Structured Literature Image Finder: Extracting Information from Text and Images in Biomedical Literature

Luís Pedro Coelho[1,2,3], Amr Ahmed[4,5], Andrew Arnold[4], Joshua Kangas[1,2,3],
Abdul-Saboor Sheikh[3], Eric P. Xing[4,5,6], William W. Cohen[1,4], and
Robert F. Murphy[1,2,3,4,6,7]

[1] Lane Center for Computational Biology, Carnegie Mellon University
[2] Joint Carnegie Mellon University–University of Pittsburgh Ph.D. Program in Computational Biology
[3] Center for Bioimage Informatics, Carnegie Mellon University
[4] Machine Learning Department, Carnegie Mellon University
[5] Language Technologies Institute, Carnegie Mellon University
[6] Department of Biological Sciences, Carnegie Mellon University
[7] Department of Biomedical Engineering, Carnegie Mellon University

Abstract. Slif uses a combination of text-mining and image processing to extract information from figures in the biomedical literature. It also uses innovative extensions to traditional latent topic modeling to provide new ways to traverse the literature. Slif provides a publicly available searchable database (http://slif.cbi.cmu.edu).

Slif originally focused on fluorescence microscopy images. We have now extended it to classify panels into more image types. We also improved the classification into subcellular classes by building a more representative training set. To get the most out of the human labeling effort, we used active learning to select images to label.

We developed models that take into account the structure of the document (with panels inside figures inside papers) and the multi-modality of the information (free and annotated text, images, information from external databases). This has allowed us to provide new ways to navigate a large collection of documents.

1 Introduction

Thousands of papers are published each day in the biomedical domain. Working scientists therefore struggle to keep up with all the results that are relevant to them. Traditional approaches to this problem have focused solely on the text of papers. However, images are also very important as they often contain the primary experimental results being reported. A random sampling of such figures in the publicly available PubMed Central database reveals that in some, if not most of the cases, a biomedical figure can provide as much information as a normal abstract. Thus, researchers in the biomedical field need automated systems that can help them find information quickly and satisfactorily. These systems should provide them with a structured way of browsing the otherwise unstructured knowledge in a way that inspires them to ask questions that they never thought of before.

Our team developed the first system for automated information extraction from images in biological journal articles (SLIF, the "Subcellular Location Image Finder," first described in 2001 [1]). Since then, we have reported a number of improvements to the SLIF system [2–4]. In part reflecting this, we are rechristening SLIF as the "Structured Literature Image Finder."

Most recently, we have added support for more image types, improved classification methods, and added features based on multi-modal latent topic modeling. Topic modeling allows for innovative user-visible features such as "browse by topic," retrieval of topic-similar images or figures, or interactive relevance feedback. Traditional latent topic approaches have had to be adapted to the setting where documents are composed of free and annotated text and images arranged in a structured fashion. We have also added a powerful tool for organizing figures by topics inferred from both image and text, and have provided a new interface that allows browsing through figures by their inferred topics and jumping to related figures from any currently viewed figure. We have performed a user study where we asked users to perform typical tasks with SLIF and report whether they found the tool to be useful. The great majority of responses were very positive [5].

SLIF provides both a pipeline for extracting structured information from papers and a web-accessible searchable database of the processed information. Users can query the database for various information appearing in captions or images, including specific words, protein names, panel types, patterns in figures, or any combination of the above.

2 Overview

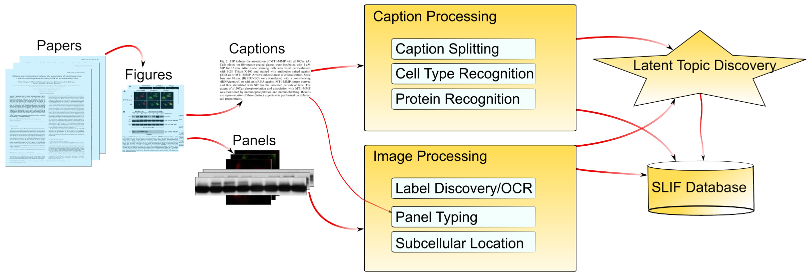


Fig. 1. Overview of SLIF pipeline

The SLIF processing pipeline is illustrated in Figure 1. After preprocessing, image and caption processing proceed in parallel. The results of these two modules then serve as input to the topic modeling framework.

