Automatic Generation of Character Groundtruth for Scanned Documents: A Closed-Loop Approach*

Tapas Kanungo

Caere Corporation
100 Cooper Court
Los Gatos, CA, 95030, USA
tapas@caere.com

Robert M. Haralick

Department of Electrical Engineering
University of Washington
Seattle, WA 98195, USA
haralick@ee.washington.edu

Abstract

Character groundtruth for scanned document images is crucial for evaluating the performance of OCR systems, training OCR algorithms, and validating document degradation models. Unfortunately, manual collection of accurate groundtruth for characters in a real (scanned) document image is not possible because (i) accuracy in delineating groundtruth character bounding boxes is not high enough, (ii) it is extremely laborious and time consuming and (iii) the manual labor required for this task is prohibitively expensive.

In this paper we present a closed-loop methodology for collecting very accurate (within a pixel error) groundtruth for scanned documents. We first create ideal documents using a typesetting language. Next we create the groundtruth for the ideal document. The ideal document is then printed, photocopied and scanned. A registration algorithm estimates the geometric transformation that registers the ideal document image to the scanned document image. Finally, groundtruth associated with the ideal document image is transformed using the estimated geometric transform to create the groundtruth for the scanned document image.

This methodology is very general and can be used for creating groundtruth for documents typeset in any language, layout, font, and style. The cost of creating groundtruth using our methodology is minimal. We use this methodology to groundtruth 33 English documents consisting of over 62000 symbols. The procedure takes approximately 5 minutes to groundtruth each page on a SUN Sparc 10. Furthermore, we use the method to groundtruth Hindi and FAX documents without any modification to our procedure. Our software will be made available to researchers shortly.

Keywords: *Groundtruth, document analysis, performance evaluation, registration, geometric transformations, image warping, FAX.*

1 Introduction

Character groundtruth for real, scanned document images is crucial for evaluating the performance of OCR systems, training OCR algorithms, and validating document degradation models. Unfortunately, manual collection of accurate groundtruth for characters in a real (scanned) document image is not possible because (i) accuracy in delineating groundtruth character bounding boxes is not high enough, (ii) it is extremely laborious and time consuming and (iii) the manual labor required for this task is prohibitively expensive.

In this paper we give a closed-loop methodology for collecting very accurate (within a pixel error) groundtruth for scanned documents. The groundtruth generated by this method, besides being directly useful for evaluating the performance of OCR systems, is crucial for validating document degradation models [8, 6].

We are unaware of any literature that uses a method similar to ours for automatically collecting groundtruth. However, lot of work on document registration has been reported in the past. Most of this literature pertains to the problem where an ideal form has to be registered to a scanned, hand-filled form. The general idea is to extract the information filled by a human in the various fields of the form. A common method is to extract features from the scanned forms and match them to the features in the ideal form [2, 1]. Unfortunately we cannot use this body of work since there are no universal landmarks that appear in each type of document.

2 Document groundtruth

Groundtruth information is essential for evaluating any document understanding system. By *groundtruth* we mean the correct location, size, font type, and bounding box of the individual symbols on the document image. More global groundtruth associated with a document image could include layout information (such as zone bounding boxes demarcating individual words, paragraphs, article and section titles, addresses, footnotes, table and figure captions, etc.) and style information (such as one column, or two columns; right justified or not; etc). The groundtruth information, of course, needs to be 100 percent accurate, otherwise the systems being evaluated will be penalized incorrectly. Having such groundtruth allows a researcher to study which factors affect the algorithm's

*This work was done when Kanungo was at the University of Washington.

performance and this in turn allows the algorithm developer to think of ways of improving the algorithm.

In our previous work [9], we presented a methodology for generating 100% accurate groundtruth for (i) ideal document images and (ii) synthetically degraded document images. The actual process was as follows. We first created noise-free document images using the LaTeX typesetting language, and extracted the groundtruth information from the DVI files generated by LaTeX. We then synthetically degraded the ideal binary document image using a local document degradation model. The the groundtruth for the synthetically degraded documents is 100% accurate, is easily generated (few seconds on SPARC 5), and does not cost anything once we have the LaTeX files.

However, if the ideal document image is printed and then scanned, the groundtruth information associated with the ideal document image is not usable for the scanned document image since the scanned document is geometrically transformed. That is, the printing and scanning process translates, scales and rotates each character on the document image, besides adding pixel noise. Thus the only alternative is to manually enter the groundtruth for the scanned documents. This task is extremely laborious, time consuming and prohibitively expensive. Furthermore, the person creating the groundtruth should be knowledgeable in the language in which the document is written.

