High Speed Page Matching

Robert Ulichney and Matthew Gaubatz, Hewlett-Packard Co., Marlboro, MA
David Rouse, Cornell University, Ithaca, NY

Abstract

In a print production system, the ability to associate each page with its original electronic form enables a number of important automation services, such as alternate-source page insertion, page/sheet routing and aspects of JDF-compliance. This work demonstrates the efficacy of a method capable of matching a scanned version of an incoming page to an original RIP, and was tested with a library of a thousand sample commercial pages. Frequency domain analysis suggests that low-resolution representations of pages can be used for matching purposes. This notion is corroborated by experimental results; a 98% correct matching is achieved for pages reduced to a bit-depth of 1-bit at only 1 dpi using a simple XOR method, and a perfect matching rate was achieved using pages represented at 16 dpi at a variety of bit depths using a sum of squared errors measurement. The method is implemented without placing any extra ink on each page.

Introduction

This paper focuses on the following question: is there an efficient method to uniquely identify each page (or sheet) on a device in a print production system? This work is motivated by the fact that the ability to associate physical media with original electronic documents addresses both present-day as well as forward-looking problems. For instance, it can be used to implement aspects of JDF compliance. It also could have a key role in distributed printing factory environments, which require high speed page identification for routing.

The XML-based Job Definition Format standard (JDF) is meant to simplify the life cycle of a print job by acting as a self-directed electronic job jacket defining all steps in the workflow. The processing instructions associated with each job (a "ticket"), the signal to begin processing, and the signal indicating processing is complete are all specified by JDF and an associated messaging format. If a compliant device is multiplexing between dynamically assigned jobs, there must be some mechanism to determine when each of these jobs is complete in order to implement JDF compliance; the work in this paper provides such a mechanism.

Lin [1] (at this conference) proposes the idea of modeling the flow of pages in a print production environment after that of the flow of packets in a network. Figure 1 shows an example of this distributed model. Finishers include mechanisms such as page coating devices, page trimmers and binding machines, and are connected via logical "Routers", depicted in yellow in the figure. Such devices do not currently exist, yet have several well-defined specifications. In order to determine how to direct a page in an automated fashion, each one must be able to associate each physical document to a page in a specific job, and therefore must include some physical mechanism to sense each page. Once this connection is established, well-understood network optimization tools can help to complete each print job efficiently, such that device utilization across multiple parallel jobs is maximized.

This paper is organized as follows. First, previous approaches to the matching problem are summarized. Next, frequency domain analysis is applied to data typical of that used to perform matching, to help determine how high speed solutions may be constructed. An algorithm is then presented based on these results. Experimental measurements and discussion conclude the paper.

The Page Matching Problem

A number of different methods are available to track physical documents. One approach popular in commerce is to include overt marks, such as bar codes, for identification. A drawback of this approach in print production is that the markings must be removed from the page. This issue can be addressed by placing the marks on the area to be trimmed and adding a cutting stage into the workflow. Along with the disadvantages of extra ink on the page and possible extra cutting, page tracking is disabled after the trim stage with this approach. The goal, however, is to use a method to identify pages based on the content of the pages themselves without adding any additional marks. The problem is thus one of matching a captured (scanned) version of the page against a rasterized version (RIP) of the source.

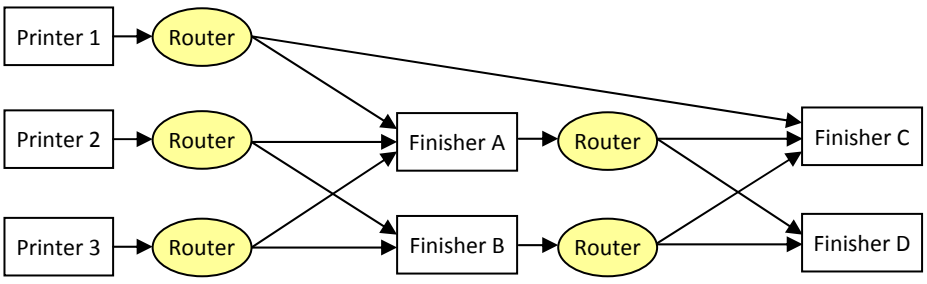

***Figure 1**. Distributed Printing model. Routers are paper path managers that contain a page identifier system.*

The topic of *image* matching is a well studied problem, a variety of solutions have been proposed. Edge-detection operators (e.g., Sobel, Prewitt, or Canny) have been combined with Hausdorff distance to create effective image comparison schemes [2]. Region-based approaches result in comparisons without the need to align the two candidate pages by detecting covariant image regions and assigning descriptors invariant to viewpoint to those image regions [3,4], and have been used in combination with the descriptor from the scale-invariant feature transform [5]. In many cases, regions are detected independently for each of the images being compared, and correspondences between regions of the two images are established based on a match of the descriptors of the regions. A reference that studies several region descriptors can be found in [6], where the descriptors were evaluated in the context of matching and recognition tasks for either the same scene or object subject to different viewing conditions. A feature descriptor for text documents based on the lengths of text lines has also been proposed [7]. It is reported to provide reliable recognition when comparing the features of the digital capture with features obtained from a collection of rasterized documents in a database.

