ABSTRACT
We propose a novel algorithm for improving the quality of binarized document images captured with a mobile device under low light conditions. In such scenarios, images captured without a flash often result in blur artifacts and poor signal-to-noise ratio, while images taken with the flash may produce information loss due to specular reflection in a localized region termed a “flash spot”. Our algorithm automatically triggers the capture of a pair of images, one with and one without flash, in rapid succession. The flash spot region (FSR) is first detected. The two images are then accurately aligned within the FSR using a multiscale alignment technique. Finally the images are binarized and fused in the vicinity of the FSR using an intelligent technique that minimizes fusion boundary artifacts. The result is a binary image that is largely identical to the binarized flash image, except within the FSR where content from the no-flash image is selectively incorporated. To our knowledge this is the first attempt to employ flash/no-flash fusion to improve binarization of mobile document images. Results show superior qualitative and quantitative performance of the proposed algorithm when compared with standard binarization applied to either the flash or no-flash image.

Index Terms—document imaging, mobile, smartphone, binarization, flash, fusion

1. INTRODUCTION
Mobile devices such as smartphones are being increasingly used to capture images of documents such as tax forms, insurance claims, bank transactions, receipts, etc. A key function in such applications is binarization that converts a color or grayscale document image to a bi-tonal (black-and-white) image. Binarization can significantly reduce image size, and is also a precursor for most optical character recognition (OCR) algorithms. While considerable strides have been made in document image binarization, significant challenges remain in the scenario of mobile capture, wherein images can be captured under a variety of environmental conditions. A common example is capture under low-light conditions that can produce images of low contrast and signal-to-noise ratio. Binarization of such images results in broken, fragmented, or connected characters that result in poor readability and OCR performance, as shown in Fig. 1(a). The use of a flash can significantly improve overall capture quality. However the strongly directed flash illumination often produces a specular reflection resulting in a “flash spot” wherein the printed content is no longer discernable, as shown in Fig. 1(b).

Figure 1(a): Non-flash image (left), and binarized output (right); Fig. 1(b) Flash image (left) and binarized output (right). Both binarized versions are zoomed around the FSR. This artifact may result in the loss of a very critical piece of information in a document (for example the monetary amount on a check).

We propose a novel algorithm that automatically captures two images – one with and one without flash in rapid succession, detects the flash spot region (FSR) in the flash image \( I_f \), aligns the two images within the FSR, and binarizes and fuses both images. The fusion step intelligently defines a modified FSR that avoids cutting through characters or words in the document. The result is a binary image that is largely identical to the binarized flash image, except within the FSR where content from the no-flash image \( I_{nf} \) is incorporated.

2. RELATED ART
There have been several efforts to enhance images captured by mobile devices such as smartphones [21, 25]. Many existing methods are applicable only to natural scene images \([16-25]\) and do not generalize well to document images, due to their very different statistical nature and purpose.
In the domain of document imaging, many researchers have explored the binarization problem as a means to both enhance and compress images [1-11]. Techniques can be broadly categorized into global [1-3] and local [4-11] techniques. In general, local methods are preferred for document images as they possess the ability to adapt to spatially varying illumination and shadow effects. All these methods focus on optimizing the binarization process on a single input image. In contrast, in our method, the capture and fusion of two images is a critical and novel contribution that improves binarization quality beyond the conventional methods.

Flash/no-flash (FnF) image fusion has garnered recent interest [16-25]. Petschnigg et al. [16] demonstrated several applications of FnF fusion for synthesizing images that are of higher quality than either of the originals. Agarwal et al. [17] compared image gradients at corresponding locations in the two images to obtain the fused image. We note that methods based on image gradients are generally not well suited for document images. Alenius et al. [18] used principal component analysis (PCA) to find optimal fusing factors for the images in order to compute the resulting image. The focus of these methods has been on transferring high-frequency detail from the flash image while preserving overall tone and color rendition from the no-flash image for natural scene settings. Our goal is to obtain a high quality document image that combines the positive attributes from flash and no-flash images while avoiding their respective pitfalls.

A critical step in any fusion algorithm is image alignment. Existing approaches can be broadly categorized into two groups (a) Direct (pixel-based) and (b) Feature based methods [14, 15]. Many techniques are developed for natural scene images [14, 15, 28]. However, document images demand extremely high alignment accuracy since small errors can have significant effect on human and machine readability. Furthermore, many approaches are too computationally expensive for a mobile implementation. Recently, for example, Varjo et al. [20] proposed an image alignment method for FnF fusion based on mutual information. While yielding high accuracy, the computational complexity of this approach is high making it infeasible for real-time fusion in mobile applications. We propose a fast and accurate technique that performs alignment only in the FSR via a multiscale cross-correlation based approach.

