A Point Matching Algorithm for Automatic Generation of Groundtruth for Document Images

Doe-Wan Kim and Tapas Kanungo
Language and Media Processing Laboratory
Center for Automation Research
University of Maryland
College Park, MD 20742
Email:{dwkim,kanungo}@cfar.umd.edu

Abstract

Geometric groundtruth at character, word, and line level is crucial for developing and evaluating optical character recognition (OCR) algorithms. Kanungo and Haralick [ICPR '96] proposed a closed loop methodology for generating character level groundtruth for rescanned image. In this article we present a robust version of their methodology. We grouped the feature points and used branch and bound algorithm on the grouped feature point set to estimate the transformation. Euclidean distance between character centroids was used as the error metric. We performed experiments on a randomly selected subset of the University of Washington dataset.

1 Introduction

Character, word, and line level geometric groundtruth is crucial for optical character recognition (OCR) algorithm development and evaluation. Such groundtruth is typically created manually and therefore is time consuming, expensive and prone to human errors.

We consider the case in which researchers already have the geometric groundtruth for a few document images but would like to use these document-groundtruth pairs to bootstrap the construction of a larger (more varied) dataset. We consider two scenarios. In the first scenario, the groundtruth for the set of original real document images is created manually, and in the second scenario, the groundtruth for the set of original synthetic document images is generated automatically. In both cases the algorithm developer would like to print, photocopy, fax and rescan the original document images and then automatically generate the geometric groundtruth for the rescanned documents.

Kanungo and Haralick [10] proposed a methodology to automatically generate groundtruth of rescanned image by estimating the transformation between two images and then transforming the groundtruth using the estimated transformation. They estimated the transformation from corresponding pairs of feature points. Figure 1 illustrates the methodology that they used for generating the groundtruth information for real images. Four

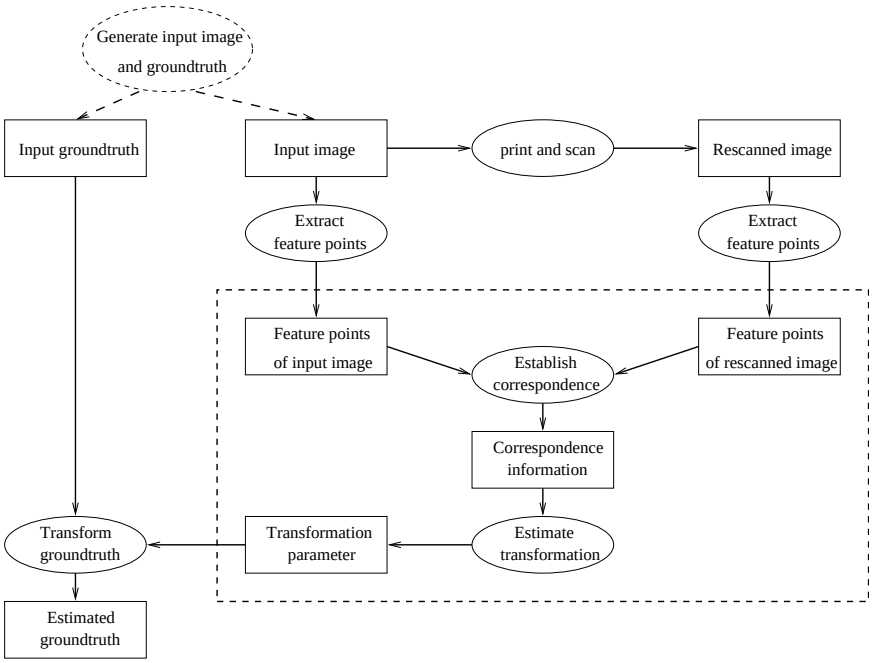


Figure 1: Automatic closed-loop methodology by Kanungo and Haralick.

corner points of the images were used as feature points to estimate the transformation. The four feature points, p_1, p_2, p_3 and p_4 were determined by the following equations.

$$p_1 = \arg\min_{a_i}(x(a_i) + y(a_i)),\ p_2 = \arg\max_{b_i}(x(b_i) - y(b_i)),$$

$$p_3 = \arg\min_{c_i}(x(c_i) + y(c_i)),\ p_4 = \arg\max_{d_i}(x(d_i) - y(d_i)),$$

where a_i, b_i, c_i and d_i are respectively the upper-left, upper-right, lower-right, and lower-left corners of the bounding boxes of each connected component in the image. Hobby [7] improved the registration by using optimization method to minimize the mismatch in the estimated transformation. He used a direct search method to find the transformation. More recently, Viard-Gaudin et al. [19] proposed a methodology to create groundtruth for handwritten document. They designed a database of online and offline handwritten data.

