Truthing for Pixel-Accurate Segmentation

Michael A. Moll, Henry S. Baird & Chang An
Computer Science & Engineering Dept, Lehigh University
19 Memorial Drive West, Bethlehem, Pennsylvania 18017 USA
E-mail: mam7@lehigh.edu, baird@cse.lehigh.edu, cha305@lehigh.edu
URL: www.cse.lehigh.edu/~mam7, www.cse.lehigh.edu/~baird

Abstract

We discuss problems in developing policies for ground truthing document images for pixel-accurate segmentation. First, we describe ground truthing policies that apply to four different scales: (1) paragraph, (2) text line, (3) character, and (4) pixel. We then analyze difficult and/or ambiguous cases that will challenge any policy, e.g. blank space, overlapping content, etc. Experiments have shown the benefit of using "tighter" zones that capture more detail (e.g., at the text line level, instead of paragraph). We show that tighter ground truth does significantly improve classification results, by 45% in recent experiments. It is important to face the fact that a pixel-accurate segmentation can be better than manually obtained ground truth. In practice, perfectly accurate pixel-level ground truth may not be achievable of course, but we believe it is important to explore methods to semi-automatically improve existing ground truth.

Keywords: *document image analysis, document content extraction, document content inventory, document content retrieval, ground-truthing, zoning, pixel accurate segmentation*

1 Introduction

In previous work [4, 5, 9, 10, 1], we have described a research program investigating versatile algorithms for *document image content extraction*, that is locating regions containing machine printed text, handwriting, photographs, etc. This program seeks to solve this problem in full generality, handling a vast variety of document and image types. While the availability of scanned document images suitable for use in such a research program has vastly increased with projects such as Google Books and other online freely available databases, the availability of any databases of ground truthed images is still very limited and lacking uniform standards. Considerable time in this research program has therefore been subjected to the discussion and cultivation of a ground truthing policy suitable to goals and difficulties of this specific problem. Many of the policy decisions made and challenges met in this program are applicable to any such project. Previous attacks on these problems are reported in [2, 3, 11, 13]. We are also strongly motivated by the work of Shafait, Keysers and Breuel in [12]. We believe it may be useful to the community to address the issues we have encountered, as well as leading a discussion of open questions that we have yet to resolve.

2 Our Ground Truth Policy

Our classifier, discussed in [4, 6, 7], is an approximation of k-Nearest Neighbors and is used to classify each pixel in a document image by assigning it a class label, such as machine print, handwriting, photograph, etc. Features are extracted from each training sample (pixel) from a small, local window of no more than 20 pixels wide. This means we are actually classifying a small window around each pixel and assigning a class label based on that window to each individual pixel. Therefore, our classification results are shape independent, that is we are not classifying rectangles or any other predefined shape. This allows the output of our classifier to handle more difficult non-rectilinear layouts, skewed pages, etc. Figure 1 shows an example of our ground truth for an image and the output of our classifier for that image.

We have carefully collected a collection of images of high diversity from a variety of online databases and through our own scanning efforts. While we seek to have a "pixel accurate" classification of an image, we immediately acknowledged that it is not feasible to manually obtain. We also realize that for other reasons discussed later that pixel accurate classification may not practically be achievable with perfection even given infinite resources. However, we believe that it is a worthwhile goal, that is an important driver of developing this technology.

We developed a web based application for our team to zone document images in PNG format, using overlapping rectangles. Unzoned pixels are not included in the training set and are ignored when scoring classifier output. We are exploring the use of more sophisticated ground-truthing tools, such as [14]. Initial experiments indicated that this much coarser and cruder ground truthing, compared to the pixel-level classification we were performing, still resulted in output that captured non-rectangular shapes and layout that the ground truth did not, seen in Figure 1.

