Connected Component Level Discrimination of Handwritten and Machine-Printed Text Using Eigenfaces

Samuel J. Pinson
Pinson Linguistic Service
1013 E. Kessler Dr.
Sultan, WA 98294
sjpinson@hotmail.com

William A. Barrett
Department of Computer Science
Brigham Young University
Provo, Utah 84602 USA
barrett@cs.byu.edu

Abstract—We employ *Eigenfaces* to discriminate between handwritten and machine-printed text at the connected component (CC) level. Normalized images of machine print CCs are treated as points in a high-dimensional space. PCA yields a reduced-dimensional character space. Representative machine print CCs are projected into character space and a local distance threshold for each representative is automatically determined. CCs are classified as machine print if they are within the local distance threshold of their closest machine print representative. Otherwise, they are classified as handwriting. Recursive character segmentation using min graph cut is used to address the problem of touching characters. Validation over a large NIST handwriting and machine print database demonstrates precision of 93.98% and 89.1% for machine print and handwriting respectively.

Keywords-Eigenfaces; Handwriting/machine print discrimination; touching character segmentation; min graph cut; NIST

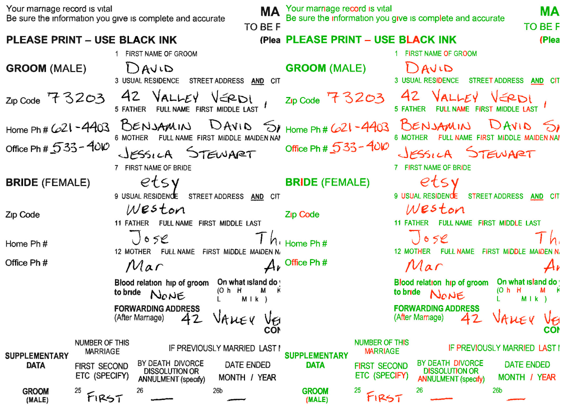
1. Introduction

Many applications in document analysis and recognition require discrimination between machine print and handwriting. Annotations may need to be recognized, recorded and removed. Handwriting may need to be extracted from forms and processed independently while machine print connected components (CCs) are passed to an OCR engine. Other documents may need to be indexed or searched based on machine print or annotated entries. The goal and contribution of this paper is to correctly label input CCs (Figure 1, left) as machine print (green) or handwriting (red), as illustrated in Figure 1, right.

Early work for discriminating between machine print and handwriting relies exclusively on stroke orientation [1]. Franke et. al train an ensemble of statistical classifiers on CC features [2]. Line straightness and symmetry features from CCs have also been used to train a neural network for character level discrimination [3].

An application for mail address blocks [4] uses a variety of features to train a neural network. Character block layout variance is used for Kanji at the text line level [5]. Discriminant functions have been applied to vertical projection profiles [6] at the line, not the CC level. In [7] several CC features extracted are used for word-level classification.

Guo and Ma use a HMM based on linguistic context [8] in conjunction with the differences between vertical projection profiles of machine print and handwriting. Zheng et al. address machine print and handwriting identification in noisy images [9] using three two-way Fischer linear classifiers. A SVM is used to isolate signatures from machine print in [12] and to identify sparse handwritten annotations occurring at arbitrary orientations in [13]. K-means clustering followed by MRF relabeling is used in [14] for segmentation of handwriting, machine print and noise with an overall recall of 96.33%. In [18] a novel approach to automatically discover features pushes error rates of handwriting and machine print to 13.8%.



Your marriage record is vital
Be sure the information you give is complete and accurate
MA
TO BE F
(Plea
PLEASE PRINT – USE BLACK INK
GROOM (MALE)
1 FIRST NAME OF GROOM
DAVID
3 USUAL RESIDENCE STREET ADDRESS AND CIT
Zip Code 73203
42 VALLEY VERDI ,
5 FATHER FULL NAME FIRST MIDDLE LAST
Home Ph # 621-4403
BENJAMIN DAVID S
6 MOTHER FULL NAME FIRST MIDDLE MAIDEN NA
Office Ph # 533-4010
JESSICA STEWART
7 FIRST NAME OF BRIDE
BRIDE (FEMALE)
etsy
9 USUAL RESIDENCE STREET ADDRESS AND CIT
Zip Code
Weston
11 FATHER FULL NAME FIRST MIDDLE LAST
Home Ph #
Jose Th
12 MOTHER FULL NAME FIRST MIDDLE MAIDEN N
Office Ph #
Mar A
Blood relation hip of groom to bride NONE
On what island do
(O h H M K L M l k)
FORWARDING ADDRESS (After Marriage) 42 VALLEY Ve
CON
SUPPLEMENTARY DATA
NUMBER OF THIS MARRIAGE
FIRST SECOND ETC (SPECIFY)
IF PREVIOUSLY MARRIED LAST
BY DEATH DIVORCE DISSOLUTION OR ANNULMENT (specify)
DATE ENDED MONTH / YEAR
GROOM (MALE) 25 FIRST 26 — 26b —

