Automated Ground Truth Data Generation for Newspaper Document Images

Thomas Strecker [1], Joost van Beusekom[2], Sahin Albayrak[1], Thomas M. Breuel[2,3]

[1]DAI Labor Technical University Berlin, Germany
[2]Image Understanding and Pattern Recognition (IUPR) Research Group
Technical University of Kaiserslautern, Germany
[3]German Research Center for Artificial Intelligence (DFKI), Kaiserslautern, Germany
{joost, tmb}@iupr.com, {thomas.strecker,sahin.albayrak}@dai-labor.de

Abstract

In document image understanding, public ground-truthed datasets are an important part of scientific work. They do not only helpful for developing new methods, but they are also a point of intersection allowing to compare the methods performance without need to implement it. For document image understanding several datasets exists, each having its own pros and cons. Generating these datasets is time consuming and costly work and therefore each existing and new dataset is valuable. In this paper we propose a way to generate a ground-truthed dataset for newspapers. The ground truth in focus is layout analysis ground truth. The proposed two step approach consists of a layout generating module and an image matching module allowing to match the ground truth information from the synthetic data to the scanned version. Using the "MyNews" system, newspaper layouts are generated using a news corpus. The output consists of a digital newspaper (PDF file) and an XML file containing geometric and logical layout information. In the second step, the PDF files are printed and scanned. Then the scanned document image is aligned with the synthetic image obtained by rendering the PDF. Finally the geometric and logical layout ground truth is mapped onto the scanned image.

1 Introduction

Public document image datasets with ground truth information play an important role, not only in the document image understanding community. They are not only useful while developing new approaches. Their major significance comes from their role as point of intersection for many methods solving the same or similar problems: public datasets allow to compare the performance of different methods without the need to implement these from scratch.

Generating such datasets is a time-consuming and costly procedure. Most often, the needed ground truth has to be created manually. The complexity of needed ground truth is also an inhibiting factor: the more detailed ground truth is needed, the more effort has to be done to generate it. This is the main reason why so little different ground-truthed datasets exist for layout analysis.

In this work we present a new approach for ground-truthed dataset generation. A layout creation system used to generate personalized newspapers is adapted to generate a set of newspapers and their corresponding ground truth. The resulting digital documents files are then printed, if needed also degenerated and then scanned again. The ground geometric and logical layout information obtained during the layout generating step is mapped to the scanned image. Finally a set of ground-truthed document images is obtained. An overview of the system can be found in Figure 1.

Section 2 gives a short overview over existing datasets and similar methods. Section 3 presents the details of the layout generating process. Then, Section 4 explains how the ground truth information is matched to the scanned images. Section 6 concludes this paper.

2 Related Work

Guyon et al. [9] give an overview over the existing datasets for optical character recognition and document understanding back in 1997. Several new datasets for page segmentation have been released since then. For several ICDAR contests, small datasets have been made available (about 20 to 40 images each), e.g. the ICDAR 2001 newspaper segmentation contest [7] or the ICDAR 2003, 2005 and 2007 page segmentation contests [3, 1, 2]. Recently, in 2005, Todoran et al. [17] released the UvA color document dataset consisting of 1014 color document images of different magazines.

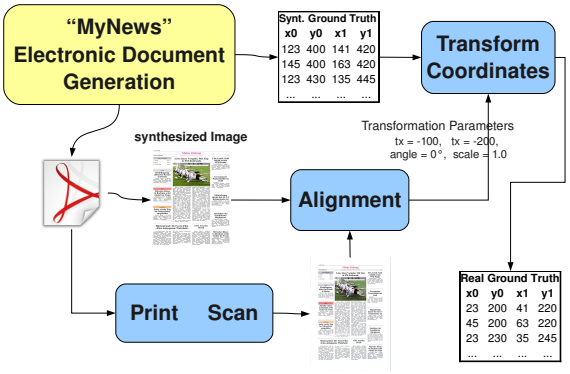

Figure 1. A digital newspaper is generated together with the ground truth. The document is printed, scanned and aligned to the synthetic image. These are then aligned. Using the resulting transformation parameters the coordinates of the ground truth components in the scanned image are computed.