In the following section we outline an automatic, closed-loop approach for generation of groundtruth for real documents. This methodology is very general and is independent of the language in which the document text is written.

3 The groundtruth generation methodology

First, the documents are typeset using LaTeX. Next these documents are converted into binary bitmap images, which are our ideal noise-free documents. When the ideal bitmap is generated from the DVI files, the corresponding groundtruth (location, bounding box, font type and size, and identity of each character) is also generated.

The ideal document image is then printed and scanned. At this point, although the groundtruth for the ideal image is known, the groundtruth for the real scanned image is unknown. However, if the exact transformation that registers the ideal and degraded images were known, we could compute the groundtruth for the real image simply by transforming the bounding box coordinates of the ideal groundtruth by the same transformation.

Thus the groundtruth creation problem now reduces to finding an appropriate transformation that models the geometric distortions the ideal document image undergoes when it is printed and then scanned. Whatever the functional form of the transformation, to estimate the parameters of the transformation we require corresponding feature points from the ideal and real images. Thus, the rough outline of the groundtruth generation method is:

1. Generate ideal document image and the associated character groundtruth.
2. Print the ideal document and scan it back.
3. Find feature points $p_1, \ldots, p_n$ and $q_1, \ldots, q_n$ in the corresponding ideal and real document images.
4. Establish the correspondence between the points p_i and q_i.
5. Estimate the parameters of the 2D transformation T that maps p_i to q_i.
6. Transform the ideal groundtruth information using the estimated transformation T.

The transformation T mentioned in the procedure above is a $2D$ to $2D$ mapping. That is $T : R^2 \rightarrow R^2$. Thus, if $(x, y) = T(u, v)$, where (u, v) is the ideal point and (x, y) is the scanned point, x in general may be a function of both u and v; and same is true regarding y.

Generation of the ideal document image and the corresponding groundtruth is achieved by our synthetic groundtruth generation software DVI2TIFF, which we described in [5] (software is available with the UW English Document Database [3]). Given a transformation T, transforming the groundtruth information is trivial – all that needs to be done is transform the bounding box coordinates of the ideal groundtruth using the transformation T. Thus, there are two main problems: finding corresponding feature points in two document images, and finding the transformation T. These problems are addressed in the following sections.

4 Geometric transformations

Assume that we are given the coordinates of feature points p_i on an ideal document image, and the coordinates of the corresponding feature points q_i on the real document image. (How these feature points are extracted and matched is described in section 6.) The problem is to hypothesize a functional form for the transformation T, that maps the ideal feature points coordinates to the real point coordinates, and a corresponding noise model. To ensure that the transformation T is the same throughout the area of the document image, we choose the points p_i from all over the document image.

The possible candidates for the geometric transformation are similarity, affine, and projective transformations:

Similarity transformation: This transformation is defined by the equation:

$$\begin{pmatrix} x_i \\ y_i \end{pmatrix} = \begin{pmatrix} a & b \\ -b & a \end{pmatrix} \cdot \begin{pmatrix} u_i \\ v_i \end{pmatrix} + \begin{pmatrix} t_x \\ t_y \end{pmatrix} + \begin{pmatrix} \eta_i \\ \psi_i \end{pmatrix} \quad (1)$$

where (u_i, v_i) is the ideal point, (x_i, y_i) is the transformed point, (η_i, ψ_i) is the noise vector, and a, b, t_x, t_y are the four similarity transformation parameters.

In the above parameterization of the similarity transformation, a represents the product of a nonnegative isotropic scale and the cosine of the rotation angle; b represents the product of the nonnegative scale and the sine of the rotation angle; t_x and t_y represent the translation in x and y directions. This parametrization is linear and unconstrained in the parameters, unlike the parametrization in terms of scale, cosine and sine of rotation angle, and translation.

Affine transformation: The affine transformation allows for rotation, translation, anisotropic scale, and shear. The functional form that maps the ideal point (u_i, v_i) onto the real point (x_i, y_i) is

$$\begin{pmatrix} x_i \\ y_i \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \cdot \begin{pmatrix} u_i \\ v_i \end{pmatrix} + \begin{pmatrix} e \\ f \end{pmatrix} + \begin{pmatrix} \eta_i \\ \psi_i \end{pmatrix}, \quad (2)$$

where (η_i, ψ_i) is the error vector and $a, b, \ldots, f$ are the six parameters of the affine transformation.