Figure 1. Discrimination between handwritten (red) and machine-printed (green) text using eigenfaces. Note touching character false positives.

Color annotation is extracted from color documents in [15] using robust feature alignment and background subtraction. Work in [16] builds on this, making use of color clustering and a decision tree to identify handwritten annotation in marked up documents containing machine print.

Muller and Herbst [11] apply *Eigenfaces* [10] to character identification by treating characters as points in a high-dimensional space that is reduced to a *Character Space* of eigenvectors using PCA. Characters are classified based on the minimum weighted Euclidean distance. The weights are proportional to the standard deviation in the direction of the eigenvector. Solli also makes use of eigenfaces for looking up fonts in a large font library [18].

We seek to overcome the limitations of previous techniques (reliance on color, manually-derived features, lengthy training or incremental learning cycles) by using Eigenfaces to exploit all relevant character components while automatically deriving CC classification thresholds.

1520-5363/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDAR.2011.280

IEEE computer society

2. Machine print and handwriting discrimination

Similar to [11], we project CCs into a hyperdimensional *character space* for classification. However, we *discriminate* between machine print and handwritten characters, rather than classifying machine print characters.

2.1 Character Space

Let $\Gamma = \mathbf{\Gamma_1}, \mathbf{\Gamma_2}, \ldots, \mathbf{\Gamma_M}$ be a set of M vectors of length N^2 derived from $N \times N$ character images (*N=64)*. The mean character is defined as $\mathbf{\Psi} = \frac{1}{M}\sum_{i=1}^{M}\mathbf{\Gamma_i}$, and the difference image of the i^{th} vector as $\hat{\mathbf{\Gamma}}_\mathbf{i} = \mathbf{\Gamma_i} - \mathbf{\Psi}$ (Figure 2).

Figure 2. Difference image, $\hat{\mathbf{\Gamma}}_\mathbf{i} = \mathbf{\Gamma_i} - \mathbf{\Psi}$, where $\mathbf{\Psi}$ = mean image.

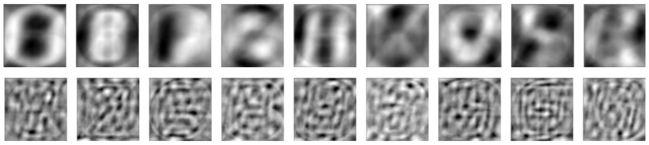


Figure 3. First and last 9 eigenvectors capture character structure, detail.

Let $A = \left[\hat{\mathbf{\Gamma}}_\mathbf{1} \ldots \hat{\mathbf{\Gamma}}_\mathbf{M}\right]$. That is, the columns of A are the difference images of the M vectors in Γ. The covariance matrix C for the set Γ is the $N^2 \times N^2$ matrix where $C = \left[\frac{1}{M}\sum_{k=1}^{M}\hat{\mathbf{\Gamma}}_\mathbf{i}\hat{\mathbf{\Gamma}}_\mathbf{j}^\mathbf{T}\right] = \frac{1}{M}AA^T$. The eigenvectors of C are found via SVD. That is, $A = USV^T$, where the eigenvectors of C are the columns of U. The eigenvector corresponding to the largest eigenvalue points in the direction of greatest variance in character space. Because relatively few of the eigenvectors span most of the variance in the set Γ, most of the remaining eigenvectors may be dropped, reducing the dimensionality of character space to 100 and making the algorithm computationally practical. The first and last nine *basis vectors* (eigenvectors – Figure 3) capture the high-level and detailed structure, respectively, of the characters along their respective axis. The complete algorithm for discrimination between machine print and handwriting is outlined in Figure 6.

