Semi-Automated Ground-Truth Data Collection and Annotation for Journal Figure Analysis

Rebekah Narum, Jie Zou, Sameer Antani
Communications Engineering Branch, National Library of Medicine, NIH, Bethesda, MD

OVERVIEW

Open-i is an online service provided by the National Library of Medicine to enable search and retrieval of abstracts and images from 1.2 million PubMed Central® articles. An important preprocessing step of building Open-i's backend is to automatically segment figures into panels and recognize panel labels, such that figure captions of individual panels can be linked to the figure panels and more precise features may be extracted. Existing panel segmentation and label recognition algorithms [1,2,3] are developed based on a tiny set of 448 figures. Due to the lack of training samples, the algorithms have to rely on many hand-crafted rules, which are not able to accommodate the large variations of the figures need to be processed.

This project creates a workflow pipeline, in an attempt to collect a significantly larger ground-truth annotated figure database efficiently. The annotation of a figure includes the style of the figure (single-panel, multi-panel or stitched multi-panel), rectangular bounding boxes of the panels, rectangular bounding boxes of the panel labels, and the panel labels. The workflow starts with running automated methods, and then the automated annotation is reviewed and fixed by humans. In order to ensure the annotation quality, a verification algorithm is developed to check the consistency of the annotations. The humans then review the suspicious annotations reported by the verification program.

FIGURE COLLECTION AND ANNOTATION

First, retrieve figures via Open-i Web Service API. Then, the program FigureDownload (as shown in the image to the right) is called.

The query string used for downloading include:
- tumor
- heart = murmur
- broken = back
- brain = neoplasm + benign
- heart = block + congenital
- cancer + tumor + benign
- ag + brain
- ag + heart = rate + low
- malaria + plasmodium + falciparum + with + complications.

The XML files are copied and pasted into the 1-image folder.

In this step, using the PanelSeg.jar program, the manual paneling corrections are checked for the following:
- That labels and panels match up
- That there are not any missing labels or panels

The individual figure names and the errors are saved to a text file (as seen in the image to the right).

Fourth, the figures must be converted to an iphotodraw formatted XML file.

The program PanelSeg.jar is called to verify the annotations and determine the label styles.

The label styles include the following:
- Single-Panel is a single figure.
- Multi-Panel has multiple figures in a given image which are clearly defined.
- Stitched-Multi-Panel also has multiple figures in a given image, but there is no definitive line separating them.

STYLE ASSIGNMENT

This is where the Style Annotation takes place.

In Panel Collection and Annotation, the main problem occurred with the PanelSegEval.jar algorithm because it did not handle certain figure types. Some of the problems are in the following types of images:
- Graphs
- Any with data charts
- Images with extra words or random letters
- Multi-panel figures with overlapping parts

As a result, the figures would have some of the following problems:
- Extra label boxes
- Wrong labeling on legends
- No labels when none were needed

In Style Assignment, the main issue was not having the style annotation boxes and labels appear on the annotated image.

Finally, the figures are moved to a completed folder and the process starts over.

REFERENCES

1. D. You, S. Antani, D. Demner-Fushman, S. Goudarnejad, G.R. Thoma, Detecting Figure-Panel Labels in Medical Journal Articles Using MRI, 997-997, Int'l Conf. on Document Analysis and Recognition, 2011
3. K.C. Santosh, S. Antani, G.R. Thoma, Stitched Multi-panel Biomedical Figure Separation, 54-54, IEEE 28th int'l Symposium on Computer-Based Medical Systems, 2015

SUMMARY

Following this workflow pipeline, we are able to collect and annotate figures efficiently. Over a period of 7 weeks, 10,262 figures are collected and ground-truth annotated by one person. With this larger dataset, a lot more rigorous evaluation can be conducted and algorithms relying more on machine learning instead of hand-crafted rules can be researched. We believe this dataset is valuable to the future R&D of figure processing.