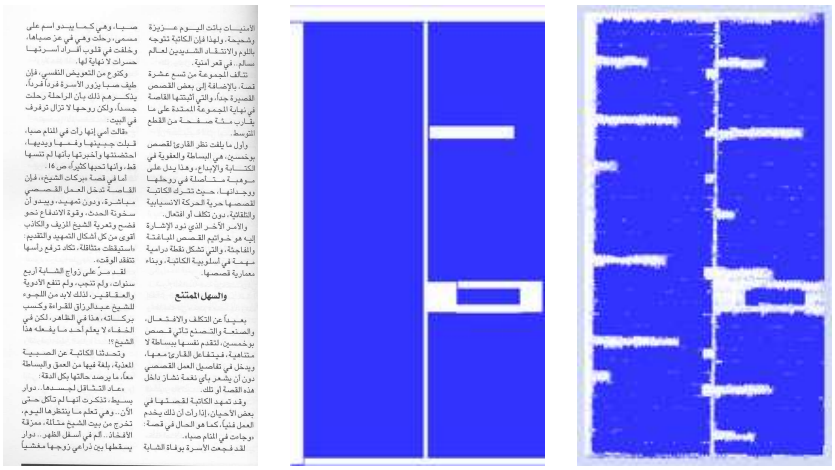

Figure 1. An example of a document image from our test set (on the left) with our ground truth for the image (in the middle) and the output from our classifier (on the right). The image is a greyscale scanned page of Arabic Machine Print. We use a tool we developed to zone the image by drawing overlapping rectangles over the different regions of content. The blue rectangles (shown in the proceedings as black) represent Machine Print. We consider the policy used to ground truth this image "loose", that is we are zoning the content at the paragraph or block level. However, our classifier labels each pixel, resulting in a more accurate representation of the layout of the text. A "tight" policy for zoning would involve drawing rectangles around each individual line of text and is discussed later in the paper.

3 Challenges and Open Issues

3.1 What to Ground Truth?

The initial discussion of our ground truth policy began with what classes we wanted to classify. In the context of our problem of document image content extraction, we started with this initial list of content types: machine printed text (MP), handwriting (HW), photograph (PH), blank (BL), line art (LA), math (MT), engineering drawings (ED), chemical diagrams (CD), maps (MP) and junk (JK). We used this list to drive a systematic collection of document images for our database, containing each content type in bitonal, greyscale and color formats, in a variety of languages (when applicable). However, for initial testing of our classifier we tested on a smaller set of content types and we realized that some of these classes were possibly subsets of others. Therefore, initial ground truthing only labeled MP, HW, PH and BL content.

As mentioned previously, manual ground truthing makes pixel-accurate ground truth infeasible, leading to a policy decision of what to classify. While we use overlapping rectangles, this also applies to any other scheme that uses polygons or any other shapes. Considering any form of text, handwritten or machine printed, the next level up from pixel accurate ground truth would be at the character level, then the line level and finally the paragraph level. Since our classifier is labeling each pixel based on a small window around each pixel, combined again with the infeasibility of manual labor, we chose not to pursue character level ground truth. Character level zoning also presents a challenge in determining where a character begins and ends, as discussed in [8]. Some of the white pixels in between and around the black pixels of a character must also be considered part of a character and sometimes these regions may overlap. We chose to ground truth at the paragraph level initially as this was the most efficient policy time wise and as we were improving the classifier this yielded acceptable results. We will discuss alternatives to this policy decision later in the paper.

3.2 Blank Space

As mentioned before, we chose to treat blank space as a unique class and therefore we must also ground truth blank space like any other content class. An initial idea was to label any pixel not explicitly "zoned" by the user in our tool as blank, however there were multiple reasons for not making this policy. Some documents may have types of content that we are currently not testing and we would like to intentionally not ground truth or there may be ambiguous areas of the document that contain multiple content types or that the user is unsure of how to label. For the purpose of training data, these pixels can be left unlabeled and will be ignored in training the classifier. Finally, since we treat blank space as an equal class to the other classes we should use the same policy for obtaining ground truthed data as we do for the other classes.

Our ground truth policy however, created some problems for our classifier in classifying blank space. At any level other than pixel accurate ground truthing, some amount of blank space will

be included in the areas zoned as other content types (i.e. the white space between lines of text, the white space inside the letter o, etc). If ground truthing is particularly sloppy or loose, this can introduce what appears to be noise in areas classified as blank space, seen in Figure 6. Experiments with our classifier show that this problem occurs most frequently with confusing blank space for handwriting and in more limited cases also for machine print. This is due to the more free form layout of handwriting samples, compared to the more uniform layout of machine print. Experiments discussed later confirm the idea that more careful, tighter ground truthing of handwriting samples, lead to mistakes in classifying blank space.

3.3 Overlapping Content

One problem that we have dealt with from the beginning of this project and have yet to find a satisfying policy for is that of how to zone areas that contain overlapping content areas, as seen in Figure 2. Part of our research goal is for our classifier to do well on images with difficult, complex layouts. This includes images that have complicated backgrounds, possibly photographs, with machine print over them. Other common forms of this problem are machine printed documents with handwriting annotations.