Projective transformation: Here the assumption is that the real image is a perspective projection of an image on a plane onto another nonparallel plane. The functional form that maps the ideal point (u_i, v_i) onto the real point (x_i, y_i) is given below.

$$\begin{pmatrix} x_i \\ y_i \end{pmatrix} = \frac{1}{w_i} \begin{pmatrix} a_1 u_i + b_1 v_i + c_1 \\ a_2 u_i + b_2 v_i + c_2 \end{pmatrix} + \begin{pmatrix} \eta_i \\ \psi_i \end{pmatrix} \quad (3)$$

where $w_i = a_3 u_i + b_3 v_i + 1$, (η_i, ψ_i) is the error vector, and $a_1, b_1, c_1, a_2, b_2, c_2, a_3, b_3$ are the eight transformation parameters. This parameterization accounts for rotation, translation and the center of perspectivity parameters.

The natural choice for noise is a Gaussian. That is, $(\eta_i, \psi_i)^t \sim N(0, \sigma^2 I)$. Furthermore, σ can be assumed to be known and a function of the image processing algorithm that is used to detect the feature points.

Each of these models can be used to fit the data. The question is which model, if any, models the transformations correctly, and how does one go about proving the hypothesis quantitatively?

In the next section, we show how to estimate the parameters of the three models. In the section that follows we show how to statistically validate/invalidate the models.

5 Estimation of geometric transformation parameters

Each corresponding point-pair provides two linear constraints on the parameters. Thus we need at least two corresponding points for estimating the similarity parameters, three corresponding points for affine, and four for projective. If we have more corresponding points than the minimum required, we can solve for the parameters of the transformation in a least squares sense, which also happens to be the maximum likelihood estimate of the parameters under the Gaussian noise model.

5.1 Similarity transformation

If there are n corresponding points, the similarity relation given in equation (1) can be written as $\mathbf{y} = \mathbf{Ap} + \mathbf{n}$ where: $\mathbf{y} = (x_1, \ldots, x_n, y_1, \ldots, y_n)^t$, the parameter vector $\mathbf{p} = (t_x, t_y, a, b)^t$, $\mathbf{n} = (\eta_1, \ldots, \eta_n, \psi_1, \ldots, \psi_n)^t$,

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & u_1 & v_1 \\ 1 & 0 & u_2 & v_2 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & u_n & v_n \\ 0 & 1 & v_1 & -u_1 \\ 0 & 1 & v_2 & -u_2 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & v_n & -u_n \end{bmatrix},$$

where (u_i, v_i) are the ideal points and (x_i, y_i) are the transformed points on the scanned image. The least squares estimate of the parameter vector is given by $\hat{\mathbf{p}} = (\mathbf{A}^t\mathbf{A})^{-1}\mathbf{y}$.

5.2 Affine transformation

If there are n corresponding points, the affine equations given in equation (2) can be rearranged $\mathbf{y} = \mathbf{Ap} + \mathbf{n}$ where: $\mathbf{y} = (x_1, \ldots, x_n, y_1, \ldots, y_n)^t$, the parameter vector $\mathbf{p} = (a, b, c, d, e, f)^t$, $\mathbf{n} = (\eta_1, \ldots, \eta_n, \psi_1, \ldots, \psi_n)^t$,

$$\mathbf{A} = \begin{bmatrix} u_1 & v_1 & 0 & 0 & 1 & 0 \\ u_2 & v_2 & 0 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ u_n & v_n & 0 & 0 & 1 & 0 \\ 0 & 0 & u_1 & v_1 & 0 & 1 \\ 0 & 0 & u_2 & v_2 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & u_n & v_n & 0 & 1 \end{bmatrix},$$

where (u_i, v_i) are the ideal points, and (x_i, y_i) are the transformed points on the scanned image. The least squares estimate of the parameter vector is given by $\hat{\mathbf{p}} = (\mathbf{A}^t\mathbf{A})^{-1}\mathbf{y}$.

