The Convergence of Iterated Classification

Chang An and Henry S. Baird

Computer Science & Engineering Dept, Lehigh University
19 Memorial Drive West, Bethlehem, Pennsylvania 18017 USA
Email: cha305@lehigh.edu, baird@cse.lehigh.edu
URL: www.cse.lehigh.edu/~cha305, www.cse.lehigh.edu/~baird

Abstract

We report an improved methodology for training a sequence of classifiers for document image content extraction, that is, the location and segmentation of regions containing handwriting, machine-printed text, photographs, blank space, etc. The resulting segmentation is pixel-accurate, and so accommodates a wide range of zone shapes (not merely rectangles). We have systematically explored the best scale (spatial extent) of features. We have found that the methodology is sensitive to ground-truthing policy, and especially to precision of ground-truth boundaries. Experiments on a diverse test set of 83 document images show that tighter ground-truth reduces per-pixel classification errors by 45% (from 38.9% to 21.4%). Strong evidence, from both experiments and simulation, suggests that iterated classification converges region boundaries to the ground-truth (i.e. they don't drift). Experiments show that four-stage iterated classifiers reduce the error rates by 24%. We also present an analysis of special cases suggesting reasons why boundaries converge to the ground-truth.

Keywords: *document content extraction, content inventory, layout analysis, convergence, shape-oblivious segmentation, uniform content classification, iterated classification*

1 Introduction

We have developed a family of algorithms for *document image content extraction*, able to find regions containing machine-printed text, handwriting, photographs, etc in images of documents [3, 5, 4, 7]. Our algorithms cope with a rich diversity of document, image, and content types. We classify individual *pixels*, not *regions*, in order to avoid arbitrary restrictions of region shapes. Previously, we achieved modest per-pixel classification accuracies (of, *e.g.*, 60–70%). Now we report accuracy of 85%.

Other researchers have attacked this problem of fine-grain classification without restricting region shape. In [8], Nicolas and Dardenne *et al* adapted and applied conditional random fields (CRF) to document image segmentation. For features, they defined two feature functions: a local feature function that takes only into account features extracted on the observed image, and a contextual feature function that takes only into account the local conditional probability densities on the label field in a neighborhood. Although they extracted features on the basis of pixels, they classified 3x3 region.

Kumar and Gupta *et al* in [10] use matched wavelets to develop the globally matched wavelet filters for text extraction, i.e. to discriminate text from nontext, of color document images. The scheme is extended for the segmentation of document images into text, background and picture components. To refine the obtained segmentation results, they exploited the contextual information by using a Markov random field (MRF) formulation-based pixel labeling scheme; and they attained MRF energy minimization using the alpha-expansion algorithm proposed in [11, 6]. Their method classifies pixels.

Our technique of iterated classification is similar in broad outline to cascading classifier [1], but with these differences: we train on the results of classification, not on the original images; and we reclassify every sample, not merely rejected samples.

Since we classify every pixel, our classifiers are similar to many image processing methods, such as mathematical morphology [9].

Our proposed trainable post-processing scheme is based on results of *document image content extraction*, and guided by the ground-truth (gt). This strategy appears to prevent the local regions which are dominated by erroneous classes from expanding, while allowing those dominated by correct class to expand slowly.[1]

The key contributions of this paper are:
- significant reductions in per-pixel error rates;
- demonstration that the methodology is highly sensitive to ground-truthing policy, and especially to precision of ground-truth boundaries;
- as refining the accurary, iterated classification continues to enforce local unformity ("purity") of regions;
- systematic exploration of the best scale (spatial extent) of fea-

[1] Before we discovered this, we trained the second stage classifier on the first stage classification results of training set, and used these training samples for all following stages of classification. This allowed local regions that are dominated by one content class to expand, whether the dominant class is correct or incorrect.

tures;
- strong evidence that iterated classification converges region boundaries to the ground-truth (they don't drift);
- analysis of reasons why boundaries converge to ground-truth.

