Improving Rendering and OCRability of Color Images for Web Publishing

Abhishek Gattani and Hareish Gur
Newgen Software Technologies Limited
gattani@gmx.net, hareish@newgensoft.com

Abstract

Multilayer color images offer a pragmatic solution for publishing paper documents on the Web as they minimize time, effort, technological barrier and bandwidth requirements that exist in doing so. To increase the usability of these images, we present a unified method for simultaneously improving rendering quality and OCRability. The solution is modeled around interpolation and smoothing methods that work independent of the primary segmentation algorithm. Defect categories and restoration algorithms for the OCR layer identified herein, help generalization of the quality improvement approaches to a large extent. Our methodology improves rendering quality by 14 per cent and the preliminary results for OCR accuracy have been promising.

1. Introduction

With increasing penetration of the Internet and declining costs of high-speed production-grade color scanners, documents are increasingly being archived, communicated, and manipulated in digital form. According to an IDC estimate [1], by 2004 there will be 19 million flatbed scanners. Most of these will be color scanners. An acquisition device such as the scanner is typically the first step to bringing existing paper into digital media. However, color scans result in huge file sizes.

One alternative is to publish the content in meta formats like PDF, HTML, etc., with suitable compression applied. The raster text regions in the scanned image have to be converted into computer understandable form by manual means or OCR. The data captured by this method is then checked for integrity. However, OCR is often limited as text regions are mingled with complex backgrounds. The benefits of this approach are reduced file sizes and increased value proposition. However, the associated costs of converting color scans into meta formats, ensuring portability, extending platform-independence and requiring technical know-how result in much smaller than desired number of Web publications of color documents.

The other alternative is to publish the document in the raster form. The document image acquired hence, serves as a universal representation unlike meta formats whose interpretations are platform and viewer dependent. Besides, originality is lost in conversions. In contrast, document image remains unchanged across platforms and distribution media are natural for interpretation and suited for user-interaction. Further, the raster approach requires negligible technological know-how and can be published as soon as acquired. The difficulties that have limited the popularity of this approach are the huge file sizes, non-symbolic raster nature of the text, poor OCR and rendering quality.

2. Compression Technologies

Compression Technologies in the past have addressed the issue of file sizes but not with abounding success. The conventional options being either completely lossless (LZW-compressed PDF/ TIF), completely lossy (JPEG-compressed PDF/ TIFF/ JPG files) or doing away with color (G3/G4-compressed binary images). An A4 size, 300 DPI resolution image with lossless compression typically occupies anything between 10 MB and 20 MB of space, while the lossy takes anywhere between 1 MB and 2 MB, with inherent 'blocky' distortion in the entire image. It is noteworthy that when a color document is scanned in binary it undergoes tremendous loss of information. Further, binary is not suited for the Web. The problems with conventional compression techniques are that they are best suited for particular types of images or portions in an image but not the image in totality.

3. Benefits of Multilayer Color Images

New Color Document Compression Technologies with a multilayer framework, such as, DjVu [2], Lura Doc are promising pragmatic solutions to the problem of file sizes while retaining supreme text quality. Studies undertaken have proved that with these Color Document Compression Technologies, a typical A4-sized office document, scanned at 300 DPI resolution, originally about 25 MB in size (uncompressed), would occupy as little as 80 KB to 120 KB a size most welcomed for distribution and exchange on the Web. They achieve this by segmenting the foreground, background, color and other parts of the document image into different layers and apply suitable encoders to the same. For instance, a lossy encoder such as JPEG or JPEG2000 is applied on a lower-resolution

background layer. Then, there is the all-important, high-resolution text layer, which is a bi-tonal image, on which lossless CCITT G4 or JBIG2 encoders are applied. Text layer color information and soft edges are stored as another low-resolution layer. Results advert to quality and file sizes have ascertained that these are the solutions for storing and communicating color scans across the Web. Another significant advantage these technologies offer is that they provide a bilevel text layer for OCR, improving the worth of this format.

3. Existing Approaches

Though, the OCR layer provides faster and improved OCR results better than directly scanned color as well as binary images, the text layer needs improvement over imperfections due to optics, spatial quantization, sampling errors, foreground- background segmentation errors, for better display quality and recognition accuracy. It is well known that OCR accuracy depends upon image quality [3]. One way in which the problem of degraded character image restoration has been approached is by surface reconstruction from multiple images [4]. The problem with [4] is that it is computationally expensive and relies on a set of degraded bilevel images of a single unknown character, which is not pragmatic to our cause. The approach followed in [9] finds and averages bitmaps of the same symbol that are scattered across a text page. Not only are there performance issues with this approach but it also fails for documents with various types of stylized fonts, as not enough samples of the same symbols are available for clustering. The approach in [5] has focused on a practical method of assessing the quality of degraded document images and restoration techniques for the same but does not stress on display quality. Our paper, instead, focuses on degradations that occur because of errors in foreground/ background segmentation.

