Chapter 9
Annotating Historical Archives of Images

Xiaoyue Wang
University of California Riverside, USA

Lexiang Ye
University of California Riverside, USA

Eamonn J. Keogh
University of California Riverside, USA

Christian R. Shelton
University of California Riverside, USA

ABSTRACT
Recent programs like the Million Book Project and Google Print Library Project have archived several million books in digital format, and within a few years a significant fraction of world’s books will be online. While the majority of the data will naturally be text, there will also be tens of millions of pages of images. Many of these images will defy automation annotation for the foreseeable future, but a considerable fraction of the images may be amiable to automatic annotation by algorithms that can link the historical image with a modern contemporary, with its attendant metatags. To perform this linking, there must be a suitable distance measure that appropriately combines the relevant features of shape, color, texture and text. However, the best combination of these features will vary from application to application and even from one manuscript to another. In this work, the authors propose a simple technique to learn the distance measure by perturbing the training set in a principled way.

INTRODUCTION
Several initiatives such as the Million Book Project and Google Print Library Project have already archived several million books in digital format, and it is believed that within a few years a significant fraction of world’s books will be online (Herwig, 2007). As Kevin Kelly recently noted, “the real magic will come in the second act, as each word in each book is cross-linked, clustered, cited, extracted, indexed, analyzed, annotated, remixed, reassembled and woven deeper into the culture than ever before” (Kelly, 2006). While
this quotation explicitly singles out text, a similar argument can be made for images. Clearly the majority of the data gleaned from scanned books will be text, but there will also be tens of millions of pages of images. Many of these images will defy automation annotation for the foreseeable future; however a considerable fraction of the images may be amiable to automatic annotation by algorithms that can link the historical image with a modern contemporary, with its attendant meta tags (Agosti et al., 2007). As a concrete example, consider Figure 1.

In this image the text annotations will surely defy even the state-of-the-art handwriting recognizers (Adamek et al., 2007), and humans, particularly those without experience in reading cursive script are also unlikely to be able to parse these words. Suppose that we segment out the individual insects and search for the most similar images on the web (for the moment, we will gloss over the technical details of how this is done). In fact we have done this, and discovered the image in Figure 2.

The image is uncannily like the query image, we can confidently assume it is the same (or closely related) species and therefore we can link the historical image to its modern counterpart to provide context and annotations to the digital archive. In this example the shape and color provided the necessary clues to the identity of unknown object. More generally, different sets of features may be useful depending on the application. For example, most Diatoms (eukaryotic algae) are colorless when viewed at the microscopic scale, but are often richly textured, as in Figure 3.

Here (rotation invariant) shape and texture allow us to link this object to a modern image, and learn of the latest research on this intriguing life form.

In this work we propose a general framework for annotating large archives of historical image manuscripts. Our work is similar in spirit to the work (Agosti et al., 2007) on the automatic discovery of relationships among images in illuminated manuscripts. The authors introduced a model of various annotations of digital contents, in forms of both text and typed links, e.g. author links. However in this work, we are focusing on the lower level primitives to support such work. We use different feature spaces such as shape, color and texture. Then we combine these similarities using appropriate weights. Our experiments show that the accuracy we can obtain is higher by using

Figure 1. A page from a scientific text published in 1849 (D’Orbigny, 1849). The heavily stylized script is difficult to read even at full resolution, however we have independently confirmed that the three insects are (left to right) Zygaena filipendulae, Acherontia atropos and Syntomis phegea.