2 Experimental Design

In the experiments reported here, we use a training set of 33 images and a distinct test set of 83 images, which are the same images we used in[2]. (The scanning resolution range from 200-400 dpi. At this moment we do not scale our features with resolution.) Together the two sets contain machine-print (MP), handwriting (HW), photograph (PH), and blank (BL) content. Each content type was zoned manually (using closely cropped isothetic rectangles, overlapped where needed to fit non-rectangular regions) and the zones were manually ground-truthed. The training data was decimated randomly by selecting only one out of every 9000th training sample.

We evaluated performance using per-pixel accuracy. This is the fraction of all pixels in the document image that are correctly classified: that is, whose class label matches the class specified by the ground truth labels of the zones. Unclassified pixels are counted as incorrect.

3 Tight Ground-truthing

By careful investigation of previous experimental results, we believe that tight ground-truth is vital to the success of post-classification. This is because in each stage a post-classifier is guided by the ground-truth to correct the errors made by its predecessor, and a loose ground-truth can cause confussion.

We rezoned and ground-truthed the training images more tightly. The effect of rezoning is illustrated in Figure 1.

4 Design of Post-classifiers

The goal of post-classification is to enforce local uniformity without imposing arbitrary region shapes. We designed a trainable post-classifier that operates on the output of the previous classifier, guided by ground truth. Note that the post-classifier also yields a per-pixel classification result for the document image. This inspired us to try *iterated* classification: a sequence of post-classifiers, each trained separately on the training-data results of the previous classifier, guided, as always, by ground truth. We will call the initial stage classifier the *first stage* classifier, the immediately following post-classifier is the called the *second stage* classifier, followed by the *third stage* classifier, etc. A diagram of iterated classification is shown in Figure 2.

Our strategy has been to extract features from small local regions, so that no single classification stage affects a large area. It's worth emphasizing that we train each of the post-classifiers separately on the results from the training set of the previous stage.

For the classification technology, we use approximate 5NN using hashed k-d trees.[4] The features for the post-classifiers are discussed[2].

5 Systematic Exploration of Scale of Features

Previouly, we extracted features from circles of radius 5 pixels. Our experiment show that the classification results are sensitive to this radius. We have explored this sensitivity over a range of scales for each classifier stage separately. The experiments show that the best scale of features changes from stage to stage, as shown in Figure 3. Guided by the classification results for the training set, we chose radius of 7 for the second stage classification, 9 for the 3rd-stage, and 7 for the 4th-stage. The differences are not always statistically significant, but it is clear that the sweet spot is somewhere between 6 and 10 pixels radius for these features.

6 Experimental Results

Experiments show great improvement on tighter ground-truth. With loose ground-truth, the error rate for the first stage of classification was 38.9%; with tight ground-truth, the error rate for the first stage of classification has decreased to 21.4%, a drop of 45%.

