Automated OCR Ground Truth Generation

Joost van Beusekom[1], Faisal Shafait[2], Thomas M. Breuel[1,2]

Image Understanding and Pattern Recognition (IUPR) Research Group
[1]Technical University of Kaiserslautern, Germany
[2]German Research Center for Artificial Intelligence (DFKI), Kaiserslautern, Germany
{joost, tmb}@iupr.net, faisal.shafait@dfki.de

Abstract

Most optical character recognition (OCR) systems need to be trained and tested on the symbols that are to be recognized. Therefore, ground truth data is needed. This data consists of character images together with their ASCII code. Among the approaches for generating ground truth of real world data, one promising technique is to use electronic version of the scanned documents. Using an alignment method, the character bounding boxes extracted from the electronic document are matched to the scanned image. Current alignment methods are not robust to different similarity transforms. They also need calibration to deal with non-linear local distortions introduced by the printing/scanning process. In this paper we present a significant improvement over existing methods, allowing to skip the calibration step and having a more accurate alignment, under all similarity transforms. Our method finds a robust and pixel accurate scanner independent alignment of the scanned image with the electronic document, allowing the extraction of accurate ground truth character information. The accuracy of the alignment is demonstrated using documents from the UW3 dataset. The results show that the mean distance between the estimated and the ground truth character bounding box position is less than one pixel.

1. Introduction and Previous Work

For the development of optical character recognition (OCR) systems ground truth data, that is a set of character images and the corresponding character codes, plays an important role. Many trainable recognizers need large amounts of training data in order to perform well on the recognition task. But also in evaluation labeled ground truth data is useful: the performance of OCR systems can be measured on character level recognition, for which character level text ground truth is indispensable.

Manually generating character level ground truth is a time consuming and costly process, as each single character has to be marked and labeled with the correct character code. Considering that a typical machine printed page contains more than 2000 characters, it is clear that generation of even a few pages of labeled character-level ground truth is expensive. This may also explain why the UW3 dataset contains 1600 document images with layout information ground truth but only 33 document images with character-level ground truth.

To overcome this problem, several approaches have been presented in literature: one approach is to use synthesized images from an electronic document, where the ground truth is available, on which an image degradation model is applied. Different degradation models have been described in literature [1]. However, there is little consensus which model for the degradation is the right one [1]. Moreover, in the document analysis community no clear consensus can be found whether synthesized images are better, worse or equivalent in quality than real data.

In literature, forced alignment has been used for generating ground truth for handwritten as well as OCR ground truth data. The idea of these methods is to force the ASCII transcription of the text line to fit to the image representing the text line using an initially trained recognizer. In Zimmermann's method for generating handwriting ground truth [12], this is done using a hidden Markov model recognizer. The path maximizing the probability for finding the word from the ASCII transcription gives the optimal cutting points. A similar approach is used by Jaeger [6]. Similar methods has also been applied to printed text. The disadvantage of this method, apart from being less accurate than manually labeled ground truth, is that a trained HMM recognizer and the transcription is needed.

The approach by Kanungo et al. [7] combines the ease of synthetic data together with real world document image degradation: they use electronic documents to extract the ground truth information (character position, size and

ASCII code). Furthermore the document is printed and scanned in again. Then the electronic document and the scanned document image are aligned, allowing to compute the positions of the characters in the scanned document image. As many scanners tend to add distortions that cannot be described by similarity transformations, a calibration step has been introduced. Kim and Kanungo [8] propose a more robust alignment method, which they test on the UW3 dataset.

Gang et al. [11] present a ground truth generating system using image degradation as well as real world data. They uses a similar global alignment approach as proposed by Kanungo et al., but without any local adaption to non linear distortions.

Our approach consists of a two step alignment: in a first step the global transformation parameters are estimated, similar to Kim and Kanungo's method. In case of similarity transformations (translation, scale and rotation), this first step is enough to obtain a good alignment. The second step does local adaptation of smaller regions of the image: nearby characters are clustered together and this cluster is then aligned a second time, starting with the global alignment parameters. This allows to adapt automatically to scanner distortions.

In Section 2 we present the overview of our approach and the global alignment step. Section 2.1 describes the global alignment and Section 2.2 the local adaption step. Section 3 describes the experimental setup and the error measure which was also used by Kim and Kanungo in [8], to measure the performance of the alignment method. Section 4 shows the results and finally Section 5 concludes this paper.

