Keyword Spotting For Cursive Document Retrieval

Patricia Keaton, Hayit Greenspan and Rodney Goodman
Department of Electrical Engineering, 136-93
California Institute of Technology
Pasadena, CA 91125
keaton,hayit,rogo@micro.caltech.edu

Abstract

We present one of the first attempts towards automatic retrieval of documents, in the noisy environment of unconstrained, multiple author, handwritten forms. The documents were written in cursive script for which conventional OCR and text retrieval engines are not adequate. We focus on a visual word spotting indexing scheme for scanned documents housed in the Archives of the Indies in Seville, Spain. The framework presented utilizes pattern recognition, learning and information fusion methods, and is motivated from human word-spotting studies. The proposed system is described and initial results are presented.

1. Introduction

In this paper, we present preliminary research on a visual word spotting indexing scheme for the archival and retrieval of scanned historical documents housed in the Archives of the Indies in Seville, Spain. An example of a digitized document is given in Fig. 1. These documents were written in cursive script by multiple authors, and are hundreds of years old (many of which date back to Columbus's era). There exists a tremendous need for scholars to constantly search and explore the contents of such archives. However, conventional OCR and text retrieval engines are inadequate for such tasks [1]. Existing OCR systems often rely upon the ability to cleanly segment the words prior to recognition. The documents in our database exhibit many problems which would certainly cause such systems to fail. We must contend with noise introduced by the photocopying and scanning processes, as well as stray marks, underlines, and overlapping words. Under these conditions perfect segmentation would be impossible. We have developed an alternative strategy for the indexing and retrieval of such documents based on learning a set of keyword signatures of particular words of interest.

Our approach applies many standard image processing techniques in the preprocessing of the documents, and the extraction of the spatial characteristics of the words. In addition, we attempt to characterize words via signatures motivated from human word spotting experiments. The recognition strategy is based upon probabilistic signature matching, in which we view the entire word globally, rather than segmenting and recognizing the individual letters of the word. We investigate the ability to use such signatures, together with advanced encoding schemes and learning, to facilitate the spotting of keywords in handwritten cursive documents.

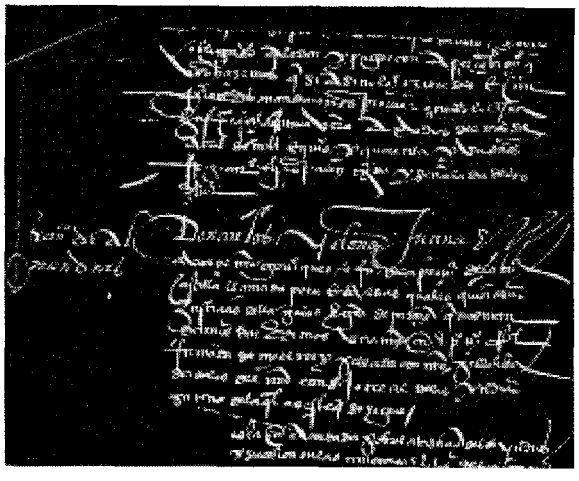

Figure 1. A sample page from the Archives.

2. Background

Very little work has been previously published on the recognition of cursive documents. Most of the work on handwritten material has focused on data collected from the U.S. Postal Service, with the emphasis being the automatic extraction and recognition of addresses and zip-code information. Recent research has focused on the general issues of line extraction, word segmentation, alignment and

0-8186-8055-5/97 $10.00 © 1997 IEEE

recognition. The differences between the postal application and our word spotting task include the availability of a priori knowledge in a restricted domain (digits, city-names, states), and the possibility of exploiting context (across words) in the recognition process.

