain. Biomedical images contain much character-sequence data (RNA, DNA, and proteins). Such image panels are readily distinguishable from one another based on character distribution, and as such lend themselves to the character-based representation and classification, similar to the images used in this work. In turn, the identified image classes can form the basis for document representation and classification.

Future work includes the application of our methods to other types of images and domains, as well as a more extensive exploration of image-text integration methods.

ACKNOWLEDGEMENT

This work was partially supported by HS's and CK's NIH Award 1R56LM01135401A1, HS's NSERC Discovery Award 298292-2009, and SI's NSF grant 0645955.

REFERENCES

[1] ABBYY Finereader for OCR, http://finereader.abbyy.com/, 2013.

[2] S. Aerts, M. Haeussler, *et al.*, "Text-Mining Assisted Regulatory Annotation," *Genome Biology*, vol. 9, no. 2, pp.2–31, 2008.

[3] S. Brady, H. Shatkay, "EpiLoc: a (working) text-based system for predicting protein subcellular location," *Proc. of the Pacific Symposium on Biocomputing (PSB08)*, pp. 604–615, 2008.

[4] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5-32, 2001.

[5] A.M. Cohen, R.T. Bhupatiraju, and W.R. Hersh, "Feature Generation, Feature Selection, Classifiers, and Conceptual Drift for Biomedical Document Triage," in *Proc. of the Text Retrieval Conf. (TREC)*, 2004.

[6] Z. Kou, W. Cohen, and R. Murphy, "Extracting Information from Text and Images for Location Proteomics," *Proc. of the 3rd ACM SIGKDD Workshop on Data Mining in Bioinformatics (BIOKDD'03)*, pp. 2–9, 2003

[7] CYRENE,
http://www.brown.edu/Research/Istrail_Lab/cyrene.php.

[8] D.Demner-Fushman, S. Antani, *et al.*, "Annotation and Retrieval of Clinically Relevant Images," *International Journal of Medical Informatics: Special Issue on Mining of Clinical and Biomedical Text and Data*, vol. 78, no. 12, pp. e59–67, 2009.

[9] R. Gonzalez, R. Woods, *Digital Image Processing*, Prentice-Hall, 2002.

[10] J. Gunjan, B. Prateek, *et al.*, "ScanDroid: Automatic Classification of Document Images on Android Mobile Devices," *International Journal of Managment, IT and Engineering*, vol. 3, no. 6, pp. 528–37, 2013.

[11] W. Hersh, A. Cohen, *et al.*, "TREC 2005 Genomics Track Overview," in *Proc. of TREC Notebook*, pp.14–25, 2005.

[12] S. Istrail, R. Tarpine, *et al.*, "Practical Computational Methods for Regulatory Genomics: A cisGRN-Lexicon and cisGRN-Browser for Gene Regulatory Networks," *Methods in Molecular Biology*, vol. 674, pp. 369–99, 2010.

[13] M. Krallinger, M. Vazquez, *et al.*, "The Protein-Protein Interaction Tasks of BioCreative III: Classification/Ranking of Aticles and Linking Bio-ontology Concepts to Full Text," *BMC Bioinformatics*, vol. 12, no. 8, pp. 3-4, 2011.

[14] T. Kuhn, T. Luong, and M. Krauthammer, "Finding and Accessing Diagrams in Biomedical Publications," *AMIA Annu Symp Proc*, pp. 468–74, 2012.

[15] T. Mitchell, *Machine Learning,* McGraw Hill, 1997.

[16] R. Murphy, M. Velliste, *et al.*, "Searching Online Journals for Fluorescence Microscope Images Depicting Protein Subcellular Location Pattern," *Proc. of the 2nd IEEE Int. Symp. on Bio-Informatics and Biomedical Engineering (BIBE01)*, pp. 119–28, 2001.

[17] H. Nishi, T. Nakada, *et al.*, "Hypoxia-Inducible Factor 1 Mediates Upregulation of Telomerase (hTERT)," *Mol. Cell. Biol.*, vol. 24, no. 13, pp. 6076–6083, 2004.

[18] D. Powers, "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation," *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37–63, 2007.

[19] C. Puppin, I. Presta, *et al.*, "Functional Interaction among Thyroid-specific Transcription Factors: Pax8 Regulates the Activity of Hex Promoter," *Mol Cell Endocrinol*, pp. 214, no. 2, pp. 117-25, 2004.

[20] Y. Qian, R. Murphy, "Recognition of Figures containing Fluorescence Microscope Images in Online Journal Articles using Graphical Models," *Bioinformatics*, vol. 24, no.4, pp. 569-76, 2008.

[21] B. Rafkind, M. Lee, *et al.*, "Exploring Text and Image Features to Classify Images in Bioscience Literature," *BioNLP Workshop on Linking Natural Language,* Association for Computational Linguistics, pp. 73–80, 2006.

[22] Y. Regev, M. Finkelstein-Landau, and R. Feldman, "Rulebased Extraction of Experimental Evidence in the Biomedical Domain: The KDD Cup 2002 (Task 1)," *SIGKDD Explor Newsl*, vol. 4, no. 2, pp. 90–92, Dec. 2002.

[23] R. Rodriguez-Esteban, I. Iossifov, "Figure Mining for Biomedical research," *Bioinformatics*, vol. 25, no. 16, pp. 2082–84, 2009.

[24] H. Shatkay, M. Craven, *Mining the Biomedical Literature,* MIT Press, 2012.

[25] H. Shatkay, N. Chen, and D. Blostein., "Integrating Image Data into Biomedical Text Categorization," *Bioinformatics*, vol. 22, no. 11, pp. e446–53, 2006.

[26] H. Shatkay, R. Narayanaswamy, *et al.*, "OCR-based Image Features for Biomedical Image and Article Classification: Identifying Documents relevant to Cis-Regulatory Elements," *ACM Conf. on Bioinformatics*, *Computational Biology and Biomedicine (BCB)*, pp. 98–104, 2012.

[27] G. Tsatsaronis, M. Schroeder, *et al.*, "BioASQ: A Challenge on Large-scale Biomedical Semantic Indexing and Question Answering," *2012 AAAI Fall Symposium Series*, 2012.

[28] I. Witten, E. Frank, "*Data Mining: Practical machine learning tools and techniques,*" Morgan Kaufmann, 2005.

[29] Xerox Rossinante,
https://pdf2epub.services.open.xerox.com/, 2013.

[30] S. Xu, J. McCusker, and M. Krauthammer, "Exploring the use of image text for biomedical literature retrieval," *Proc. of the AMIA Annu Symp*, p.1186, 2008.

[31] H. Yu, F. Liu, B.P. Ramesh, "Automatic Figure Ranking and User Interfacing for Intelligent Figure Search," *PLoS One*, vol. 5, no. 10, pp. 129-83, 2010.

[32] J. Zar, *Biostatistical Analysis*, Prentice Hall, 1999