2. Alignment for Ground truth extraction

The proposed approach for OCR ground truth generation uses electronic documents as a starting point. The electronic document is used three times: once, for printing out a paper version of the document; once for generating a synthetic image of the document and once for extracting the character bounding boxes and the corresponding ASCII code.

The print out is scanned in again. The synthetic image is used for aligning the scanned image to the synthetic one. Using the transformation parameters obtained by the alignment, the position of bounding boxes of the characters can be computed based on the ground truth obtained from the electronic document. An illustration of the method can be found in Figure 1.

As scanner degradations can be quite arbitrarily (e.g. stretching or squeezing of the page), a global alignment technique based on similarity transforms will just be accurate only if neither the printer nor the scanner add any distortions.

To overcome this limitation our alignment is composed of two steps: first a global alignment is used to get the estimate of the global similarity transformation parameters. Second, a local adaption of the alignment parameters is done. Starting from the global parameters, a narrow search space around these parameters is searched for parameters aligning the selected subregion better. This allows to adapt to non uniform distortions as they are produced through scanning and/or printing.

2.1. Global Alignment

The global alignment used for our approach was presented in more detail in our previous work [10]. In the following a short overview of the global alignment will be given.

The alignment of two document images aims at identifying the transformation parameters that allow to overlay both images. For this purpose we use an optimal branch-and-bound search algorithm, called RAST [2] (Recognition by Adaptive Subdivision of Transformation Space). This method allows finding the globally optimal parameters describing the transformation needed to align both images.

The quality function used for document image alignment is defined as the number of model points matching an image point under the error bound ϵ.

RAST algorithm uses a branch-and-bound search for quickly finding an global optimum. It uses a priority queue containing 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 sub space is smaller than a given threshold.

For applying RAST, an initial parameter space (also called transformation space) has to be defined. 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 upper bound quality 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.

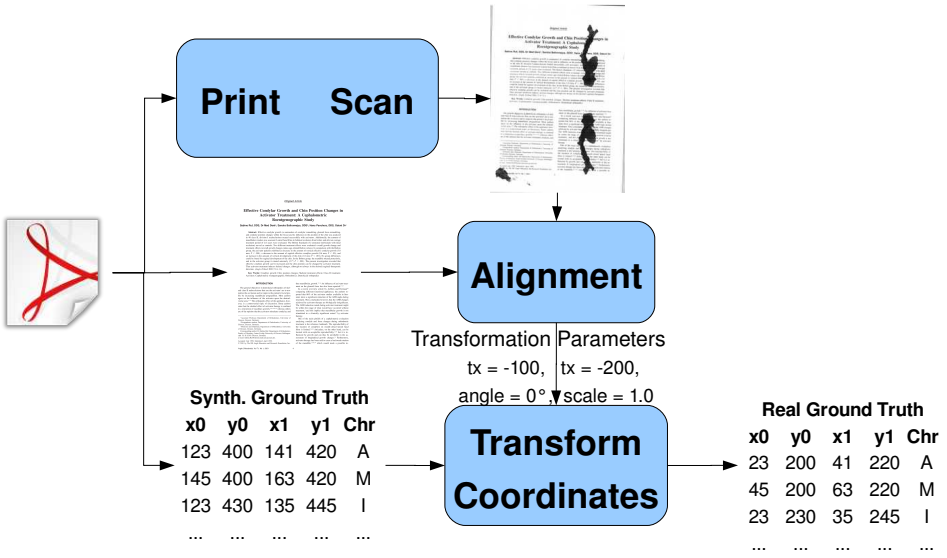

Figure 1. An overview of the method: from the digital document (PDF in our case) a print out, a synthetic image and the ground truth information is extracted. The printed version is scanned and then aligned with the synthetic image. The transformation parameters obtained during alignment are used to compute the positions of the ground truth in the scanned document image.

If the distance d, defined as $d = min_{g \in G_R(m), b \in B} g - b$ is less than a threshold ϵ, the quality of the parameter subspace is incremented. A more detailed description of RAST can be found in [2, 3].