More recently, algorithms have been developed for document image analysis applications, which seek to transform document images into symbolic form for modification, storage, retrieval, reuse and transmission [8]. Page segmentation strategies describe either the geometric or functional layout of a page, and may provide a means of comparing a scanned document to a rasterized document. Like the region-based comparisons, a descriptor corresponding to the document layout is generated to serve as a signature for a document. Page segmentation algorithms for text documents (with and without Manhattan layouts) have also been previously evaluated [9], yielding recommendations for pairing the algorithms tested with particular applications and document types. These cases, however, are all limited to text documents rather than general commercial pages.

One desirable quality for a high speed *page* matching algorithm is that it takes advantage of constraints associated with the print matching problem to produce an efficient solution. Although there is a rich history of research on the *image* matching problem, most techniques are quite unsuitable for high speed implementation due to multi-pass complexity and the resolutions expected for reasonable results. Indeed, many of the works named above are effective, but essentially can be used to solve more difficult problems, i.e., where lighting distortion is not very controlled, or where views of scene are from different angles, or even when items in a scene are in different possible configurations. The proposed approach includes a simple approach that is constructed based on two principles: (1) frequency domain analysis of the RIP data suggests that each such digital page can be represented as a thumbnail, and (2) the fact that scanned pages are obtained in a controlled environment, requiring little adjustment to alignment or lighting.

Frequency Domain Analysis of Page Data

The following procedure was used to help determine possible efficient page representations useful for the matching problem, where the variability of RIP page data was measured as a function of frequency. A collection of 1085 typical commercial PDF pages of sample test data was assembled. Pages included various catalogues as well as technical journals; samples of the pages used are shown in Figure 2.

***Figure 2**. Sample Commercial Pages Used in this Study.*

For each page, a horizontal and vertical frequency spectrum was created using the average per-row or per-column discrete Fourier transforms (DFTs), that is, the average 1-D DFT was computed along both dimensions. The horizontal and vertical standard deviation of the mean spectral amplitude as a function of frequency is shown in Figure 3.

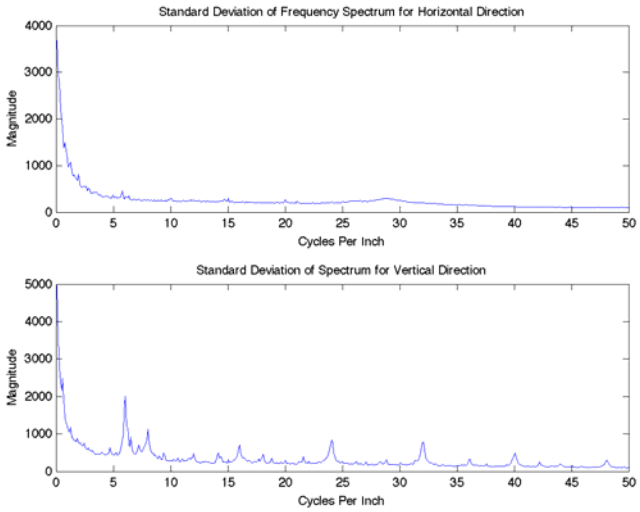

***Figure 3**. Standard Deviation of Averaged line Spectra.*

Even though the text-dominant content of these page images contain considerable high frequency information, most of the differences are in the very low frequency regions below a few cycles/inch. This concentration of differences suggests that the pages can possibly be matched successfully matched successfully using fairly low frequency information. It is interesting to note that the harmonics appearing at eight cycles/inch in the vertical spectra are due to the preponderance of 9-point 8-lines-per-inch type in the sample pages. Notice that in order to capture the information represented by these harmonics in a minimal manner, it is necessary to sample the image at a resolution of at least 16 dpi. Furthermore, because these harmonics are only present in the