In this paper, we present a robust method for matching point sets by using all the feature points available. The dashed rectangle in Figure 1 is the module that is being replaced by the algorithm described in this paper.

2 Problem definition

Given an image and its groundtruth information, we wish to generate groundtruth for the image which is a transformed (scanned, photocopied, fax-ed, etc.) version of the

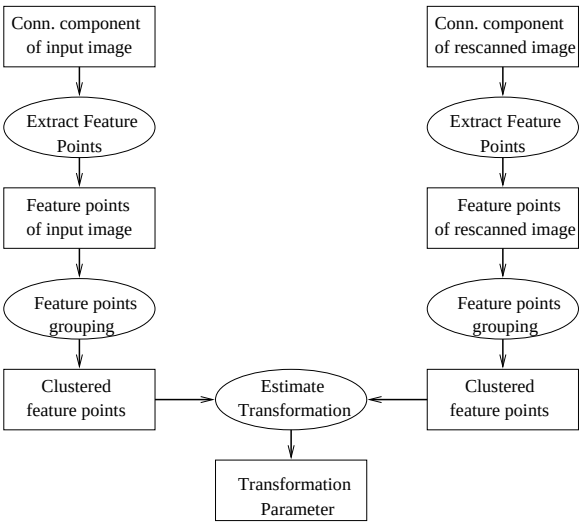


Figure 2: The automatic registration methodology.

original image. The basic idea is to estimate the transformation between two images and transform the groundtruth information using the estimated transformation. Figure 2 shows the illustration of this procedure.

3 The automatic groundtruthing methodology

First we extract the connected components of the original and transformed images. The number of connected components in a typical document image is 1000∼5000, which makes the running time of estimation too large. To reduce the complexity of problem, we group connected components instead. These groups are approximately at the word level. As a result of grouping, the number of feature points to be considered is reduced to about $20 \sim 25\%$ of its original size. We explain the procedure of this feature point grouping in Section 4.1.

Using the two feature point sets, one from the original image and the other from the transformed image, we estimate the transformation by using branch and bound algorithm. We consider the partial Hausdorff distance [8] as the similarity measure. Given point sets A and B, and parameter k, the partial Hausdorff distance is defined as:

$$H_k(A, B) = k^{th}_{a \in A} \min_{b \in B} dist(a, b).$$

4 The matching algorithm

We need to find the correspondence and the transformation between two point sets. There are two major steps in the matching procedure: (i) feature point grouping and (ii) branch and bound algorithm.

```
begin
    for all b ∈ B
        root(b) ← b
        for all b' ∈ NN^k(b)
            put (b, b') into PQ
    pair (b, b') ← pair with smallest distance of PQ
    while distance of (b, b') < τ
    do
        if root(b) ≠ root(b')
        then for all b'' with root(b')
                    root(b'') = root(b)
    end
end
```

Figure 3: The feature point grouping algorithm.

