An Image Processing Method that Enables Efficient Document Management and Reproduction in a Distributed Working Environment

Mu Qiao, Shutterfly Inc., 2800 Bridge Parkway, Redwood City, CA 94065

Abstract

With distributed office environment and virtual interaction becoming more dominant, a original document could be modified digitally and marked on printed hard copy by multiple individuals. Software document management solutions are available to manage digitally modified document from multiple users. However, no convenient method exists for hand marked document especially for tracking, archiving, composing and reproducing documents that has been individually modified but from a common original source. In this paper we propose a method that enables preserving multiple users generated modifications made to a common original document. One single composite document is generated. The composite document that can be re-produced without loss of quality contains the original document and all user markings. In the composite document, the original document content is preserved and user marking can also be indexed and converted to other format using technologies such as optical character recognition. The new document is also much smaller in terms of file size for the ease of transmission and archiving without loss of quality for both the original content and the user applied modifications.

Introduction

Digital documents like Power Point^{TM}slides are digitally generated and modified by user on a physical printout with markings. Normally, consumers of such documents would mark on the document with their own comments, drawings, and suggestions in various colors and most likely in hand writing on the printout. Currently, the most common way to electronically store these hand marked document is to scan the document with consumer's markings and save the scanned files as an image file (JPEG or PDF). Such static capture of the electronic original document with hand writing presents three major challenges to document management: 1) large file size of the scan comparing to the electronic original for both storage and transmission; 2) difficulty in performing search in the scan comparing to the electronic original; and 3) reduction in image quality comparing with the electronic original and potential printing artifacts if the scan is re-printed. In this paper, we propose a novel scheme to face the three challenges and eliminate them from the root cause. This paper proposes a more sophisticated process on the scanned file with the knowledge of the digital original so that a new digital file can be generated and stored instead [1].To generate the new file, the paper proposes a method to decompose the static document (digital scan) that contains consumer's markings. The static document is decomposed into two parts: 1) information represented by the electronic original and 2) document consumer's own markings on the electronic original after it was printed. The consumer markings are stored as a static image representing any difference between the electronic original and the static document. The static document that is both large in file size and contain inferior image quality due to scanning of printed document is replaced by the new file that is the combination of electronic original and the consumer markings static image. By incorporating the electronic original, the new representation of the static document is 1) smaller in file size, 2) easier to perform search and indexing on, and 3) has minimal loss of image quality in reprinting.

Overall System

The traditional document archive system of printed document can be described as Figure 1. The digital original document is printed and then marked by user. A scanner is used to capture the static image of the entire document and the scan is archived.

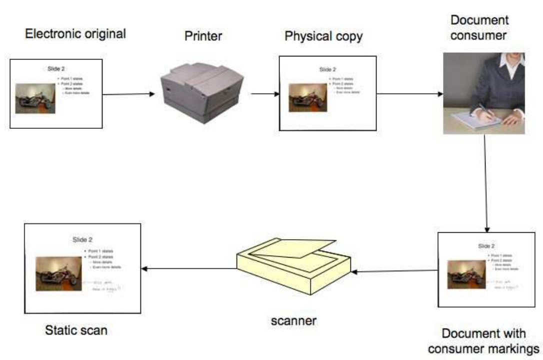


Figure 1. *Traditional Document Archive Work Flow*

The work flow that this paper propose is outlined in Figure 2. The proposed work flow is performed on the static scan and shows the procedures for generating the digital file that contains the electronic original and a static image of the consumer markings. Red dashed circle indicates the location of the consumer markings. The first step is to align the static scan with the electronic original. An image registration algorithm is used to achieve alignment between the two images. The aligned scan image has the same size as the electronic original. Both the aligned scan image and the electronic original are feed to the static image decomposition algorithm to isolate the consumer markings. Static image decomposition algorithm performs the following operations: 1) Convert both aligned scan image and electronic original to gray scale im-