2.2 Automatic Selection of Local Distance Threshold, θ_i

Eigenfaces uses a single global threshold for face recognition. However, because machine-printed characters cluster much more tightly than handwriting, we introduce an algorithm to automatically select local distance thresholds, θ_i for each representative machine print *template*, $\mathbf{\Gamma'_i}$. We do this by computing the relative density of machine print and handwriting surrounding each $\mathbf{\Gamma'_I}$ (Figure 4). We used Adobe Illustrator to create 5,957 representative templates of fonts and styles of characters, numbers, punctuation and symbols.

To discriminate between an unknown CC, $\mathbf{U}$, we project its difference image, $\hat{\mathbf{U}} = \mathbf{U} - \mathbf{\Psi}$, into character space to get $\mathbf{U'}$. We determine the Euclidean distance from $\mathbf{U'}$ to the projection of each $\mathbf{\Gamma'_i}$. If the projection of $\mathbf{U'}$ lies sufficiently close, $\mathbf{U}$ is considered machine print, otherwise handwriting.

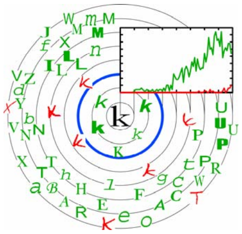


Figure 4. Radial Density (inset) for determining threshold θ_i (blue circle).

A user-supplied global target for machine print precision, *P,* is needed to determine θ_i. We begin with a large set of machine print connected components, A, and a large set of handwritten connected components, B, meant to represent the distribution of its machine print or handwriting representative in character space.

A = 273,286 machine print CCs (26 fonts, normal, bold, italic)
B = 599,724 handwriting CCs, 2100 writers (NIST SD19 DB, 1st 4 parts)

The CCs in these sets are used to determine the CC radial density (number of proximate CCs) of machine print and handwriting with respect to each $\mathbf{\Gamma'_i}$.

If we let $v(r)$ be the volume of a hypersphere of radius r in character space, then the machine print CC density for the representative $\mathbf{\Gamma'_i}$ at a radial distance r is:

$$\rho_i(r) = \frac{1}{v(r)}\sum_{j=1}^{\|A\|}\begin{cases} 1 \text{ if } \|\mathbf{\Gamma}'_\mathbf{i} - \mathbf{A}'_\mathbf{j}\| \leq r \\ 0 \text{ otherwise} \end{cases} \qquad (1)$$

where $\|A\|$ is the cardinality of the set A. Similarly for the handwriting CCs. The summation in Equation 1 yields the number of machine print connected components within distance r of $\mathbf{\Gamma'_i}$ divided by the volume, $v(r)$, of character space under consideration.

We use P = 98%, to automatically select θ_i. We quantize character space about each $\mathbf{\Gamma'_i}$ into b concentric hyper-spheres (Figure 4). The radius of the outermost hypersphere, R, which defines the local neighborhood of each $\mathbf{\Gamma'_i}$, is empirically determined. θ_i is the last (quantized) radial distance, r, before which P drops below 98%. In other words, it is the greatest distance for which it, and all smaller distances, satisfy the target machine print precision (98%).

2.3 Calculating Local Machine Print Precision

Local machine print precision, $P_i(r)$, for a given $\mathbf{\Gamma'_i}$ and θ_i yields the fraction of CCs classified as machine print that were actually machine print (Figure 5).

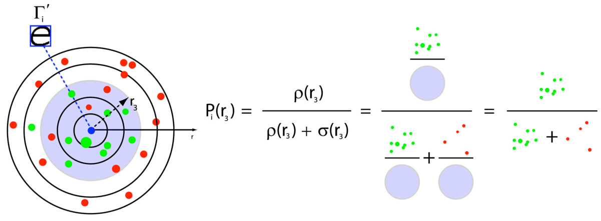


Figure 5. Local Machine Print precision, $P_i(r)$ for given $\mathbf{\Gamma'_i}$, θ_i.