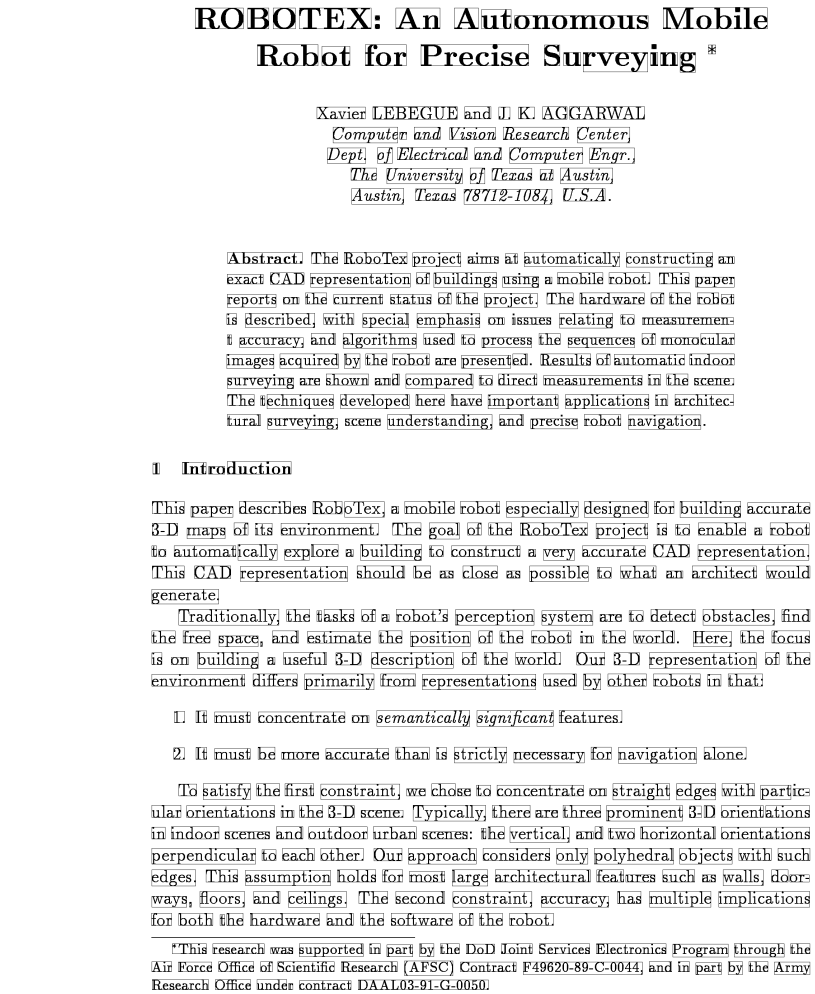
4.1 Feature point grouping

To reduce the size of the problem, we group connected components to word token level. Let B be the set of bounding boxes, $NN^k(b)$ be the k nearest neighbors of bounding box b, PQ be a priority queue, and τ be a threshold, and $root(b)$ be the root of b, which is initialized to be b. The key of the priority queue is the distance between the bounding boxes in the pair. The pair with the smallest distance appears on the top of the queue. In selecting the threshold, we used the threshold selection method by Kittler and Illingworth [11]. Figure 3 illustrates this grouping algorithm.

Figure 4 is an example of image overlaid with bounding boxes of the grouped connected components. This sample image contains 2127 characters, and 442 groups. We can see that these groups are approximately at the word level. Grouping takes less than 10 seconds per image when run on Sun Ultra-Sparc 5 with clock speed 361.2 MHz.

4.2 Feature matching of document image using branch and bound algorithm

We now outline the branch and bound algorithm for feature point matching. Let T be the range of affine transformation, and ϵ be the error bound. The basic approach of branch and bound algorithm is as follows: for a given T, we first compute the upper and lower bound of similarity. Next, a priority queue is constructed such that the element that has the largest size comes on top of the queue. In the iteration, we pick up the largest element out of the priority queue, and see if its lower bound of similarity is better than the current best similarity. If not, we simply kill that element and proceed to the next largest element. Otherwise, we compute the upper bound and check if it is better than the current best similarity. If it is, (i) we update the best similarity to be the upper bound of current element, (ii) update the best transformation, (iii) split the element into two parts along the longest side,and (iv) insert both new elements into the priority queue. This process is iterated until we achieve the target similarity or there is no more element to



ROBOTEX: An Autonomous Mobile Robot for Precise Surveying *

Xavier LEBEGUE and J. K. AGGARWAL
Computer and Vision Research Center,
Dept. of Electrical and Computer Engr.,
The University of Texas at Austin,
Austin, Texas 78712-1084, U.S.A.

Abstract. The RoboTex project aims at automatically constructing an exact CAD representation of buildings using a mobile robot. This paper reports on the current status of the project. The hardware of the robot is described, with special emphasis on issues relating to measurement accuracy, and algorithms used to process the sequences of monocular images acquired by the robot are presented. Results of automatic indoor surveying are shown and compared to direct measurements in the scene. The techniques developed here have important applications in architectural surveying, scene understanding, and precise robot navigation.

