Towards the Yale Image Finder 2.0:
A New and Improved Biomedical Image Search Engine
(Continued from Fall ’14)

Luke Hansen (Computer Science & Mathematics ’15),
with Mate Nagy (PhD Candidate, CB&B),
and Professor Michael Krauthammer (Yale Med. School)

Dept. of Computer Science Advisor: Professor Holly Rushmeier
Dept. of Mathematics Reader: Professor Ronald Coifman

February 4, 2015

1 Motivation

In the ever-increasing domain of biomedical publications, there are fundamental information retrieval questions: How can scientists most efficiently access the data, experiments, medical results, and protocols they need to continue their research? While powerful text-based search engines exist (for example, Google Scholar and Entrez), there are fewer image-based engines, and all of these would benefit from an overhaul in functionality. Because of the wealth of information presented in biomedical images (including graphs, gels, microscope slides, brain scans, x-rays, and many more), and because oftentimes this information is not fully and/or explicitly declared in the text, biomedical image search engines should prove a powerful tool for information retrieval in the field.

2 Background

The Yale Image Finder (YIF) is a publicly accessible tool developed by the Krauthammer lab to retrieve open-access biomedical images and their associated articles. The web app specifically searches for text embedded directly into such images, as well as for text in an image’s caption, title, and abstract.

While the YIF is an incredibly useful tool for retrieving images based on their embedded or associated text, the ideal image finder would also search by type of image (graph, gel, x-ray, etc.), as well as by features within the image itself (specific type of gel, body part x-rayed, etc.). Furthermore, an optimal image finding tool would include a ‘related images’ feature that could return figures similar to an image of interest. From a survey of 31 scientists
and medical practitioners, we concluded that such a related images feature would be quite useful for: retrieving related experiments and techniques; validating or comparing results; finding higher caliber images or literature (e.g., a better designed experiment or one from a more familiar author or institution); and overall literature reviews.

The goal of my year-long CS 490 project is to develop the latter of these two ideal image finder components: a robust related content feature, based on embedded or associated text, as well as image analysis. (Note that many of the platforms we develop for such a related images tool can also be directly applied to a primary search for an image as well.)

3 Specifics

During the Fall ’14 term, I worked on optimizing related content retrieval based solely on embedded or associated text. For a discussion of those results, please refer to my final report from that term.

During the first month of the Spring ’15 term, I have worked on caching the text-based related content we developed last term. We noticed that queries for related images required more than a few seconds to process; so together with Mate Nagy, I developed a multi-threaded platform to cache the top 100 related images of each image in our corpus (of 1.5 million images). The speedup in queries for related content is tremendous.

The remainder of Spring ’15 will be spent developing a feature analysis framework to improve related content browsing. As a first step, Mate Nagy implemented a Haralick feature analysis system. In a 1973 paper, Haralick et. al. present 14 statistics that attempt to describe the texture and tone of an image. Having calculated these 14 statistics for each image in our corpus, we can now additionally search for related images by similar Haralick features – a chart and an x-ray, for example, have significantly different textures, which should be manifested in some or all of their respective Haralick features.

Another important image processing technique we plan to implement this term is Fourier analysis. The Fourier transform is a transformation from the “spatial” domain (e.g. \( (x, y, z) \) coordinates, where \( x \) and \( y \) are coordinates of a pixel, and \( z \) is the pixel’s value) to the “frequency” domain (e.g. \( (f, A) \) coordinates, where \( f \) is the frequency of a sinusoid associated with the Fourier series decomposition of the image signal, and \( A \) is the amplitude associated with that sinusoid). For each image, we can thus perform a Fourier transform, and use similar frequency/amplitude pairings as another criterion of relatedness – again, a chart and an x-ray have significantly different periodicities, which should be manifested in the Fourier transforms of the images.

In the final stage of feature extraction, we plan to implement a Convolutional Neural Network (CNN). Such networks have been described previously (most notably by Yann LeCun from NYU’s Courant Institute), and have been used for such applications as handwriting recognition, biological object recognition, and facial detection. Studying and working out the details of such a CNN for our applications will be an important part of my project this term; we may subsequently implement such a CNN under an energy-based architecture for efficiency.
4 CS 490 Deliverables

- Java code for caching related images by text-based queries (already completed this term)

- Java, C, Python, and/or Matlab code for implementing our feature analysis systems (Fourier and CNN)

- A written report describing in detail the final pipeline for related content browsing, along with results from applying this pipeline to numerous test cases (and potentially from applying it to an entire small corpus for which we hand-label related images)

- An oral report at the end of the term to the mathematics department as part of the CS & Math senior requirement. This report will detail the mathematical basis for our clustering and feature analysis pipeline.

References