It is clear that diversified and perfect ground truth will by definition not be generated in a total automatic process. Nevertheless some research has been done concerning automated generation of ground truth for different document image understanding tasks.

For Optical Character Recognition (OCR) several approaches have been presented to generate ground truth. Forced alignment is being used to generate the data from line- or word-level transcription in [21, 12]. Kanungo et al. [13, 14] present a closed-loop approach using document image alignment for automatic generation of OCR ground truth. In our previous work [19] we extended this method with more robust alignment allowing to cope with distortions introduced during printing and scanning process.

In the field of page segmentation ground truth generation, Héroux et al. [11] presented a system generating synthetic images with corresponding ground truth.

3 Layout Generation

3.1 MyNews System

The layouts are generated by the "MyNews" system. This system is part of a project which examined the personalization of user interfaces and used the idea of personalized newspapers as one example. Users can choose from a set of news sources and topics in order to create their personal newspaper. Contents from these sources can be collected either via a web extraction framework, web services or FTP and is subsequently transformed into an internal XML format. These XML files are the input for the layout algorithm.

The goal was to provide a daily newspaper with articles especially relevant for a user laid out in a way which resembles as much as possible the look-and-feel of traditional newspapers. Therefore, the system uses a style guide which is described in the following section.

3.2 Style Guide

The basic concept of the layout is the employment of a 4-by-16 grid structure (cf. Figure 2. Grids are a well-known design principle[1,2] and most newspapers use it in a more or less strict way (exceptions are typically found in the yellow press).

When looking at models of aesthetics (cf. [15], [10]) it is clear, that the use of a grid naturally optimizes the alignment, regularity and uniformity and separation criteria of the presented measures.

Based on the grid, articles may only occupy a rectangular area of connected cells. While the text of articles is broken into lines which may not span more than one column, article headlines and media may span any number of columns. Depending on the number of columns a headline is laid out in different sizes but never hyphenated unless a hyphen is already present in the text; article text is laid out in a single size and hyphenation is applied when needed. After breaking the text of a paragraph into lines, whitespace is inserted to simulate justified alignment of the text.

If an article contains media elements the best one is chosen based on the desired width of media and aspect ratios. This ensures that media elements retain their original aspect ratio and are not skewed.

After each part (headline, article text and media) has

[1] http://www.smashingmagazine.com/2008/02/11/award-winning-newspaper-designs/

[2] http://poynteronline.org/column.asp?id=47&aid=37529

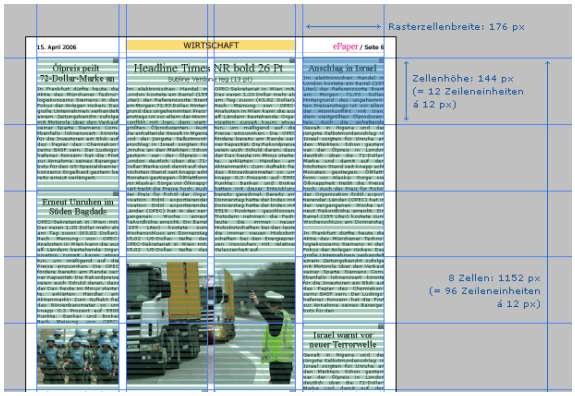

Figure 2. Grid Structure for Pages

been laid out, they are arranged to form a layout variant for the article. Because an article may be laid out spanning any number of columns, and with or without media elements in different positions, each article can be laid out in several variants instead of only one. The procedure for selecting variants from the pool and placing them on a page is done with an optimization algorithm which is described in the next section.

3.3 Optimization Procedure

As was shown in [16], different optimization algorithms can be applied to the task of selecting and placing laid out articles on a grid which is essentially a Cutting & Packing problem ([20]).