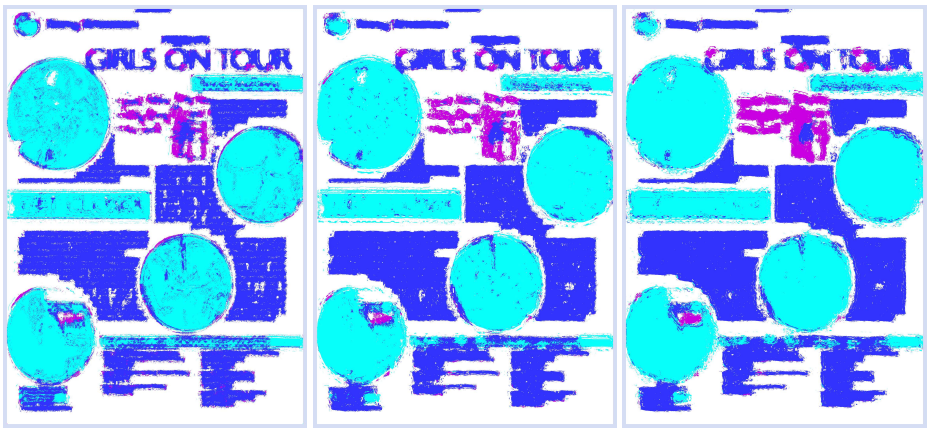
Our results are illustrated in Figure 1 and Figure 6. Each figure contains nine images of four types: (a) the original image; classification images from stage one using loose ground-truth (b), classification images using tight ground-truth from stages one (c), two (d), three (e), and four (f); and three *mask* images for MP(g), PH(h), and HW(i) content classes. In each of these two figures, the original images are shown on the upper left (the original images are full color, but are printed in this Proceedings as grey-level). The results of classication are shown in (b)-(f), as classification images where the content classes are shown in color: machine print (MP) in dark blue (printed as dark grey), handwriting (HW) in red (printed as medium grey), photographs (PH) in light bluegreen (printed as light grey), and blank (BL) in white (printed as white). (In this Proceedings, the distinction between MP and HW may be hard to see.)

Figure 1 shows results on a color image of a sports magazine page containing complex non-rectilinear regions. With tight ground-truth, the per-pixel error of the first-stage classifier is 22.9%; Figure 1(b) shows the result obtained with loose ground-truth: the per-pixel classification error of the first-stage classifier is 36.7%. Note that BL regions are mixed with PH pixels, MP and PH regions are mixed with HW pixels, HW regions are mixed with MP pixels. Figure 1(c) shows the result obtained with tight ground-truth: the per-pixel error of the first-stage classifier is 22.9%; compared to Figure 1(b), BL regions are much purer, MP and PH regions have less HW pixels in them, but HW regions are mixed with more MP pixels. the error of the second-stage classifier is 17.2%; the error of the third-stage classifier is 15.4%; and error of the fourth-stage classifier is 14.4%.

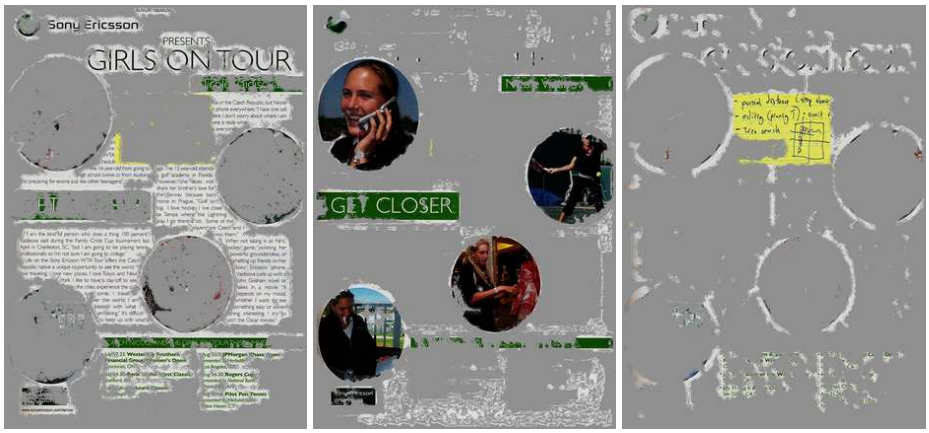
Figure 6 shows results on a color image of a movie magazine page containing complex non-rectilinear regions. With loose ground-truth, the per-pixel classification error of the first-stage classifier is 32.5%. And the background is mixed with HW. With tight ground-truth, the per-pixel error of the first-stage classifier is 25.2%; the error of the second-stage classifier is 18.9%; the error of the third-stage classifier is 17.7%; and error of the fourth-stage classifier is 17.7%.