3. FLASH/NO-FLASH BINARIZATION ALGORITHM

A pair of FnF images of a document is captured with a mobile camera. In a preprocessing step, the borders of the document are detected, the document region is cropped, and a homographic (perspective) transform is applied to map the document region to a rectangle of the correct aspect ratio using standard image processing techniques. The result is an image as shown in Fig. 1. The flow diagram of the proposed algorithm is shown in Fig. 2. The sections below elaborate on each of the main steps.

\[
\begin{align*}
& I_f \\
& I_{nf} \\
& \text{FSR detection} \\
& \text{FSR content detection} \\
& \text{Sufficient differential content?} \\
& \text{Yes} \\
& \text{No} \\
& \text{FSR Alignment} \\
& \text{Binarize } I_f \\
& \text{Image Fusion} \\
& I_f, FSR(I_{nf}) \\
& q_{fsr}
\end{align*}
\]

Figure 2. Flow diagram of proposed flash/no-flash fusion algorithm. FSR stands for flash spot region.

3.1 Flash spot detection

In order to localize the FSR, we employ a two-pass window based approach and count the number of pixels within a sliding window with brightness greater than a threshold. This threshold is the \( p^{th} \) percentile of brightness values over the entire flash image. In the first pass \( p \) is set to a high value (99\(^{th}\) percentile) to capture very bright windows and in the second pass, \( p \) is set to a lower value (98\(^{th}\) percentile) and a window is selected only if it is adjacent to a window in the first pass. This helps in accurately localizing the FSR while being computationally efficient. In the next step we employ additional logic within the FSR to ensure that two criteria are met: i) there is significant loss in content in the flash image and ii) there is sufficient content in the no-flash image that can be transferred to the flash image. To this end, we compute a gradient-based image content measure proposed in [12] for each window within the FSR of both \( I_f \) and \( I_{nf} \). Horizontal and vertical gradients \( G_x \) and \( G_y \) are computed and accumulated into an \( N \times 2 \) matrix \( \mathbf{M} \), where \( N \) is the number of pixels in the window. Next the Singular Value Decomposition (SVD) of \( \mathbf{M} \) is computed. If \( s_1 \) and \( s_2 \) are singular values (\( s_1 \geq s_2 \) ) of \( \mathbf{M} \), we compute a measure of image content using the Eqn. 1:

\[
q = \frac{s_1 - s_2}{s_1 + s_2} s_1
\]

The q scores are summed within the FSR to produce a final measure \( q_{fsr} \). In order to perform fusion we require that \( q_{fsr} \) within \( I_f \) is less than a first threshold, while \( q_{fsr} \) within \( I_{nf} \) is greater than a second threshold. Both thresholds are determined empirically.

3.2 Image alignment

For effective image fusion, it is critical that the FnF images are aligned to a fraction of a line width. A multi-resolution alignment technique is used to do this quickly and
accurately only within the FSR (Fig. 3). Since the dynamic ranges of \( I_f \) and \( I_{nf} \) can be very different, histogram equalization is first applied on both images. The two images are then subsampled successively by a factor of 2 to produce a series of images of decreasing resolution.

Image alignment is first carried out at the lowest resolution. The process (described in detail below) involves finding the optimal warping function on \( I_{nf} \) that maximizes the cross-correlation between \( I_f \) and \( I_{nf} \). The warping function from a given resolution is applied to \( I_{nf} \) at the next higher resolution to produce \( I_{nf,a} \). This step is critical to reducing computational cost, as it greatly prunes the search space of the optimization problem. The process is repeated as the module works its way towards aligning the highest resolution images, as shown in Fig. 3.