In [2], Manmatha, et al. introduce the topic of "keyword spotting" in single authored archived manuscripts. Their framework entails segmenting the document into words, and then matching actual word images with each other to create equivalence classes. Each class consists of multiple instances of the same word; the words in these classes can be used for indexing the document. A given word image is used as a template and matched against all other word images. This is repeated for every word in the document. Matching is based on entire words, rather than segmenting the words further into individual letters or connected components. The experiments performed match 2 pages from 2 authors with a given input word, and produce as output a ranked set of best-matched candidate words.

In this paper, we also focus on the matching of *words* in a given historical document. However, we avoid the page segmentation problem by incorporating a focus-of-attention module (described below), to identify candidate locations prior to performing the word-level matching. The main differences from the above referenced work are that we do not limit ourselves to documents written by a single author, instead we attempt keyword matching in a multi-authored domain exhibiting high variability within each keyword class; and the word matching is performed using a feature-space representation (rather than in the raw pixel domain).

3. The System

The system we propose is composed of several modules, as outlined below.

Focus-of-attention module: This module involves normalized cross-correlation of the document image with a set of keyword prototypes (templates) which have been extracted from a training set of documents. A set of candidate locations is extracted, with the different locations ranked by correlation strength. The locations of the top correlation peaks are then passed along to the preprocessing stage. An example is presented in Fig. 2. Computational considerations involved in the normalized cross-correlation procedure are derived in Appendix A.

Preprocessing module: Having isolated the regions of interest, various preprocessing routines are applied to regularize the appearance of words lying within the attentional window, and to remove noise.

Estimation of Word Zones: The word image is subdivided into three zones: Upper, Middle, and Lower, using projection analysis. The horizontal projections of each row in the image consist of a simple running count of the "on" pixels in each column. The resulting histogram is used to estimate of the top-line, center-line, and base-line of the word [3]. The height of the middle zone is a good estimate of the lower-case characters' height, while the upper and lower bounds provide information about the maximum and minimum escalations called the ascenders and descenders, respectively.

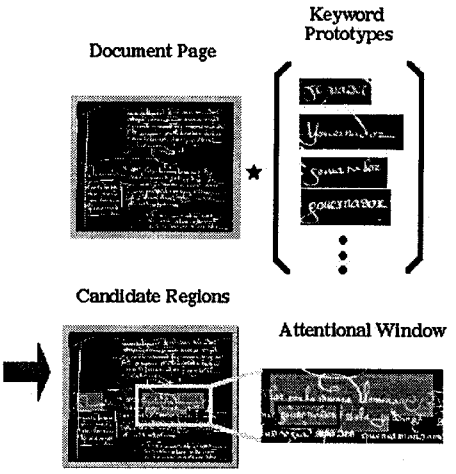

Figure 2. Focus-of-Attention module: Normalized cross-correlation of the document image with a set of keyword prototypes is performed to identify candidate locations.

Filtering of Stray Marks: Connected component analysis (or eight-way connectivity) is used next to determine the number of connected stroke regions in the word image. For each connected region extracted, the "bounding box" is found, which enables the computation of location, dimension, and centroid information. Using this information, components found to be far away from the top-line and base-line of the word are estimated to be stray marks, and thus deleted. An example of an input image and the filtered output, following zoning and stray-mark removal, is shown in Fig. 3.

Skeletonization: For feature extraction, it is necessary to first skeletionize the word image. In this step we remove extra pixels to produce a thinned image of the word. The process of thinning by successive deletion is much like that of erosion: the pixels to be removed are marked and are removed in a second pass. This is repeated until there are no more redundant pixels, at which point the remaining pixels are those belonging to the skeleton of the object. The skeleton remaining must possess the following properties: 1. Thinned regions should be one pixel wide. 2. The pixels comprising the skeleton should lie near the center of a cross section of the region. 3. Skeletal pixels must form the same number of regions as those of the original binary image. A