Figure 4 gives the representation of total error rate as a function

(a) test image (b) 1st stage classification (with loose GT) (c) 1st stage classification (with tight GT)



(d) 2nd stage classification (e) 3rd stage classification (f) 3rd stage classification



(g) MP masked (h) PH masked (i) HW masked

Figure 1. Illustration of the dramatic improved result of tighter ground-truth. A document image with a complex non-rectilinear page layout, contains content of MP, HW, PH and BL. The original image (a) is in full color (printed in this Proceedings as grey-level). In the classification results (b)-(f), machine print (MP) is dark blue (printed as dark grey), handwriting (HW) red (printed as medium grey), photographs (PH) light blue-green (printed as light grey), and blank (BL) white. Tighter ground-truth drops the error rate of this image from 36.7% to 22.9%, a drop of 38%. The final MP, PH and HW masks extract their content types well, as shown in (g)-(i). except for some small patches of HW misclassified as MP, and some small patches of PH misclassified as MP or HW.

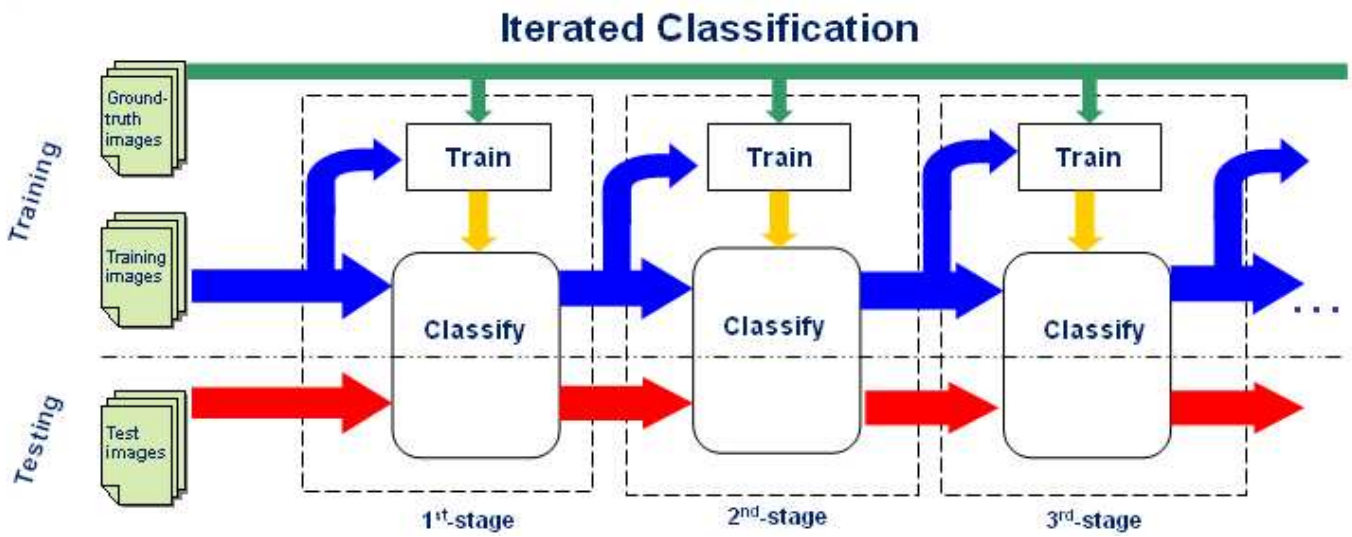


Figure 2. The methodology of iterated-classification. The same ground-truth is passed to every training phase. Classification results are passed from one classifier to its successor for training and classification. Take the 2nd-stage classifier for example, the classifier is trained on ground-truth and the first-stage classification results of the training iamges, and classifies both the training and test images. Note that each classifier is, in general, different from one another.

