EVALUATING SPATIAL CORRESPONDENCE OF ZONES IN DOCUMENT RECOGNITION SYSTEMS

Michael D. Garris

National Institute of Standards and Technology
Gaithersburg, MD 20899 USA

Published in IEEE's Int. Conf. on Image Processing, Washington D.C., October 1995.

ABSTRACT

This paper introduces scoring methods developed to automatically assess the performance of document recognition systems; specifically, to evaluate the spatial correspondence of zones produced by a document segmentor. Two different approaches are discussed. The first approach (based on zone overlap and nearest-neighbors) is better applied to merged zones, whereas the second approach (based on zone alignments) is better applied to nested zones (such as those found in tables and graphs). Definitions of coverage and efficiency error are presented, and scoring results on real system output is provided that validates the usefulness of these methods to compare different document recognition algorithms. Currently, no standard testing procedures exist for measuring and comparing algorithms within a complex document recognition system. Scoring methods, like the ones introduced in this paper, serve as design and validations tools, expediting the development and deployment of document analysis technology for system developers and end users.

1. INTRODUCTION

The ability to automatically assess the performance of document recognition systems will expedite the development and the deployment of document analysis technology. Currently, no standard methodology exists for measuring and comparing algorithms within a complex document recognition system. If automatic performance measurements and procedures were available, they could be used as a design tool that would enable system developers to efficiently evaluate new ideas and algorithms.The results from standard testing methods can also be used by end users of the technology as a validation tool to compare the performance of different document recognition products. Without these measurements and procedures in place, no performance baseline can be derived in an unbiased way from within the document analysis community, leaving both developers and users of this technology with the difficult task of filtering through published system statistics that are often inconsistent and at times obscure.

The National Institute of Standards and Technology (NIST) has spent considerable effort in establishing standardized evaluation methods for optical character recognition (OCR) systems [1], [2] and text retrieval systems [3], [4]. As a result, a number of widely used and accepted databases [5], [6], research publications [7], and software packages [8], [9] have been developed. It is anticipated that a similar contribution can be made by NIST to the document analysis community.

This paper presents research primarily focussed on evaluating the spatial correspondence of zones produced by document segmentors. Zones include document structures such as columns and paragraphs of text, titles, headings, footnotes, page number, mathematical equations and chemical formulae, logos, tables, graphs, drawings, and pictures. The first goal of this research is to develop measures and procedures that evaluate how well a recognition system detects and then determines the proper position and extent of each zone on a document's page. To do this, a test set of document pages has been collected and digitized, and the zones on each page have been marked using a computer-assisted labeling tool. These *reference* zones are stored as ground truth and later compared with the location of *hypothesis* zones produced as output from the segmentor of a document recognition system.

At least two types of observations should be measured from the zone comparisons. The first determines the *coverage*, in other words how well the hypothesis zones cover the reference zones. The second observation is the *efficiency*, which measures how many hypothesis zones were detected as compared to how many zones really exist on the document's page. These two observations complement one another. For example, one document recognition system may achieve high zone coverage while reporting nearly the same number of hypothesis zones as reference zones. A second system may also achieve high zone coverage, however it could do so by fragmenting the reference zones into many smaller hypothesis zones. Clearly, the first system's performance is desirable, and the behavior of the second system should be avoided. Without the measure of efficiency, the two system's would appear to be performing with the same level of accuracy when compared according to coverage alone. Acknowledging this relationship, techniques for measuring zone coverage and efficiency were developed and tested.

2. POLYGONS VS. PIXELS

When measuring the coverage of a reference zone with a hypothesis zone, one could implement a pixel painting routine, where pixels mutually contained in both zones are painted one color, and pixels contained in one zone (but not in the other) are painted another color. The pixels of the various colors could then be counted and used to derive a coverage value. While this would provide measurements at the finest possible detail, there are nearly 15 million pixels in one of our full page test images scanned at 15.75 pixels per millimeter (400 pixels per inch). To label, store, search,

This work was supported by the Central Intelligence Agency under Interagency Agreement 403325.

and compile statistics at the pixel level would be unnecessarily costly and cumbersome. Therefore, operations on larger multipixel objects, such as polygons, are more desirable. Using polygons makes statistical compilations more efficient and record keeping more manageable while still permitting distinctions to be drawn between different system performances.

