Improved Blur Detection of Historical Document Images Using a Neural Network

Ben Baker
FamilySearch
bakerb@familysearch.org

ABSTRACT

At the 2012 Family History Technology Workshop, the author presented a novel method for blur detection of historical document images [1]. After a few improvements from the original paper, this method was eventually incorporated into FamilySearch operations. However, camera operators using the feature found that this method indicated too many false positives and the feature was no longer used for several years.

Since the original paper was published, FamilySearch has continued to capture millions of digital images annually using digital cameras at sites throughout the world. Blurriness due to an out of focus camera and/or motion during capture remain the top image quality problems requiring expensive rework. Therefore, priority was given to again examine a method for automatically detecting blurry and out of focus images.

This paper documents using the metrics from the previous blur detection method as input features to a neural network to produce improved classification accuracy over the previous method. The level of false positives has been vastly reduced and overall accuracy has increased dramatically. Specifically, the weighted percent of images being correctly identified as "good" or "bad" has increased from about 87% to over 97% and the weighted percent of "bad" images caught has increased from under 58% to nearly 83%. This new implementation is currently being tested with camera operators and feedback is directing potential further improvements.

1. INTRODUCTION

Since the original paper was published in 2012, several improvements were made to the original metric and an implementation was released as part of the DCam digital camera capture software used to capture images at FamilySearch. Despite correctly identifying many blurry and out of focus images, camera operators found the number of false positives was too high and slowed image capture too much, so the experimental feature was abandoned until 2017.

1.1 Image Audit Quality Control Failures

Analysis presented in the 2012 paper showed that 60.7% of all image audit quality control failures were due to blurring and focus issues. This analysis was repeated for the 2016 calendar year to find a majority of images failing image audit remain due to these types of failures. (Table 1)

Table 1 – 2016 Image Audit Quality Control Failures

Failure Reason	Percent of Images
Blurring and Focus Issues	51.4%
All Other Issues Combined	48.6%

Because images that fail image audit require expensive rework to correct the issues, it remains desirable to have an effective and automatic way to identify blurry and out of focus images. In addition, FamilySearch has also been investigating automated methods for other image audit quality control checks to reduce the need for human auditing of images, so manpower can be directed to higher value tasks.

1.2 Implementation and Improvements

In the original paper, it was suggested that an overall blur measure of 1.44 could correctly identify 81% of failed images and 84% of passing images. While an implementation to compute this measure existed in the FamilySearch imaging library, it was known that this simple characterization was insufficient for camera operations.

The author changed teams within FamilySearch shortly after publication of the original paper in 2012 and was unable to continue work on this project. However, in 2013 Alan Cannaday and others made several improvements and completed an experimental implementation in FamilySearch's imaging library and DCam digital camera capture software. [2]

1.2.1 M-Shift Addition to Logistic Function

The first improvement was to introduce M-shift to the logistic function to better approximate edges. This was because the original implementation assumed edges are centered across pixel lines which is not true in many cases. The introduction of the variable M from the general logistic function reduced the average mean squared error by 68.4%. [2] (Equation 1)

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

1.2.2 Gaussian Mixture Model

The second improvement made to the original metric was to train a Gaussian Mixture Model (GMM) to better fit the data. This model used the horizontal and vertical blur metrics and divided the images into sets based on size. This GMM could then make a pass/fail classification decision based on the horizontal and vertical blur

measures for a given image. [2] A visualization of the GMM distribution for 11 to 16-megapixel images is show in Figure 1.

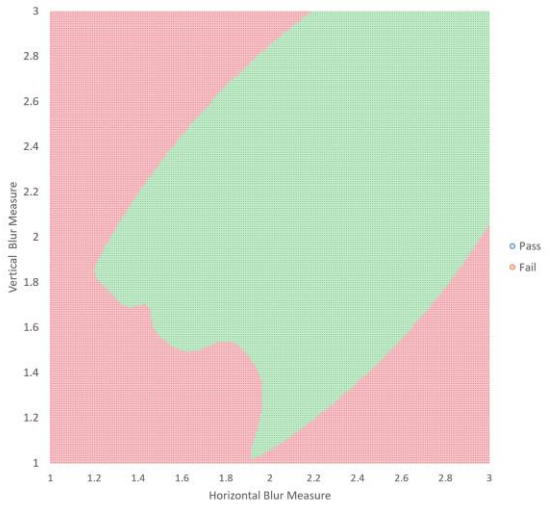

Figure 1 – Example visualization of GMM distribution

1.2.3 Implementation Results

An implementation with these two improvements was integrated with the DCam digital camera capture software in 2013 as an experimental feature. Camera operators stopped using this feature because they found too many images that were not blurry or out of focus were flagged as such (false positives).