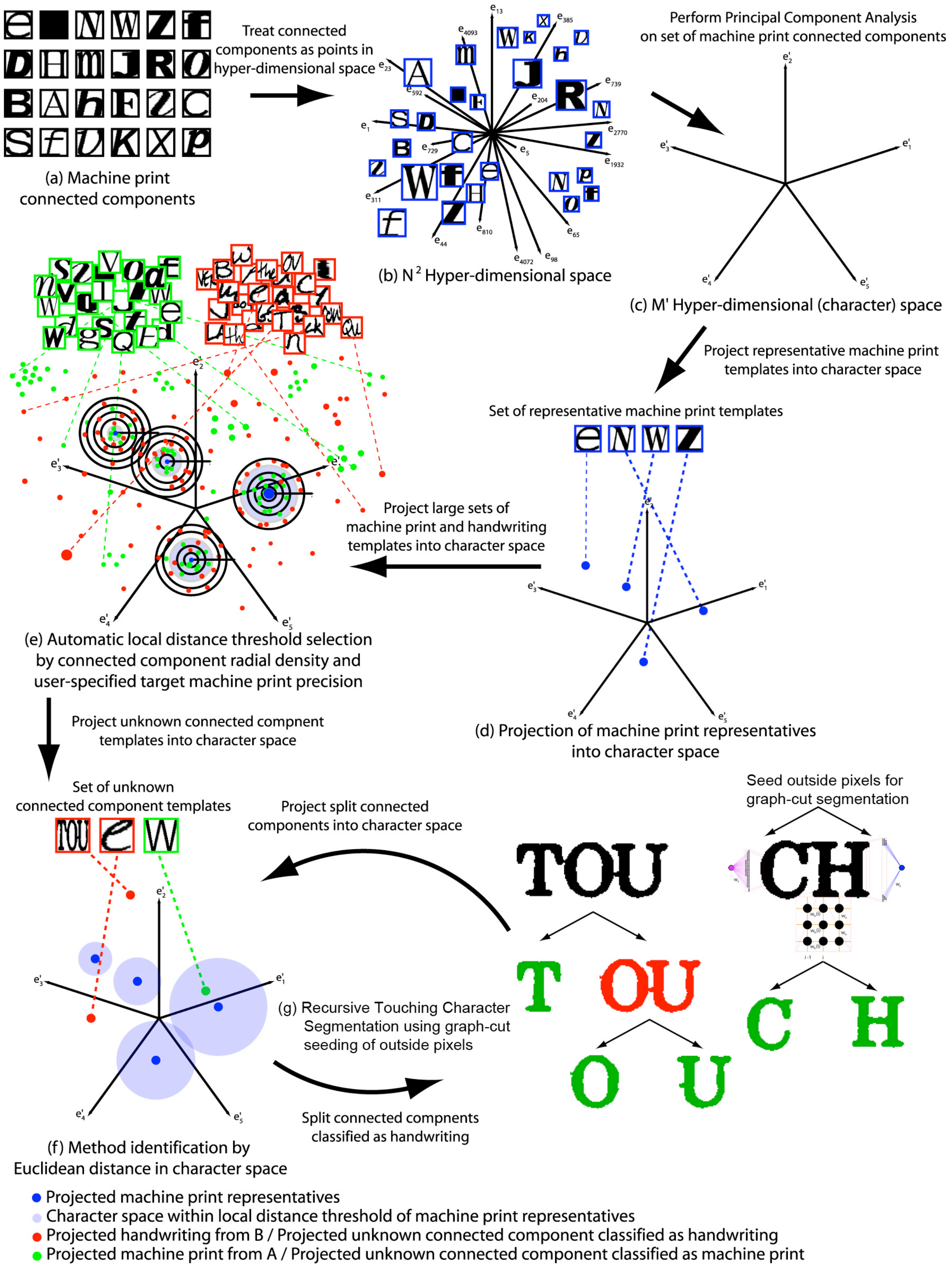

Treat connected
components as points in
hyper-dimensional space
(a) Machine print
connected components
(b) N^2 Hyper-dimentional space
e_{13}
e_{385}
e_{4093}
e_{23}
e_{592}
e_{739}
e_{204}
e_1
e_{2770}
e_{729}
e_5
e_{1932}
e_{311}
e_{810}
e_{65}
e_{44}
e_{4072}
e_{98}
Perform Principal Component Analysis
on set of machine print connected components
e'_2
e'_3
e'_1
e'_4
e'_5
(c) M' Hyper-dimensional (character) space
Project representative machine print
templates into character space
Set of representative machine print templates
(d) Projection of machine print representatives
into character space
Project large sets of
machine print and handwriting
templates into character space
(e) Automatic local distance threshold selection
by connected component radial density and
user-specified target machine print precision
Project unknown connected compnent
templates into character space
Set of unknown
connected component templates
Project split connected
components into character space
Seed outside pixels for
graph-cut segmentation
TOU
CH
T
OU
O
U
C
H
(g) Recursive Touching Character
Segmentation using graph-cut
seeding of outside pixels
Split connected compnents
classified as handwriting
(f) Method identification by
Euclidean distance in character space
● Projected machine print representatives
● Character space within local distance threshold of machine print representatives
● Projected handwriting from B / Projected unknown connected component classified as handwriting
● Projected machine print from A / Projected unknown connected component classified as machine print