To simplify matters, it was determined that zones would be represented by their tightest bounding rectangle aligned with the image's raster grid. Using rectangles, the position of a zone is represented by an (x, y) pixel location within the image, and the extent of a zone is represented by a pixel width and height. This provides for a very compact encoding, such that all measurements are computed on zone rectangles (boxes) rather than the 15 million original pixels. Once boxes were chosen to represent zones, it became necessary to develop two types of measurements (distance and similarity) for comparing rectangles.

2.1. Box Distance

A distance measure was developed for determining how far apart any two boxes are from each other in terms of pixels. The distance chosen was the length of the intervening line segment shown in Fig. 1. This segment lies along the line connecting the center points of the two boxes. As the two boxes increasingly overlap the distance becomes negative, preserving continuity at the crossover boundaries. The minimum distance (the most negative value) occurs when the two boxes are perfectly centered on each other (aligned top-to-bottom, left-to-right). This distance measure is easily derived using simple geometry and is inexpensive to compute.

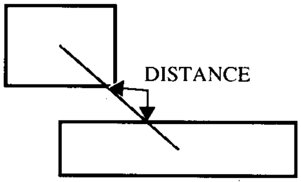


Fig. 1. Distance measured between two heterogeneous rectangles.

2.2. Box Similarity

A similarity measure was developed for determining how relatively close (in terms of size and shape) any two boxes are to each other. The difference between two boxes can be measured by using their corresponding widths and heights as coordinate points and calculating a normalize Euclidean distance between them according to (1). This measures the length of the diagonal of the shaded region shown in the right-most illustration of Fig. 2. The maximum possible normalized length of this diagonal is $\sqrt{2}$, so the difference measure can be converted to a similarity measure according to (2).

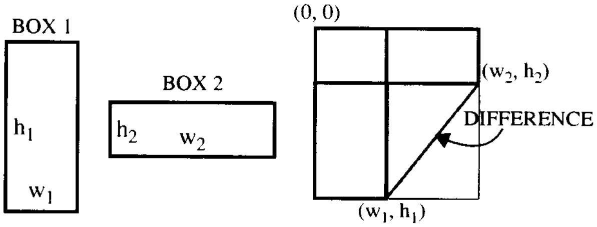


Fig. 2. Relative similarity measured between two rectangles.

$$\mathrm{diff} = \sqrt{\left(\frac{w_1 - w_2}{\max(w_1, w_2)}\right)^2 + \left(\frac{h_1 - h_2}{\max(h_1, h_2)}\right)^2} \qquad (1)$$

$$\mathrm{sim} = \frac{\sqrt{2} - \mathrm{diff}}{\sqrt{2}} \qquad (2)$$

3. ERROR MEASURES

By applying the concepts of coverage and efficiency in combination with box distance and similarity, several different scoring procedures were developed and tested using real document segmentor results. The goal of these methods is to automatically assess the quality of the segmentation without reconstructing the precise pathology, in other words, not trying to identify exactly how zones were split, merged, inserted, and deleted. This is important due to the fact that these events can occur in varying degrees and in combination with each other, making the pathology ambiguous and extremely difficult to derive after that fact. The principle exists that, as the quality of the system's output degrades, the ability to accurately derive the pathology becomes increasingly difficult.

Two components of zone coverage error are illustrated in Fig. 3. The darker gray area, referred to as *underage*, is the portion of the reference zone that is not covered by the system's hypothesis zone. The lighter gray area, referred to as *overage*, is the portion of the hypothesis zone that does not cover any part of the reference zone.

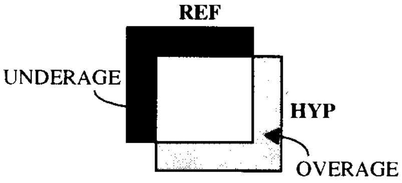


Fig. 3. Two sources of zone coverage error.