Analysis done in conjunction with this paper found that this implementation did not solve the problem of images of blank pages, only correctly identified about 91% of in focus images and only caught less than 58% of "bad" images. (Table 2)

Table 2 – GMM Implementation Results

Classification	Percent Correctly Identified
Blank	23.19%
In Focus	90.97%
Horizontal Blur	48.6%
Vertical Blur	50.25%
Out of Focus	76.86%

In this paper, "good" images are defined as Blank or In Focus images and "bad" images are blurry in either direction or out of focus and should require recapture. Precision and recall like scores are computed against a test set, where values are weighted according to the class distribution in captured images.

The weights used were derived from frequency of occurrence in audit results. (Table 3) Since vertical and horizontal blur are not differentiated by image auditors, the frequency of blurry images was simply divided in half. The percent of blank images was estimated from a large sample of images, manually counting the number of blank pages and providing an estimate across all images.

Table 3 – Frequency of Occurrence of Image Classes

Classification	Frequency
Blank	5.0%
In Focus	94.5601%
Horizontal Blur	0.15257%
Vertical Blur	0.15257%
Out of Focus	0.13476%

A precision-like score is computed as the weighted percent of correctly identified images. In addition, a recall-like score is computed as the weighted percent of blurry and out of focus images that were correctly identified as "bad". These two measures are combined into an F-score to compare different algorithms. These measurements serve as the baseline for the initial implementation. (Table 4)

Table 4 – Precision and Recall for GMM Classifier

Precision-like score	87.44%
Recall-like score	57.79%
F-score	69.59%

2. BLUR DETECTION METHODOLOGY

Since it is desirable to classify images into more than two classes, an alternate approach must be used. It was decided to use a multi-layer perceptron neural network with a single hidden later with metrics from the blur detection algorithm as input features. Images are classified into one of five classes:

1. Blank page
2. In focus
3. Horizontal blur
4. Vertical blur
5. Out of focus

Horizontal and vertical blur are separate classes because the blur metrics used as inputs differentiate between these two types. For example, an image with a low horizontal blur measure and a high vertical blur measure is often blurry in the horizontal direction and vice versa.

2.1 Input Features

Because metrics from the blur detection scheme did a reasonable job of discriminating blurry and out of focus images from in focus images, it was decided to use these metrics as the basis for inputs into the neural network.

Blur metrics are computed in the horizontal and vertical directions as well as combined. Additional metrics were produced in the original implementation and found to increase accuracy in the resulting neural network. The following seven metrics are computed in both directions and combined for a total of 21 inputs.

1. Blur measure
2. Edge count
3. Standard deviation of the sharpness of all edges
4. Mean squared error when fitting to the logistic function
5. Standard deviation of the squared errors of all edges
6. Mean contrast across all edges
7. Standard deviation of the contrast across all edges

Images were also subdivided into 16 subsections and blur metrics computed within each section. Adding the blur measure and edge count in horizontal, vertical and combined directions for each subsection results in an additional 96 inputs. These additional inputs improved overall accuracy. It is believed this is because often motion blur manifests itself in localized areas of an image.

Finally, three more metrics were added as inputs to further discriminate the classes for a total of 120 inputs. The standard deviation of pixel intensities was particularly helpful at improving the classifier's ability to correctly detect images of blank pages.

1. Number of color channels in the image (1 or 3)
2. Total number of pixels
3. Standard deviation of pixel intensities throughout the image

2.2 Neural Network Topology

The Waikato Environment for Knowledge Analysis (weka) was used to train and test the neural network. Weights from the resulting neural network were used in the FamilySearch imaging library which was integrated into FamilySearch's DCam software used as capture time.

The topology of the neural network had the 120 inputs as described previously. A fully connected neural network with a single hidden layer of 62 nodes was used. This number of hidden nodes comes from the default 'a' option in weka which uses the number of (attributes + classes) / 2.

The sigmoid activation function was used in the neural network. Five output nodes corresponding to each desired class are used as output with the values normalized to sum to 1 and the highest value indicating the predicted class.

