Digital Microfilm Frame Detection

Christopher Nelson, Heath Nielson, and Shane Hathaway
The Church of Jesus Christ of Latter-Day Saints
Email: {nelsoncb, nielsonhe, hathawaysd}@ldschurch.org

Document Images – From Microfilm to Digital

For decades many genealogical documents have been archived onto microfilm. A single microfilm roll can contain thousands of documents and possesses a storage life of over two hundred years. While microfilm serves as a great archival medium, accessibility to the archived documents is still limited. To view an archived document, an individual must first have physical access to the microfilm. If the microfilm is not available, a copy is made and shipped, incurring an expense of time and materials. Once received, the microfilm can only be viewed using specialized equipment and, even with access to the microfilm, the desired document must still be found by winding through the roll of film in a linear search.

In recent years, the use of digital images as a media for distributing documents has grown tremendously. Preserving documents as digital images greatly increases their accessibility. With the increased ubiquity of the Internet, documents can be viewed from any location throughout the world with access. As a result, the expense of creating and shipping microfilm copies is eliminated. Additionally, digital document images can be viewed on any ordinary computer, thereby eliminating the need for specialized equipment. Searching operations are also improved as the granularity of indexing moves from microfilm rolls to the documents themselves.

With over 2.3 million rolls of microfilm, the Granite Mountain Records Vault has started investigating ways to improve the quality and efficiency of the otherwise tedious process of digitizing large portions of its vast collection. To this end, a new process is being evaluated. This process uses many novel approaches to facilitate the conversion of document images from microfilm to digital images in a semi-automated environment. This maximizes the production of digital images while still maintaining a high level of quality. To facilitate this automation, an automated process for detecting documents is applied to large digital representations of microfilm rolls. This process is known as "automated frame detection".

Ribbon File – Digital Microfilm

During scanning, microfilm rolls are digitized into a digital "ribbon". The ribbon data is stored to disk in an image pyramid format. This format consists of eight copies of the ribbon stored in sequence. Each copy of the ribbon, known as a "level", represents the ribbon at a different resolution. Level 0 is the digital film at the scanned resolution. Each successive level is one-half the resolution of the previous (see *Figure 1*). These ribbon files are not compressed and range in size from 20-30 gigabytes for an 8-bit grayscale scan.

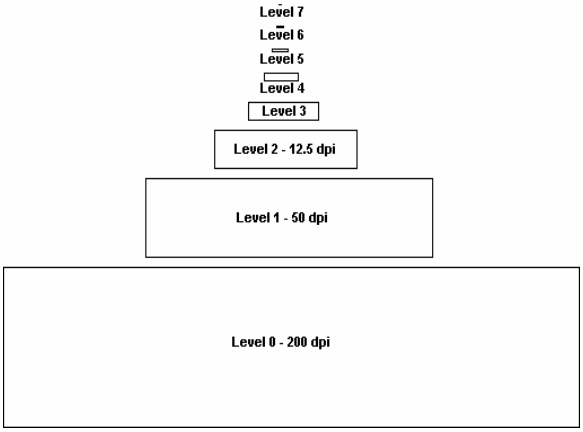

Figure 1 - Image Pyramid Format

The higher levels of the image pyramid are essentially "smoothed" versions of the original ribbon. Using these higher levels allows the frame detection to operate much faster and more effectively since there's less data to process and noise has been reduced. That being the case, frame detection currently runs on Level 4 of the ribbon file. A small segment of a ribbon file is shown in *Figure 2*.

Figure 2 - Ribbon Sample

Frame Detection – Boxing the Documents

Frame detection is the process of identifying the documents in a ribbon and placing boxes (known as "frames") around them. Given the wide variety of microfilms in existence, this operation can be trivial or extremely complex. The automated frame detection process consists of three steps: 1) generating a ribbon profile, 2) creating a threshold level, and 3) detecting horizontal frame edges.