The first step in image processing is to split the image into its panels, then identify the type of image in each panel. If the panel is a fluorescence micrograph image (FMI), the depicted subcellular localization is automatically identified [1]. In addition, panel labels are identified through optical character recognition, and scale-bars, if present, are identified. Annotations such as white arrows are removed.

In parallel, the caption is parsed and relevant biological entities (protein and cell types) are extracted from the caption using named entity recognition techniques. Also, the caption is broken up into logical scopes (sub-captions, identified by markers such as "(A)"), which will be subsequently linked to panels.

The last step in the pipeline aggregates the results of image and caption processing by using them to infer underlying themes in the collection of papers. These are based on the free text in the caption, on the annotated text (i.e., protein and cell type names are not processed as simple text), and the image features and subcellular localization. This results in a low-dimensional representation of the data, which is used to implement retrieval by example ("find similar papers") or even interactive relevance feedback navigation.

Access to the results of this pipeline is provided via a web interface or programatically with SOAP queries. Results presented always link back to the full paper for user convenience.

3 Caption Processing

A typical caption, taken from [6], reads as:

> S1P induces relocalization of both **p130Cas** and **MT1-MMP** to peripheral **actin**-rich structures. (**A**) **HUVEC** were stimulated for 15 min with 1 μM S1P and stained with polyclonal **MT1-MMP** [...]. (**B**) Cells were stimulated with S1P as described above [...]. Scale bars are **10μm**.

We have highlighted, in bold, the pieces of information which are of interest to SLIF: The text contains both a global portion (the first sentence) and portions scoped to particular panels (marked by "(A)" and "(B)"). Thus the caption is broken up into three parts, one global, and two specific to a panel. In order to understand what the image represents, SLIF extracts the names of proteins present (p130Cas, MT1-MMP,...), as well as the cell line (HUVEC) using techniques described previously. Additionally, SLIF extracts the length(s) of any scale bars to be associated with scale bars extracted from the image itself.

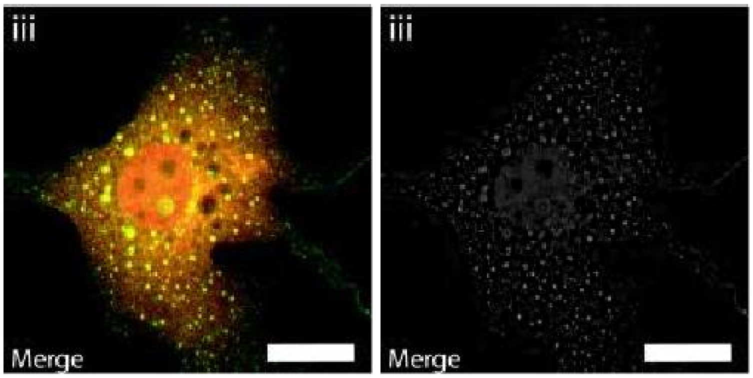
The implementation of this module is described in greater detail elsewhere [2, 4, 5, 7].

4 Image Processing

4.1 Figure Splitting

The first step in our image processing pipeline is to divide the extracted figures into their constituent components, since in majority of the cases, the figures are comprised of multiple panels to depict similar conditions, corresponding analysis, etc. For this purpose, we employ a figure-splitting algorithm that recursively finds constant-intensity boundary regions to break up the image hierarchically. Large regions are considered panels, while small regions are most likely annotations. This method that was previously shown to perform well [1].