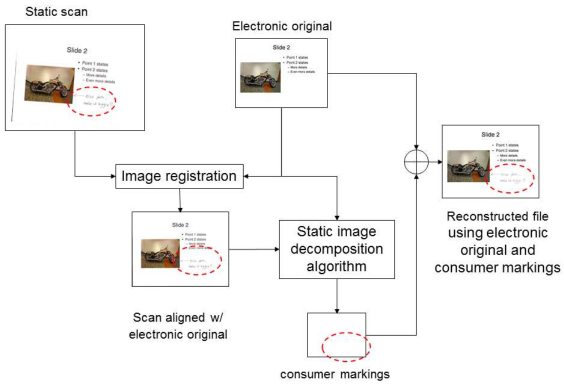


Figure 2. *Work Flow of the Proposed Method*

ages with luminance preserved. 2) Adjust the white area's color in scan image to match the digital value of 255. 3) Perform edge dilation on the electronic original image. 4) Detect any image pixel in the aligned scan image that contain information at the location where there is none in the electronic original. 5) Extract consumer additions using the mask to generate consumer markings as they appear in the scan. The customer markings will be merged with the electronic original to create a new file that can achieve the three benefits mentioned above.

Static Scan Decomposition Method

Static Image, the scan of the printed document with user markings, is first compared with the electronic original in Figure 2. Image registration method [2] is the pre-process on the static image to resize and align the static image to the electronic original. The re-aligned static image and the electronic original are used to generate a consumer marking mask that is used to decompose the static scan into the user markings and original information presented in the digital original. Each step of the decomposition method is described in detail below.

1) Color image conversion to grayscale: Assume the input image is in RGB color space. The conversion to grayscale operation discards the hue and saturation value of the each image pixel but preserve the luminance information. If R,G, and B represent the red, green and blue value of a image pixel, then the luminance value of the pixel is calculated as in Equation 1:

$$L = 0.2989 * R + 0.5870 * G + 0.1140 * B \tag{1}$$

Both static scan and the electronic original are converted to luminance.

2) Adjust the white and black to account for scanner variation: To account for scanner variation on the luminance value and also to the digital original, an efficient operation is applied. Let S represent the luminance channel of the scanned image that we obtain from Step 1 above. $S_{max} = MAX(S)$, and $S_{min} = MIN(S)$ are the maximum and minimum values of the luminance channel respectively. We apply two threshold values Th_w and Th_b representing white and black percentage thresholds, respectively. Assign any pixel that has value greater than Th_w of the luminance range to 255 (WHITE) and 0 (BLACK) if less than Th_b . The white adjustment for pixel i with luminance value S_i can be expressed as Equation 2:

$$S'_i = \begin{cases} 255 & , S_i > Th_w * LR \\ 0 & , S_i < Th_b * LR \\ S_i & , \text{otherwise.} \end{cases} \tag{2}$$

where luminance range $LR = S_{max} - S_{min}$.

3) Edge dilation on the electronic original image: Edge dilation is performed on the electronic original to account for scanner blurring and any small mis-registration between the aligned scan image and the electronic original. The edge dilation is square shape dilation with 2 pixel dilation on each side of the edge as shown in Figure 3.

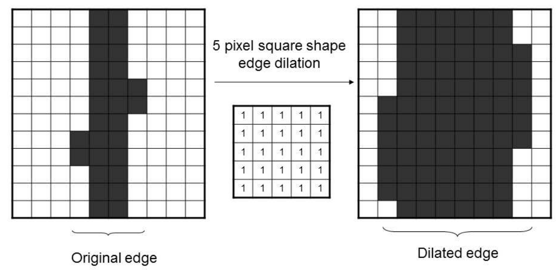


Figure 3. *Edge Dilation on Electronic Original*

4) Difference pixel detection between the white adjusted scan image and the dilated electronic original: $Scan_{adjusted}$ is the re-aligned, white adjusted scan image and $Img_{dilated}$ is the dilated electronic original image. First, both images are converted to " negative" by:

$$ScanImg' = 255 - Scan_{adjusted} \tag{3}$$

$$ImgOrig' = 255 - Img_{dilated} \tag{4}$$

We assume the information contained in scanned image is the super set of the information in the electronic original image and user markings, therefore, at any location where there is information in the scanned image but not in the original image is treated as a location could contain information that the document consumer added.