As image points we choose the centers of connected components, as they are relatively stable and easy to compute. In order to speed up the computation of the upper bound for the quality, a filtering step is added before the branch-and-bound search: to avoid comparing bounding boxes that are not similar at all, Fourier descriptors for the contour of the connected components have been extracted [5], describing the shape of the connected component. In order to be invariant to scale and rotation, the images of the connected components are downscaled to a fixed size and the phase is discarded to obtain rotation invariance for the Fourier Descriptors. For each connected component only the 50 most similar image points are considered for the quality estimation. The value of 50 was chosen manually and showed to work quite well for standard documents.

2.2. Local Alignment

After obtaining initial parameters $tx_i, ty_i, s_i, \alpha_i$ from the global alignment procedure, local adaption is done. First, the model points are clustered according to their local neighborhood. This clustering can be implemented in different ways, if two conditions are satisfied:

- "Locality": the selected local constellation of model points should extend to an area small enough to adapt to the local distortions.
- "Local uniqueness": the selected local constellation of model points should be unique given the parameter search space. This means that there is only one set of parameters transforming the local model points to the image points while obtaining maximum quality.

The two extrema of the clustering are thus:

- one cluster containing all model points: this will not adapt to any local distortions
- one cluster per model point: this may lead to alignments where the local neighborhood constellations of the model points may be broken up, thus leading to a wrong assignment.

For our experiments on local adaption we used a nearest neighbor approach for clustering: for a randomly selected cluster center, all model points are added that are within a certain distance of the cluster center, e.g. all points that are closer than 200 pixels.

Given a reasonable clustering, each cluster can now be adapted locally. Principally, the same method as for the global matching can be used, with the following differences:

- *Search space*: using the transformation parameters

from the global alignment, a new search space is defined in narrow bounds around these parameters.

- *Feature points*: instead of taking all components from both images, only the components belonging to the currently analyzed cluster are taken. From the second image, only the components that are within the area that is covered by the new search space have to be considered.

Instead of taking centers of connected components as image points, edges are extracted from the clusters. This leads to more robust results than just taking one point per connected component, as connected components are sensitive to noise. The alignment is done using RAST for matching edges instead of single points. The edges are detected using Canny edge detector [4].

For each cluster we obtain a new set of adapted transformation parameters $tx_a, ty_a, s_a, \alpha_a$ which is used to compute the ground truth position of characters in the scanned image.

In Figure 2 a real world example can be found showing the effect of local adaptation. An example document from the UW3 dataset was printed and scanned by a commercial flatbed scanner. The result of the global alignment and the result after local adaptation can be seen.

3. Evaluation and Error Measure

In order to test the global matching and for our results to be comparable with previous work by Kim and Kanungo [8], we follow their evaluation approach. The University of Washington data set [9] contains among other document images 33 document images together with character level ground truth consisting of the bounding box coordinates and the ASCII code of the character in the bounding box. Ten images where chosen randomly and transformed using the following parameters: For the first test without rotation:

- $X_t = \{-50, 0, 50\}$
- $Y_t = \{-50, 0, 50\}$
- $S = \{0.65, 0.8, 1.0, 1.2, 1.35\}$

For the second test with rotation:

- $(X_t, Y_t) = \{(0, 0), (50, 0), (100, 0)\}$
- $R = \{0, 1, 3\}$
- $S = \{0.65, 0.8, 1.0, 1.2, 1.35\}$

The initial search space is given by:

- $X_t = [-100, 100]$
- $Y_t = [-200, 200]$
- $R = [-10, 10]$
- $S = [0.6, 1.4]$

The error measure is the Euclidean distance between the centers of the ground truth bounding boxes and the aligned bounding boxes. It is defined as follows: given the bounding box coordinates x_l, y_l, x_h, y_h. The coordinates of the ground truth boxes after applying the ground truth transformation parameters t_x, t_y, s, α are given by $x_{lg}, y_{lg}, x_{hg}, y_{hg}$. The center of the transformed bounding box is given by $x_{cg} = \frac{x_{lg}+x_{hg}}{2}$ and $y_{cg} = \frac{y_{lg}+y_{hg}}{2}$. Applying the estimated transformation on the ground truth bounding box leads to coordinates $x_{le}, y_{le}, x_{he}, y_{he}$ with center coordinates $x_{ce} = \frac{x_{le}+x_{he}}{2}$ and $y_{ce} = \frac{y_{le}+y_{he}}{2}$. The Euclidean distance is then be computed by $d = \sqrt{(x_{ce} - x_{cg})^2 + (y_{ce} - y_{cg})^2}$.