Generating the Ribbon Profile

To start off, the system must find the left and right edges of each document. To accomplish this, a vertical profile is generated. One way to look at a ribbon is a grid of pixels, each with a grayscale "intensity value" for each pixel (255 means white, 0 means black). To create a profile, simply add all the intensity values in each column and plot the result as shown in *Figure 3*. Because white pixels have a greater value than black pixels, a high profile value indicates a detected document, whereas a low profile value indicates background.

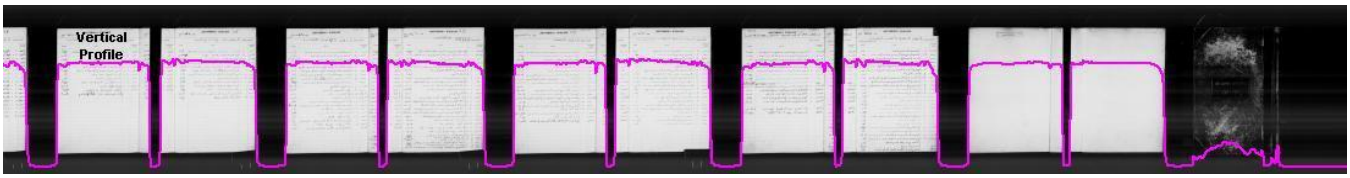

Figure 3 - Vertical Ribbon Profile

Creating the Threshold

Now that the profile has been generated, a "threshold" is needed to guide the frame detection. A threshold is a dividing line between the high profile values that identify a document area and the low profile values that identify a column as background. Every point where the threshold intersects the profile marks a left or right frame edge. *Figure 4* shows the optimum threshold (the lower line) for the displayed ribbon segment. This threshold is determined by following three steps: 1) generate an average minimum profile, 2) snap the threshold to local minimum values, and 3) adjust the threshold to allow for gradual changes in the profile.

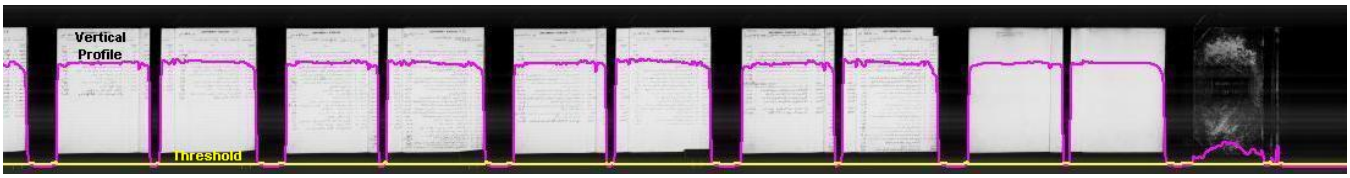

Figure 4 - Dynamic Threshold

Generating the average minimum profile

To create an average minimum profile, a "sliding window" equal to twice the height of the ribbon is used. For each column (treated as the center of the sliding window as shown in *Figure 1.5*), the smallest profile value detected inside this sliding window is saved off as a threshold value. This allows the threshold to rise or fall as the ribbon darkens or lightens.

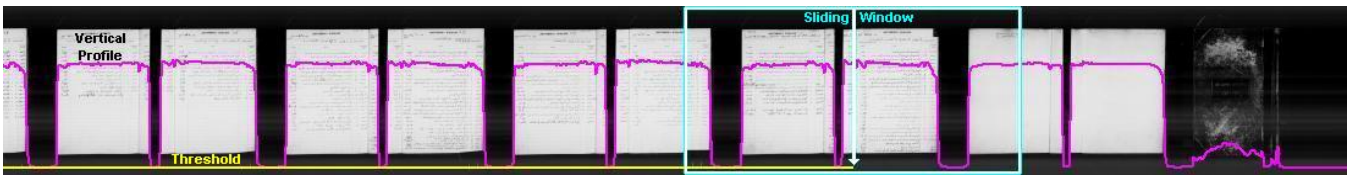

Figure 5 - Sliding Window

Snapping the threshold to local minimum values

Once this initial estimate for a threshold has been generated, the threshold is modified at any location where the ribbon profile and the threshold come very close to each other. By bringing the threshold above the ribbon profile at these locations, the threshold identifies most of the obvious sections of background on the ribbon.

