A Data Set of Annotated Historical Maps

Ravi Chande, Dylan Gumm, and Jerod Weinman
Department of Computer Science
Grinnell College
Grinnell, Iowa 50112
weinman@grinnell.edu

May 29, 2013

1 Introduction and scope

This document describes a data set designed for testing the performance of optical character recognition algorithms on text in scanned historical map images. Thirty maps from the nineteenth and early twentieth centuries were chosen from the David Rumsey Map Collection.[1] We focus on maps of the United States because the U.S. Board of Geographic Names[2] has free lists of all named geographic entities in the United States, both current and historical. Most maps are of individual states, though some are regional and one is of the entire U.S.; most feature little handwritten text. The original MrSid files were converted into uncompressed TIFF images for annotation and recognition.

The images, annotations, and example processing code are available from `http://www.cs.grinnell.edu/~weinman/research/maps.shtml` and `http://digital.grinnell.edu`.

2 XML data format

We annotated maps as four primary nested entities: `map`, `window`, `label`, and `word`.

Each XML file contains one `map` entity. The sole attribute of a map entity is `src`, which indicates the filename (but not path) of the TIFF image the XML file describes. The filename (sans extension) is the unique portion of the originating MrSid file's URL on the Rumsey website.

If an image has more than one geographic coordinate frame (e.g., an inset map), the `map` entity contains two or more `window` entities to separate the sets of `label`s for each frame of reference. If only one coordinate frame is present in the image, no `window` entity is used.

A `label` entity annotation exists for each distinct piece of text on a map. A `label` entity has a `text` attribute for the ground truth label, which could include the full name of a city (e.g., "Des Moines"), body of water (e.g., "Rock Creek"), county, or even the text in a map's legend. A map entry should have as many label entries as there are labels on the map.

If a map text label is associated with an object at an image location denoted by the map, a `point_location` entity (with numerical attributes `x` and `y`) within the `label` annotation indicates the image coordinate of the label. Cities, mines, and mountain peaks often feature such a point location.

Each label's text is manually segmented into words. We annotate each space-separated word in the text of a `label` with a `word` entity. Each `word` entity is nested within a `label` and has a `text` attribute for the word string (which may have punctuation, but no spaces).

Several entities nested within `word` contain the information necessary to create a tight bounding box around that portion of the image. A word's baseline is stored as a series of `point` entities within a single `baseline` entity. The order of these points follows the letters of the word in reading order from left to right, regardless of the word's absolute orientation on the map. The left and right word boundaries are represented as `point` entities inside `leftbound` and `rightbound` entities. Image coordinates used to measure x-height or capital letter height are stored in `x_height_points` or `caps_height_points` entities. This data is recast as a series of `x_height` and `caps_height` entities, summarized by `average_x_height` and `average_caps_height` entities, each with a single

[1] `http://davidrumsey.com`

[2] The U.S. BGN is part of the federal U.S. Geographical Survey created to "maintain uniform geographic name usage throughout the Federal Government". See `http://geonames.usgs.gov`.

numerical `height` attribute. If no instances of capital or lowercase points can be found, the averages are given as `NaN` values.

All this data allows us to draw tight bounding boxes around words and eventually normalize them.

3 Annotation methods

Each map was manually annotated by first marking the label on the map, marking its point location if one existed, and finally marking all of the words that made up that label. Each word was annotated individually with its baseline, lowercase letter heights, capital letter heights, its left and right boundaries, and the text of the word itself. Linear baselines were marked under the first and last characters of words and curved baselines were approximated by marking multiple characters within the curve. Points for the lower and upper case letters were marked at the topmost point of those respective letters. The left and right boundaries were marked on the outermost pixel of the leftmost and rightmost letters.

The annotation program calculated the letter heights by projecting the point above a letter onto the closest baseline line segment.

4 Example

The following is an abbreviated example containing one label for a map.

```
<map src="D0042-1070007.tiff">
        <label text="Grinnel">
                <point_location x="4063.9419" y="4014.7475">
                </point_location>
                <word text="Grinnel">
                        <leftbound>
                                <point x="4105.0485" y="4005.9115"></point>
                        </leftbound>
                        <rightbound>
                                <point x="4226.4475" y="4007.4482"></point>
                        </rightbound>
                        <baseline>
                                <point x="4118.11" y="4020.51"></point>
                                <point x="4224.91" y="4022.05"></point>
                        </baseline>
                        <x_height height="11.0653"></x_height>
                        <x_height height="10.6729"></x_height>
                        <x_height height="12.0049"></x_height>
                        <average_x_height height="11.2477"></average_x_height>
                        <x_height_points>
                                <point x="4139.62" y="4009.75"></point>
                                <point x="4165.75" y="4010.52"></point>
                                <point x="4204.93" y="4009.75"></point>
                        </x_height_points>
                        <caps_height height="22.2821"></caps_height>
                        <average_caps_height height="22.2821"></average_caps_height>
                        <caps_height_points>
                                <point x="4118.11" y="3998.23"></point>
                        </caps_height_points>
                </word>
        </label>
</map>
```

Acknowledgments

Development of this resource was made possible with the support of Grinnell College and HHMI.