3.1. First Approach

To assess the amount of underage and overage error, reference zones must be matched to their representative hypothesis zones. The first approach computes the total amount of area overlapped by reference and hypothesis zones. The total area of overlap is subtracted from the combined area of all the reference zones to measure the underage, and the total area of overlap is subtracted from the combined area of all the hypothesis zones to measure the overage. Coverage error is then computed according to (3), where *refarea* is the combined reference area. In this way, the matching of reference to hypothesis zones is based on overlap.

To measure efficiency error, each hypothesis zone is matched to its nearest reference zone using the definition of box distance defined above. The nearest reference zones are chosen, such that more than one hypothesis zone can be matched to a single reference zone. If a reference zones ends up having more than one hypothesis zone pointing to it, then the hypothesis zones, minus one, are tallied as an *insertion*. If a reference zone ends up with no hypothesis zones pointing to it, then a reference zone is tallied as a *deletion*. Efficiency error is then computed according to (4), where *refnum* is the total number reference zones that could have possibly been recognized by the system.

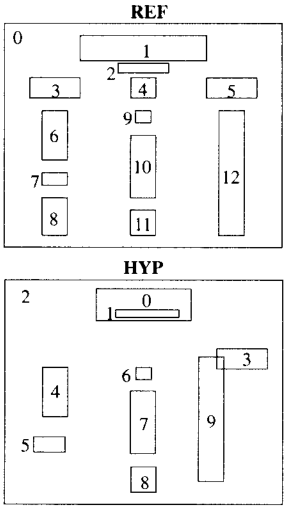
$$\text{coverage} = \frac{\text{underage} + \text{overage}}{\text{refarea} + \text{underage} + \text{overage}} \quad (3)$$

$$\text{efficiency} = \frac{\text{deletions} + \text{insertions}}{\text{refnum} + \text{deletions} + \text{insertions}} \quad (4)$$

Several deficiencies were noted when the scores produced by this approach were analyzed. It is possible for multiple hypothesis zones to overlap with a single reference zone. This can result in the combined area of overlap to exceed the actual area of the reference zone, making the underage error negative. It is unclear as to how this situation should be handled. Perhaps the underage and overage should be calculated as the intersection of the union of all overlapping areas, rather than the subtraction of all overlapping areas. But this would not assess penalties for the redundancy among the overlapping hypotheses.

3.2. Second Approach

A more fundamental problem is that overlapping hypotheses may in fact be legitimate nestings of document structures. Such nestings are common when decomposing tables and graphs. A second approach was developed to handle cases where document structures overlap one another. To do this a different method of matching reference zones to hypothesis zones was created.



ALIGNMENTS ref, hyp
0, 2
1, 0
10, 7
6, 4
11, 8
9, 6
5, 3
7, 5
12, 9
2, 1

3, 4, 8 Deleted

Fig. 4. Zone alignments for a simulated table.

A box alignment algorithm was implemented that determines a one-to-one mapping of reference to hypothesis zones based on box distance, similarity, and overlap. The technique matches the most likely candidates first, attempting to disambiguate assignments based on these three criteria. Once a one-to-one mapping is derived, those reference to hypothesis links that exceed certain tolerances of distance and similarity are removed. Reference zones not mapped to any hypothesis zone are tallied as deletions, whereas hypothesis zones not mapped to any reference zone are tallied as insertions. An example of zone alignments produced on a simulated table is shown in Fig. 4.

Using this second approach, the area of deleted zones contributes to underage error and inserted zones contribute to overage error. The aligned reference / hypothesis pairs contribute to the underage and overage factors based on their overlap (just as in the first approach), and coverage error is computed according to (3). Given the new definitions of deleted and inserted zones, efficiency error is calculated according to (4).