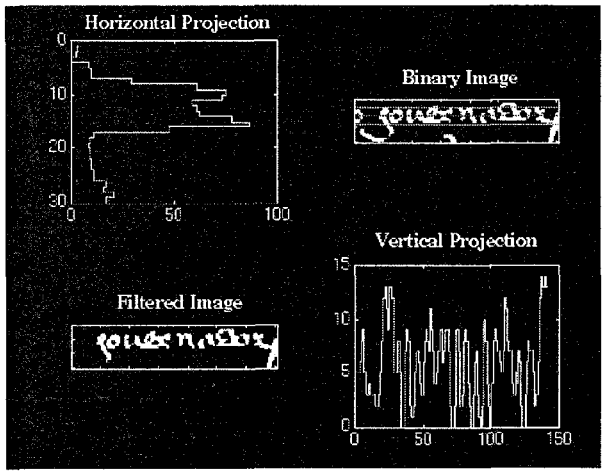

Figure 3. Example of an input image and the filtered output, following zoning and stray-mark removal.

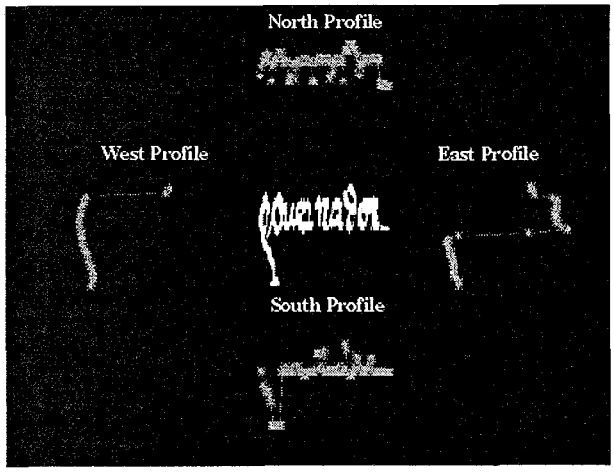

Figure 4. Profile Signature Extraction

number of methods were implemented and tested, and the Zhang-Suen [4] method was found to perform the best (see Appendix B).

Feature-extraction module: We utilize a combination of global and local features in the form of profile signatures and morphological cavities to characterize a word. Human word spotting experiments have shown that global shape information is one of the most important cues we use to distinguish words. The general shape of a word may be approximated using simplified *profile signatures*. Vertical projection analysis is used to determine the upper (North) and lower (South) profiles of the word. West and East profiles are generated by horizontal projection analysis and are used to detect descenders and ascenders. Fig. 4. shows an example of the profile signatures extracted for the word

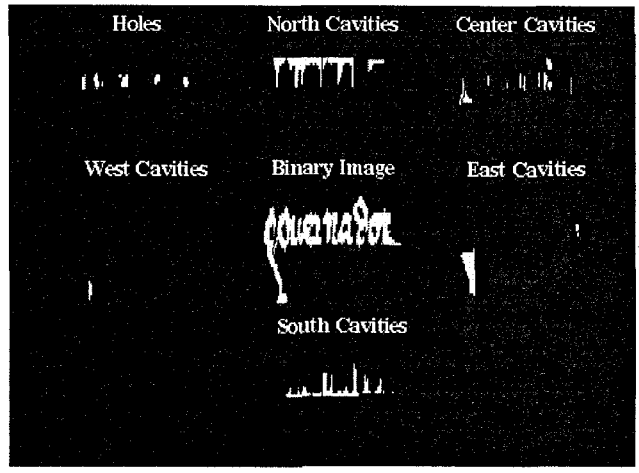

Figure 5. Morphological Cavity Feature Extraction

governador.

Cavity features are essentially the gaps between the strokes of the word. They capture local variations in the word, which are useful in discriminating words having the same general shape. There are six cavity feature types: east, west, north, south, center, and hole. A cavity is a region of points bounded by the character stroke on at least three sides (named by the side on which they are not bounded). A hole is a region completely bounded, while a center cavity is surrounded on all four sides, but is not a hole. A morphological algorithm [5] is used to compute the cavity features using combinations of dilations or smears, in different directions and intersections, and generating six feature images, as shown in Fig. 5 (see Appendix C).