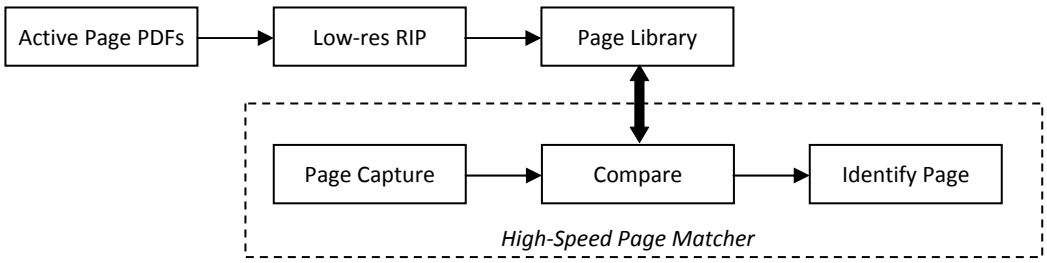

Figure 4. *Page Matching System.*

vertical direction, it stands to reason that an asymmetric sampling scheme could potentially be used to capture most of the information needed to differentiate between images.

Proposed Thumbnail-Based Solution

Because the variability of the image data seems to be captured by relatively low frequency information, it is reasonable to consider thumbnails as a possible representation for efficient matching. Such a choice inherently takes advantage of the constraints associated with the print production environment in which pages would be imaged. Referring back to the distributed model in Figure 1, for example, each routing device captures an image of each received page. Presumably, the lighting involved during this procedure is completely controllable, and the use of thumbnails only decreases the need for performing any kind of alignment.

The proposed page matching algorithm is straightforward, and system implementing it is depicted in Figure 4. At RIP time, a reduced complexity representation (i.e., a thumbnail) of each page is generated and stored in a Page Library. A reduced bit rate is achieved by a down-sampling operation (filtering, decimation) followed by a reduction in bit-depth. Each captured page is converted to a thumbnail representation, by a similar process, then is compared with an electronic record of the RIP pages (also thumbnails) active in the system. The page in this record that best matches the newly captured page is then associated with that physical document. In order to make this process efficient, or even implementable, the representation of each page must be compact, further motivating the choice of thumbnails, provided that they are small enough. As speed is a premium, thumbnail comparison strategies requiring a minimum amount of computation are of interest.

Since registration is not a serious issue for reasons above, this application allows the possibility that much simpler comparison algorithms than those previously proposed will be sufficient for strong performance. In fact, as experimental testing in the next section indicates, a sum-of-squared-errors (SSE) based approach is entirely appropriate. Since speed of execution is increased as the number pixels decreases, it is important to quantify matching performance as a function of thumbnail data rate to better understand the speed-accuracy tradeoff.

Matching Performance

The High-Speed Page Matcher in Figure 4 was tested as follows. Each page from the collection of 1085 was RIPed, printed and scanned. The RIPed pages were treated as the source library, and were compared with each captured page using the SSE algorithm to select the best match. This comparison was repeated where both the source and scan data were reduced by (1) scaling down the resolution using bilinear interpolation, and (2) reducing the bit depth.

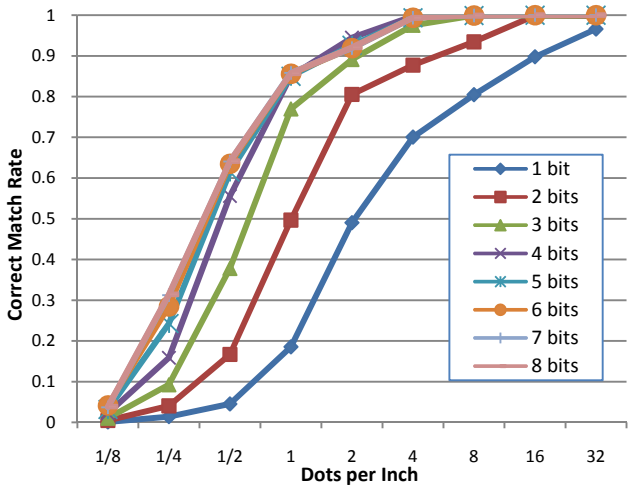

Figure 5*. Sum of Square Errors Page Matching Performance.*

Figure 5 shows the result of our experiment for bit depths from 1 to 8 per pixel. Resolutions ranged from 32 dpi to 1/8 dpi (1 pixel every 8 inches). Note that a 100% correct match rate is reached for 4 or more bits at 8 dpi and above, and that at 16 dpi, almost all bit depths yield perfect performance. This result corroborates the intuition developed from the frequency analysis which suggests that in a sense, much of the variability can be captured by a 16 dpi thumbnail. All tests were replicated simulating varying degrees of motion blur yielding little change in performance, especially at low resolutions, which makes sense, since moving a 200 dpi image by a couple pixels will not have much of an effect on the 16 dpi thumbnail, for example.