5.3 Projective transformation

If there are n corresponding points, the projective transformation equations given in equation (3) can be rearranged as $\mathbf{y} = \mathbf{Ap} + \mathbf{Wn}$ where: $\mathbf{y} = (x_1, \ldots, x_n, y_1, \ldots, y_n)^t$, the parameter vector $\mathbf{p} = (a_1, b_1, c_1, a_2, b_2, c_2, a_3, b_3)^t$, $\mathbf{n} = (\eta_1, \ldots, \eta_n, \psi_1, \ldots, \psi_n)^t$,

$$\mathbf{A} = \begin{bmatrix} u_1 & v_1 & 1 & 0 & 0 & 0 & -u_1x_1 & -v_1x_1 \\ u_2 & v_2 & 1 & 0 & 0 & 0 & -u_2x_2 & -v_2x_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ u_n & v_n & 1 & 0 & 0 & 0 & -u_nx_n & -v_nx_n \\ 0 & 0 & 0 & u_1 & v_1 & 1 & -u_1y_1 & -v_1y_1 \\ 0 & 0 & 0 & u_2 & v_2 & 1 & -u_2y_2 & -v_2y_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & u_n & v_n & 1 & -u_ny_n & -v_ny_n \end{bmatrix},$$

the diagonal weight matrix $\mathbf{W} = \mathrm{diag}(w_1, \ldots, w_n, w_1, \ldots, w_n)$, (u_i, v_i) are the ideal points, (x_i, y_i) are the transformed points on the scanned image, and $w_i = a_3 u_i + b_3 v_i + 1$.. The weighted least squares estimate of the parameter vector is given by $\hat{\mathbf{p}} = (\mathbf{A}^t\mathbf{W}^{-2}\mathbf{A})^{-1}\mathbf{W}^{-1}\mathbf{y}$. Since the w_i's are initially unknown, we can solve estimate $\mathbf{p}$ iteratively: initialize $\hat{\mathbf{p}} = \mathbf{0}$ and then in each iteration compute $\mathbf{W}$ using the estimate of $\mathbf{p}$ from the previous iteration.

6 Finding corresponding feature points

In a document image with text, figures and mathematics, there are no universal feature points in the interior of the document that are guaranteed to appear in each type of document. However, most documents have a rectangular text layout, whether they are in one-column format or in two-column format.

We use the upper-left (UL), upper-right (UR), lower-right (LR), and lower-left (LL), corners of the text area as feature points.

The four feature points, $p_1, \ldots, p_4$, are detected on the ideal image as follows.

1. Compute the connected components in the image.
2. Compute the upper-left (a_i), upper-right (b_i), lower-right (c_i), and lower-left (d_i) corners of the bounding box of each connected component.
3. Find the four feature points using the following equations:

$$
\begin{aligned}
p_1 &= \arg\min_{a_i}(x(a_i) + y(a_i)), \\
p_2 &= \arg\max_{b_i}(x(b_i) - y(b_i)), \\
p_3 &= \arg\max_{c_i}(x(c_i) + y(c_i)), \\
p_4 &= \arg\min_{d_i}(x(d_i) - y(d_i)).
\end{aligned}
$$

   The above equations compute the upper-left (p_1), upper-right (p_2), lower-right (p_3), and lower-left (p_4).

The above algorithm is also used to compute the corresponding four feature points $q_1, \ldots, q_4$ on the real image. Since noise blobs can appear in a real image, we check that the bounding box sizes of the components are within a specified tolerance before establishing correspondence between p_i and q_i. A transformation T can be estimated using the corresponding points $p_1, \ldots, p_4$ and $q_1, \ldots, q_4$ by the methods described in section 5.

7 Registration results on scanned images

Unfortunately, none of the three geometric transformations described in sections 4 and 5 model the transformation very accurately. That is, the real points are displaced from ideal transformed points in some nonlinear fashion. The mismatch must arise from nonlinearities in the optical and mechanical systems. Note that these nonlinearities could be either in the printer or the photocopier or the scanner or in any combination of the three units.

In figure 1 we show a subimage of a scanned image with the groundtruth (character bounding boxes) overlaid. We see that there is a lot of error. This error is not systematic over the entire page.

To confirm the fact that there are nonlinearities in the printing-photocopying-scanning processes, we set up a calibration experiment and performed a statistical test to prove the point that the similarity, affine and projective transforms alone do not model the transformation correctly.