4. A Unified Approach

More attention in the literature has been given to OCR accuracy improvements than to display quality. The problems of recognition accuracy and display quality have been addressed separately in the literature, despite the fact that these aspects are deeply interrelated. For instance, while designing algorithms for improving recognition accuracy, it is imperative to assess effects on display quality, and vice versa. Here we propose a unified model for improvements in display quality and recognition accuracy. We first identify the OCR layers defect categories, then design a simple classifier and put forward a series of restoration algorithms addressing each of the defect categories.

4.1. Improving Display Quality

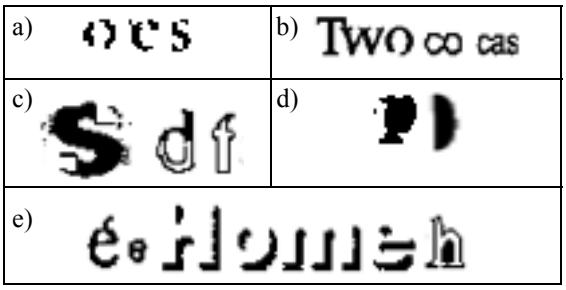
Since the final rendered image is constructed on the fly after merging component layers, the choice of the scaling algorithm becomes critical to display quality. There were visible losses in rendering quality when we programmed three layers in a simple viewer that mapped each of input layers to the desired resolution (the same can be verified with 3-layered PDF and in Acrobat Reader). We applied combinations of standard image interpolation algorithms [6][7] namely, Bicubic, Bilinear and Nearest Neighbor to these component layers. The results obtained have led to an interpolation model suited for scaling multilayer color images. In the proposed multilayer color images, text layer is maintained at scan resolution whereas the background and color layers are typically low-resolution layers. When up scaling the color image on or above the scan resolution, each component layer is scaled to the desired resolution and then merged together. However, while down scaling, the low-resolution layers are first up scaled to scan resolutions, merged, and then down scaled to the desired resolution. In this approach, when all the up scaled layers are first merged and then down scaled, sharp transitions between foreground and background are minimized as these transitions get blurred during down scaling. Instead, if individual layer were first down scaled and later merged, it was observed that foreground-background transitions sharpened and defect highlightened. In this model, each layer uses its custom-scaling algorithm for up scaling and down scaling.

Another important rendering aspect is developing a smoothing algorithm to remove the zigzag effect from text. We use Hobby's Smoothing algorithm [8], which minimizes the number of inflection points of a polygon-approximated character image. The algorithm provides impressive improvements in display quality but does not significantly improve OCR results as it minimizes the amplitude and frequency of the edge function, which does not affect today's OCR engines. The algorithm does suffer from certain drawbacks. A piecewise linear approximation of the character boundary is made due to which a sudden elevation or depression is minimized but not eliminated. An alternative solution is approximate portions of boundaries as curves. Further, inflection points are not considered for turns made in the same direction. On practical investigation we concluded this assumption is not correct because such turns also need suppression. We are in the process of developing a variation of the algorithm to overcome the aforementioned shortcomings. In Fig. 1, color image on the left is the improved image.

"Figure 1. Comparing Rendering Quality"

4.2. Improving Recognition Accuracy

Since the foreground-background segmentation algorithm focuses at color retention with the optimally best character boundary, there is always a conflict of interests between foreground and background leading to segmentation errors. At certain places, text is mingled with textured or low contrast background hence making the segmentation more prone to errors. Upon investigation, we found the following defects categories of the text layer that adversely affect OCR accuracy.



"Figure 2. OCR Defect Categories"

4.2.1. Broken Characters (See Fig 2a)**:** The problem of this class includes characters broken up in one or more components. These pose serious challenges to OCR recognition accuracy. If the page is a light photocopy or scanned at a high threshold, more characters are likely to fall in this class. The process of rebuilding character components starts with a grouping algorithm for clustering characters fragments and then a filling algorithm keeping in mind display quality. The grouping algorithm uses a set of heuristics such as:

a. *Symmetry*: Machine printed characters are known be symmetrical, if not the complete character then parts of it. Fragments are grouped if they increase the degree of symmetry.
b. *Overlapping fragments*: Character fragments not qualifying as individual characters with overlapping bounding boxes are grouped together.
c. *Line Grid*: Knowing the top, middle, under and base lines of text, using horizontal projection profiles characters fragments lying outside the region of consideration are rejected.

Character fragments are weighted according to the above heuristics and are grouped if their weighted average clears the cut-off. These components need to be joined together with smooth curves approximating the ideal character artwork keeping in mind display quality. From piecewise linear approximation of these components we can obtain the points from which the required curve passes. Tangents at these points yield a Hermite curve for joining the character components.

4.2.2. Joint Characters (See Fig 2b)**:** Our approach exploits the fact that the complete text layer is available to our analysis whereas for OCR engines this might not be necessarily true. We address this problem by grouping together individual characters into words then analyzing the vertical projection profile. Local derivates and auto-correlation are used to estimate the dominant character spacing; the profile obtained is smoothed to sharpen the peaks; non-maximum suppression is applied to locate the peaks; fourier analysis is applied to get the period of the typewriter grid and characters are segmented so that the resultant centers of gravity align with the typewriters grid. If the text is printed with proportional spacing, segmentation is difficult, since number of enclosed characters and positions for cutting demonstrate high variability. The approach works fine when text is typed with fixed character spacing.