4.2 "Ghost" Detection



(a) Color image (b) Blue channel

Fig. 2. Example of a ghost image. Although the color image is obviously a two-channel image (red and green), there is a strong bleed-through into the blue component.

Fmi panels are often false color images composed of related channels. However, due to treatment of the image for publication or compression artifacts, it is common that an image that contains one or two logical colors (and is so perceived by the human reader), will have signal in all three color channels. The extra channel, we call a "ghost" of the signal-carrying channels. Figure 2 illustrates this phenomenon.

To detect ghosts, we first compute the white component of the image, i.e., the pixel-wise minimum of the 3 channels. We then subtract this component from each channel so that the regions with homogeneous intensities across all channels (e.g. annotations or pointers) get suppressed. Then, for each channel, we verify if its 95%-percentile pixel is at least 10% of the overall highest pixel value. These two values were found empirically to reject almost all ghosts, with a low rate of false negatives (a signal carrying channel that has less than 5% bright pixels will be falsely rejected, but we found the rate of false positives to be low enough to be acceptable). Algorithm 1 illustrates this process in pseudo-code.

Algorithm 1: Ghost Detection Algorithm

```
White := pixelwise-min(R,G,B)
M := max( R−White, G−White, B−White )
foreach $ch \in (R,G,B)$ do
    Residual := ch−White
    sort pixels from Residual
    if $95\%$ percentile pixel $< 10\% M$ then
        ch is a ghost
```

4.3 Panel Type Classification

SLIF was originally designed to process only FMI panels. Recently, we expanded the classification to other panel types, in a way similar to other recent systems [8–10].

Panels are classified into one of six panel classes: (1) FMI, (2) gel, (3) graph or illustration, (4) light microscopy, (5) X-ray, or (6) photograph. To build a training set for this classification problem, while minimizing labeling effort, we used empirical risk reduction, an active learning algorithm [11]. We used a libSVM-based classifier as the base algorithm. In order to speed up the process, at each round, we labeled the 10 highest ranked images plus 10 randomly selected images. The process was seeded by initially labeling 50 randomly selected images. This resulted in ca. 700 labeled images.

The previous version of SLIF already had a good FMI classifier, which we have kept. Given its frequency and importance, we focused on the *gel* class as the next important class. Towards this goal, we define a set of features based on whether certain marker words appeared in the caption that would signal gels[8] as well as a set of substrings for the inverse class[9]. A classifier based on these boolean features was learned using the ID3 decision tree algorithm [12] with precision on the positive class as the function being maximized. This technique was shown, through 10 fold cross-validation, to obtain very high precision (91%) at the cost of moderate recall (66%). Therefore, examples considered positive are labeled as such, but examples considered negative are passed on to a classifier based on image features. In addition to the features developed for FMI classification, we measure the fraction of variance that remains in the image formed by the differences between horizontally adjacent pixels:

[8] The positive markers were: *Western*, *Northern*, *Southern*, *blot*, *lane*, *RT* (for "reverse transcriptase"), *RNA*, *PAGE*, *agarose*, *electrophoresis*, and *expression*.

[9] The negative markers were: *bar* (for bar charts), *patient*, *CT*, and *MRI*.

$$h(I) = \frac{\mathrm{var}(I_{y,x-1} - I_{y,z})}{\mathrm{var}(I_{y,x})}. \tag{1}$$

Gels, consisting of horizontal bars, score much lower on this measure than other types of images. Furthermore, we used 26 Haralick texture features [13]. Images were then classified into the six panel type classes using a support vector machine (SVM) based classifier. On this system, we obtain an overall accuracy of 69%.

Therefore, the system proceeds through 3 classification levels: the first level classifies the image into FMI or non-FMI using image based features; the second level uses the textual features described above to identify gels with high-precision; finally, if both classifiers gave negative answers, an SVM operating on image-based features does the final classification.

4.4 Subcellular Location Pattern Classification

Perhaps the most important task that SLIF supports is to extract information based on the subcellular localization depicted in FMI panels.

