Blur Detection for Historical Document Images

Ben Baker
FamilySearch
bakerb@familysearch.org

ABSTRACT

FamilySearch captures millions of digital images annually using digital cameras at sites throughout the world. The top image quality problem encountered during this image capture process is blurriness due to an out of focus camera and/or motion during capture. Several automatic measurements of digital image blur exist, but have been found to not be very accurate at correctly classifying blurry images of historical documents.

A new blur detection scheme is presented in this paper that leverages the characteristics of historical documents to more effectively classify blurry images. The proposed technique attempts a best fit of edge data to the logistic function, resulting in a growth rate parameter that can be used as a measurement of image blur. Experimental results demonstrate the effectiveness of the proposed measurement at accurately detecting image blur.

1. INTRODUCTION

Two primary methods for obtaining digital images of historical documents are used at FamilySearch. The first method creates digital images from scanning microfilm at the Granite Mountain Records Vault. The second method employs digital cameras at locations throughout the world. This paper only considers the images captured via digital cameras, although the same method may apply to images scanned from microfilm as well.

During capture, operators of digital cameras manually focus the camera to ensure as sharp of images as possible. Normal capture conditions create a very shallow depth of field and a relatively long exposure time.

These capture conditions sometimes lead to situations where images become blurry during capture. Some of the potential causes of image blur include:

- **An out of focus camera** – Improper focus calibration by the camera operator may lead to suboptimal focus.
- **Distance from camera to document changed** – May be caused by moving the camera or document. Includes slight changes due to turning the pages of a book. Operators periodically reevaluate camera focus since changes in distance may have moved the document outside the desired depth of field.
- **Motion during capture** – Camera or document movement may occur, including the settling of a page in a book after it is turned.

After images have been captured, a statistical sample audit mechanism is employed for quality control of camera captured images. Human auditors check images for a variety of abnormalities to ensure high quality and readable images are published. Images that fail audit are sent back to the camera operators for rework and may be recaptured to correct problems.

The internal standard for image blur to which images are held to during the image audit is that no more than two transitional pixels may exist in any direction across a high contrast boundary between foreground and background [1].

During the 2011 calendar year, the majority of image quality failures discovered during image audit was due to image blur during image capture. Table 1 shows a high level breakdown of the reasons images failed during image audit.

Table 1 – 2011 Image Audit Quality Control Failures

Failure Reason	Percent of Folders
Blurring and Focus Issues	60.7%
All Other Issues Combined	39.3%

Because of the high percent of quality control failures due to image blurriness, it is very desirable to obtain a measure to accurately identify the blurriness of an image automatically. Such a measure could be utilized at capture time to greatly reduce the time taken to recapture blurry images and increase overall image capture throughput.

Several existing blur measures were attempted to solve this problem, but were deemed unsuitable for the types of images processed at FamilySearch. The blur detection scheme by Wong in [2] did a poor job of selecting relevant line segments with its Hough transform based method and gradient calculation often gave inconsistent results on historical documents.

Other edge-based blur measures were attempted, but as Chung, Chang, Wang and Chen pointed out in [3] the method of Marziliano, Dufaux, Winkler and Ebrahimi [4] often computed incorrect edge widths, plus it was very sensitive to the thickness of the text in the document and the spatial resolution it was captured at.

A novel method for blur detection using singular value decomposition (SVD) was proposed by Su, Lu and Tan [5]. This method was implemented, but found to be overly affected by the amount of text within an image.

All of these measures were designed to work with natural scene images. However, historical document images possess the unique traits of being relatively bitonal in nature with dark foreground text on a lighter page background with documents framed in an image in relatively regular ways. Therefore, development of a

novel blur measure to detect image blur was undertaken to leverage these traits to work more effectively on historical document images. The proposed measure is described in the sections that follow.

2. PROPOSED BLUR DETECTION SCHEME

At a high level, the algorithm can be summarized by the following steps. Any color images are converted to grayscale prior to performing these steps.

1. Crop images to remove page borders from consideration.
2. Perform edge detection to find all horizontal and vertical edges meeting certain criteria.
3. Iterate to find the best fit of edge data to the logistic function.
4. Compute measures of horizontal, vertical and overall blur by averaging the growth rate values obtained for each applicable edge.

2.1 Cropping

Images are cropped to remove page borders from consideration. This is because there are often characteristics along these edges that make them unsuitable for measuring blur. For example, historical documents often have book bindings and/or worn page edges which interfere with accurate estimates of blur.