The image warping function at a given resolution is derived as follows. \( I_f \) is split into a set of blocks, and the displacement vector \( S_i \) of the \( i^{th} \) block is found that maximizes the normalized cross correlation (NCC) [13] between that block and the corresponding block in \( I_{nf} \). Specifically, to find the peak in the NCC, a block in \( I_f \) is shifted in direction \( S = [S_x, S_y]^T \) over a specified range. The cross-correlation function \( CC(S) \) is calculated as in Eqn. 2, where \( V_f(X) \) and \( V_{nf}(X) \) are the brightness values respectively of \( I_{nf} \) and \( I_f \) at pixel location \( X \), and \( N \) is the number of pixels in the region.

\[
CC(S) = \frac{1}{N} \sum_x V_{nf}(X) V_f(X + S)
\]  
(2)

This quantity is divided by a normalization factor derived from Eqn. 3 and 4 to obtain the NCC in Eqn. 5. Normalization takes into account linear brightness variations between the flash and no-flash images.

\[
P_m = \sum_x \frac{V_{nf}(X) W_m(X)}{N}
\]
(3)

\[
P_t(S) = \sum_x \frac{V_f(X + S) W_t(V_f(X + S))}{N}
\]
(4)

\[
NCC(S) = \frac{CC(S)}{\sqrt{P_m P_t(S)}}
\]  
(5)

\( S_i \) is then the maximizer of NCC(\( S \)) for the \( i^{th} \) block. Next, a global quadratic warping function \( F(X_i, W) \) is fitted to the \( S_i \). Here \( X_i \) is the center pixel of the \( i^{th} \) block, and \( W \) is the parameter vector for a 2-D quadratic function that is obtained by minimizing a weighted fitting error as follows. For the \( i^{th} \) image block, we compute a 2×2 symmetric curvature matrix \( H \) of function \( CC(\) \) at the peak location \( S_i \) from its 8-neighbors. \( H \) indicates how fast the cross-correlation drops off as one move away from location \( S_i \). Next, let \( E_i \) denote the displacement fitting error given by:

\[
E_i = [E_{ix} E_{iy}]^T = S_i - F(X_i, W)
\]  
(6)

We weight this error with the curvature to obtain a cost function as follows:

\[
Q(i) = E_{ix}^2 H_{xx} + E_{iy}^2 H_{yy} + 2 E_{ix} E_{iy} H_{xy}
\]  
(7)

where \( H_{xx} \), \( H_{yy} \) and \( H_{xy} \) are elements of \( H \). Note that if \( H \) exhibits strong curvature in a given block, this means \( CC(S) \) is highly sensitive to small misregistration errors, and hence \( Q \) is assigned a larger weight in that block. The overall weighted fitting error for the image is given by \( Q = \sum_i Q(i) \). Finally, \( W \) is obtained by minimizing \( Q \) via standard least squares. An additional iterative procedure is used to perform outlier rejection towards a more robust fit, which we do not elaborate on given space constraints.

### 3.3 Image binarization

The aligned flash spot region of \( I_{nf} \) and \( I_f \) is binarized in the next step. In principle any binarization technique can be used. In our work we use the algorithm proposed by Sauvola [4]. In this approach the binarization threshold is determined in a local adaptive fashion for each pixel, given by Eqn. 8:

\[
T = m \times \left[ 1 + k \times \left( \frac{S}{R} - 1 \right) \right]
\]
(8)

where \( k \) is a user defined parameter, \( m \) and \( s \) are respectively the pixel mean and the standard deviation within a window of size \( w \times w \) centered on the current pixel and \( R \) is the dynamic range of standard deviation. Parameters \( k \) and \( w \) are tuned heuristically for optimal subjective image quality, and separately for flash and no-flash images.

### 3.4 Image fusion

In the final step, we fuse the content in \( I_{nf} \) with the FSR in the \( I_f \). In order to improve the quality of fused image, we first refine the boundary of FSR to avoid any splitting of characters and words in the image. Abrupt changes in the structural characteristics of words may be perceived as degradation of text, and may lead to a poor reading experience by user. We perform a morphological operation...
(opening [27]) to connect characters to form words, and obtain connected-components (CCs) in the \( I_{nf} \). We then find the CCs which are only partially included in the FSR and extend the FSR to include all such components entirely. In order to extend the FSR we expand the FSR obtained in first step (section 3.1) to a larger region before alignment and binarization. Fig. 4 shows a sample result from our dataset.

**Figure 4.** A sample result showing flash-binarized image (left) and fused-binarized image (right).

### 4. EXPERIMENT RESULTS

The aforementioned algorithm was implemented as an iOS application running on an iPhone 5. We present both quantitative and qualitative experimental results below.