To provide training data for pattern classifiers, we hand-labeled a set of images into four different subcellular location classes: (1) *nuclear*, (2) *cytoplasmic*, (3) *punctate*, and (4) *other*, following the active learning methodology described above for labeling panel types. The active learning loop was seeded using images from a HeLa cell image collection that we have previously used to demonstrate the feasibility of automated subcellular pattern classification [14].

The dataset was filtered to remove images that, once thresholded using the methods we described previously [14], led to less than 80 above-threshold pixels, a value which was empirically determined. This led to the rejection of 4% of images. In classification, if an image meets the rejection criterion, it is assigned into a special *don't know* class.

We computed previously described field-level features to represent the image patterns (field-level features are features that do not require segmentation of images into individual cell regions). We added a new feature for the size of the median object (which is a more robust statistic than the previously used mean object size). Experiments using stepwise discriminant analysis as a feature selection algorithm [15] showed that this was an informative feature. If the scale is inferred from the image, then we normalize this feature value to square microns. Otherwise, we assume a default scale of $1\mu m$/pixel.

We also adapted the threshold adjacency statistic features (TAS) from Hamilton et al. [16] to a parameter-free version. The original features depended on a manually controlled-two-step binarization of the image. For the first step, we use the Ridler–Calvard algorithm to identify a threshold instead of a fixed threshold [17]. The second binarization step involves finding those pixels that fall into a given interval such as $[\mu - M, \mu + M]$, where μ is the average pixel value of the above-threshold pixel and M is a margin (set to 30 in the original paper). We set M to the standard deviation of the above threshold pixels.[10] We call these *parameter-free* TAS.

On the 3 main classes (Nuclear, Cytoplasmic, and Punctate), we obtained 75% accuracy (as before, reported accuracies are estimated using 10 fold cross-validation and the classifier used was an svm). On the four classes, we obtained 61% accuracy.

4.5 Panel and Scope Association

As discussed above, figures are composed of a set of panels and a set of subimages which are too small to be panels. To associate panels with their caption pointers (e.g., identifying which panel is panel "A" if such a mention is made in the caption), we parse all panels and other sub-images using optical character recognition (OCR). In the simple case, the panel contains the panel annotation and there is a one-to-one match to annotations in the caption. Otherwise, we match panels to the nearest found in-image annotation.

5 Topic Discovery

The previous modules result in panel-segmented, structurally and multi-modally annotated figures: each figure is composed of multiple panels, and the caption of the whole figure is parsed into scoped caption, global caption, and protein entities. Each scoped caption is associated with a single panel and the global caption is shared across panels and provides contextual information. Given this organization, we would like to build a system for querying across modality and granularity. For instance, the user might want to search for biological figures given a query composed of key words and protein names (across-modality), or the user might want to retrieve figures similar to a given panel (across-granularity) or a given other figure of interest. In this section, we describe our approach to address this problem using topic models.

Topic models aim towards discovering a set of latent themes present in the collection of papers. These themes are called topics and serve as the basis for visualization and semantic representation. Each topic k consists of a triplet of distributions: a multinomial distribution over words β_k, a multinomial distribution of protein entities $\varOmega_k$, and a gaussian distribution over every image feature s, $(\mu_{k,s}, \sigma_{k,s})$. Given these topics, a graphical model is defined that generates figure f given these topics (see [18] for a full description). There are two main steps involved in building our topic model: inference and learning. In learning, given a set of figures, the goal is to learn the set of topics $(\beta_k, \varOmega_k, \{\mu_{k,s}, \sigma_{k,s}\})$ that generates the collection using Bayesian inference [18]. On the other hand, given the discovered topics and a new figure f, the goal of inference is to deduce the latent representation of this figure $\theta_f = (\theta_{f,1} \cdots \theta_{f,k})$, where the component $\theta_{f,k}$ defines how likely topic k will appear in figure f. Moreover, for

[10] Other methods for binarizing the image presented by Hamilton et al. are handled analogously.

each panel p in figure f, the inference step also deduces its latent representation: $\theta_{f,p} = (\theta_{f,p,1} \cdots \theta_{f,p,k})$. In addition, from the learning step, each word w and protein entity r can also be represented as a point in the topic space: $\theta_w = (\beta_{1,w}, \cdots, \beta_{k,w})$ and $\theta_r = (\Omega_{1,r}, \cdots, \Omega_{k,r})$.