7.1 The calibration experiment

We now describe a controlled experiment that was conducted to confirm the fact that the geometric transformation that occurs while printing and scanning documents is not similarity or affine or perspective. First we create an ideal calibration image consisting of only '+' symbols arranged in a grid. We print

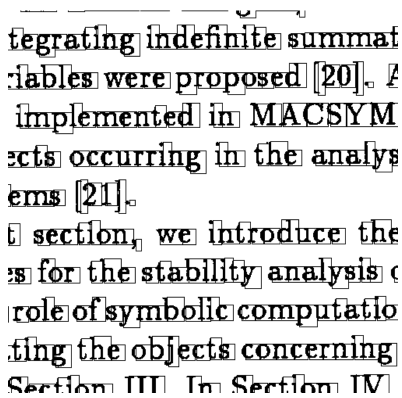


Figure 1: A scanned image with groundtruth overlaid. In this case perspective transform was used to register the ideal document image to the scanned image. It can be seen that there is large error in groundtruth.

this document and then scan it back. The crosses in the ideal image are then matched to the crosses in the scanned image. This set of corresponding points are then used to estimate the geometric transform parameters. The sample mean and sample covariance matrix of the registration error vectors are then computed. Since the population mean and population covariance matrix of the error vectors can be theoretically derived, we can test whether the theoretically derived distribution parameters are close to the experimentally gathered sample parameters.

In the next subsection we provide the details of the calibration data gathering process. In the subsequent subsection we give the details of the statistical hypothesis testing procedure.

7.2 Protocol for calibration experiment

The ideal image for calibrating the printer-photocopier-scanner process is created as follows. First a grid of equally spaced "+" symbols is arranged on a 3300×2500 binary image. The vertical and horizontal bars of the "+" symbol are 25 pixels long and 3 pixels thick. The number of symbols on each row and column of the grid are 23 and 30, respectively.

The ideal image is then printed and scanned. The intersection points of the two bars of the "+" symbols are used as the calibration points. The calibration points are detected by a morphological algorithm: first the image is closed with a 3×3 square structuring element. Next, two images are created by opening the closed image with a vertical and horizontal structuring elements, respectively. Calibration points on the scanned image are detected by binary-anding these two images. A connected component algorithm is then run on the image with the detected calibration points. The centroids of the connected components are used as the coordinates of the calibration points. The calibration points in the ideal image are known since the ideal calibration image is created under experimenter's control.

To estimate the projective transform, four feature points are first detected using the algorithm described in section 6. Next, we estimate the projective transform parameters from the ideal and real points (correspondences are known since we order the four points in a counter clockwise order, starting with the upper left feature point, and assume that the orientation of the page is unchanged). The estimated transform parameters are then used to project all the ideal points. An exhaustive search is conducted to establish correspondences between the projected ideal calibration points and real calibration points. That is, for each projected ideal point, we find the closest real point, and assume the two points match. This is done by a brute-force $O(n^2)$ algorithm. Since n is of the order of 1000, the computation required is of the order 10^6, which takes approximately three seconds on a Sparc 2. Next, for each calibration point we compute the registration error vector, which is the displacement vector between the real calibration point and the projected calibration point. The maximum error we attain is with ± 4 pixels in each coordinate.

In Figure 2(a) we show a subimage of the scanned calibration document. The detected calibration points are shown in Figure 2(b). In Figure 2(c) the ideal calibration points are transformed using the estimated projective transformation and overlaid on the real calibration points. A scatter plot of the error vectors is shown in Figure 3.

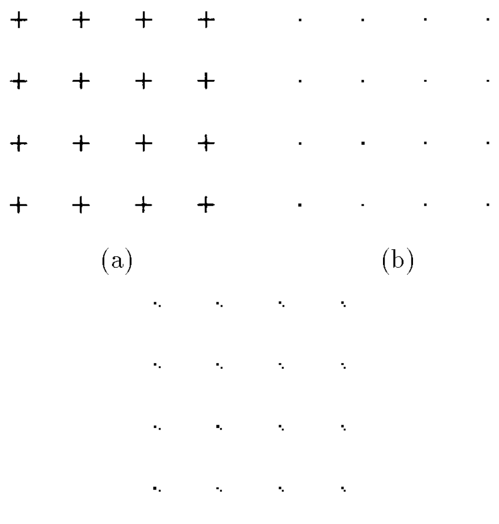


Figure 2: (a) A subimage of the scanned calibration document. The detected calibration points are shown in (b). (c) The ideal calibration points are transformed using the estimated projective transformation and overlaid on the real calibration points.