	3	5	7	9	11	13
2nd-stage	0.162	0.158	0.148	0.151	0.166	0.174
3rd-stage	0.144	0.141	0.138	0.137	0.143	0.160
4th-stage	0.136	0.135	0.133	0.134	0.141	0.150

Figure 3. Error rates for training set of each stage using different scale of features, that is in radius of 3, 5, 7, 9, 11 and 13 pixels. Guided by the classification results for the training set, we chose radius of 7 for the second stage classification, 9 for the 3rd-stage, and 7 for the 4th-stage.

of stages of classification. The post-classifiers reduce the error rate by 23.4%.

7 Post-classifiers' Performance Analysis

One of our previous experiments shows that the post-classifiers reduce the per-pixel classification errors by 23%, running a four-stage classification on 83 test images. Another experiment with fewer test images shows that per-pixel errors can fall monotonically for even eight stages. We notice that, as uniformity is improved in local regions, boundaries tend to remain stationary – that is they do not drift. This observation leads us to try to prove there exist iterated classifiers that are guaranteed to converge to ground-truth boundary.

We begin the investigation by simulating that an image contains two content-classes, say MP and BL, and we have a classifier trained and tested on this image. The ground-truth and first-stage classification result for this image are shown schematically in Figure 5(a)-(b); MP pixels are colored black, BL pixels are colored white. In Figure 5(a)-(d), t_g marks the horizontal coordinate of the boundary in the ground-truth and t_r marks the horizontal coordinate of the boundary in the classification results for the image. In Figure 5(c)-(f), gray regions represent the discrepancies between ground truth and the classification results for the training image.

Given the ground-truth and results of the first-stage classifier, we can analyze how the second-stage classifier performs. Recall that features are extracted within a local window (a circle of radius R) centered on the target pixel.

7.1 Analysis of the Second-stage Classifier

We start by analyzing the case where the width of the discrepancy is greater than R, i.e. $t_r - t_g > R$, as shown in Figure 5(e). For the post-classifers, we consider one feature that we have been using in experiments: the number of BL pixels within the right half of the feature extraction window. Recall that all features are extracted from the results of classification.

Consider these different cases of a target pixel based on its ground-truth class, labeled class from classification results, and the number of BL pixels within the right half of the feature window.

Case I: Target pixel is ground-truthed MP, classified MP, and contains no BL pixels whithin the right half of its feature window.

Case II: Target pixel is ground-truthed BL, classified BL, and all pixels whithin the right half of its feature window are BL.

Case III: Target pixel is ground-truthed BL, classified MP, and contains no BL pixels whithin the right half of its feature window.

Case IV: Target pixel is ground-truthed BL, classified MP, and contains at least one BL pixel whithin the right half of its feature window.

For pixels that fall outside the discrepancy region, the classification is obvious: pixels in case I, *i.e* those in the black region in Figure 5(c), are still labeled MP; pixels in case II, *i.e* those in the white region in Figure 5(c), are still labeled BL.

For pixels within the discrepancy (ground-truthed BL but classified MP by the first-stage classifier), part of them will be correctly classified using the feature, as follows:

If the right half of its feature extraction window contains any BL pixels – case IV – the target pixel is then classified BL, because its feature value is different from that of pixels in case I. For example: in Figure 5(e), the pixel centered on circle *b* and *c* is labeled BL. If the right half of its feature extraction window contains no BL pixel – case III – the pixel is still classified MP because its feature value is the same as that of pixels in case I. For example: in Figure 5(e), the pixel centered on circle *a* is labeled MP. Pixels that are less than R pixels left from the boundary t_r are in case IV, and are thereby are labeled BL.

After the second-stage classification, the horizontal coordinate of the resulting boundary would be $t_r - R$, which moves towards ground-truth boundary t_g by a distance of R pixels.

7.2 Analysis of Classifiers that Follow the Second-stage Classifier

As long as the width of the discrepancy is greater than R, each following classifier must behave the same as the second-stage classifier and cause the boundary to move again towards t_g by R.