**Table 1** Number of OCR character errors per page for different ambient light levels. 25 FnF pairs in each category.

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Std. dev.</th>
<th>95(^{th}) Percentile</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Low light</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Flash</td>
<td>141.9</td>
<td>70.8</td>
<td>251.0</td>
</tr>
<tr>
<td>No-flash</td>
<td>320.4</td>
<td>187.8</td>
<td>745.0</td>
</tr>
<tr>
<td>Fused</td>
<td><strong>99.8</strong></td>
<td><strong>55.7</strong></td>
<td><strong>191.0</strong></td>
</tr>
<tr>
<td><strong>Medium light</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Flash</td>
<td>86.0</td>
<td>42.6</td>
<td>172.0</td>
</tr>
<tr>
<td>No-flash</td>
<td>98.8</td>
<td>34.7</td>
<td>145.0</td>
</tr>
<tr>
<td>Fused</td>
<td><strong>78.4</strong></td>
<td><strong>35.7</strong></td>
<td><strong>138.0</strong></td>
</tr>
<tr>
<td><strong>Bright light</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Flash</td>
<td>75.4</td>
<td>36.8</td>
<td>125.0</td>
</tr>
<tr>
<td>No-flash</td>
<td><strong>32.0</strong></td>
<td><strong>29.1</strong></td>
<td><strong>89.0</strong></td>
</tr>
<tr>
<td>Fused</td>
<td>38.3</td>
<td><strong>20.5</strong></td>
<td><strong>81.0</strong></td>
</tr>
</tbody>
</table>

#### 4.1 OCR analysis

Sixteen documents comprising of single page each with different fonts and layouts (single column, two-column, etc.) were created and printed on A4 sized paper using a variety of matte and glossy inks and media. Using the iPhone, 75 FnF image pairs were captured in different lighting conditions from very low light to bright light. Additionally, we obtained a fused image for each pair of flash/no-flash capture whenever the quality requirements for fusion were met. Using ABBYY Finereader [26], we performed OCR and obtained the number of character errors per page for binarized versions of the three cases: flash, no-flash and fused. Table 1 shows the mean, standard deviation and 95\(^{th}\) percentile of the number of character errors per page. As shown, our approach is most advantageous under low light condition where it significantly outperforms both flash and no-flash outputs.

#### 4.2 Alignment error

To evaluate alignment accuracy, we computed the weighted root-mean square error (RMSE) of displacement vectors between corresponding features in \( I_f \) and \( I_{nf} \) for the cases of no alignment, and multiscale alignment. To provide a comparison with another frequently cited approach, we also evaluated the alignment based on mutual information proposed in [20]. The mean RMSEs obtained for our dataset were 0.72, 0.10 and 0.08 respectively for no alignment, mutual information, and our approach. The latter proves to be competitive with the state-of-art. As a point of reference, an RMSE of 0.08 corresponds approximately to a 2-pixel error in a 3000×3000 image, and is hard to visually discern, as exemplified in Fig. 4.

#### 4.3 Computational speed

On average, our approach takes 0.08 seconds for flash-spot detection and 0.22 seconds to align and fuse a pair of images of dimension 3200×2400 on an iPhone 5. For the same set of images, the implementation proposed in [20] takes an average of 15 seconds for alignment alone.

#### 4.4 Subjective preference

Fifteen subjects with normal vision were asked to participate in a qualitative evaluation experiment. Five original document images representing a variety of text content and formats were selected, captured with the iPhone 5, and processed through three paths i) flash + binarization; ii) no-flash + binarization; and iii) FnF fusion + binarization. For each original, the three binary images were shown on a Dell laptop display; and subjects were asked to assign a preference score from 3 (most preferred) to 1 (least preferred) based on overall quality and readability of the text. Ties were not allowed. Subjects were permitted to scroll and zoom into the images using a standard image browser. Mean preference scores of 1.44 for flash, 1.66 for no-flash and 2.89 for fused output were obtained. Out of the total of 75 rankings, the fused image was ranked first 69 times.

### 5. CONCLUSION

We presented a novel algorithm for producing high-quality binarized document images via flash/no-flash fusion. Experiments indicate that the technique provides significantly improved document quality especially under very dim light, where OCR error is reduced by approximately 25%. Computational cost is greatly reduced by limiting most of the operations within the FSR, thus enabling real-time execution on a smartphone. A critical component of the algorithm is an accurate and fast multiscale image alignment technique. Future directions include applying post-processing techniques to harmonize the structure and morphology of text characters from flash and no-flash images; and the use of more formal techniques to evaluate qualitative performance via reading and comprehension tasks.
6. REFERENCES