This results in a unified space where each figure, panel, word and protein entity is described using a point in this space which facilitates querying across modality and granularity. For instance, given a query $q = (w_1, \cdots, w_n, r_1, \cdots, r_m)$ composed of a set of text words and protein entities, we can rank figures according to this query using the query language model [19] as follows:

$$\begin{aligned} P(q|f) = \prod_{w \in q} P(w|f) \prod_{r \in q} P(r|f) &= \prod_{w \in q} \Big[\sum_k \theta_{f,k} \beta_{k,w} \Big] \prod_{r \in q} \Big[\sum_k \theta_{f,k} \Omega_{k,r} \Big] \\ &= \prod_{w \in q} \Big[\theta_f \odot \theta_w \Big] \prod_{r \in q} \Big[\theta_f \odot \theta_r \Big] \qquad (2) \end{aligned}$$

Equation 2 is a simple dot product operation between the latent representations of each query item and the latent representation of the figure in the induced topical space. The above measure can then be used to rank figures for retrieval. Moreover, given a figure of interest f, other figures in the database can be ranked based on similarity to this figure as follows:

$$sim(f'|f) = \sum_k \theta_{f,k} \theta_{f',k} = \theta_f \odot \theta_{f'} \qquad (3)$$

In addition to the above capabilities, the discovered topics endow the user with a bird's eye view over the paper collection and can serve as the basis for visualization and structured browsing. Each topic f summarizes a theme in the collection and can be represented to the user along three dimensions: top words (having high values of $\beta_{k,w}$), top proteins entities (having high values of $\Omega_{k,r}$), and a set of representative panels (panels with high values of $\theta_{f,p,k}$). Users can decide to display all panels (figures) that are relevant to a particular topic of interest [18, 5].

6 Discussion

We have presented a new version of SLIF, a system that analyzes images and their associated captions in biomedical papers. SLIF demonstrates how text-mining and image processing can intermingle to extract information from scientific figures. Figures are broken down into their constituent panels, which are handled separately. Panels are classified into different types, with the current focus on FMI and gel images. FMIs are further processed by classifying them into their depicted subcellular location pattern. The results of this pipeline are made available through either a web-interface or programmatically using SOAP technology.

A new addition to our system is latent topic discovery which is performed using both text and image. This is based on extending traditional models to handle the structure of the literature and allows us to customize these models with domain knowledge (by integrating the subcellular localization looked up from a database, we can see relations between papers using knowledge present outside of them).

Our most recent human-labeling efforts (of panel types and subcellular location) were performed using active learning to extract the most out of the human effort. We plan to replicate this approach in the future for any other labeling effort (e.g., adding a new collection of papers). Our current labeling efforts were necessary to collect a dataset that mimicked the characteristics of the task at hand (images from published literature) and improve on our previous use of datasets that did not show all the variations present in real published datasets. These datasets are available for download from the SLIF webpage (http://slif.cbi.cmu.edu) so that they can be used by other system developers and for building improved pattern classifiers.

6.1 Acknowledgments

The SLIF project is currently supported by NIH grant R01 GM078622. L.P.C. was partially supported by a grant from the Fundação Para a Ciência e Tecnologia (grant SFRH/BD/37535/2007).