Several embellishments are needed to improve this second approach. First, contributions to underage and overage areas should probably be normalized according to the relative size of their contributing zones. This will keep small amounts of error on large zones from dominating errors on smaller zones. Second, the current alignment algorithm is strictly one-to-one, so that a hypothesis zone cannot map to more than one reference zone. This is undesirable when the segmentor merges several reference zones together as one large hypothesis zone. In this case, the alignment maps the hypothesis to only one of the merged reference zones, and the remaining reference zones are tallied as deletions which may cause inflated penalties to be assessed. Some notion of merging should be accounted for in the alignment procedure, however it may not be possible to disambiguate merged zones from nested zones. If this is true, then the first approach is better applied to merged zones, whereas the second approach is better applied to nested zones.

4. RESULTS

A simplistic segmentor was implemented that divides a document image into non-overlapping tiles (32 by 32 pixels in size). The number of black pixels in each tile is counted and thresholded, and the image is down-sampled by replacing each tile in the image with a single pixel (white if the number of black pixels is sufficiently low and black if sufficiently high). Connected components within the down-sampled image are then extracted and hypothesis zones are generated as the bounding box around each connected component. This segmentor is prone to errors, but it is extremely inexpensive to compute, and it can be embellished to improve performance. Many alternative approaches to segmenting documents exist [10],[11], but keep in mind that the primary goal here is to measure segmentor errors.

The document segmentor was run across 100 test pages extracted from *NIST Special Database 20* [12], a publicly available database of 104 science and technical (S&T) documents containing 23,468 binary scanned pages. (Other document databases are available such as the *English Document Image Database* from the University of Washington [13].) The 100 test pages were marked with reference zones using a computer-assisted labeling tool, and the document segmentor's hypothesis zones were scored against the reference zones using several different techniques. The scatter-plots shown in Fig. 5 and Fig. 6 graph the scores from the 100 test pages processed by two different segmentors. The individual page scores in these plots were derived using the second scoring approach described above, with the x-axis representing coverage error and the y-axis representing efficiency error. Those points closest to the graph's origin represent pages that were most accurately segmented, and segmentation quality decreases as points progress away from the origin.

The segmentor used to generate the scores in the first plot used down-sampled tiles and derived zones by drawing bounding boxes around the resulting connected components along tile boundaries. The second segmentor did the same down-sampling, only the bounding boxes were drawn tightly around the data within the tiles, rather than along the tile boundaries. Therefore, the first segmentor's zones were (by design) loosely defined causing more coverage error than the second segmentor. The difference in performance can be seen by comparing the two graphs. There tends to be a horizontal shifting of results in the direction of decreased coverage error, and there are more scores in the second plot closer to the origin than in the first.

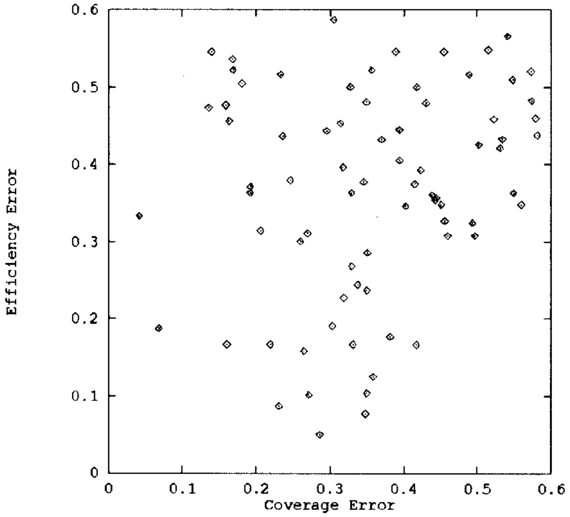


Fig. 5. Scoring results using noisy segmentor.

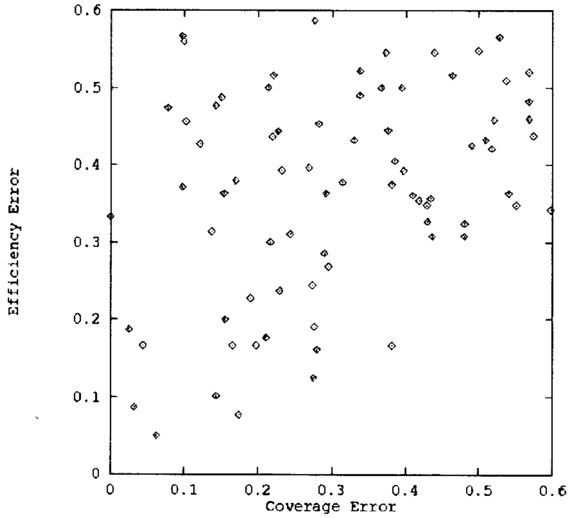


Fig. 6. Scoring results using more accurate segmentor.