Figure 1 illustrates an example of how the algorithm crops the image to only consider the inner area of the document where the text is. This area is the prime area of consideration because the sharpness of the text is ultimately what leads to better readability of the document. Blur of textual information may make it difficult to read and is therefore a condition to be avoided.

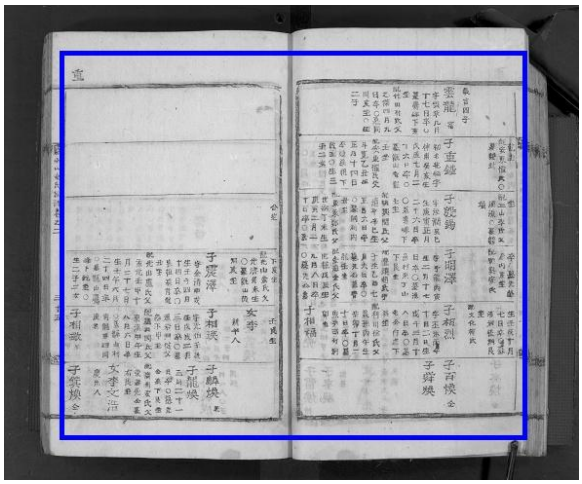

Figure 1 – Illustration of Cropping to Remove Page Borders

2.2 Edge Detection

After the image has been cropped to remove page borders from consideration, edges are detected using the Sobel edge detection algorithm in both the horizontal and vertical directions

It is simpler to only consider horizontal and vertical edges during this step. This also correlates with how human auditors typically currently measure transitional pixels via observation.

While only considering horizontal and vertical edges eliminates some edges from consideration that may be good candidates, this method finds more relevant edge segments than the method by Wong in [2] and is much faster.

Edges must meet three criteria that are controlled by parameters to be considered:

1. Minimum length of line segment perpendicular to the edge
2. Minimum contrast from highest to lowest intensity
3. Monotonically increasing/decreasing across the edge

Figure 2 illustrates edges that were detected in a sample image. As long as the minimum length of line segments is fairly small, the algorithm can find edges across irregular handwriting text.

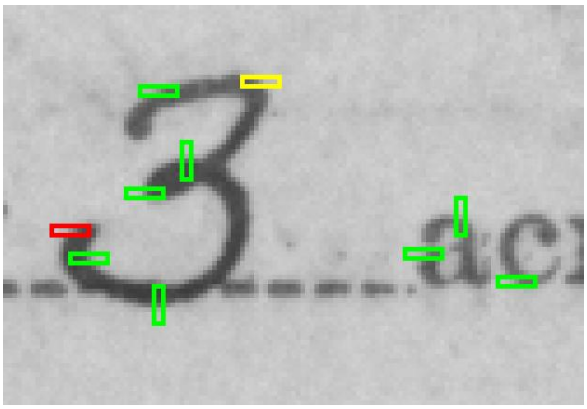

Figure 2 – Edges Detected Within a Sample Image

The minimum contrast criterion eliminates low-contrast edges that adversely affect the results. As stated in the digital image specification used by FamilySearch [1], high contrast boundaries between foreground and background are the best places to measure blur.

Only edges that are monotonically increasing or decreasing are considered to eliminate issues when an edge crosses over another edge or has noise that makes a blur measurement more difficult.

2.3 Best Fit to the Logistic Function

Other blur measurements such as those by Marziliano et al. [3] and Chung et al. [4] consider both sides of the edge when measuring image blur. Due to the high variability of thickness in text and spatial resolution of historical document images, the decision to use only a single sided edge or G-step structure edge as described by Tong, Li, Zhang and Zhang in [6] was made.

Typically, historical documents consist of dark foreground text on a lighter background and the transition is expected to be fairly uniform from background to foreground. The logistic function is

a sigmoid curve that is often used to model natural phenomena such as population growth. It is visualized as an "S-shaped" curve and models the expected transition from background to foreground. Therefore, after edges have been obtained, intensity values across the edge can be best fit to the logistic function to obtain an estimate of blur.

The generalized logistic function is defined as:

$$Y(t) = A + \frac{K-A}{\left(1+ Qe^{-B(t-M)}\right)^{1/v}} \quad (1)$$

where A is the lower asymptote and K is the upper asymptote. Other parameters include Q which determines the value at Y(0), B which is the growth rate of the curve, M which is the time of maximum growth if Q=v and v which affects near which asymptote maximum growth occurs.