1 Introduction

This paper describes RoboTex, a mobile robot especially designed for building accurate 3-D maps of its environment. The goal of the RoboTex project is to enable a robot to automatically explore a building to construct a very accurate CAD representation. This CAD representation should be as close as possible to what an architect would generate.

Traditionally, the tasks of a robot's perception system are to detect obstacles, find the free space, and estimate the position of the robot in the world. Here, the focus is on building a useful 3-D description of the world. Our 3-D representation of the environment differs primarily from representations used by other robots in that:

1. It must concentrate on *semantically significant* features.

2. It must be more accurate than is strictly necessary for navigation alone.

To satisfy the first constraint, we chose to concentrate on straight edges with particular orientations in the 3-D scene. Typically, there are three prominent 3-D orientations in indoor scenes and outdoor urban scenes: the vertical, and two horizontal orientations perpendicular to each other. Our approach considers only polyhedral objects with such edges. This assumption holds for most large architectural features such as walls, doorways, floors, and ceilings. The second constraint, accuracy, has multiple implications for both the hardware and the software of the robot.

*This research was supported in part by the DoD Joint Services Electronics Program through the Air Force Office of Scientific Research (AFSC) Contract F49620-89-C-0044, and in part by the Army Research Office under contract DAAL03-91-G-0050.

Figure 4: Sample document image overlaid with the bounding boxes of the grouped connected components.

```
begin
    construct and initialize PQ with given T
    while PQ size ≠ 0 and best_similarity > ε
    do
        T ← next element in PQ
        compute lower bound of similarity for T
        if lower bound of T > best_similarity - ε
        then kill this cell and proceed to the next one
        compute upper bound of similarity for T
        if upper bound of T < best_similarity
        then update best_similarity and transformation
        split T into T1 and T2
        insert T1 and T2 into PQ
    end
end
```

Figure 5: Branch and bound algorithm for feature point matching.

be processed in the queue. In computing the upper and lower bound of given range of transformation, we use the kd-tree based nearest neighbor searching algorithm by [1, 5]. The algorithm is illustrated in figure 5. More detail of branch and bound algorithm can be found in [8, 14].

5 Error metric

For the analysis of experimental result, we need to define an error criterion. Let G be the set of groundtruth elements $g_i, i = 1, \cdots, N$,, where N is the number of characters in image. Typically, g_i is a tuple: $g_i = (x_i, y_i, w_i, h_i, f_i) \in R \times R \times R^+ \times R^+ \times \mathcal{F}$, where, x_i, y_i are the x and y coordinates of the upper-left corner of character-level bounding box, w_i, h_i are the width and height of that bounding box, and f_i is the font information. Let θ and $\hat{\theta}$ denote the true and estimated transformations respectively. We can get the groundtruth for rescanned image by transforming G using the estimated transformation. Then we can define G^θ and $G^{\hat{\theta}}$ to be the set of transformed groundtruth elements as follows:

$$G^\theta = T^\theta(G) \text{ with elements } g_i^\theta = (x_i^\theta, y_i^\theta, w_i^\theta, h_i^\theta, f_i^\theta)$$
$$G^{\hat{\theta}} = T^{\hat{\theta}}(G) \text{ with elements } g_i^{\hat{\theta}} = (x_i^{\hat{\theta}}, y_i^{\hat{\theta}}, w_i^{\hat{\theta}}, h_i^{\hat{\theta}}, f_i^{\hat{\theta}}).$$

We can compute g_i^θ and $g_i^{\hat{\theta}}$ as follows:

$$(x_i^\theta, y_i^\theta)^t = T^\theta(x_i, y_i)^t, (x_i^{\hat{\theta}}, y^{\hat{\theta}})^t = T^{\hat{\theta}}(x_i, y_i)^t.$$

To define w_i^θ, h_i^θ and $w_i^{\hat{\theta}}, h_i^{\hat{\theta}}$, let u_i, v_i be the x and y coordinates of lower-right corner of bounding box.