Figure 6. Discrimination between machine print and handwritten text: (a) machine print connected components (CCs) (b) projection of CCs into N^2 hyper-dimensional space (c) PCA to create reduced dimensional character space (d) projection of machine print representatives in character space (e) automatic local distance threshold selection from large sets of machine print and handwriting templates and user-specified precision (f) classification as machine print or handwriting (g) min graph cut recursive touching character segmentation by seeding source/sink with outside pixels.

3. RESULTS

The NIST SD19 database consists of 3,699 handwriting forms such as in Figure 7. A certain partition of NIST SD19, *hsf 4*, with samples from 500 different writers has been designated by NIST for the purposes of reporting OCR results. We use *hsf 4* to evaluate our method.

After registering the images and removing the form boxes and name field, each connected component is classified as machine print (green) or handwriting (red) (Figure 7). Extremely small connected components, which do not provide enough information to make a confident classification, are not considered.

The NIST SD08 database includes 360 binary images of machine print characters. Three styles (normal, bold, and italic), six fonts (Courier, Helvetica, New Century Schoolbook, Optima, Palatino, and Times Roman), and ten point sizes (4, 5, 6, 8 10, 11, 12, 15, 17, 20) are represented. Two images of randomly selected characters for each style, font, and size, were used to test our classifier.

The confusion matrix in Table 1 summarizes our results. The upper left entry means that when we predicted a connected component to be machine print, we were right 98.21% of the time. Similarly, the lower right entry means when we classified a connected component as handwriting, we were right 71.05% of the time.

For a connected component to be classified as machine print it must lie within the local distance threshold of the nearest representative machine print template (Figure 4). High machine print precision implies that local distance thresholds are not too loose, else handwriting connected components would be mistaken for machine print.

In general, there is a tradeoff between machine print precision and handwriting precision. However, in this case the low handwriting precision, 71.052%, has a specific cause: small machine print characters get mistaken for handwritten characters. The average height of machine print characters is 19.72 pixels with 40% of them ≤ 20 pixels high, and over 93% ≤ 23 pixels in height.

As the size of connected components diminishes, so do many of the distinguishing features between machine print and handwriting. To test this assertion we replaced the machine print with a similar but larger font (about 40 pixels high) from NIST SD08. The results, shown in Table 2, show 18% improvement in the precision of handwritten characters while maintaining ~94% precision with machine print.

	Predicted Machine Print	Predicted Handwriting
Actually Machine Print	98.21%	1.8%
Actually Handwriting	28.95%	71.05%

Table 1. NIST SD19 *hsf 4* confusion matrix. Over 98% machine print precision, but lower handwriting precision due to font size.

	Predicted Machine Print	Predicted Handwriting
Actually Machine Print	93.98%	6.02%
Actually Handwriting	10.9%	89.1%

Table 2. NIST SD19 *hsf 4* adjusted confusion matrix from replacing machine print in NIST SD19 with a larger font (~40 pixels high) from NIST SD08.

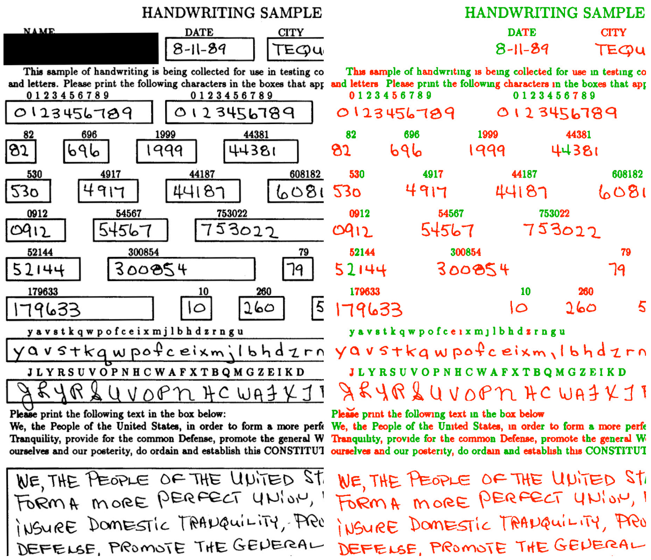

HANDWRITING SAMPLE
HANDWRITING SAMPLE

Figure 7. NIST SD19 handwriting sample form. Left: original. Right: labeled – machine print (green), handwriting (red). Note false positives from touching machine print characters.