When the the width of the discrepancy is smaller than R, i.e. $t_r - t_g < R$, we must consider more cases, as follows:

Case V: Target pixel is on boundary t_g, ground-truthed MP, classified MP, and contains a number, say B, of BL pixels whithin the right half of its feature window.

Case VI: Target pixel is ground-truthed MP, classified MP, and contains more than one but less than B of BL pixels whithin the right half of its feature window.

Case VII: Target pixel is within the discrepancy, ground-truthed BL, classified MP, and contains more than than B of BL pixels whithin the right half of its feature window.

Pixels that fall outside the discrepancy are classified in this way: pixels in cases I, V and VI are still labeled MP; pixels in case II are still labled BL.

Pixels within the discrepancy will be classified BL: all of them are in case VII, and their feature values are different from that of ground-truthed MP pixels in cases I, V and VI, therefore the classifier must classify them BL. This is illustrated in Figure 5(f): the center pixel of circle *a* lies on the left boundary of the discrepancy area will be classified MP, following its ground-truthed content; circle *b* has more BL pixels in its right half than circle *a* does, therefore the center pixel of *b* can be discriminated and classified BL; for the same reason, pixels in the discrepancy, but not on its left boundary, are to be classified BL. Consequently, the boundary in the classification result moves towards t_g by $t_r - t_g$, the boundary of the classification result has converged to ground-truth.

Simulation shows the same behavior as the analysis above suggests. We simulated a discrepancy of 174 pixels wide, and a feature extraction (circular) window of radius 20. For the first eight stages, the boundary moved left by 20 pixels in each stage of classification. At the ninth stage, the boundary moved left by 14 pixels, which converged exactly to the ground-truth boundary.

In summary, analysis of special cases, experiments and simulations, behave as the classifiers appear to do. That is, with proper choice of features and guidance by the ground-truth, there exists a sequence of post-classifiers that refine the obtained results and force them to converge to ground-truth. This implies that the post-classifiers can converge linear boundaries oriented at any direction