$$u_i = x_i + w_i,\ v_i = y_i + h_i$$
$$(u_i^\theta, v_i^\theta)^t = T^\theta(u_i, v_i)^t, (u_i^{\hat{\theta}}, v^{\hat{\theta}})^t = T^{\hat{\theta}}(u_i, v_i)^t$$
$$w_i^\theta = u_i^\theta - x_i^\theta, h_i^\theta = v_i^\theta - y_i^\theta$$

$$w_i^{\hat{\theta}} = u_i^{\hat{\theta}} - x_i^{\hat{\theta}},\ h_i^{\hat{\theta}} = v_i^{\hat{\theta}} - y_i^{\hat{\theta}}.$$

Also, we assume that $f_i^{\theta} = f_i^{\hat{\theta}} = f_i$. The Euclidean distance between centroid of corresponding bounding boxes $delta_i$ is defined as:

$$\delta_i = \|Centroid(g_i^{\theta}), Centroid(g_i^{\hat{\theta}})\|.$$

Then, the mean and maximum error measures for an image can be defined as follows.

$$\rho_{mean}(G^{\theta}, G^{\hat{\theta}}) = \frac{1}{N}\sum_{i=1}^{n}\delta_i$$
$$\rho_{max}(G^{\theta}, G^{\hat{\theta}}) = \max_i\{\delta_1, \cdots \delta_N\}$$

6 Experimental methodology and protocol

Experiment is performed on the University of Washington data set [16]. This dataset contains journal images with corresponding character level geometric groundtruth. We performed the experiment on 450 images. These images were generated by transforming 10 images randomly selected from UW dataset by 45 different transformations. The rotation angle R was set at zero and the scale S and translation X_t, Y_t parameters were selected from the following sets. We are currently conducting the experiments where the angle is a variable.

$$S = \{65\%, 80\%, 100\%, 120\%, 135\%\},$$
$$X_t = \{-50, 0, 50\},\ Y_t = \{-100, 0, 100\}.$$

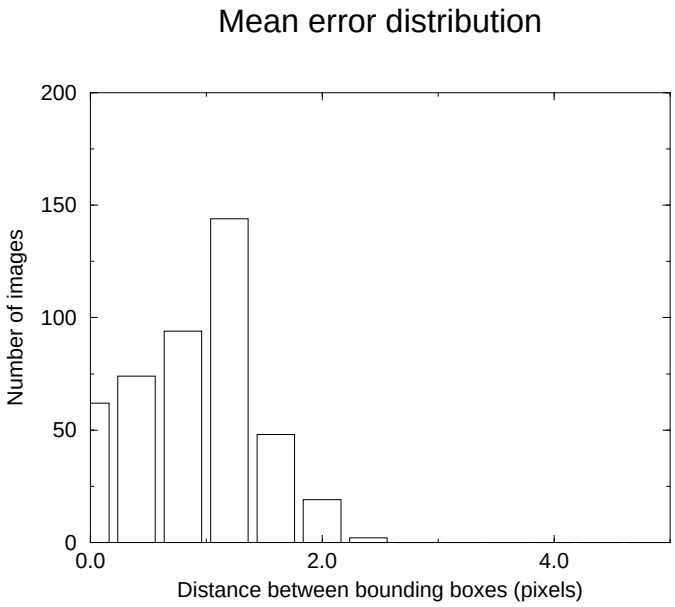
7 Results and discussions

To analyze the results, we generate the histogram of estimation errors. As discussed in Section 5, we calculate $\rho_{mean}(G^{\theta}, G^{\hat{\theta}})$ and $\rho_{max}(G^{\theta}, G^{\hat{\theta}})$ for each image pair. Let O be the set of images, T be the set of transformations, Δ be the width of the range, I be set of transformed images, and $\mathcal{G}$ be the set of groundtruth elements G_i. The histograms of mean and maximum error, $H_{mean}(k; O, T, \Delta)$ and $H_{max}(k; O, T, \Delta)$, are defined as follows.

$$H_{mean}(k; O, T, \Delta) = \|\{i \in I \mid \frac{(k-1)\Delta}{2} < \rho_{mean}(G_i^{\theta}, G_i^{\hat{\theta}}) \leq \frac{(k+1)\Delta}{2}\}\|$$

$$H_{max}(k; O, T, \Delta) = \|\{i \in I \mid \frac{(k-1)\Delta}{2} < \rho_{max}(G_i^{\theta}, G_i^{\hat{\theta}}) \leq \frac{(k+1)\Delta}{2}\}\|$$