3. EXPERIMENTAL RESULTS

3.1 Training

Training of the neural network was performed on 2,331 images gathered from the output of human image auditors and further refined into a truth set. Results from the training are shown below (Table 5)

Table 5 – Training Results

Classification	Percent Correctly Identified
Blank	99.36%
In Focus	97.19%
Horizontal Blur	86.86%
Vertical Blur	80.86%
Out of Focus	81.14%

3.2 Testing

For testing of the neural network's ability to correctly predict image classes, an additional 775 labeled images were used. (Table 6)

Table 6 – Testing Results

Classification	Percent Correctly Identified
Blank	88.24%
In Focus	96.39%
Horizontal Blur	71.43%
Vertical Blur	73.58%
Out of Focus	55.56%

Computing an F-score in the same manner as described in Section 1.2.3 results in the following values. (Table 7)

Table 7 – Precision and Recall on Test Set

Precision-like score	95.86%
Recall-like score	67.31%
F-score	79.09%

While the neural network performs significantly better than the GMM implementation, there is still a fair amount of error. However, the above results only consider when the correct class is predicted. If the predicted class is considered correct when it matches a "good" designation for Blank and In Focus images or a "bad" designation for Horizontal Blur, Vertical Blur or Out of Focus, the results improve significantly. (Table 8)

Table 8 – Testing Results Identifying "Good"/ "Bad" Images

Classification	Percent Correctly Identified as "good" or "bad"
Blank	94.12%
In Focus	97.53%
Horizontal Blur	83.93%
Vertical Blur	84.91%
Out of Focus	79.63%

When comparing this binary classification against the previous GMM based approach, results are improved across the board, particularly for Blank images, which was the primary source of false positives previously. (Table 9)

Table 9 – Percent Correctly Identified Comparison

Classification	GMM Implementation	Neural Network Implementation
Blank	23.19%	94.12%
In Focus	90.97%	97.53%
Horizontal Blur	48.6%	83.93%
Vertical Blur	50.25%	84.91%
Out of Focus	76.86%	79.63%

Similarly, the precision, recall and F-score for the neural network implementation is significantly better. (Table 10)

Table 10 – Precision and Recall Comparison

	GMM Implementation	Neural Network Implementation
Precision-like score	87.44%	97.30%
Recall-like score	57.79%	82.95%
F-score	69.59%	89.55%

3.3 Production Implementation

Although this neural network implementation has been integrated into the DCam camera captures software, the feature is still behind an experiment and not available to all camera operators by default yet. The current implementation is such that the workflow of the operator is stopped until they correct the problem. While the blur detection can be computed quickly enough to not slow down the desired rate of 3 seconds per image captured, the interruption of an operator's workflow is costlier. Alternatives are being investigated for UI and workflow changes to alleviate this disruption.

Qualitative feedback from selected camera operators indicates that while this method of blur detection is significantly better than before, it is perhaps still not good enough. Camera operators have reported that certain classes of images such as typewritten documents with bleed out and documents with blue or purple rubber stamps often still confuse the classifier.

It should be noted that each camera operator often captures hundreds or even thousands of images per day. Therefore, the current false positive rate of about 3% means the operator will be falsely alerted on 3 out of every 100 images captured. Also, certain types of collections containing problematic images result in even higher false positive rates.

4. CONCLUSIONS AND FUTURE WORK

This paper has demonstrated an improved blur detection scheme that is more effective detecting image blur and focus problems than the previous one. It shows great promise for automatic image quality assessment and to reduce costly rework.

Several possible improvements or alternate solutions may further improve the ability to automatically detect blurry and out of focus images, including:

1. More labeled images for training and testing.

   Additional images have been and continue to be gathered from image quality audit to build up the truth set. However, these images have not yet been used to train an improved classifier because the author no longer works on the imaging team responsible for this feature.

2. Add one or more classes to deal with bleed out, rubber stamps or other problematic images.

   The addition of a blank page class and features to detect it greatly improved the ability to detect those images and the same could likely be done with these types of images.

3. Using a deeper neural network with a different activation function such as Relu instead of the sigmoid function.

   The author attempted using deeper neural networks than the single hidden layer one used, but found the results did not improve significantly. This was likely due to the use of the sigmoid activation function which has the vanishing gradient problem.

4. Using a Convolutional Neural Net (CNN) based approach

   Since blur detection is performed on images, a Convolutional Neural Network (CNN) may be well suited to detecting the difference between blurry, out of focus and in focus images.

5. REFERENCES

[1] *Blur Detection for Historical Document Images*. **Ben Baker.** [Online] https://fhtw.byu.edu/static/conf/2012/baker-blur-fhtw2012.pdf

[2] *Findings Paper for Research Relevant to Edge Sharpness for Historical Documents During 2013*. **Alan Cannaday.** FamilySearch Research Team Internal Publication.