Intensity values are normalized to the 0-255 range to reduce the effect of contrast differences across different edges. Doing this means that the lower asymptote A = 0 and the upper asymptote K = 255. For simplicity, parameters Q, M and v are set to default values Q = 1, M = 0 and v = 1, respectively. Using these parameter values results in the simplified logistic function equation:

$$Y(t) = \frac{255}{1+ e^{-Bt}} \quad (2)$$

The simplified equation permits solving for the growth rate parameter B across several values of t. The growth rate represents the "slope" of the curve with lower values corresponding to a more gradual increase of values and higher values a steeper increase. See Figure 3 for example curves for values of B from 1.0 to 2.0.

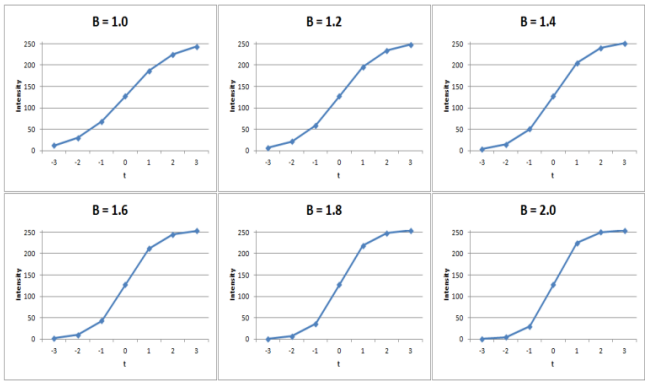

Figure 3 – Logistic Curves at Various Growth Rates

The intention of solving for the growth rate is to determine the spread of the edge or how many transitional pixels are present across the edge, thus providing a measurement of the blurriness of each edge.

To solve for the growth rate B, an iterative method is employed that seeks to find the growth rate where the sum of squared error (SSE) from the logistic curve is minimized.

Edges are rejected whose minimum SSE is higher than a specified threshold. This is so edges that do not match the expected function do not affect the overall results.

As an example, consider the edge highlighted in Figure 4. The intensity values of the edge are first normalized to the 0-255 range. Iteration finds that a growth rate of 1.90 is the best fit of the normalized intensities to the expected intensities from Function 2.

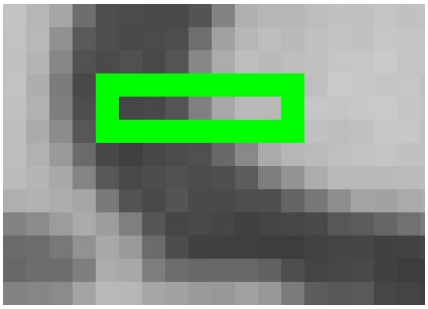

Figure 4 – Example Edge Within an Image

Table 2 shows the data used to calculate the blur measure for this edge. The fit to the logistic function was fairly close, with a sum of squared error of only 65.78. This close fit can be visualized in the graph in Figure 5.

Table 2 – Data Used to Measure Blur of the Edge in Figure 4.

t	Measured Intensities	Normalized Intensities	Expected Intensities (B=1.90)	Squared Error
-3	71	0	0.85	0.72
-2	73	4.47	5.58	1.22
-1	85	31.32	33.18	3.47
0	129	129.74	127.50	5.00
1	168	216.97	221.82	23.51
2	185	255	249.42	31.13
3	185	255	254.15	0.72

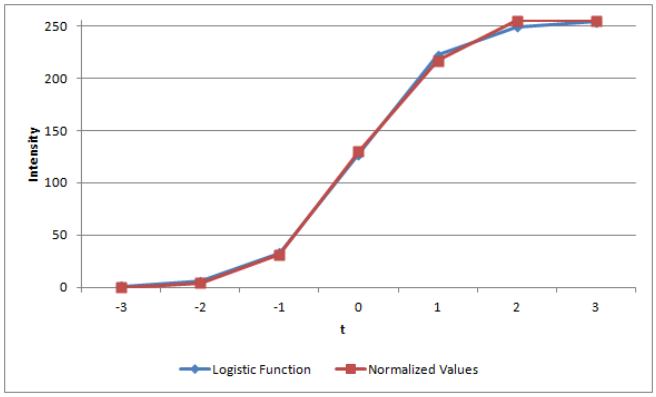

Figure 5 – Graph Showing Best Fit of Edge in Figure 4

2.4 Compute Blur Measures

After the growth rate has been calculated for each edge in the image, blur measures may be computed for the image. This is done by computing the mean of the growth rates in each direction and for all edges. For example, the horizontal blur measure β_h is computed by adding the growth rates for all horizontal edges and dividing by the number of horizontal edges, n_h.