Adjusting the threshold for gradual slopes

The last step allows for a gradual change in the background color by adjusting the threshold when the background color slowly changes (by gradually getting darker or lighter). At each point where the threshold lies above the profile, the neighboring profile values are checked as well. If the difference between these neighboring profile values is small (as is the case is a gentle slope), the threshold is placed above these profile values. At this point, the process repeats itself.

Now that the threshold has been generated, marking the left and right border of each frame is the simple process of finding where the threshold and profile intersect (filtering out obvious noise in the process). This effectively labels a number of "vertical frame regions" as shown in *Figure 6*.

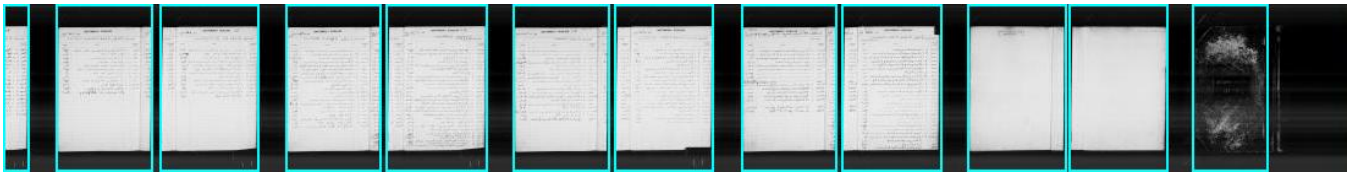

Figure 6 - Vertical Frame Regions

Detecting Horizontal Frame Edges

Now that the left and right borders for each frame have been identified, all that remains is to find the top and bottom of each frame (or frames if they are stacked on top of each other). This process for consists of four steps: 1) generating ribbon profiles, 2) setting thresholds, and 3) selecting the best results.

Generating ribbon profiles

To detect the top and bottom of each document, ribbon profiles are needed again, only this time rows are used instead of columns. A few cleaning operations are also needed to remove horizontal rows of light pixels which sometimes occur on scanned microfilm ribbons. For this operation two profiles are needed: the standard pixel intensity value and the variance (the standard deviation of a row of pixels squared). Just as in the earlier steps, a profile is compared with its respective threshold to mark the edges of each frame. Both profiles are processed in this stage and the one with the best results is actually used to detect frame edges.

Setting thresholds

Unlike the "sliding window" strategy used earlier, thresholds used in this stage are created through histogram analysis. The histogram used at this stage plots how often a specific profile row value appears in the vertical regions. *Figure 7* shows one of these histograms. Profile rows with low values (dark pixels or low variance) make the narrow peaks on the left higher, whereas rows with high profile values make the peaks on the right larger.