Another error measure we used is the mean and the maximum difference of the transformation parameters. For each document, the difference between the ground truth parameters and the estimated parameters is computed. For each parameter (translation in x and y direction, rotation angle and scale factor), the mean and the maximum of the absolute differences are shown.

As currently no ground truth is available for testing the local adaption, we give only an example of the effect of the local adaption. Initial experiments and visual inspection of the results show that the here proposed local adaption is able to adapt to distortions introduced by scanning devices. Application on carefully camera-captured documents is also possible, as far as the distortions are small enough that it can be approximated by translation, rotation and scale on a cluster level.

4. Results

The histogram of the distances between ground truth bounding boxes and the estimated bounding boxes for the test without rotation can be found in Figure 3. It can be seen that the overall estimation of the parameters is quite accurate, as all estimates are within 1.5 pixels distance to the ground truth box. The mean distance is about 0.65 pixels and the maximum distance is about 1.09 pixels.

The same can be concluded for the histogram of the distances between ground truth bounding boxes and the estimated bounding boxes for the test with rotation in Figure 4. The mean distance is about 0.68 pixel, the maximum distance 1.77 pixel. The mean and maximum distance are slightly higher than for the test without rotation, but still most are below one pixel distance.

Compared to the results of Kim and Kanungo [8], there is some change to the better concerning the test without rotation, as nearly all of the components where aligned within