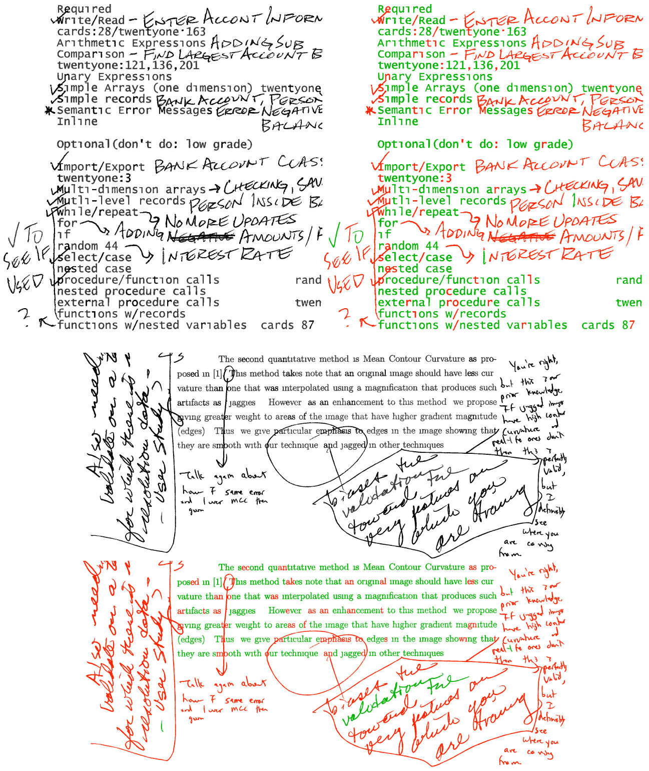


Figure 8. Handwritten annotation and machine print discrimination captures annotations although some annotations cause false positives.

Figures 1 and 8 are representative of results achieved for discriminating between machine print and handwriting. Most handwriting false positives result from touching characters or annotations that artificially connect components. Others result from too much similarity between machine print and handwriting representatives or insufficient discrimination of the eigenfaces algorithm.

3.1 Handwriting Accuracy by Writer

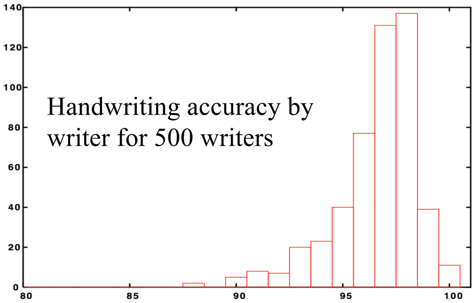


Figure 9. Handwriting accuracy by writer. Each bin in the histogram shows the number of writers for which the labeled accuracy was achieved.

Figure 9 shows the fraction of handwriting CCs correctly classified as handwriting for 500 writers. The worst accuracy for any single writer was 88.24%, the best 100%, with an average accuracy of 96-97%.

3.2 Touching Character Segmentation

About 13% of the handwriting false positives are touching machine print characters. To address this we created a recursive touching character segmentation algorithm using min graph cut to split touching characters by seeding source/sink with outside pixels (Figure 6). Application of the algorithm to 1,056 touching characters consisting of every pairwise combination of letters, in all combinations of upper and lower case, improved classification from 4.07% (without segmentation) to 89.19% (Table 3).

	Without Segmentation	With Segmentation
Machine print precision	4.07%	89.19%

Table 3. 4.07% machine print precision over 1,056 touching character CCs. Recursive segmentation increased precision to 89.19%.

4. CONCLUSION AND FUTURE WORK

Based on a user-supplied global target precision, and automated local threshold detection, discrimination of machine print from handwriting is performed with high precision (94%) over a large data set while maintaining handwriting precision at 89%. Automated selection of principle component features [18] could possibly increase precision while greatly reducing the dimensionality of the character space. Training on handwritten CCs and improved segmentation of touching characters and annotations (Fig. 8) would also likely improve precision/recall to match or exceed levels reported in [14] while competing with error rates reported in state-of-the-art approaches [18].