This analysis along with other verification tests were conducted, and the results of numerous pages were visually inspected and compared. As a result, it was determined that these proposed measures and procedures (even with their limitations) do in deed capture the performance of the recognition system in an automated way.

5. CONCLUSION

This paper introduces scoring methods developed to automatically assess the performance of document recognition systems. Advantages and disadvantages of two different approaches were discussed. The first approach (based on zone overlap and nearest-neighbors) is better applied to merged zones, whereas the second approach (based on zone alignments) is better applied to nested zones (such as those found in tables and graphs). Definitions of coverage and efficiency error were presented, and scoring results on real system output were provided that validated the usefulness of these methods to compare different document recognition algorithms.

Currently, there are no standard methods to measure the performance of complex document recognition systems. The techniques proposed in this paper require further refinement as discussed and should be applied to a much larger set of documents and other types of document segmentors. This work only examines the issue of evaluating the spatial correspondence of zones. Future work should incorporate assessing the performance of zone identification tasks, such as measuring a system's ability to correctly identify the contents of a zone as being comprised primarily of text, graphics, math, etc..

REFERENCES

[1]. R. A. Wilkinson, J. Geist, S. Janet, P. J. Grother, C. J. C. Burges, R. Creecy, B. Hammond, J. J. Hull, N. W. Larsen, T. P. Vogl, C. L. Wilson, "The First Census Optical Character Recognition Systems Conference," NIST Internal Report 4912, August 1992.

[2]. J. Geist, R. A. Wilkinson, S. Janet, P. J. Grother, B. Hammond, N. W. Larsen, R. M. Klear, M. J. Matsko, C. J. C. Burges, R. Creecy, J. J. Hull, T. P. Vogl, C. L. Wilson, "The Second Census Optical Character Recognition Systems Conference," NIST Internal Report 5452, May 1994.

[3]. D. K. Harman, "The First Text REtrieval Conference (TREC-1)," NIST Special Publication 500-207, March 1993.

[4]. D. K. Harman, "The Second Text REtrieval Conference (TREC-2)," NIST Special Publication 500-215, March 1994.

[5]. C. L. Wilson and M. D. Garris, "Handprinted Character Database, *NIST Special Database 1*", NIST Technical Report and CDROM, April 1990.

[6]. M. D. Garris and R. A. Wilkinson, "Handwritten Segmented Characters Database, *NIST Special Database 3*," NIST Technical Report and CDROM, February 1992.

[7]. M. D. Garris, "Methods for Evaluating the Performance of Systems Intended to Recognize Characters from Image Data Scanned from Forms," NIST Internal Report 5129, February 1993.

[8]. M. D. Garris, J. L. Blue, G. T. Candela, D. L. Dimmick, J. Geist, P. J. Grother, S. A. Janet, and C. L. Wilson, "NIST Form-Based Handprint Recognition System," NIST Internal Report 5469 and CDROM, July 1994.

[9]. M. D. Garris and S. A. Janet, "Scoring Package Release 1.0," NIST Technical Report and CDROM, October 1992.

[10].R. G. Casey and G. Nagy, "Document Analysis - A Broader View," *1st ICDAR*, Saint-Malo, pp. 839-849, 1991.

[11].R. M. Haralick, "Document Image Understanding: Geometric and Logical Layout," *in Proc. IEEE Computer Soc. Conf. on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, pp. 385-390, June 1994.

[12]. P. J. Grother, "Scientific and Technical Document Database, *NIST Special Database 20*," NIST Technical Report and CDROM, April 1995.

[13].R. M. Haralick, "English Document Image Database I," University of Washington Technical Report and CDROM, August 1993.