4. Keyword Signatures

We define *keyword signatures* as the collection of features which characterize a keyword, and which allow its matching with candidate words from the documents. In this work, we focus on *holistic* signatures, that characterize the entire word, without breaking it into individual letters or strokes. The features we use are motivated from studies of human visual word-spotting [6]. Examples of such features include the relative frequency of ascenders and descenders in the word; their relative coordinates in the word; and intra-word gaps. These features are captured by the profile and cavity information extracted. Recall that descenders are the portions of a character that fall below the base-line of a word. Their presence results in strong valleys in the South Profile. Ascenders are the portions of a character that fall above the top-line of a word, which are detected as peaks in the North profile.

Profile Encoding: One dimensional transform methods are used to encode the North and South profiles as well as their difference, into feature vectors that are suitable for matching. The desired transform should concentrate the energy associated with the profile signatures into as few coefficients as possible. A variety of methods were tried such as: Discrete Cosine Transform (DCT); Fast Fourier Transform (FFT); Real Cepstrum; Median Coarse Coding; and the Haar, Mallat, and 4-PT Daubechies Wavelet Transforms. All encoding methods were applied to the three profiles, with the top 20 coefficients of each retained for matching (producing a 60-dim. feature vector). We found that the DCT performed best due to the fact that successive values of the profile signatures were often highly correlated. This is particularly true for the lower profile which tends to exhibit a flat response except in areas where a descender exists (see Fig. 4). In this case, the DCT is known to perform very close to the optimal KLT - Karhunen-Loeve Transform. This is particularly true for the lower profile which tends to exhibit a flat response except in areas where a descender exists (see Fig. 4). In addition, the DCT handles signals with trends, which occur when there exists an ascender or descender.

Cavity Encoding: Graph-based models in which the relative 2D spatial arrangement is preserved, are used to encode the cavity features, as well as the descender and ascender information. The feature attributes incorporated into the graph are the type, size, and relative location of each feature. Fig. 6 shows two examples of graphical models: the data graph associated with the attentional window (top), and the prestored keyword graph (bottom).

Keyword Signature Matching: The process of matching the keyword signatures to the signatures generated from the data contained in the attentional window is shown in Fig 6. First, the DCT encoded profile signatures are matched. We experimented with using a number of measures of similarity including: Cross-correlation; Cosine similarity; Euclidean distance; City-block distance; Minkowski distance; and Dynamic Time Warping of raw profiles. We found that the best classification performance was achieved using a K-nearest-neighbor (K = 5) classifier with the Minkowski distance (r = 4):

$$d_{mink}(\mathbf{x},\mathbf{y}) = (\sum_{i=1}^{n} |y_i - x_i|^r)^{1/r} \qquad (1)$$

The profile matching score is incorporated into the keyword signature graph as an additional feature.

Probabilistic graph matching based on Bayesian evidential reasoning is used to find the best match between the keyword signatures, and those generated by words lying in the candidate regions. At this time, the matching algorithm only aggregates positive evidence in the form of correspondences between the graphs. Fig. 7 shows the information