4.2.3. Hollow and Partially Hollow Characters (See Fig 2c)**:** OCR engines fail to recognize hollow or partially hollow characters components. We have developed a linear boundary-tracing algorithm to capture their outline. The algorithm developed is robust to noise artifacts, separates inner and outer boundary segments, calculates the character stroke width and density, all in a single pass over the character image. The algorithm links run-lengths in a given scan row and cross-link these with overlapping run-lengths in adjacent rows. The cross-linked map of run lengths obtained give the character boundary using the following algorithm. Given a character image I[m,n], with m rows and n columns. Let $M_{[i(0..k-1),j(0..m-1)]}$ be the i^{th} run-length of the j^{th} row and k is a variable quantity denoting the number of run-lengths in that row, then

$$(\text{Outer boundary})_j = \cup [f_s(M_{[0,j]}), f_e(M_{[k-1,j]})]$$

$$(\text{Inner Boundary})_j = \cup_{a=0\ldots k-2} [\cup [f_e(M_{[a,j]}), f_s(M_{[a+1,j]})]]$$

$\forall$ f_s and f_e are functions returning the start and end of the run-length respectively. The average character stoke width

is total number of black pixels that form the boundary divided by the perimeter of the boundary segment. The regions between outer to corresponding inner boundaries are then filled row by row.

4.2.4. Filled Characters (See Fig 2d)**:** The problem of this class includes characters whose inner boundaries are completely filled. We obtained the boundary using the algorithm described in 4.2.3 and analyzed the output for atypical stroke widths. The outer boundaries of the affected regions are used for approximating the inner boundary that got lost due to filling.

4.2.5. Other Characters (See Fig 2e)**:** Reverse Text, Shaded Text, 3D Text, etc. also pose problems to recognition accuracy. Since, their percentage error in test images were the lowest, restoration algorithms are yet to be developed.

5. Results and Conclusion

Twelve human subjects rated the improvements in rendering quality of ten different color images at resolutions of 200 and 300 DPI, based on the following quality parameters (on a scale of 0 to 10): -

Parameters	Weights	Points for evaluation
Binary Text-Appearance	40%	Non broken, Crisp boundary, OCRability
Multicolor Text Appearance	10%	Color information of text
Reverse Text	10%	Text on colored background
Natural Photo/Scenery	30%	Clarity, Blurring effect, Chequered effect, Patches
Graphs, Plots and Engg. Drawing	10%	Light and dark shades, small data indexes, clarity of lines and curves.

"Table 1. Quality Parameters"

The weighted average was taken and it was observed that the proposed methodology improved rendering quality by 14%. The above set is characterized by good office documents covering all types of artifact parameters advert to OCR. The classifier implementation is not complete at this stage. Preliminary results, although without automated classifier and complete implementation of restoration algorithms, have been encouraging. The specifications of the OCR results is by no means complete as it can be highly influenced by the success of the classifier algorithm, artifacts coverage, physical condition of paper documents and the sample set size, which were too small at present. However, it can be confidently stated that the multilayer color document compression proposed overcomes bandwidth constraints for Web publishing thereby corroborating our research direction of improving OCRability and rendering for these formats. Future work continues for our unified approach.

6. References

[1] "About Infoimaging" at *http://www.kodak.com/US/plugins/acrobat/en/corp/infoImaging/infoimaging.pdf*

[2] Patrick Haffner, Yann Le Cun, Léon Bottou, Paul Howard, Pascal Vincent. "Color Documents on the Web with DjVu," *Proceedings of the International Conference on Image Processing*, vol 1, pp 239-243, Kobe, Japan, October 1999

[3] S. V. Rice, J. Kanai, and T. A. Nartker. "An evaluation of OCR accuracy." In *Information Science Research Institute, 1993 Annual Research Report*, pages 9–20. University of Nevada, Las Vegas, 1993.

[4] John D. Hobby and Henry S. Baird, "Degraded Character Image Restoration," *Proceedings of the Fifth Annual Symposium on Document Analysis and Image Retrieval*, pp. 233--245, 1996.

[5] M. Cannon, J. Hochberg, and P. Kelly. "QUARC: A Remarkably Effective Method for Increasing the OCR Accuracy of Degraded Typewritten Documents". In *Proceedings of the 1999 Symposium on Document Image Understanding Technology*, pp. 154-158, Annapolis, MD, May 1999.

[6] S. B. Kang. "A survey of image-based rendering techniques" In *VideoMetrics, SPIE* Vol. 3641, pages 2–16, 1999

[7] Hill, D.L.G.; Batchelor, P.G.; Holden, M.; Hawkes, D.J, "Medical image registration", *Physics in Medicine and Biology*, vol.46, pp. R1-45, 2001

[8] John D. Hobby, "Polygonal Approximations that Minimize the Number of Inflections," *Proceedings of the Fourth Annual ACM}-SIAM Symposium on Discrete Algorithms*, 1993.

[9] John D. Hobby and Tin K. Ho, "Enhancing Degraded Document Images via Bitmap Clustering and Averaging," *ICDAR'97*, pp. 394--400, 1997.