Our policy has been to try to as tightly as possible zone the foreground pixels (the MP over the PH, the HW over the MP) before labeling the background pixels. However, since we are not adopting a pixel-level ground truth policy this has the potential of introducing some "noise"' pixels to the ground truth for that class. Current experiments have not shown any serious problems with this policy for the classifier, however more experiments should be conducted using training sets consisting of much larger amounts of overlapping data. An alternative policy would be to assign two class labels to overlapping areas. No experiments have been completed with this policy yet.

3.4 Machine Print in Photographs

A special form of the above problem is specifically how to handle machine print and photographs when they overlap. The above mentioned example of a magazine article with a photograph as background with a story printed over it or a caption on a photograph seems straight forward. We try to tightly zone the MP and then zone the PH around it. However, a unique case is that of a photograph that contains machine print. For example, an image taken from a handheld camera of a street sign or even a newspaper article with a photograph of a football player showing his name on his jersey, shown in the image on the right in Figure 2. While the case of the street sign in the image obtained from a digital camera seems straight forward, to label the text as machine print it quickly becomes less clear if the street sign is not the focus of the photograph or the case of the newspaper article with a photograph. In this case we consistently do not label the text as machine print.

3.5 Difficult Shapes

We chose to use overlapping rectangles for our zoning to make implementation of our zoning tool simpler, as well as simplifying the zoning process for the user. Many of the documents we collected to train and test on contain difficult, non-rectangular layouts, shown in Figure 3. Even with a tool for zoning that uses polygons instead of simple rectangles would have an imperfect representation of the actual layout in the ground truth. The policy we use for these areas are trying to capture as much of the detail and as little noise as possible using many small rectangles. This is unfortunately a very time consuming process for the person doing the zoning, and at best is still imprecise. An alternative in our research program is to leave images like this out of the training set, as our classifier does not learn from the layout of a page, but from the content of a page. However, this is obviously not an acceptable policy for all research. This also creates an evaluation problem that will be discussed later, as it will force pages with these difficult layouts to be scored worse than they should be using some evaluation metrics.

3.6 Subsets of Content Types

In our research program, we eventually hope to be able to distinguish between some content types that are naturally subsets of each other. The handling of this problem is largely dependent on the application the ground truthed data is being used for. For example, the content type machine print, can have subtypes such as math, chemical diagrams, some elements of engineering drawings, etc. A policy decision must be made on how to ground truth images that contain this content and for our initial experiments this was to simply ground truth it all as machine print. However, in the future this may require the reground truthing of pages that contain these subsets of MP. One possibly policy would be to initially ground truth subtypes as specifically as possible, such as mathematical equations as MT and later map them back to MP for experiments if experiments are not currently using that content subtype.

3.7 Line Art

Our initial experiments originally considered only 4 content types: HW, MP, PH and BL. Eventually we expanded to a fifth content type and began experimenting with line art (LA). An initial problem in zoning line art was deciding what type of line art we wanted to test on. We noticed that we had collected what loosely could be considered two different types of LA, seen in Figure 4. The first type are drawings made by hand that can be very simple or when complex look very similar to photographs. The second type are things like diagrams, technical drawings, etc. We realized that these two types should probably be considered as two different types of classes as initial experiments containing both in the training set as LA, resulted in a nearly complete failure to recognize any LA. This also led to reconsidering our larger list of content types we were collecting for future experiments as the second form of line art seemed to have subtypes such as engineering drawings, chemical diagrams, etc.

Experiments treating both of these as two separate types of content also revealed a new problem with the ground truthing of line art in the form of technical drawings. These types of images frequently contain large amounts of blank space, and also machine print. We have yet to reach a suitable policy for how to handle this content type.

Figure 2. Examples of some difficult images to zone that include Machine Print overlapping other content types. The full-color image on the left features a background which is actually a photograph, and is not just a solid color. The machine printed text overlaps different regions of the photograph in different sizes and colors. The middle full-color image shows a simpler form of overlapping text where a caption in machine print overlaps a photograph. In this case we ground truth the text as MP, not PH. The image on the right illustrates the difficult case of text within a photograph, shown as the school name on the jersey of the football player. For this case, we consistently and arguably do not zone the text as MP.

Figure 3. The full-color image on the left shows a background which can partly be considered blank space and partly photograph with complex boundaries. There is also some gutter noise on the left side of the image. The full-color image on the right is particularly difficult as it shows a weather map with multiple types of overlapping content.