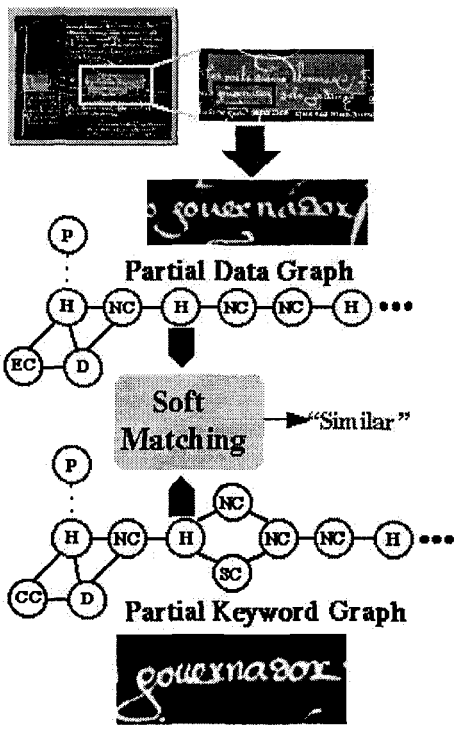

Figure 6. Keyword signature matching, where ?C-Cavity (H,N,S,E,W,C), D-Descender, A-Ascender, P-Profile Coefficients.

fusion process under an evidential reasoning framework. We treat the features extracted as information sources (S_i), and view the correspondences between the data graph and the keyword graph as evidence (E_i) pointing to the hypothesis that the word in the attentional window is an instance of the input keyword (H).

Bayesian theory [7] uses an "Odds-Likelihood Ratio" formulation of Bayes' rule to aggregate the evidence from multiple sources. The *a priori* odds, $O(H)$, of a given hypothesis, H, is related to its *a priori* probability, $P(H)$, by the following relations:

$$O(H) = \frac{P(H)}{P(\neg H)} \qquad (2)$$

and

$$P(H) = \frac{O(H)}{1 + O(\neg H)} \qquad (3)$$

where $\neg H$ means "not H". Thus a hypothesis with a probability of 0.2 has odds of 0.25 (or "4 to 1 against"), and a hypothesis that is absolutely certain (i.e., has a probability of 1) will have infinite odds. The likelihood of the evidence E_i, given that the hypothesis H is true, is:

$$L(E_i|H) = \frac{P(E_i|H)}{P(E_i|\neg H)}. \qquad (4)$$

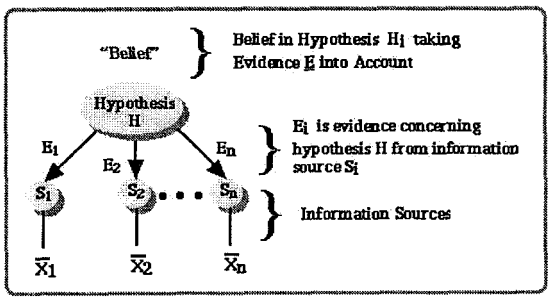

Figure 7. Information fusion through evidential reasoning.

with $P(E_i|H)$ derived from the statistics of the training set. The formula for updating the odds (i.e., the a posteriori odds) of a hypothesis H, given the evidence observed, E_i, is:

$$O(H|E_1, E_2, ..., E_n) = O(H) \prod_{i=1}^{n} L(E_i|H) \quad (5)$$

Statistical independence of the evidence sources is assumed, with an initial *a priori* odds of 1 to 1 for each hypothesis H or $P(H)$ uniformly distributed. The "belief" or *a posterior* probability for a hypothesis is:

$$P(H|E_1, E_2, ..., E_n) = \frac{O(H|E_1, E_2, ..., E_n)}{1 + O(H|E_1, E_2, ..., E_n)} \quad (6)$$

The classification is chosen to be that hypothesis H having the greatest probability given the accumulated evidence.

5. Experimental Results

The database we currently hold has on the order of 100 document pages, which were scanned in at 150 dpi from photocopying copies of manuscripts contained in the Archives of the Indies in Seville, Spain. We were also provided with a set of meaningful "keywords" to be used as indices. Initial experiments focused on the recognition of two keywords: *governador* and *provincia* which appeared in 66 of the documents. We present results using three types of feature inputs: 1. Profile features only. 2. Cavity features only. 3. Combination of both profile and cavity features. All classification percentages reported were computed with no rejection threshold applied to the matching confidence. Table 1 shows the K-nearest-neighbor (K = 5) classification results obtained using only the profile signatures (60 dimensional DCT encoded feature vector), with the Minkowski (r = 4) similarity metric. The overall classification accuracy achieved was **82.58%**. For the cavity experiment, the six cavity feature maps were coarsely encoded to generate a 36 element vector. Table 2 shows the K-nearest-neighbor (K = 5) classification results obtained using only the encoded cavity feature maps, with the a cosine similarity metric. The overall classification accuracy achieved was **78%**. Finally, the probabilistic graph matching of keyword signatures which include both cavity and profile information, produced the best result of **91.7%** (see Table 3). The increase in the percentage of correctly classified words can be attributed to the fact that the feature sets exhibit some error independence (i.e., they are complementary feature sets). That is, the incorrect classifications produced by one set of features, become the correct classifications of the other set, which is exploited by the Bayesian evidential reasoning framework.