For the generation of the presented pages we used a version of the ϵ-approximate relative greed algorithm which provides predictable results of good quality at a fair computational complexity.

Greedy algorithms for layout work by placing the next item at the next available location. Therefore, it is the task of the algorithm designer to take care of the detection of feasible locations for an item and the ordering of items because the algorithm vitally depends on these. In our case the order of the items is determined by their "density", i.e. the ratio of their value and their weight ($d_i = \frac{v_i}{w_i}$). In the case of generating newspaper pages we define the weight and value of an article layout to be the area it occupies and the contribution to the total value, respectively.

The total value of a layout is then computed with the formula given in [16] which considers the relevance of the original article, the coverage of the layout and the contribution to the coverage of the page, i.e.

$$f(i) = r_i + c_i + \frac{w_i * h_i}{WH - \sum_{\substack{j \in \mathcal{B} \\ j \neq i}} w_j h_j}$$

where i is the layout to score, w_i and h_I are the width and height of the item, W and H are the number of rows and columns of the page, and $\mathcal{B}$ is the set of items already placed on the page. The total page score is then computed as the sum over all item values.

The placement strategy is an extension of the basic bottom-left heuristic ([6]) which allows skipping areas into which no item fits. Figure 3 a, b and d show the computation of the next available area in different situations; c shows the case when skipping of the next free area (dark) creates a new placement option for large items.

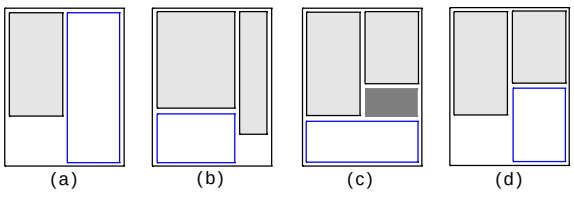

Figure 3. Strategy used for placing articles

The optimization algorithm finally works in two steps: At first, it iterates over all items and places them in the first available position and fills the page by recomputing the densities of the remaining items and selecting the one with the highest density for placement until no item can be placed anymore. The second step consists of selecting the best of the achieved layouts and returning it.

4 Ground Truth Data Mapping

In order to obtain more realistic data, the generated electronic documents are printed, degraded analogously if wanted and scanned again. The degradation step can be used to generate data containing a specific kind of problem, e.g. using thin paper to increase the bleed-through effects.

The next steps consist of mapping the ground truth for the electronic document to the scanned images.

- Step 1: generate synthetic images of the electronic document
- Step 2: compute transformation parameters to align synthetic and scanned document
- Step 3: compute positions of ground truth elements in the scanned image using the transformation parameters obtained in step 2.

Step 1 can be solved using standard software for rendering electronic documents, in our case PDF files, to image files.

Step 2 is solved using a method described in our previous work [18]. A short overview will be given in Section 4.1.

In step 3 the coordinates of the ground truth elements are transformed to the coordinates in the scanned image and so all information needed to generated the ground truth for the scanned images is generated.

4.1 Document Image Alignment

The method used for aligning two document images has been described in more detail in our previous work [18]. In the following a short overview of the alignment method will be given.

Step 2 consists of finding the transformation parameters tx and ty (translation in x and y direction), s (scale) and α (rotation angle) that align both images so that they are superimposed.

For finding the optimal parameters, we use an optimal branch-and-bound search algorithm, called RAST [4]

(Recognition by Adaptive Subdivision of Transformation Space). The quality function optimized by the branch-and-bound search is defined as the number of model points matching an image point under the error bound ϵ.

At start, the algorithm is initialized with the whole parameter space, also called transformation space. Let $[tx_{min}, tx_{max}] \times [ty_{min}, ty_{max}] \times [a_{min}, a_{max}] \times [s_{min}, s_{max}]$ be the initial search space, where tx stand for translation in x direction, ty translation in y direction, a for the rotation angle and s for the scale.

