Document Segmentation using Pixel-Accurate Ground Truth

Chang An, Dawei Yin and Henry S. Baird

Department of Computer Science & Engineering
Lehigh University
Bethlehem, PA 18015 USA
{cha305, day207}@lehigh.edu, baird@cse.lehigh.edu

Abstract—We compare methodologies for trainable document image content extraction, using a variety of ground-truth policies: loose, tight, and pixel-accurate. The goal is to achieve pixel-accurate segmentation of document images. Which ground-truth policy is the best has been debated [1], [2], [3], [4], [5], [6]. “Loose” truth is obtained by sweeping rectangles to enclose entire text blocks etc, and can be an efficient manual task. “Tight” truth requires more care, and more time, to enclose individual textlines. Pixel-accurate truth, in which only foreground pixels are labeled, can be obtained by applying the PARC PixLabeler [7] tool; in our experience this tool was as quick to use as loose truthing. We have compared the accuracy of all three truthing policies, and report that tight truth supports higher accuracy than loose truth, and pixel-accurate truth yields the highest accuracy. We have also experimented on morphological expansions on pixel-accurate truth, by expanding sets of foreground pixels morphologically, and report that expanded pixel-accurate truth supports higher accuracy than pixel-accurate truth.

Keywords—document content extraction; layout analysis; iterated classification; pixel-accurate

I. INTRODUCTION

In previous work [8], [9], [10], [11], we have described a family of algorithms for document image content extraction (DICE), able to find regions containing machine-printed text, handwriting, photographs, etc in images of documents. Our algorithms cope with a rich diversity of documents, image formats, image quality, languages, etc. We classify individual pixels, not regions, in order to avoid arbitrary restrictions on region shapes.

We developed a trainable iterated classification technology [12], [13] using a sequence of classifiers, each trained separately on the training-data results of the previous classifier, guided always by the same ground truth. Previous experiments have shown that iterated classifiers can significantly reduce per-pixel error rates and enforce local uniformity (“purity”) of regions. This strategy appears also to shrink clusters of wrongly classified pixels, while allowing regions of correctly classified pixels to expand. Thus it produces high-accuracy (more precisely, high recall), pixel-accurate segmentations.

The best methods for obtaining ground truth to support pixel-accurate segmentations is an open question. Previous discussions are reported in [2], [3], [4], [5]. We are also strongly motivated by the work of Shafait, Keysers and Breuel in [6]. In [1] we argued that the tighter (the more pixel-accurate) the truth, the better the resulting classifiers, and we showed this in large-scale experiments comparing “loose” to “tight” truth. A PARC research team agreed [7] with this observation, and built an efficient manual truthing tool, PixLabeler. We have applied this to our full-color and greyscale images, and compared it to earlier truthing tools we used. In section II, we introduce the method of iterated classification. In section III, we briefly discuss the ground-truth policies. In section IV, we describe morphological expansions on pixel-accurate ground truth. In section V, we describe the experimental design. In section VI, we present the experimental results. In section VII, we discuss the results. In section VIII and IX, we present conclusion and future work.

II. DESIGN OF ITERATED CLASSIFICATION

The goal of iterated classification is to enforce local uniformity without imposing arbitrary region shapes. We designed a trainable post classifier that operates on the output of the DICE 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, i.e. DICE classifier, the first stage classifier, the immediately following post-classifier is called the second stage classifier, followed by the third stage classifier, etc. A diagram of iterated classification is shown in Figure 1.

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 [10]. The features for the post-classifiers are discussed [12].

III. GROUND-TRUTH POLICIES

In this section, we briefly discuss the differences among three ground-truth policies.
We have developed a web-based user interface to zone document images in PNG format, using overlapped rectangles. Using this, we can capture loose and tight ground truth. Loose ground truth is obtained by sweeping rectangles to enclose entire block of a particular content type. This policy inevitably encloses blank space that is inter-column, inter-paragraph, at the end of a paragraph. Therefore, loose ground-truthing is an efficient manual task. Tight ground truth requires more care to enclose individual textlines, sometime even hand-written strokes or large letters. As a result, tens of times more rectangles, need to be swept. Pixel-accurate truth, in which only foreground pixels are labeled, is obtained by applying the PARC PixLabeler [7] tool; in our experience this tool was faster and easier than loose truthing methods.

IV. MORPHOLOGICAL EXPANSIONS OF PIXEL-ACCURATE TRUTH

To better understand the effectiveness of pixel-accurate ground truth, we generated morphological expansions on it. We expanded foreground pixels by applying morphological dilation operations [14] with a circular disk structuring element, that is, background pixels within a distance $d$ of a foreground pixel are labeled as the same class as the foreground pixel. We have generated four morphological expansions on pixel-accurate ground truth, using Matlab®, labeled by radii, in pixels, of the disks: $d=1, d=2; d=4$, and $d=8$.