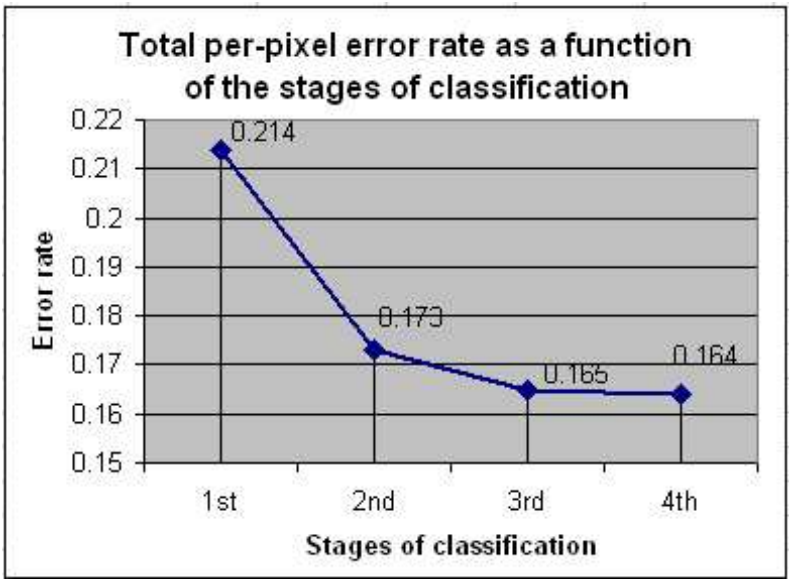
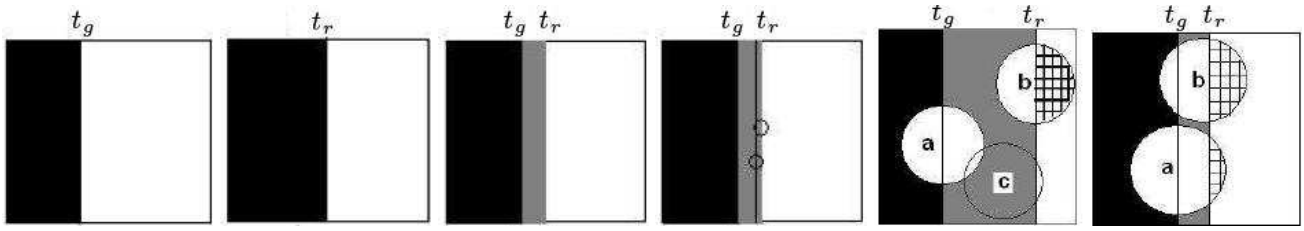
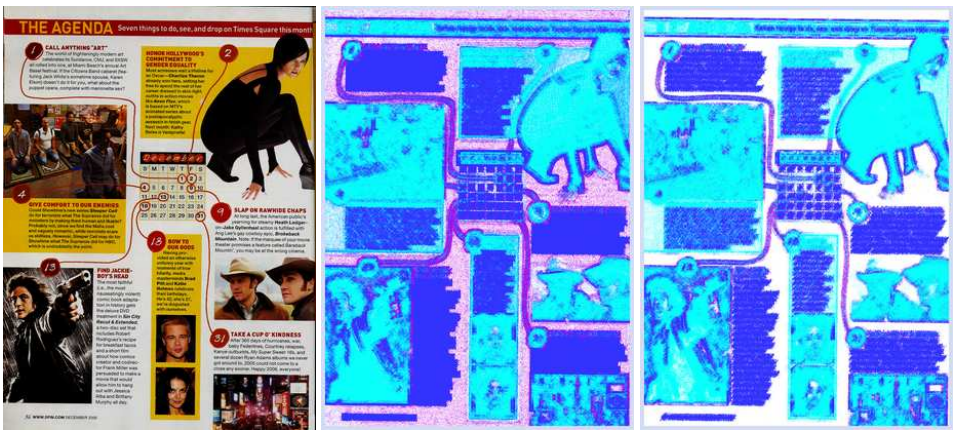


Figure 4. Total error rate averaged over the larger test set, in a function of the stages of classification. After four stages of classification, the error rate has fallen from 0.214 to 0.164, a drop of 24%.

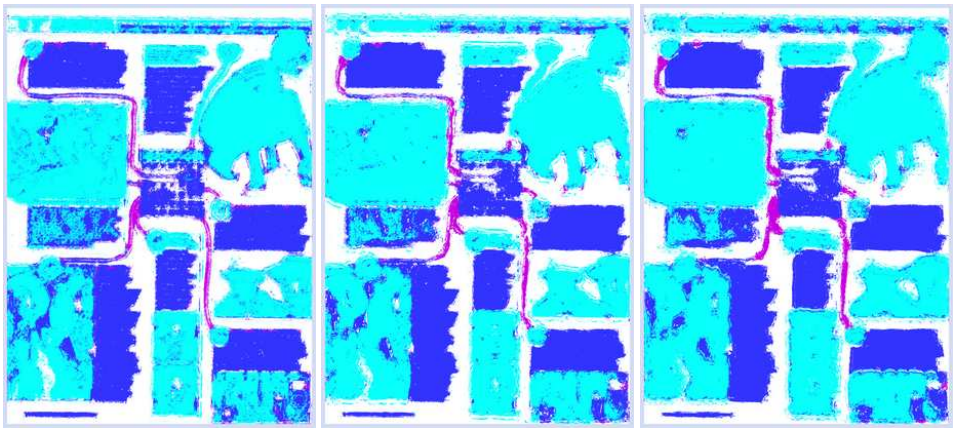


(a) ground-truth (b) 1st-stage classification result of the image (c) discrepancy between ground-truth and 1st-stage classification result, shows in gray (d) feature extraction windows and new boundary (e) the width of the discrepancy is greater than the radius of feature extraction windows (f) the width of the discrepancy is smaller than the radius of feature extraction windows

Figure 5. Analysis of convergence of iterated classification. Black represents MP, white represents BL. In figure (c)-(e), the discrepancies between ground-truth and the classification results for the image are colored gray. In figure (e) and (f), circles represent feature extraction windows. In each figure, t_g marks the horizontal coordinate of the boundary in the ground-truth and t_r marks the horizontal coordinate of the boundary in the classification results for the image. Initially, $t_g < t_r$.