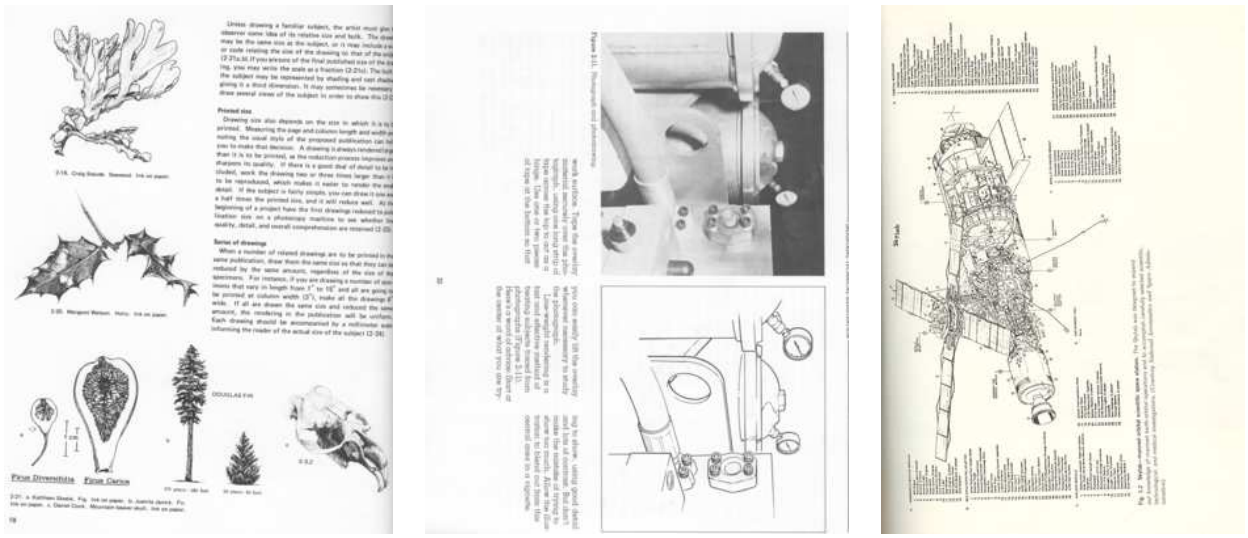

Figure 4. Examples of different types of Line Art. The image on the left illustrates multiple Line Art segments that take the form of hand drawn sketches. The center and right show Line Art in the form of technical or engineering drawings. These images are particularly difficult as the areas that contain the Line Art also contain other content types like machine print and large areas of blank space.

A third problem with line art is what to consider document image objects like paragraph dividers and horizontal rules. While we have seen in other ground truth policies the creation of a new content type for layout elements like this, we have not yet included a new content type for them in our research program.

3.8 Junk and Noise

We initially included junk as a possible content class, however we have not yet attempted to systematically collect samples of junk. Junk can be thought of as salt and pepper noise in a scanned image, other artifacts from scanning or faxing, margins and other edge effects, scribbles, gutter noise, etc. There is an endless amount of content that can be included in this category from any number of sources. In some subtle cases, it may be safe to just consider it to be blank space. However insignificant it may seem, there are still a very large number of document images that contain some example of this and our current policy is to ignore any significant areas of junk and noise.

4 Ground Truthing Can Distort Evaluation

Our initial experiments with our classifier brought to our attention a very simple, yet significant fact that was initially overlooked. For evaluation of our classifier we were using as a metric the per-pixel accuracy. That is, comparing each pixel in our classified output to the content type assigned to that pixel in the ground truth. Given that we are not using a per pixel ground truthing policy, this results in the per-pixel evaluation metric being pessimistic. Even with a crude ground truthed version of a paragraph of text, our classifier more accurately captures the layout of the text that the ground truth, shown in Figure 1. However, as a result of our ground truth our classifier is actually penalized by this metric for not labeling the content like the ground truth, which is clearly subjectively worse.

We have developed an alternative evaluation metric, that considers the content inventories of an image [10] that appears to be less pessimistic in evaluating classifier performance when used with a ground truthing policy like ours, and also can be a useful tool for an end user in browsing diverse image collections. This metric considers the ratios of the amount of each content type classified in an image and is discussed further in [10].