The 1-bit case is particularly interesting because the SSE algorithm can be equivalently implemented with a bitwise XOR operation. A considerable difference in matching performance is observed based on the means by which 1-bit quantization is carried

out. Figure 6 shows the performance for three bit-depth reduction strategies: choosing the value of the MSB of each pixel, quantizing each pixel based on the mid-point between the minimum and maximum pixel values, and quantizing each pixel as being above or below the mean observed pixel value. These results are notable due to of the compactness of representation. The mean-based threshold 1-bit XOR method achieved 98% correct matches at only 1 dpi.

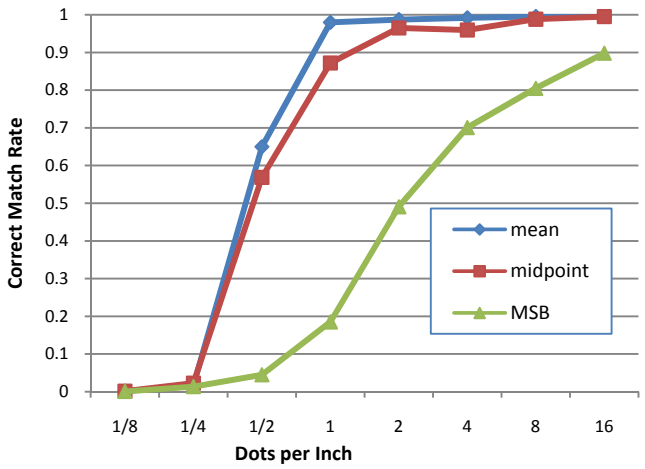

***Figure 6**.1-bit XOR Page Matching Performance.*

Conclusion

A page matching algorithm was presented for use with commercial printed page data. The tradeoff between matching accuracy and data rate was quantified for a large sample collection of typical commercial pages. Results suggest that simple and inexpensive capture devices could be potentially used. Motion blur, for instance, could be leveraged to low pass filter in one dimension.

Effective page matching is achieved with a very simple XOR algorithm at 1dpi. At 1 dpi an 8x10 inch area is represented with 80 bits. If an image can be represented with only 64 bits, then each page representation can fit into a single 64-bit register which has considerable efficiency implications. Almost all processors include an XOR and a population instruction; in this situation, the core of a page matching algorithm is reduced to two machine instructions per tested candidate. Clearly there are cases where matching will fail at this low resolution (for example, when variable data pages that differ by only a few characters are compared). This image-based matching approach can be combined with other factors, such as expected groups of pages in series (in time) or in parallel (N-up 2-sided sheets), to collectively contribute to high accuracy page identification.

Future work will involve characterizing the efficiency of asymmetric, spatially adaptive sampling methods designed to improve performance with very similar pages that differ in one particular location. The considerable change in performance due to the nature of the threshold process show in Figure 6 illustrates gains that can be achieved with non-linear processing. Alternative adaptive thresholding and compression techniques that yield compact, accurate comparisons are other items to be explored.

References

[1] I. Lin, E. Hoarau, J. Zeng, and G. Dispoto, "Proposal for Next Generation Print Infrastructure: Gutenberg-Landa TCP/IP," in *NIP25, 25th Internatinoal Conf. on Digital Publishing Technologies*, Sep. 20-24, Louisville, KY, 2009.

[2] D. Huttenlocher, G. Klanderman, and W. Rucklidge, "Comparing images using the hausdorff distance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 15, no. 9, pp. 850-863, Sep. 1993.

[3] D. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91-110, 2004.

[4] K. Mikolajczyk, et al., "A comparison of affine region detectors," *Int. Journal of Computer Vision,* vol. 65, no. 1/2, pp. 43-72, 2005.

[5] D. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the Eighth Conference on Computer Vision*, 1999, p. 1150–1157.

[6] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 10, p. 1615–1630, Oct. 2005.

[7] M. Lamming and W. Newman, "Activity-based information retrieval: Technology in support of personal memory," in *Proceedings of the 12th World Computer Congress, vol 3*, Madrid, 1992, pp. 68-81.

[8] G. Nagy, "Twenty years of document image analysis in PAMI," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 1, p. 38–62, Jan. 2000.

[9] F. Shafait, D. Keysers, and T. Breuel, "Performance evaluation and benchmarking of six page segmentation algorithms," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 30, no. 6, pp. 941-954, 2008.

Author Biography

Robert Ulichney received a Ph.D. from MIT in electrical engineering and computer science. He was with Digital Equipment Corp. from 1978-1998 where he focused on image and video implementations for both hard copy and display products. From 1998-2002 he was with Compaq's Cambridge Research Lab where he led a number of research efforts in video and image processing. Since joining HP Labs in 2002, Bob's research activities have included image enhancement, camera-projector systems and steganographic halftoning.