$$\beta_h = \frac{\sum_{i=1}^{n_h} B_i}{n_h} \quad (3)$$

Similarly, the vertical blur measure β_v is calculated as the mean of $n_v$vertical edge growth rates.

$$\beta_v = \frac{\sum_{j=1}^{n_v} B_j}{n_v} \quad (4)$$

An overall blur measure is calculated as the sum of all growth rate values divided by the total number of edges.

$$\beta_{overall} = \frac{\sum_{i=1}^{n_h} B_i + \sum_{j=1}^{n_v} B_j}{n_h + n_v} \quad (5)$$

3. EXPERIMENTAL RESULTS

3.1 Manually Blurred Images

To illustrate how well the algorithm performs at detecting image blur, an image was selected and derivative images produced manually with varying levels of blur using GIMP.

First, Gaussian blur was applied to the original image, using increasing blur radii to produce more severe blur. Table 3 shows that as the blur radius increases, the blur measures decrease as expected.

Table 3 – Effect of Gaussian Blur

	β_h	β_v	$\beta_{overall}$
Original Image	1.79	1.85	1.83
Gaussian Blur (radius = 1.0)	1.60	1.65	1.63
Gaussian Blur (radius = 2.0)	1.36	1.40	1.38
Gaussian Blur (radius = 3.0)	1.24	1.27	1.26
Gaussian Blur (radius = 4.0)	1.15	1.18	1.17
Gaussian Blur (radius = 5.0)	1.09	1.13	1.12

Motion blur of increasing strength was also applied to the image. The motion blur was applied in both horizontal and vertical directions. Table 4 shows that as the unidirectional blur length increases, the blur measure in that direction decreases accordingly.

Table 4 – Effect of Horizontal and Vertical Motion Blur

	β_h	β_v	$\beta_{overall}$
Original Image	1.79	1.85	1.83
Horizontal Blur (length = 2)	1.64	1.80	1.72
Horizontal Blur (length = 5)	1.11	1.61	1.29
Horizontal Blur (length = 10)	1.00	1.53	1.49
Vertical Blur (length = 2)	1.76	1.71	1.72
Vertical Blur (length = 5)	1.58	1.10	1.27
Vertical Blur (length = 10)	1.43	1.00	1.37

These results suggest that the measures can detect both uniform and unidirectional blur between images with the same contents.

3.2 Production Images

To further validate the algorithm's ability to differentiate between blurred and non-blurred images in production, a test set was created from images selected from day to day operations. The test set consists of 100 images that human auditors failed for blur/focus issues and a set of 100 randomly selected images that passed image audit.

The blur measures were then computed across all images in the data set. Visualization of the overall blur measure $\beta_{overall}$ for the images in the data set is shown in Figure 6.

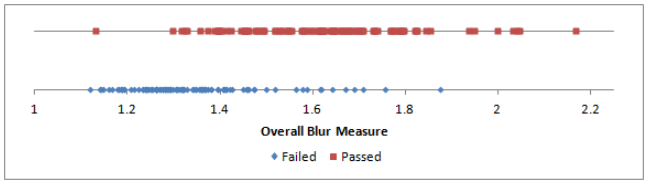

Figure 6 – Distribution of Images in Test Data Set

A general separation between the images that passed audit and those that failed is evident, but there is also a lot of overlap. Using an overall blur measure of 1.44, 81% of the failed images have an overall blur measure less than this and 84% of the images that passed image audit have values above this.

Some of the reasons for the wide distributions include a large variety of characteristics within the images and differences in auditor classifications. Certain types of images give the measurement trouble including blank or nearly blank documents, inherently blurry text and low contrast images. Future improvements on the algorithm should handle these cases more effectively.

Another explanation for some of the results was a small number of edges used for the measurement skewing the result in a particular direction. An extreme example of one outlier in the images that passed audit was a completely blank page where no edges were found and the measure could not be computed at all.

One observation that was made while examining the images was that motion blur seems to affect readability more than an image being out of focus. It is therefore useful to examine the β_h and β_v measures independently. Figure 7 plots the horizontal and vertical blur measures for images in the test data set. Points a great distance from a slope of 1 in the graph potentially have motion blur.