7.3 Statistical tests

Since the estimated parameters of the models are functions of real point coordinates, which are random variables, the estimated parameters are random variables. The distribution of estimated parameters can be derived in terms of the assumed distribution of the noise in the real point coordinates. To confirm that the geometric transformation model and the noise model are valid, we test whether or not the theoretically derived distribution of the estimated parameter vector is

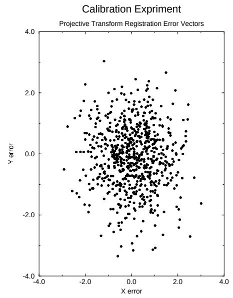


Figure 3: A scatter plot of the error vectors computed between real calibration points and projected ideal calibration points.

the same as that computed empirically. If either the geometric transformation model or the noise model is incorrect, the test for equality of the empirically computed distribution and the theoretically derived distribution will not pass. Furthermore, instead of testing the distribution of the estimated parameters, we can test the distribution of the residual error, which in turn has a known distribution. For details please see [5, 7, 4].

8 Dealing with nonlinearities

As we saw, because of the nonlinearities, the groundtruth bounding boxes for the characters in a scanned image are not correct. Our solution to this problem is very simple. We first transform the ideal document image using the perspective transformation. The groundtruth associated with the ideal image is also transformed using the estimated perspective transform parameters. Next, each character in the perspective transformed image is locally translated and matched (using Hamming distance) with the same size subimage in the scanned document image. Thus, if the the nonlinearity gave rise to a $(2, 3)$ translation error in pixels, our template matching process would give the best match (minimum Hamming distance) when the translation is $(2, 3)$. The size of the search window is decided by the calibration experiment. If the error vectors are large, the search window has to be made large. This local search process gives us a highly accurate groundtruth, and the errors are within a pixel.

9 Dealing with outliers

At times, when two very similar characters (for example two 'i's, or one 'i' and one 'l') are close to each other, the template matching process can match the perspective transformed character to the wrong scanned character. This typically happens if we use a large search window size. This means that the error translation vector associated the wrongly matched character will be off. We use robust regression for detecting and correcting such outliers. Briefly the procedure is as follows. Once the error vectors are computed, we fit a piece-wise bilinear function to the x and y translation errors associated with characters in a small area of the image. We assume that within this small area the error vectors do not vary much. Thus the robust regression detects the outliers error vector and estimated function is used to estimate the correct error vector. Details can be found in [5, 10].

10 Experimental protocol

10.1 Data collection

The ideal document is a LaTeX formatted document. The IEEE Transaction style is used for typesetting the document. The corresponding ideal binary image and character ground truth is created using the DVI2TIFF software. The ideal document is created at 300×300 dots/inch resolution and the size of the binary document in pixels is 3300×2550. This document is printed using a SparcPrinter II. Next, the original printed document is photocopied five times using a Xerox photocopier - once at the normal setting, twice with darker settings, and twice with lighter settings. Finally the five photocopied documents are scanned using a Ricoh scanner. The scanner is set at 300×300 dots/inch resolution. The rest of the scanner parameters are set at normal settings. The scanned binary image is of size 3307×2544.

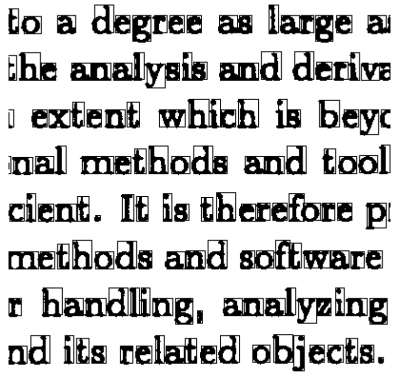
10.2 Protocol for generating real groundtruth

Once the real scanned documents have been gathered as described in the previous section, we use the registration algorithm, described in section 1 to i) transform the ideal binary documents so that it registers to the scanned document and ii) to create the groundtruth corresponding to the scanned document. The transformed groundtruth also forms the groundtruth for the transformed ideal document. The local nonlinearities of the transformation are accounted for by searching in a local neighborhood for a good match between the ideal character symbol and the real character symbol. The local template match window size is determined by the calibration experiment we performed earlier. Since the maximum error in the registration is ± 4 pixels, we used a window with $-7 \leq \Delta x, \Delta y \leq 7$. The groundtruth generated by our algorithm is highly accurate.