Next, the parameter space is divided into two parts. The quality of these parts, also called parameter subspaces, is computed. Let $B = \{b_1, \ldots, b_N\} \in R^2$ be the set of image points of the scanned image and $M = \{m_1, m_M\} \in R^2$ the set of image points of the synthetic image, also called "model points" (in order to stick to the original notation of the RAST algorithm). For each model point m, a bounding rectangle $G_R(m)$ can be computed using the transformation space to be searched. This rectangle represents the possible positions where a model point m may be transformed to, using all possible transformations from the current transformation subspace. If the distance $d = min_{g \in G_R(m), b \in B} ||g - b||$ is less than a threshold ϵ, the quality of the parameter subspace is incremented. A more detailed description of RAST can be found in [4, 5].

RAST uses a priority queue containing the parameter subspaces in order of their upper bound quality. The subspace with highest priority is divided into two new subspaces, by splitting it into two parts of equal size. For each part, the new upper bound quality is determined and both subspaces are added into the priority queue. These steps are repeated until a stopping criterion is met. In our case the method stops when the size of the remaining parameter subspace is smaller than a given threshold.

Centers of connected components are used as image points, as these are relatively stable and easy to compute. A filtering step is added before the branch-and-bound search to speed up the computation of the upper bound of the quality: to avoid comparing bounding boxes that are not similar at all, Fourier descriptors for the contour of the connected components have been extracted [8], describing the shape of the connected component. In order to be invariant to scale and rotation, the connected components are downscaled to a fixed size and the phase is discarded to obtain rotation invariance. For each connected component only the n most similar image points are considered for the quality estimation. The value of $n = 50$ was chosen manually and proved to work well for standard documents.

5 Discussion

In Figure 4 a few examples of generated layouts are shown. It is clear that using this automated method a general dataset covering a wide variety of all possible (and sometimes weird) newspaper layouts cannot be created. However, we plan to generate a dataset that should be released soon [3] and future improvements of the layout algorithm can lead to even more general data sets. This will hopefully present a good starting point for evaluation of different layout analysis and page segmentation methods on newspaper data.

6 Conclusion

In this paper we presented a new approach for ground-truthed dataset generation. Using an automatic layout producing system for personalized newspapers, synthetic ground-truthed images are generated. In the second step these are printed, degenerated (if needed) and scanned in again. A document image alignment technique is used to compute the transformation needed to map the ground truth from the generation step to the scanned image. Logical as well as geometrical ground truth for layout analysis is obtained. Optionally also OCR ground truth can be generated. It is planned to generate a set of ground-truthed pages, based on texts by dpa[4] and images freely available on wikimedia commons[5], and to make them available for research purposes.

References

[1] A. Antonacopoulos, D. Bridson, and B. Gatos. Page segmentation competition. In *Proc. of the 8th Int. Conf. on Document Analysis and Recognition*, pages 75–79, Washington, DC, USA, 2005. IEEE Computer Society.

[2] A. Antonacopoulos, B. Gatos, and D. Bridson. Page segmentation competition. In *Proc. of the 9th Int. Conf. on Document Analysis and Recognition*, pages 1279–1283, Washington, DC, USA, 2007. IEEE Computer Society.

[3] A. Antonacopoulos, B. Gatos, and D. Karatzas. Icdar 2003 page segmentation competition. In *ICDAR '03: Proceedings of the Seventh Int. Conf. on Document Analysis and Recognition*, pages 688–692, Washington, DC, USA, aug 2003. IEEE Computer Society.

[4] T. M. Breuel. A practical, globally optimal algorithm for geometric matching under uncertainty. *Electronic Notes in Theoretical Computer Science*, 46:1–15, 2001.

[5] T. M. Breuel. Implementation techniques for geometric branch-and-bound matching methods. *Computer Vision and Image Understanding*, 90(3):258–294, 2003.