Handwritten annotations are effectively identified without the aid of color, although annotations that touch machine print cause false positives (Figure 8). Automated seeding of strokes in the character segmentation algorithm (Figure 6) could be used to decouple machine print from annotations. Grayscale, lexical and typographic context might also improve segmentation of touching characters.

In general, machine print that is falsely labeled as handwriting occurs infrequently in the midst of correctly labeled machine print. (See Figures 1, 7 and 8.) These errors could be corrected contextually by inspecting local thresholds of the nearby machine print CCs to discover a better global fit at the word or sentence level. Elimination of font and style templates that are globally distant from the connected components in a document might also improve results.

REFERENCES

[1] T. Umeda and S. Kasuya, “Discriminator between handwritten and machine-printed characters,” http://patft.uspto.gov, 1990.

[2] J. Franke and M. Oberlander, “Writing style detection by statistical combination of classifiers in form reader applications,” in Document Analysis and Recognition, 2nd Intl. Conference, pp. 581–584, 1993.

[3] K. Kuhnke, L. Simoncini, and Z. Kovacs-V, “A system for machine-written and hand-written character distinction,” 3rd Intl. Conf on Document Analysis and Recognition (Vol. 2), pp. 811–814, 1995.

[4] S. Violante, R. Smith, and M. Reiss, “A computationally efficient technique for discriminating between hand-written and printed text,” pp. 1–7, 1995.

[5] K. Fan, L. Wang, and Y. Tu, “Classification of machine printed and handwritten texts using character block layout variance,” Pattern Recognition, vol. 31, no. 9, pp. 1275–1284, 1998.

[6] E. Kavallieratou, S. Stamatatos, and H. Antonopoulou, “Machine-printed from handwritten text discrimination,” Intl. Workshop on Fontiers in Handwriting Recognition, pp. 312-316, 2004.

[7] F. M. Wahl, K. Y. Wong, and R. G. Casey, “Block segmentation and text extraction in mixed text/image documents,” in Computer Graphics and Image Processing, vol. 20, pp. 375–390, 1982.

[8] J. Guo and M. Ma, “Separating handwritten material from machine printed text using hidden markov models,” in Document Analysis and Recognition, 6th Intl. Conference, pp. 439–443, 2001.

[9] Y. Zheng, H. Li, and D. Doermann, “Machine printed text and handwriting identification in noisy document images,” Pattern Analysis and Machine Intelligence, IEEE Transactions, vol. 26, no. 3, pp. 337–353, 2004.

[10] M. Turk and A. Pentland, “Face recognition using eigenfaces,” in Computer Vision and Pattern Recognition, IEEE Computer Society Conference, pp. 586–591, 1991.

[11] N. Muller and B. Herbst, “The use of eigenpictures for optical character recognition,” in Pattern Recognition, 14th Intl Conf, pp. 1124–1126, 1998.

[12] M.S. Shirdhonkar and Manesh B. Kokare, “Discrimination between Printed and Handwritten Text in Documents,” in IJCA Special Issue on “Recent Trends in Image Processing and Pattern Recognition” RTIPPR, pp. 131-134, 2010.

[13] S. Chanda, K. Franke and U. Pal, “Structural handwritten and machine print classification for sparse content and arbitrary oriented document fragments,” in *SAC'10*, pp. 18-122, March 22-26, 2010.

[14] X. Peng, S. Setlur, V. Govindaraju, R. Sitaram and K. Bhuvanagiri, “Markov Random Field Based Text Identification from Annotated Machine Printed Documents,” 10th Intl. Conference on Document Analysis and Recognition (ICDAR 2009), pp. 431-435, 2009.

[15] T. Nakai, K. Kise, and M. Iwamura, “A Method of Annotation Extraction from Paper Documents Using Alignment Based on Local Arrangements of Feature Points,” 9th Intl. Conference on Document Analysis and Recognition (ICDAR 2007), pp. 23-27, 2007.

[16] A. Mazzei, F. Kaplan and P. Dillenbourg, “Extraction and Classification of Handwritten Annotations,” in ACM 978-1-60558-843-8/10/09, UbiComp, pp. 1-7, 2010.

[17] M. Solli and R. Lenz, “FyFont: Find-your-Font in Large Font Databases,” in SCIA 2007, LNCS 4522, pp. 432–441, 2007.

[18] S. Wang, H. Baird, and C. An, “Document Content Extraction Using Automatically Discovered Features,” in 10th Intl Conf on Document Analysis and Recognition, pp. 1076-1080, 2009.