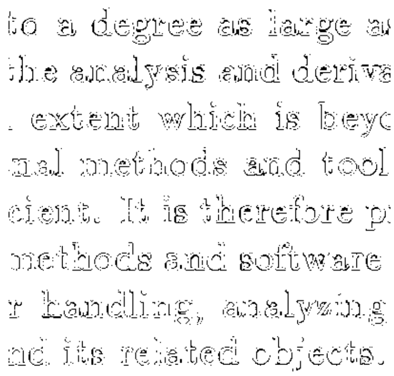
11 Results and discussion

In this section we show few sample output of our automatic groundtruth generation algorithm. A subimage of the scanned image with the overlaid bounding box is shown in Figure 4. An exclusive or-ed image of the real scanned document and the registered ideal document is shown in Figure 4. From the exclusive or-ed image we can see that the registration of the ideal image to the scanned image is extremely accurate. The time taken for this procedure on a SUN Sparc 10, is 5 minutes.

In figure 5 we show automatically generated groundtruth for a FAXed document. In this case the ideal bitmap was generated on the computer and then printed. The printed document was then FAXed and the FAX output was scanned using a Ricoh scanner. It is interesting to note that in many cases even though the scanned documents are highly degraded, our algorithm produces the correct groundtruth.



(a)



(b)

Figure 4: Groundtruth for real documents. (a) shows a subimage of a document with the estimated bounding boxes of each character. (b) shows the result of exclusive-OR between the real document and the registered ideal document.

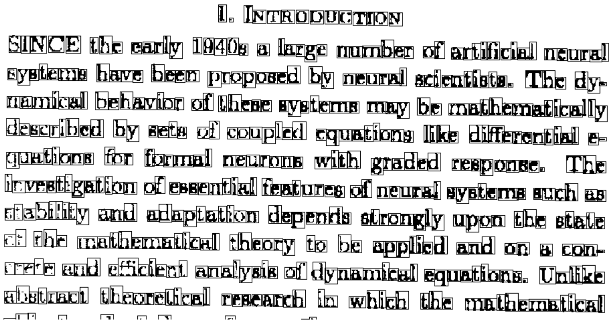
Finally in figure 6 we show a Hindi document written in Devanagari script. The document was typeset in LaTeX using macros made available by Frans Velthuis (velthuis@rc.rug.nl). We can see that our methodology is general enough to handle documents in any language. We have also used this methodology to groundtruth Arabic and Music documents [5].

In addition, we used the groundtruth generation software to groundtruth 33 English document pages consisting of over 62000 symbols. The algorithm takes about five minutes to groundtruth each page on a SUN Sparc 10. Some of these documents had numerous mathematical equations.

Few of the limitations of our algorithm are: (i) it is sensitive to the feature points that are used for registration, (ii) if the scanned image is from a bound book, our procedure will not perform well, (iii) the population of documents one can generate by printing and scanning ideal documents is a subset of the population of document images in the real world.

12 Summary

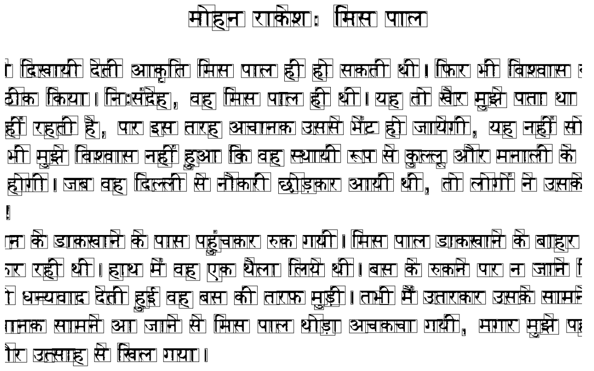
In this paper we presented a closed-loop method for producing character groundtruth for real document



1. INTRODUCTION

SINCE the early 1940s a large number of artificial neural systems have been proposed by neural scientists. The dynamical behavior of these systems may be mathematically described by sets of coupled equations like differential equations for formal neurons with graded response. The investigation of essential features of neural systems such as stability and adaptation depends strongly upon the state of the mathematical theory to be applied and on a concrete and efficient analysis of dynamical equations. Unlike abstract theoretical research in which the mathematical

Figure 5: A subimage of a FAXed document with the groundtruth overlaid. Notice that the characters in the bottom left of the image are hardly legible. Manual groundtruth for these type of documents would be prone to errors. In contrast, our software has produced correct groundtruth without any problem.