We have 450 transformed images, for which groundtruth is estimated. The histograms of error distribution of this image set are shown in Figure 6. We set Δ to be 0.4 pixel. From this result, we can see that the estimated groundtruth is close to the true groundtruth with less than 3 pixels of mean error and 5 pixels of maximum error. The mean of mean error is 1.09 pixels, and mean of maximum error is 2.16 pixels. The estimation takes $10 \sim 15$ minutes per image when run on Sun Ultra-Sparc 5 with clock speed 361.2 MHz. We faxed and rescanned an image, and run our feature matching algorithm to produce the groundtruth for this image. Figure 7 shows the faxed image overlaid with the estimated groundtruth.



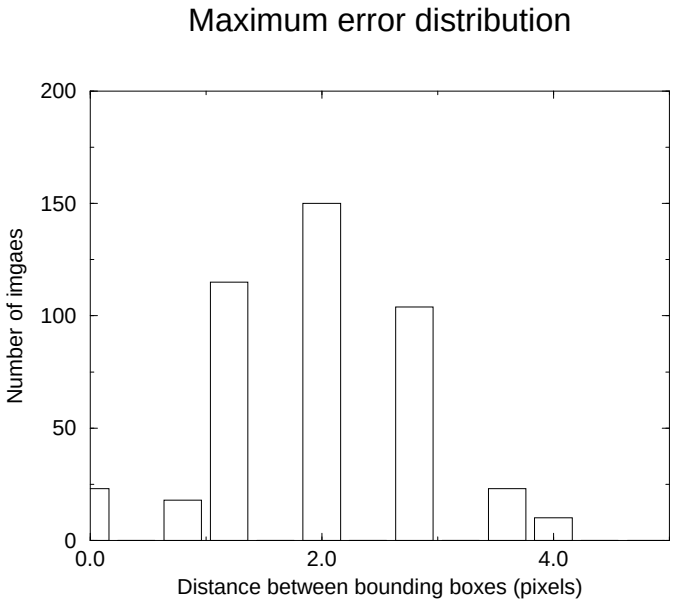


Figure 6: Distribution of mean and maximum errors.

ROBOTEX: An Autonomous Mobile Robot for Precise Surveying *

Xavier LEBEGUE and J. K. AGGARWAL
Computer and Vision Research Center,
Dept. of Electrical and Computer Engr.,
The University of Texas at Austin,
Austin, Texas 78712-1084, U.S.A.

Abstract. The RoboTex project aims at automatically constructing an exact CAD representation of buildings using a mobile robot. This paper reports on the current status of the project. The hardware of the robot is described, with special emphasis on issues relating to measurement accuracy, and algorithms used to process the sequences of monocular images acquired by the robot are presented. Results of automatic indoor surveying are shown and compared to direct measurements in the scene. The techniques developed here have important applications in architectural surveying, scene understanding, and precise robot navigation.

1 Introduction

This paper describes RoboTex, a mobile robot especially designed for building accurate 3-D maps of its environment. The goal of the RoboTex project is to enable a robot to automatically explore a building to construct a very accurate CAD representation. This CAD representation should be as close as possible to what an architect would generate.

Traditionally, the tasks of a robot's perception system are to detect obstacles, find the free space, and estimate the position of the robot in the world. Here, the focus is on building a useful 3-D description of the world. Our 3-D representation of the environment differs primarily from representations used by other robots in that:

1. It must concentrate on *semantically significant* features.
2. It must be more accurate than is strictly necessary for navigation alone.

To satisfy the first constraint, we chose to concentrate on straight edges with particular orientations in the 3-D scene. Typically, there are three prominent 3-D orientations in indoor scenes and outdoor urban scenes: the vertical, and two horizontal orientations perpendicular to each other. Our approach considers only polyhedral objects with such edges. This assumption holds for most large architectural features such as walls, doorways, floors, and ceilings. The second constraint, accuracy, has multiple implications for both the hardware and the software of the robot.

*This research was supported in part by the DoD Joint Services Electronics Program through the Air Force Office of Scientific Research (AFSC) Contract F49620-89-C-0044, and in part by the Army Research Office under contract DAAL03-91-G-0050.