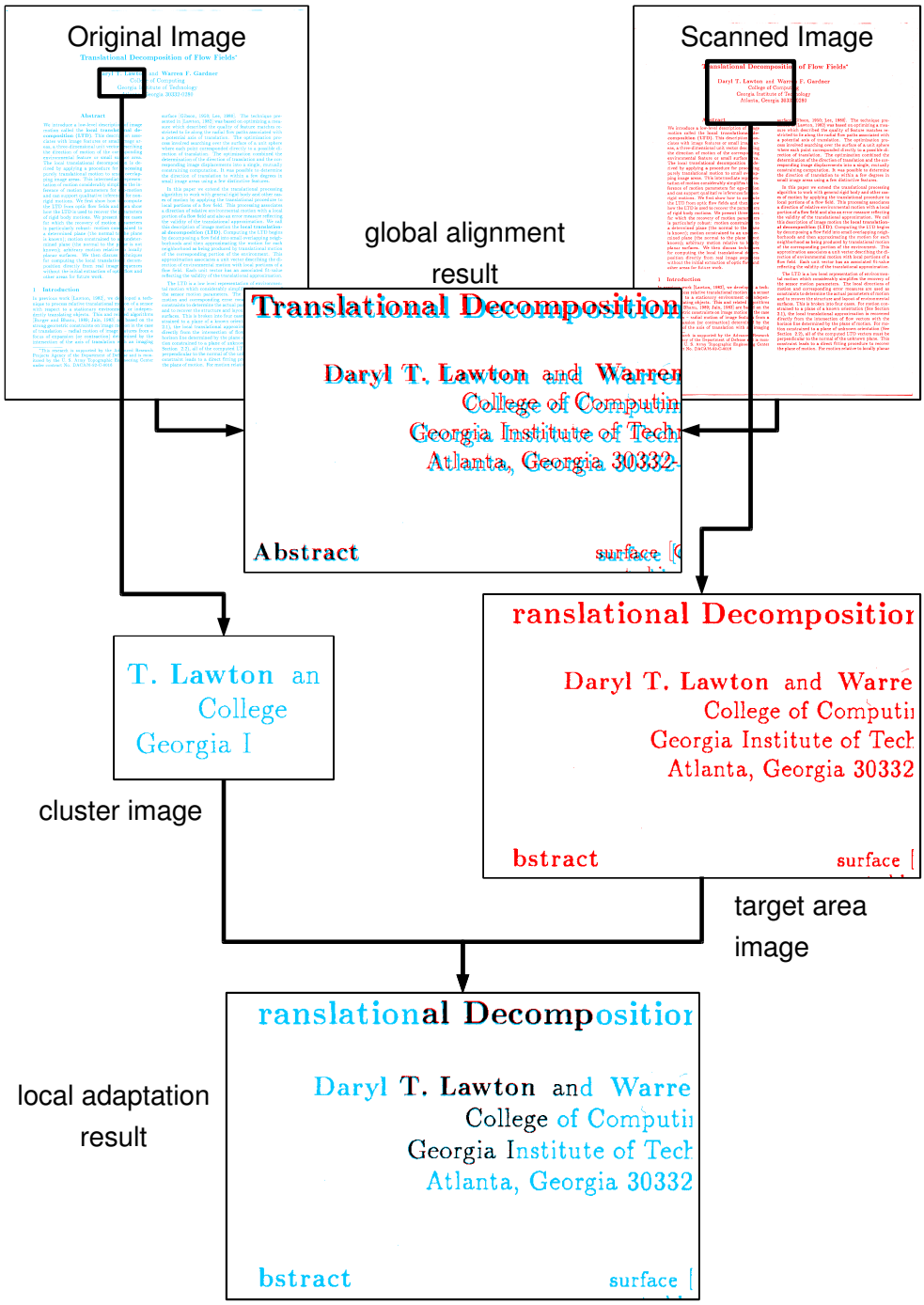

Figure 2. Example of local adaptation. The document was scanned using a commercial flatbed scanner. The original image is colored in blue, the scanned one in red. Black pixels are used to show overlapping pixels (positions that are black in both images). Due to local distortions introduced by the scanner, the global alignment is not perfect. Local adaption is used to compensate for these distortions. The cluster image is realigned locally to the target area. The new resulting aligned image is shown at the bottom.

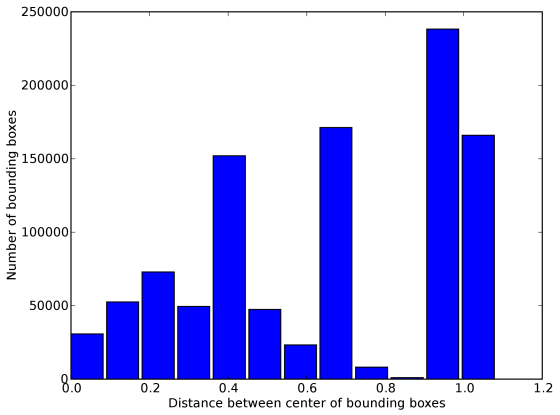

Figure 3. Results for the alignment test without rotation. It can be seen, that the maximum error is below 1.1 pixel, and the mean distance is about 0.7

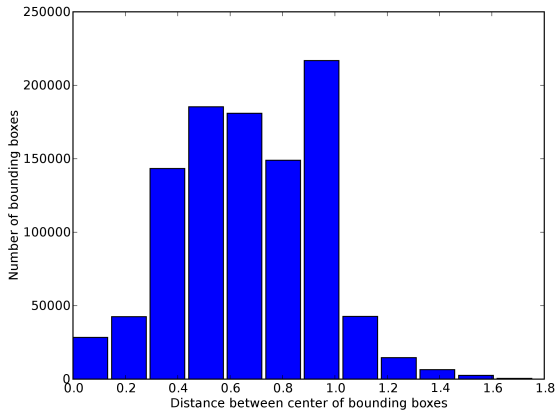

Figure 4. Results for the alignment test with rotation. It can be seen, that the maximum error is below 1.8 pixel, and the mean distance is about 0.7

one pixel accuracy: maximum error reported by Kim et Kanungo is about 4.0, whereas the maximum error of our approach is around 1.8 pixel. The major improvement can be seen on the test with rotation: our method is nearly as accurate as on the test without rotation, whereas Kim and Kanungo's method leads to distances of as much as 200 pixels for even small rotation angles.

In Table 1 the mean and maximum differences for the transformation parameters can be found. It can be seen that the estimates are very close to the ground truth transformation parameters.