5 The Effect of Tighter Zoning

Given the problems encountered with using a non-pixel-level ground truthing policy for a pixel level classifier, we began to experiment with using a tighter ground truthing policy to try and reduce errors to improve overall classification. As discussed and illustrated before, our initial ground truthing policy was designed to drive development of the classifier and running experiments with very large numbers of training and test images. This required a ground truthing policy that was not extremely labor intensive and relatively simple for new people in our lab to adopt. However, as performance of the classifier became more stable and test set sizes started growing less slowly, we realized one area of our program that could potentially lead to great increases in performance was our ground truthing policy.

As an experiment, we took our most recent training and test sets, which had been ground truthed in our original, relatively "loose" standards seen in Figure 5, and rezoned them much more tightly. As mentioned, previously our zoning could be thought of as being done on the paragraph level and the new policy reduced this to a sentence level. In particular, we were very careful to label handwriting content much more tightly than before as that class had previously presented the most errors in classification and introduced the most noise into classifying blank content. Figure 6 shows a common improvement seen in most of our classified images zoned with the "tighter" policy. The most noticeable results across the entire test set was a much "cleaner" classification of blank space for most images, more often classification of machine print on the line level, rather than paragraph level, and better classification of handwriting. However, this policy still does not attempt to ground truth all blank space between lines of text as blank space. For the entire test set, the overall per-pixel error rate dropped by 45% from 38.9% to 21.4%.

6 Future Work

A recurring theme in most of the problems and challenges discussed in this paper, is the use of non pixel-level ground truthed images for pixel level classification. However, we believe that this is an important issue for any research program. Obviously, it is nearly impossible, in both complexity and time to obtain pixel accurate ground truth by manual zoning. A second issue with ground truthing, is that it along with feature selection, now remain as the only unautomated parts of our research program. It seems that an ideal ground truth policy, at least for our research purposes, would be an automated, pixel-accurate ground truthing mechanism. The early successes of our classifier lead us to believe that this may not be completely unrealistic.

We are working on a system using our current "best" classifier to build a semi-automated, bootstrapped pixel-level ground truthing system. The system will run our classifier on a previously unseen image to be included in our training set. The output of the classifier will then be manually inspected for areas that subjectively appear to have been classified accurately, as mentioned before traditional evaluation metrics that count per-pixel accuracy are naturally pessimistic in our system. These areas will then be manually selected for addition to the training set before classifying new unseen images. Hopefully, with little manual intervention we will be able to quickly expand the size of training set and radically improve its quality, without needing to manually hand zone every image.

Acknowledgements

We thank Sui-Yu Wang for her assistance in ground-truthing images. We are also grateful for stimulating conversations with Thomas Breuel about this topic. We also thank the referees who read this paper for their insightful comments and suggestions.

References

[1] C. An, H. Baird, and P. Xiu. Iterated document content classification. In *Proc., IAPR 9th Int'l Conf. on Document Analysis and Recognition (ICDAR07)*, Curitiba, Brazil, September 2007.

[2] A. Antonacopoulos, B. Gatos, and D. Bridson. Icdar2007 page segmentation competition. In *Proc., IAPR 9th Int'l Conf. on Document Analysis and Recognition (ICDAR07)*, Curitiba, Brazil, September 2007.

[3] A. Antonacopoulos, D. Karatzas, and D. Bridson. Ground truth for layout analysis performance evaluation. In *Proceedings., 7th IAPR Document Analysis Workshop (DAS'06)*, Nelson, New Zealand, February 2006.

[4] H. S. Baird, M. A. Moll, and C. An. Document image content inventories. In *Proc., SPIE/IS&T Document Recognition & Retrieval XIV Conf.*, San Jose, CA, January 2007.

[5] H. S. Baird, M. A. Moll, J. Nonnemaker, M. R. Casey, and D. L. Delorenzo. Versatile document image content extraction. In *Proc., SPIE/IS&T Document Recognition & Retrieval XIII Conf.*, San Jose, CA, January 2006.

[6] M. R. Casey. *Fast Approximate Nearest Neighbors*. Computer Science & Engineering Dept, Lehigh University, Bethlehem, Pennsylvania, May 2006. M.S. Thesis; PDF available at `www.cse.lehigh.edu/~baird/students.html`.

[7] M. R. Casey and H. S. Baird. Towards versatile document analysis systems. In *Proceedings., 7th IAPR Document Analysis Workshop (DAS'06)*, Nelson, New Zealand, February 2006.

[8] G. Kopec and P. Chou. Document image decoding using Markov source models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI–16:602–617, June 1994.