मोहन राकेश: मिस पाल

दिखायी देती आकृति मिस पाल ही हो सकती थी। फिर भी विश्वास
ठीक किया। निःसंदेह, वह मिस पाल ही थी। यह तो खैर मुझे पता था
हीं रहती है, पर इस तरह अचानक उससे भेंट हो जायेगी, यह नहीं सो
भी मुझे विश्वास नहीं हुआ कि वह स्थायी रूप से कुल्लू और मनाली के
होगी। जब वह दिल्ली से नौकरी छोड़कर आयी थी, तो लोगों ने उसके
!
न के डाकखाने के पास पहुंचकर रुक गयी। मिस पाल डाकखाने के बाहर
र रही थी। हाथ में वह एक थैला लिये थी। बस के रुकने पर न जाने
धन्यवाद देती हुई वह बस की तरफ मुड़ी। तभी मैं उतरकर उसके साम
ानक सामने आ जाने से मिस पाल थोड़ा अचकचा गयी, मगर मुझे प
र उत्साह से खिल गया।

Figure 6: A subimage of a Hindi document in Devanagri script with the groundtruth overlaid.

images. The method starts by generating ideal noise-free document images using a document typesetting software like LaTeX. These binary document images are printed, photocopied, and then scanned. Feature points are extracted from the ideal and the scanned document images, and their correspondences established. We showed that the similarity, affine and projective transformations alone cannot be used to represent the transformation between the ideal and the scanned documents. This fact was confirmed by using test images specially designed for calibration, and verifying that the statistical distribution of the registration error is not what the theory predicts. The local nonlinearities that exist can be accounted for by performing a local template match using the ideal character as the template, and searching a small neighborhood in the real image for the best match. The size of the local search neighborhood is decided by the calibration experiment. The calibration experiment gives us the maximum deviations that can occur between the ideal feature points after they have been transformed using the estimated transformation and the feature points on the scanned image. We used this methodology to groundtruth 33 documents consisting of over 62000 symbols. The procedure took approximately 5 minutes to groundtruth each page on a SUN Sparc 10. Furthermore, we used the method to groundtruth Hindi documents without any modification to our procedure.

Acknowledgements

Kanungo would like to thank Bhabuthosh Chanda, Ken Thornton, Harpreet Sawhney, B. K. P. Horn, Sinan Karasu and Mindy Bokser for comments and discussion.

References

[1] R. G. Casey and D. R. Ferguson. Intelligent forms processing. *IBM Systems Journal*, 29(3):435–50, 1990.

[2] D. S. Doermann and A. Rosenfeld. The processing of form documents. In *Proc. of Int. Conf. on Document Analysis and Recognition*, pages 497–501, Tsukuba, Japan, October 1993.

[3] R. M. Haralick, I. Phillips, et al. U.W. English Database I, 1994.

[4] T. Kanungo. MVNTEST: software for multivariate hypothesis testing, 1995. STATLIB anonymous ftp: lib.stat.cmu.edu:/general/mvntest.gz.

[5] T. Kanungo. *Document Degradations Models and a Methodology for Degradation Model Validation.* PhD thesis, University of Washington, Seattle, WA., 1996.

[6] T. Kanungo, H. S. Baird, and R. M. Haralick. Validation and estimation of document degradation models. In *Proc. of Fourth Annual Symposium on Document Analysis and Information Retrieval*, Las Vegas, Nevada, April 1995.

[7] T. Kanungo and R. M. Haralick. Multivariate hypothesis testing for gaussian data: Theory and software. Technical Report ISL Tech. Report: ISL-TR-95-05, University of Washington, Seattle, WA, Oct 1995.

[8] T. Kanungo, R. M. Haralick, H. S. Baird, W. Stuetzle, and D. Madigan. Document degradation models: Parameter estimation and model validation. In *Proc. of Int. Workshop on Machine Vision Applications*, Kawasaki, Japan, December 1994.

[9] T. Kanungo, R. M. Haralick, and I. Phillips. Non-linear local and global document degradation models. *Int. Journal of Imaging Systems and Technology*, 5(4), 1994.

[10] G. Wolberg. *Digital Image Warping.* IEEE Computer Society Press, Los Alamitos, CA, 1990.