Recognizing Patient Names in Handwritten Clinical Notes

Bethany Percha
Stanford University
Stanford, CA
blpercha@stanford.edu

1. Introduction & Background

Debate about the future of the U.S. health care system currently pervades the halls of politics and the front pages of major newspapers. One thing almost everyone can agree on, however, is that medicine in the United States could benefit from more widely and consistently adopting electronic forms of patient record-keeping, like electronic medical record (EMR) and prescription ordering systems. With the passage of 2009's HI-TECH Act, physicians can receive up to $40,000 as an incentive for using EMR systems that meet certain criteria, and the use of these systems is on the rise. However, there are still some major roadblocks to their universal adoption.

One major barrier to EMR adoption is the fact that so many physicians' offices and hospitals continue to use paper records. When a health organization decides to make the switch to an EMR, all of the past patient records must somehow be introduced into the system, either by scanning or manual data entry. Currently the easiest way to do this is simply to fax in all of the old records. Administrators must then manually assign individual documents to the appropriate patient name, document date, and document type. Sometimes optical character recognition (OCR) can be used to identify patient names and other metadata in the document, but many documents are handwritten. What is more, documents pertaining to a particular patient may contain handwriting from many different individuals, which thwarts traditional OCR techniques in which a classifier is trained to recognize handwriting from a particular person.

For the past several weeks, I have been working with a Silicon Valley startup, ElationEMR, to solve this very problem; that is, we would like to automatically assign handwritten documents to the appropriate patient file when they are faxed in from a new client. OCR works very well for assigning typewritten documents, but because of the variety of handwriting styles and document layouts present in the handwritten documents, we are unable to reliably assign those to the correct patient. One advantage in our case (relative to matching arbitrary names to arbitrary documents) is that we already have a list of all possible patient names for a given practice; that is, we know the "sample space" of names we can find in the documents. So I suspect this is really more of a pattern-matching problem – similar to recognizing and identifying a human face in a composite image, for example – than a task for OCR. For my CS231A final project, I would like to see if it is possible to match patient names to documents using techniques similar to those used for object recognition in composite images. I will develop and test the algorithm on a set of authentic patient documents from ElationEMR, searching only for patient first names to ensure patient privacy.

2. Objective

To design a system for automatically assigning faxed, handwritten clinical documents to the correct patient record.

3. Data

The training set will consist of two hundred or more faxed images of handwritten clinical documents from the records of ElationEMR, an electronic medical record startup here in Silicon Valley. This training set will include documents describing approximately ten different patients. The handwritten patient name will be present in each document.

4. Methods

The first subtask of this project will be extracting features relevant to classifying different documents into individual patient names. This is a very similar problem to that of recognizing individual faces in composite images, so feature extraction algorithms similar to those used to recognize faces could be applied. For example, I could first use PCA ("Eigenfaces") or LDA ("Fisherfaces") on names that have been cropped out of the larger documents to see what features are most relevant to distinguishing different names from each other, even if those names are written by different people.

The next step will be locating the names within the larger image. To do this, I will need to use techniques similar to those outlined in Viola and Jones's 2001 paper on robust real-time object detection. I expect that my final algorithm for patient name identification will be similar to theirs, but I plan to explore more recent literature on object detection in composite images over the course of this project.

5. Evaluation

Evaluation for this project will first consist of cross-validation on the training set, then evaluating the algorithm's performance on a hold-out test set. I anticipate holding out approximately 20-30 images with different handwritten patient names during the original training process, then evaluating the algorithm's performance on those later.