V. EXPERIMENTAL DESIGN

We have previously compared the effectiveness of loose and tight ground truth in iterated classification [13], and found that tight truth reduces per-pixel classification errors by 45% (from 38.9% to 21.4%). Now we add experiments on pixel-accurate ground truth and its morphological expansions. We use a training set of 33 images and a distinct test set of 83 images, which are the same images we used in [13]. Together the two sets contain machine-print (MP), handwriting (HW), photograph (PH) and blank (“don’t care”) (BL). The training data was decimated randomly by selecting one out of every 9000th training sample.

We evaluated performance using per-pixel accuracy, precision and recall. Per-pixel accuracy is the fraction of all pixels in the document image that are correctly classified. Unclassified pixels are counted as incorrect. Precision is defined as the number of pixels correctly classified as belonging to a positive class divided by the total number of pixels classified as belonging to the positive class. Recall is defined as the number of pixels correctly classified as belonging to a positive class divided by the total number of pixels that actually belong to the positive class.

For each truthing policy, we trained a classifier using that kind of ground truth, then tested, and evaluated performance on the same type of ground truth.

VI. EXPERIMENTAL RESULTS

In this section, experiments comparing different ground-truth policies are presented.

The results of loose, tight and pixel-accurate truth are shown in the form of per-pixel error rate in Figure 2(a), which indicates that in each of the classification stages, total error rate (averaging over all classes) decreases as the truth goes from “loose” to “tight”, and finally to pixel-accurate (via PixLabeler). Notice that the drop of error rate from tight to pixel-accurate is less significant than that from loose to tight. The figure also show that the error rate decreases monotonically as a function of stages. For the fourth stage, the difference between error rates of tight and pixel-accurate is...
moderate. It seems like a clear win for pixel-accurate ground truth.

However, a closer look at each content class reveals problems. For one thing, the overall error rate drops only slightly (from 16% to 14%) from tight to pixel-accurate. More seriously, some classes suffer a catastrophic fall in recall in the final move to pixel-accurate ground truth: for examples, handwriting recall drops from 70% to 2%.

The results of precision and recall for morphological expansions on pixel-accurate truth, along with loose, tight and pixel-accurate, are shown in Figure 2(b), Figure 3(a) and Figure 3(b). In all the three figures, pixel-accurate truth is labeled as “d=0”.

In Figure 2(b), the recall on handwriting (HW) increases as the pixel-accurate truth expanded, and reaches the highest with tight truth. The recall on machine-print (MP) and photograph (PH) also increases as pixel-accurate truth is expanded (i.e. d=1,2,4,8). Figure 2(b) also suggests that one of the pixel-accurate truth’s morphological expansions, labeled as “d=8”, yields results similar to tight truth.

In Figure 3(a), we can see that some morphological expansions yield higher precision on machine-print (MP) and photograph (PH) than pixel-accurate ground truth. Note that precision on handwriting (HW) increases monotonically from pixel-accurate truth to the morphological expansion of “d=8”. Figure (a) also suggests that the morphological expansion of “d=8” is nearly equivalent to tight truth on MP and PH. In Figure 3(b), notice that the number of MP pixels increases the most in both quantity and percentage, and HW the least.

**VII. DISCUSSION**

The results suggest that something about pixel-accurate ground truth can damage particular classes. Several things could have caused the damage: (1) imbalance in the training...
set (handwriting pixels were significantly fewer than others); (2) confusion between foreground pixels and background pixels that are inter-character, inter-word and inter-line, etc; and (3) bad fit with the features (e.g. the radius of the feature extraction window is badly chosen).

Figure 2(b), Figure 3(a) and Figure 3(b) show that a moderate increase (less than eight times) of ground-truthed HW pixels leads to a significant improvement (more than 23 times) of recall (from 2% to 46%).

Our algorithms extract features from a circular window (of radius pixels) centered on target pixels. It is possible that the background pixels (inter-character and inter-word) have similar feature values as foreground pixels, e.g. handwriting. When loaded to hashed bins, such background pixels would be hashed with foreground pixels in the same bin, thereby causing classification mistakes.

VIII. CONCLUSION

We have compared performance among three ground-truth policies, loose, tight and pixel-accurate, along with morphological expansions on pixel-accurate truth. Experiments suggest that pixel-accurate ground truth can be captured for high-contrast document images as easily as loose ground truth, and can improve overall accuracy. However, some individual classes might suffer catastrophically. We conclude that for some classes, especially handwriting, some “background” pixels need to be included in ground truth along with foreground pixels.

IX. FUTURE WORK

As discussed in section VII, the feature-extraction window size may be crucially large compared to pixel-accurate ground truth. Experiments with more dilations of pixel-accurate truth might reveal the relation between them. Expanding foreground pixels to include background pixels that are inter-character, inter-word and inter-line can produce another set of morphological expansions on pixel-accurate ground truth. And a careful investigation of these morphological expansions could help explain confusions between foreground and background pixels. Also we think it would be interesting to see what happens when a classifier trained on one ground truth policy is evaluated by a different policy – if such an experiment is not unethical!

REFERENCES