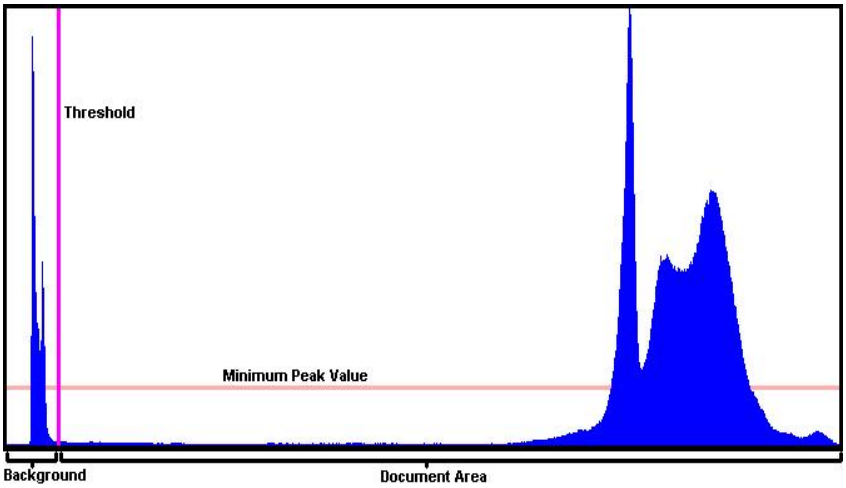

Figure 7 - Histogram Thresholding

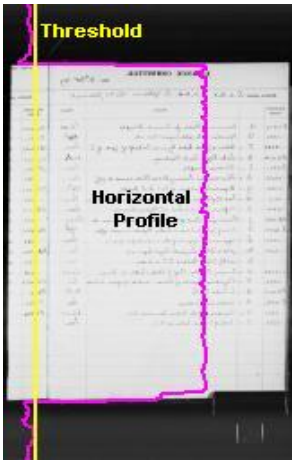

Figure 8 – Horizontal Threshold

To set this threshold, the system labels peaks as being larger than a "minimum peak value". After finding all the peaks (including where they begin and end), nearby peaks are grouped together. Lastly, the threshold is set to the right edge of the leftmost group of peaks. *Figure 8* shows a graphical example of a row-based profile and its respective threshold.

Selecting the best results

Once the threshold and profiles are finished, frame detection is this is a trivial matter of marking frame borders at the points where the profile and threshold cross. These results (created by the standard profile and the variance profile) are then compared. Whichever process created the larger frames is selected and used to identify the top and bottom edges for each frame. These frames are then examined and small noise frames (any frame which is less than 1/3 the height of any frame above or below it) are deleted. This process is repeated for each of the "vertical frame regions", thereby finishing the frame detection operation. The results of the frame detection are shown in *Figure 9*.

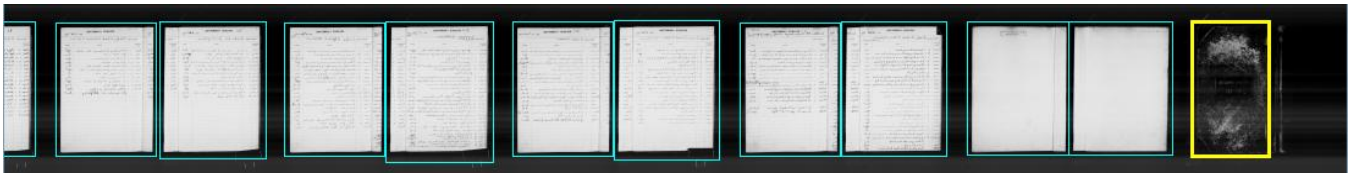

Figure 1.9 Frame Detection Results

Although automated frame detection works for most microfilm rolls, there is no guarantee that every frame in a ribbon will be placed perfectly. Given the wide variety of microfilm rolls and document types, orientations, and characteristics, perfect frame detection for every condition is very unlikely. Fortunately, for most of the microfilm rolls tested (over two hundred), automated frame detection identifies more than 99% of the documents accurately. Additionally, this frame detection strategy works well on rolls which give current "detect as you scan" frame detection problems (such as variable-sized documents, near-overlapping frames, and changes in microfilm quality partway through the scan). By improving frame detection, both the quantity and quality of digitally represented microfilm documents can be improved.

Future Directions – The Quest for the Perfect Frame

Despite all the success of the current frame detection strategy, there is much more work to be done. Overlapping documents, for example, are particularly difficult to detect with the current system. There are, of course, other unusual conditions where improved frame detection could make the system more effective.

Additionally, there is more work that can be in a post-processing stage. Statistics can be created about each frame allowing for more sophisticated operations such as frame quality analysis, automated frame adjustment, document classification, and form registration. Documents which have been correctly framed can be copied from the ribbon and saved off in a wide variety of image formats. Any data which can be gained at this point about the frame's contents can be very useful for indexing and storage operations.