[3] When the dataset is ready the link will be added here.

[4] The Deutsche Presse Agentur was kind enough to grant us permission to use news articles for our corpus and for research purposes.

[5] We are grateful for all the people who provided the images and made them available for free. All images used in the corpus have been nominees for the "Picture of the Year" awards of wikimedia commons, proving they are outstanding works of photography.

Figure 4. Examples of generated layouts.

[6] B. Chazelle. The bottom-left bin-packing heuristic: An efficient implementation. *IEEE Transactions on Computers*, 32(8):697–707, 1983.

[7] B. Gatos, S. Mantzaris, and A. Antonacopoulos. First int. newspaper segmentation contest. In *Proc. of 6th Int. Conf. on Document Analysis and Recognition*, pages 1190–1195, Los Alamitos, CA, USA, 2001. IEEE Computer Society.

[8] G. H. Granlund. Fourier Preprocessing for Hand Print Character Recognition. *IEEE Trans. on Computers*, C–21(2):195–201, 1972.

[9] I. Guyon, R. M. Haralick, J. J. Hull, and I. T. Phillips. *Data Sets for OCR and Document Image Understanding Research*, pages 779–799. World Scientific, Singapore, Singapore, 1997.

[10] S. J. Harrington, J. F. Naveda, R. P. Jones, P. Roetling, and N. Thakkar. Aesthetic measures for automated document layout. In *DocEng '04: Proceedings of the 2004 ACM symposium on Document engineering*, pages 109–111, New York, NY, USA, 2004. ACM.

[11] P. Heroux, E. Barbu, S. Adam, and E. Trupin. Automatic ground-truth generation for document image analysis and understanding. In *Proc. of the 9th Int. Con. on Document Analysis and Recognition*, pages 476–480, Washington, DC, USA, 2007. IEEE Computer Society.

[12] S. Jaeger, S. Manke, J. Reichert, and A. Waibel. Online handwriting recognition: the npen++ recognizer. *Int. Journal on Document Analysis and Recognition*, 3(3):1433–2833, jun 2001.

[13] T. Kanungo and R. M. Haralick. An automatic closed-loop methodology for generating character groundtruth for scanned documents. *IEEE Trans. Pattern Anal. Mach. Intell.*, 21(2):179–183, 1999.

[14] D.-W. Kim and T. Kanungo. Attributed point matching for automatic groundtruth generation. *Int. Journal on Document Analysis and Recognition*, 5(1):47–66, 2002.

[15] D. C. L. Ngo, L. S. Teo, and J. G. Byrne. Modelling interface aesthetics. *Information Sciences*, 152:25–46, June 2003.

[16] T. Strecker and L. Hennig. Automatic layouting of personalized newspaper pages. In *OR '08: Proceedings of the International Conference on Operations Research*, Augsburg, Germany, September 2008. accepted for publication.

[17] L. Todoran, M. Worring, and A. W. M. Smeulders. The uva color document dataset. *Int. Journal on Document Analysis and Recognition*, 7(4):228–240, 2005.

[18] J. van Beusekom, F. Shafait, and T. M. Breuel. Image-matching for revision detection in printed historical documents. In *DAGM 2007, Pattern Recognition, 29th DAGM Symposium*, volume 4713 of *Lecture Notes in Computer Science*, pages 507–516, 2007.

[19] J. van Beusekom, F. Shafait, and T. M. Breuel. Automated ocr ground truth generation. In *8th IAPR Workshop on Document Analysis Systems*, pages 111–117, Nara, Japan, sep 2008.

[20] G. Wascher, H. Hau[ss]ner, and H. Schumann. An improved typology of cutting and packing problems. *European Journal of Operational Research*, 127(3):1109–1130, December 2007.

[21] M. Zimmermann and H. Bunke. Automatic segmentation of the IAM off-line database or handwritten english text. *Proc Int. Conf. on Pattern Recognition*, 04:40035, 2002.