References

1. Murphy, R.F., Velliste, M., Yao, J., Porreca, G.: Searching online journals for fluorescence microscope images depicting protein subcellular location patterns. In: BIBE '01: Proceedings of the 2nd IEEE International Symposium on Bioinformatics and Bioengineering, Washington, DC, USA, IEEE Computer Society (2001) 119–128
2. Cohen, W.W., Wang, R., Murphy, R.F.: Understanding captions in biomedical publications. In: KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, New York, NY, USA, ACM (2003) 499–504
3. Murphy, R.F., Kou, Z., Hua, J., Joffe, M., Cohen, W.W.: Extracting and structuring subcellular location information from on-line journal articles: The subcellular location image finder. In: Proceedings of IASTED International Conference on Knowledge Sharing and Collaborative Engineering. (2004) 109–114
4. Kou, Z., Cohen, W.W., Murphy, R.F.: A stacked graphical model for associating sub-images with sub-captions. In Altman, R.B., Dunker, A.K., Hunter, L., Murray, T., Klein, T.E., eds.: Proceedings of the Pacific Symposium on Biocomputing, World Scientific (2007) 257–268
5. Ahmed, A., Arnold, A., Coelho, L.P., Kangas, J., Sheikh, A.S., Xing, E.P., Cohen, W.W., , Murphy, R.F.: Structured literature image finder: Parsing text and figures in biomedical literature. Journal of Web Semantics (in press) (2009)
6. Gingras, D., Michaud, M., Tomasso, G.D., Bliveau, E., Nyalendo, C., Bliveau, R.: Sphingosine-1-phosphate induces the association of membrane-type 1 matrix metalloproteinase with p130cas in endothelial cells. FEBS Letters **582**(3) (2008) 399 – 404

7. Kou, Z., Cohen, W.W., Murphy, R.F.: High-recall protein entity recognition using a dictionary. Bioinformatics **21** (2005) i266–i273
8. Geusebroek, J.M., Hoang, M.A., van Gernert, J., Worring, M.: Genre-based search through biomedical images. In: Proceedings of 16th International Conference on Pattern Recognition. Volume 1. (2002) 271–274 vol.1
9. Shatkay, H., Chen, N., Blostein, D.: Integrating image data into biomedical text categorization. Bioinformatics **22**(14) (2006) e446–453
10. Rafkind, B., Lee, M., Chang, S., Yu, H.: Exploring text and image features to classify images in bioscience literature. In: Proceedings of the BioNLP Workshop on Linking Natural Language Processing and Biology at HLT-NAACL, Morristown, NJ, USA, Association for Computational Linguistics (2006) 73–80
11. Roy, N., Mccallum, A.: Toward optimal active learning through sampling estimation of error reduction. In: Proc. 18th International Conf. on Machine Learning, Morgan Kaufmann (2001) 441–448
12. Mitchell, T.M.: Machine Learning. McGraw-Hill (1997)
13. Haralick, R.M.: Statistical and structural approaches to texture. Proceedings of the IEEE **67** (1979) 786–804
14. Boland, M.V., Murphy, R.F.: A neural network classifier capable of recognizing the patterns of all major subcellular structures in fluorescence microscope images of HeLa cells. Bioinformatics **17**(12) (2001) 1213–1223
15. Jennrich, R.: Stepwise Regression & Stepwise Discriminant Analysis. In: Statistical Methods for Digital Computers. John Wiley & Sons, Inc, New York (1977) 58–95
16. Hamilton, N., Pantelic, R., Hanson, K., Teasdale, R.: Fast automated cell phenotype image classification. BMC Bioinformatics **8**(1) (2007) 110
17. Ridler, T., Calvard, S.: Picture thresholding using an iterative selection method. IEEE Trans. Systems, Man and Cybernetics **8**(8) (August 1978) 629–632
18. Ahmed, A., Xing, E.P., Cohen, W.W., Murphy, R.F.: Structured correspondence topic models for mining captioned figures in biological literature. In: Proceedings of The Fifteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2009), New York, NY, USA, ACM (2009) 39–47
19. Ponte, J.M., Croft, W.B.: A language modeling approach to information retrieval. In: SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, New York, NY, USA, ACM (1998) 275–281