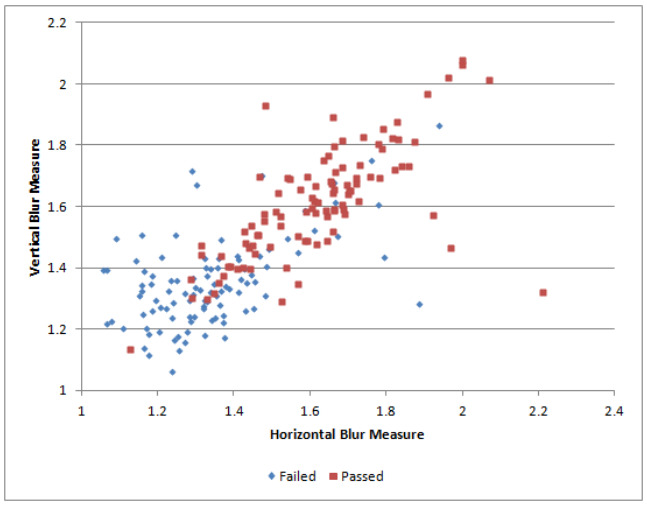

Figure 7 – Horizontal and Vertical Blur Measures of the Test Data Set

Some of the outliers are not really due to motion blur, but are due to an inability of the algorithm to find many relevant edges. For example, the image with β_h = 2.21 and β_v = 1.32 is an image with a blank page with rule lines. The rule lines resulted in a strong signal with many good edges in one direction and a weak signal with very few edges in the other.

3.3 Trend Over Time

One other item that was examined was the change in the blur measure as images are being captured. Some documents exhibited definitive trends and jumps. It is believed that the slight upward or downward trends are due to slight changes in the sharpness of the image due to the change in distance from the camera as pages are turned in a book. The jumps are believed to be when the camera operators readjust the focus settings periodically.

As an example, Figure 8 shows a slight upward trend in the first third of the images captured, then a large jump and a slight downward trend, followed by another jump and a slight downward trend. It is believed that this blur measure can be used during capture to assist camera operators in knowing when the documents they are imaging have fallen out of focus and whether their periodic adjustments have helped or hurt the overall sharpness of the images being captured.

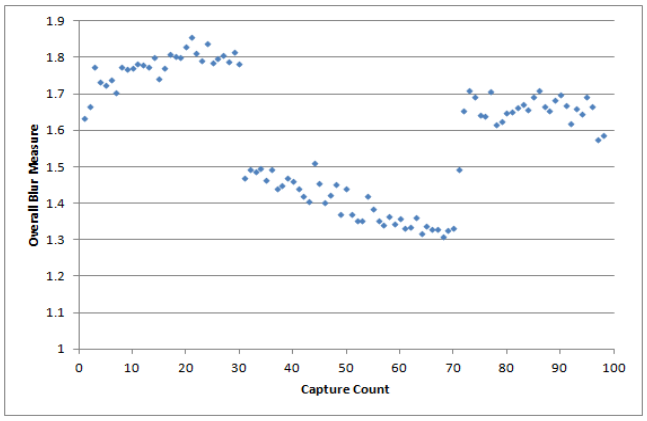

Figure 8 – Changes in Overall Blur Measure Over Time

4. CONCLUSIONS AND FUTURE WORK

This paper has demonstrated a blur detection scheme that is effective in many cases at detecting image blur in a wide variety of historical document images. The potential to identify slight changes in the sharpness of images as they are being captured has also been established. While promising, further refinement is in order to make the algorithm more effective and useful on a large scale.

For example, better detection of blank and nearly blank documents would help improve the robustness of the algorithm. Development of a true ground truth data set and additional testing with camera operators and auditors on production images would also help refine the algorithm.

Utilizing additional parameters of the generalized logistic function to more accurately fit more edges may also help obtain better classification by getting closer fits of edge data to the logistic function, resulting in more accurate estimates of blur.

5. REFERENCES

[1] Family History Department Digital Image Specification. Version 3.32 [Internal Standard]

[2] *Image Blur Detection via Hough Transform.* **Kerry Wong** [Online] http://www.kerrywong.com/2009/06/19/image-blur-detection-via-hough-transform-i/

[3] *A No-Reference Perceptual Blur Metric.* **Pina Marziliano, Frederic Dufaux, Stefan Winkler and Touradj Ebrahimi.**

[4] *An Edge Analysis Based Blur Measure for Image Processing Applications.* **Yun-Chung Chung, Shyang-Lih Chang, Jung-Ming Wang and Sei-Wang Chen.**

[5] *Blurred Image Region Detection and Classification.* **Bolan Su, Shijian Lu and Chew Lim Tan.**

[6] *Blur Detection for Digital Images Using Wavelet Transform.* **Hanghang Tong, Mingjing Li, Hongjiang Zhang and Changshui Zhang.**