[9] M. Moll and H. Baird. Segmentation-based retrieval of document images from diverse collections. In *Proc., SPIE/IS&T Document Recognition & Retrieval XIV Conf.*

[10] M. Moll and H. Baird. Document content inventory and retrieval. In *Proc., IAPR 9th Int'l Conf. on Document Analysis and Recognition (ICDAR07)*, Curitiba, Brazil, September 2007.

[11] I. Philips, S. Chen, J. Ha, and R. Haralick. English document database design and implementation methodology. In *Proceeding of the 2nd Annual Symposium on Document Analysis and Retrieval*, pages 65–104, UNLV, USA, 1993.

[12] F. Shafait, D. Keysers, and T. M. Breuel. Pixel-accurate representation and evaluation of page segmentation in document images. In *Proc., IAPR 18th Int'l Conf. on Pattern Recognition (ICPR20006)*, Hong Kong, China, August 2006.

[13] S. Simske and M. Sturgill. A ground-truthing engine for proofsetting, publishing, re-purposing and quality assurance. In *Proceedings of the 2003 ACM Symposium on Document Engineering (Doc Eng'03)*, pages 150–152, Grenoble, France, 2003.

[14] B. A. Yankikoglu and L. Vincent. Pink panther: A complete environment for ground-truthing and benchmarking document page segmentation. *Pattern Recognition*, 31(6), 1998.

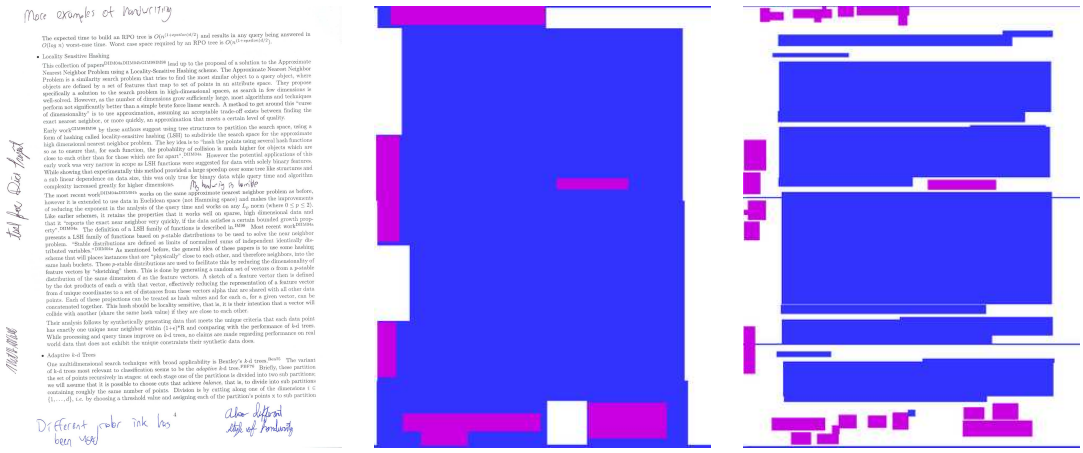

Figure 5. A sample image from our training set is shown on the left. It is an image of a technical article annotated with handwriting. The image in the center shows our original zoning of the image, using our "loose" paragraph level zoning policy (which also includes some errors from the manual zoning process). The blue (black in the proceedings) pixels are machine print and the purple (lighter gray) pixels are handwriting. The image on the right shows the zoning using the new "tighter" policy, which attempts to more closely represent the actual layout of the text (such as the jagged edges of paragraphs and more tightly zoning handwriting). The results of this policy change in classification output can be seen in Figure 6

Figure 6. An example of the effects of the "tighter" zoning policy on the output of our classifier. This is for an image of a newspaper article with machine print text, handwriting and a line art like section in the lower right hand corner that is classified as a photograph (in this experiment, line art was not included as a possible content type). The middle image shows the output of the classifier from the training set that was zoned with the original "looser" zoning policy that zoned the text areas at the paragraph level. The image on the right shows the output from the training set using the "tighter" zoning policy. For this image, and many others, the original classifier had many problems distinguishing handwriting from blank space resulting if the flawed classification of blank space with many errors (seen as the pink pixels in the middle image which represents handwriting). The new output from the "tighter" zoned images are largely free of this problem. Also, the new output does a much better job of classifying the handwriting content in the image (this may be difficult to see when printed in black and white; in the middle image the handwriting pixels are more of a mixture of two content types and in the image on the right they are more uniformly only labeled as handwriting pixels).