5) Extract consumer markings from scanned image: Any pixel location detected in Step 4 with luminance value difference greater than Th_m out of 255 maximum is treated as the a location in the consumer markings mask. The mask indicates the location on the scanned image where there are document consumer markings. Mask satisfies Equation 5:

$$Mask(i,j) = \begin{cases} 1 & , ScanImg'(i,j) - ImgOrig'(i,j) > Th_m \\ 0 & , ScanImg'(i,j) - ImgOrig'(i,j) \leq Th_m \end{cases} \tag{5}$$

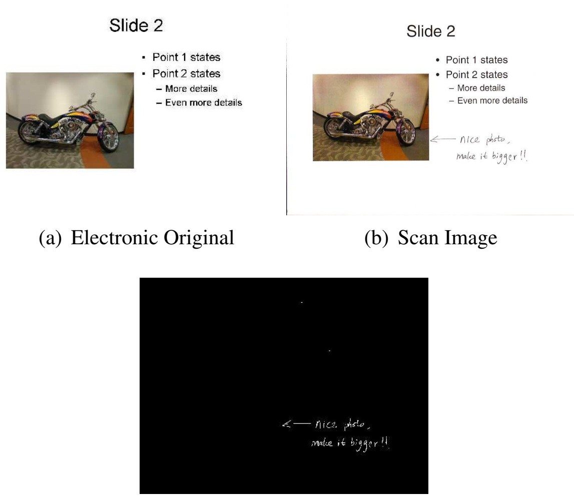
where (i, j) represents pixel location (i, j) in the scan. The consumer additions are extracted by applying the mask to the scanned image:

$$UserMark(i,j) = Mask(i,j) \cdot Scan_{adjusted}(i,j) \tag{6}$$

where $Scan_{adjusted}$ is the re-aligned scanned RGB image and the mask from Step 5 is applied to all three color channels.

Results

We present the images at various steps in the proposed method in Figure 4 and Figure 5 represents the combination of the electronic original and the consumer markings extracted from the scanned image. In this paper, we use Microsoft® Power Point files as original files. We chose $Th_w = 95\%$, $Th_b = 5\%$ and $Th_m = 100$. For Slide3.PPT in Figure5, the scanned file is 714 KB in JPEG format with 150 dots per inch (dpi) scan resolution. Electronic original is 73KB, and the consumer markings image file is 137 KB. We achieved significant reduction in file size: 70% reduction rate (from 714KB to 210KB) with no perceived image quality reduction and no object in the newly combined file is halftoned. The final combined file compared with the scan as shown in Figure 6 shows no image quality degradation from the electronic original and therefore eliminates any possibility for artifacts and image quality reduction comparing with printing the static scanned document. The final combined file is in the same format as the electronic file, in this case Power Point, thus preserving all objects in the electronic file. This enables the same search and indexing abilities that are available to the original format. The static scanned document does not enable these features.