(a) test image (b) 1st stage classification (with loose GT) (c) 1st stage classification (with tight GT)



(d) 2nd stage classification (e) 3rd stage classification (f) 3rd stage classification

(g) MP masked (h) PH masked (i) HW masked

Figure 6. A magazine page with a complex non-rectilinear page layout, containing content of MP, PH and BL. The original image (a) is in full color. The results of classification are shown (b)-(f). The final MP, PH and HW masks are shown in (g)-(i). The MP mask extracts its content class well, except for three patches misclassified PH. This error is possibly due to the lack of training sample of MP written in red color on a yellow background. For curvature preservation, notice the small red circles containing numbers: their curvature changes slightly.

to ground-truth. We conjecture that for all region shapes, whose radius of curvature is bounded below, there exists a similar training methodology such that all boundaries converge to ground-truth. Some of the experiments show that the post-classifiers also converged on regions with small radii of curvatures. For example, in Figure 6 the small red circles containing numbers are preserved.

We also conjecture that to converge to ground-truth, the number of post-classifiers needed is proportional to the width of the discrepancies, and inversely proportional to the radius of the feature extraction window.

8 Discussion and Future Work

We are pleased to report that the overall per-pixel error rate drops by more than 45% through tighter ground-truthing, even on a large and diverse test set; the post-classifiers contine to drop the error rate by 24%. We believe there is room for further improvement.

We are working to prove or disprove that the sequence of post-classifiers converge to ground-truth in real problems. If such post-classifiers exist, at least how many features and training samples are necessary for the classification?

Acknowledgements

The data base and much of the software architecture is due to Michael Moll. We are grateful for insights and encouragement offered by Michael Moll, Jean Nonnemaker, and Sui-Yu Wang. We acknowledge the continually helpful advice and cooperation of Professor Dan Lopresti, co-director of the Lehigh Pattern Recognition Research laboratory.

References

[1] E. Alpaydin and C. Kaynak. Cascading classifiers, 1998.

[2] C. An, H. S. Baird, and P. Xiu. Iterated document content classification. In *Proc., IAPR 9th Int'l Conf. on Document Analysis and Recognition (ICDAR2007)*, Curitiba, Brazil, September 2007.

[3] 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.

[4] 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`.

[5] 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.

[6] V. Kolmogorov and R. Zabih. What energy function can be minimized via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI–26:147–159, Feb 2004.

[7] M. A. Moll and H. S. Baird. Document content inventory & retrieval. In *Proc., IAPR 9th Int'l Conf. on Document Analysis and Recognition (ICDAR2007)*, Curitiba, Brazil, September 2007.

[8] T. P. S. Nicolas, J. Dardenne and L. Heutte. Document image segmentation using a 2d conditional random field model. In *Proc., Int'l Conf. on Document Analysis and Recognition (ICDAR2007)*, pages 407–411, September 2007.

[9] J. Serra and J. Soille. *Mathematical Morphology and its Applications to Image Processing*. Kluwer Academic Publishers, Dordrecht, 1994.

[10] N. K. S. C. Sunil Kumar, Rajat Gupta and S. D. Joshi. Text extraction and document image segmentation using matched wavelets and mrf model. *IEEE Transaction on Image Processing*, IP–16.

[11] O. V. Y.Boykov and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI–23:1222–1239, November 2004.