Table 1. The mean and maximum difference between the estimated and the ground truth transformation parameters are given for the test without and with rotation.

	tx	ty	angle	scale
mean w/o rot	0.44	0.44	0.00	0.00
max w/o rot	0.81	0.77	0.01	0.00
mean with rot	0.32	0.49	0.00	0.00
max with rot	0.97	1.06	0.02	0.00

5. Conclusion and Future Work

In this paper we presented an improved way of generating OCR ground truth from real world data without manual labeling. Starting from an electronic document, it extracts ground truth, consisting of bounding boxes and the ASCII code of the characters. Then, a synthetic image is generated and aligned to the scanned image of the same document. This alignment is done using RAST, a robust branch-and-bound search algorithm. In a first step a global transformation is computed allowing to align the images within one-pixel accuracy if the scanned image is not distorted by non-similarity transformations. As many scanners add non-similarity transformation distortions to the digitized document, a local alignment method adapts the transformation parameters by locally aligning clusters of nearby connected components. This allows to compute the positions of the ground truth components with high enough accuracy. The superior performance of the method has been shown on UW3 dataset. In future work it would be important to also test the performance of the local adaption technique. Therefore ground truth data needs to be generated. Furthermore, extracting ground truth from PDFs is not trivial and has also some drawbacks, e.g. no logical data is contained. Other electronic document formats may be more suitable for the task of ground truth generation.

An online demo of our approach can be found under *http://demo.iupr.org/ocr-gt-gen/ocr-gt-gen.php*. Note, that this demo currently only uses the global alignment and thus does not adapt to distortions other than those covered by the similarity transform. An example image of the output of the demo can be found in Figure 5.

References

[1] H. S. Baird. State of the art of document image degradation modeling. In *IAPR Workshop on Document Analysis Systems*, Rio de Janeiro, Brazil, Dec. 2000.

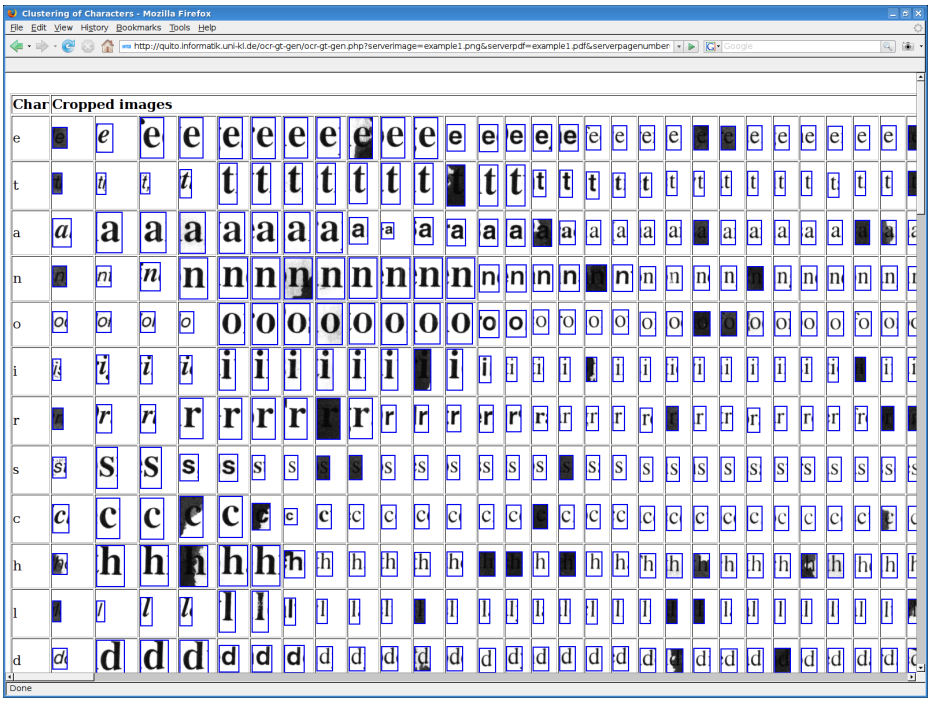

Figure 5. A screenshot of the online demo of the proposed approach.

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

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

[4] J. Canny. A computational approach to edge detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 8(6):679–698, 1986.

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

[6] 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.

[7] 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.

[8] 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.

[9] I. T. Phillips. User's reference manual for the UW english/technical document image database III. Technical report, Seattle University, Washington, 1996.

[10] J. van Beusekom, F. Shafait, and T. M. Breuel. Image-matching for revision detection in printed historical documents. In *DAGM07*, pages 507–516, 2007.

[11] G. Zi and D. Doermann. Document image ground truth generation from electronic text. *17th Int. Conf. on Pattern Recognition*, 02:663–666, 2004.

[12] 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.