Figure 7: Estimated groundtruth of faxed image.

8 Conclusions

We proposed an improvement over the automatic groundtruthing algorithm proposed by Kanungo and Haralick. We used feature point grouping to reduce the complexity of problem. Then we used the branch and bound algorithm on the grouped feature point sets to estimate the transformation between two images. To analyze the experimental result, we defined the error metric to be the Euclidean distance between the centroids of corresponding characters. On the 450 experimental trials, the estimated groundtruth boxes have a mean error of 1.09 pixels and a maximum error of 2.16 pixels. Further reduction in groundtruth location error can be achieved by using the local matching algorithm described in Kanungo and Haralick [9, 10]. We plan to conduct further experiments (i) to study the performance of our matching algorithm as a function of the rotation angle, and (ii) to quantitatively compare our algorithm to that proposed by Kanungo and Haralick [9, 10], and Hobby [7].

Acknowledgments

We would like to thank Melissa Holland and Jeff DeHart of the Army Research Laboratory and Steve Dennis and Glenn Van Doren of the Department of Defense for funding this work.

This research was funded in part by the Department of Defense under Contract MDA 9049-6C-1250, Lockheed Martin under Contract 9802167270, the Defense Advanced Research Projects Agency under Contract N660010028910, and the National Science Foundation under Grant IIS9987944.

References

[1] J. L. Bentley. Multidimensional binary search trees used for associative searching. *Communications of the ACM*, 18:509–517, 1975.

[2] L. G. Brown. A survey of image registration techniques. *ACM Computing Surveys*, 24:325–376, 1992.

[3] R. G. Casey and D. R. Ferguson. Intelligent forms processing. *IBM Systems Journal*, 29:435–450, 1990.

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

[5] J. H. Friedman, J. L. Bentley, and R. A. Finkel. An algorithm for finding best matches in logarithmic expected time. *ACM Transactions on Mathematical Software*, 3:209–226, 1977.

[6] R. Haralick and L. Shapiro. *Computer and Robot Vision*. Addison-Wesley, Reading, Mass., 1992.

[7] J. D. Hobby. Matching document images with ground truth. *International Journal on Document Analysis and Recognition*, 1, 1998.

[8] D. P. Huttenlocher, G. A. Klanderman, and W. J. Rucklidge. Comparing images using the Hausdorff distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15:850–863, 1993.

[9] T. Kanungo and R. Haralick. Automatic generation of character goundtruth for scanned documents : A closed loop approach. In *Proc. of IAPR International Conference on Pattern Recognition*, pages 669–675, Vienna, Austria, August 1996.

[10] T. Kanungo and R. Haralick. An automatic closed-loop methodology for generating character groundtruth for scanned documents. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21:181–183, 1999.

[11] J. Kittler and J. Illingworth. On threshold selection using clustering criteria. *IEEE Transactions on Systems, Man, and Cybernetics*, 15:652–655, 1985.

[12] D. E. Knuth. *TEX: the program*. Addison-Wesley, Reading, Mass., 1988.

[13] L. Lamport. *LATEX: a document preparation system, 2nd edition*. Addison-Wesley, Reading, Mass., 1994.

[14] D. Mount, N. Netanyahu, and J. LeMoigne. Efficient algorithms for robust point pattern matching and applications to image registration. *Pattern Recognition*, 32:17–38, 1999.

[15] J. A. Nelder and R. Mead. A simplex method for function minimization. *Computer Journal*, 7:308–313, 1965.

[16] I. Phillips. *Users' Reference Manual*. CD-ROM, UW-III Document Image Database-III.

[17] R. Sedgewick. *Algorithms in C*. Addison-Wesley, Reading, Mass., 1990.

[18] V. Torczon. PDS: Direct search methods for unconstrained optimization on either sequential or parallel machines. Technical Report CRPC-TR92206, Rice University Center for Research on Parallel Computation, 1992.

[19] C. Viard-Gaudin, P. M. Lallican, S. Knerr, and P. Binter. The IRESTE On/Off(IRONOFF) dual handwriting database. In *Fifth International Conference of Document Analysis and Recognition*, pages 455–458, Bangalore, India, September 1999.