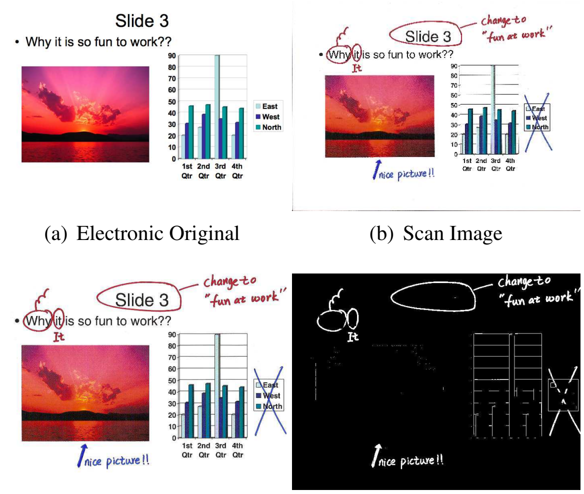
(a) Electronic Original (b) Scan Image

(c) Consumer markings mask

Figure 4. *Static Scan Decomposition of Simple Markings*

Conclusion

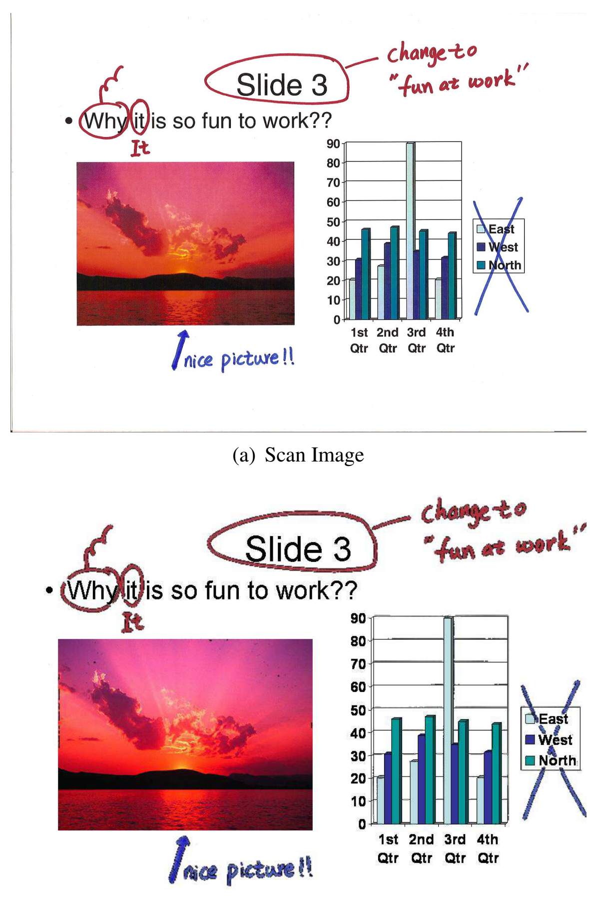
This paper proposed a more sophisticated process on the scanned file with the knowledge of the digital original so that a new digital file can be generated and archived instead of the scan. To generate the new file, the paper proposed a method to decompose the static document (digital scan) that contains consumer's markings. The static document is decomposed into two parts: 1) information represented by the electronic original and 2) document consumer's own markings on the electronic original after it was printed. The consumer markings are stored as a static image representing any difference between the electronic original and the static document. The static document that is both large in file



(a) Electronic Original (b) Scan Image

(c) Re-aligned Scan Image (d) Consumer Markings Mask

Figure 5. *Static Scan Decomposition of Complex Markings*



(a) Scan Image

(b) Reconstructed Document

Figure 6. *Static Scan vs. Reconstructed Document*

size and contain inferior image quality is replaced by the new file that is the combination of electronic original and the consumer markings static image. By incorporating the electronic original, the new representation of the static document is 1) smaller in file size, 2) easier to perform search and indexing on, and 3) has minimal loss of image quality.

References

[1] M. Qiao, *Preserving User Applied Markings Made To A Hardcopy Original Document.* United States Patent Application No. 20110141521.

[2] K. Chandu, E. Saber, W. Wencheng, *A Mutual Information Based Automatic Registration and Analysis Algorithm for Defect Identification in Printed Documents.* IEEE International Conference on Image Processing 2007, vol.3, no., pp.III-449-III-452, Sept. 16 2007-Oct. 19 2007.

Author Biography

Mu Qiao received his Bachelor, Master and PhD from School of Electrical and Computer Engineering, Purdue University in 2002, 2004 and 2008 respectively. He currently works in Shutterfly Inc., Redwood City, California as an Imaging Science Engineer. His research interests are Web 2.0 image processing and analysis, variable data printing, and color management.