keyword	correct	error
governador	52	14
provincia	57	9

Table 1. Classification results on two keywords, using DCT encoded profile information. Overall correct classification is 82.6%.

keyword	correct	error
governador	42	24
provincia	61	5

Table 2. Classification results on two keywords, using coarsely encoded cavity information. Overall correct classification is 78.0%.

It is evident that the *"word spotting"* task we have set out to tackle is a difficult one, even in this presented two-word case. The scenario of multiple-author documents produces large variability in the in-class word characteristics. While across classes, similar length words exhibit strong similar-

keyword	correct	error
governador	57	9
provincia	64	2

Table 3. Classification results on two keywords, using keyword signatures and Bayesian inferencing. Overall correct classification is 91.7%.

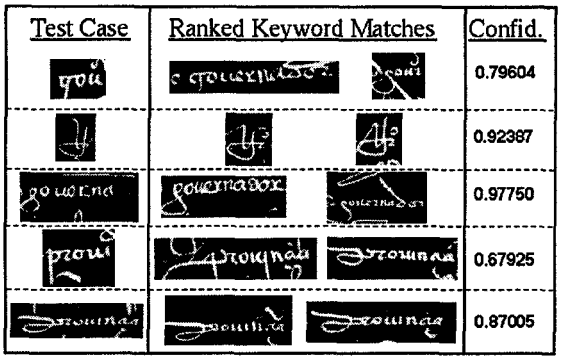

Figure 8. Correctly identified words (2 keyword experiment). For each test case (left column), the ranked keyword matches are given along with the matching confidence associated with the first keyword (leftmost center column).

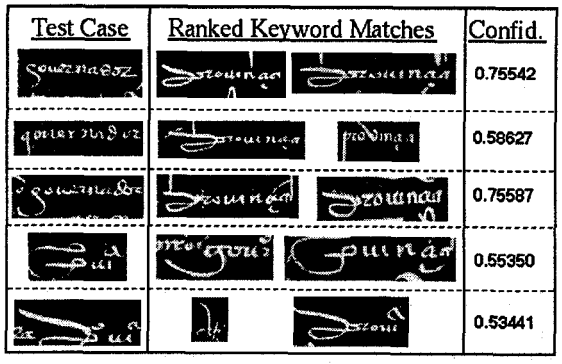

Figure 9. Incorrectly identified words (2 keyword experiment).

ity, as in this two-word case where both words are initialized with a descender. Even so, our preliminary results are comparable to the results presented in the single-authored (2-document case) [2].

Examples of correctly classified samples and incorrectly classified samples are presented in Fig. 8 and Fig. 9 respectively. For each test case, the ranked keyword matches are given along with the matching confidence associated with the first keyword (leftmost center column). It is interesting to note the correspondence between the percentage value and the visually perceived similarity between words. A high confidence level is present for the correctly classified cases, with a much reduced level associated with the incorrectly classified ones. As mentioned earlier, no rejection threshold is applied. Incorporating a rejection threshold of 60%, for example, would have enabled the removal of many of the misclassified cases shown, thereby leading to a much higher percentage level.

In our graph-matching scheme, we only consider node correspondences as evidence, and do not penalize for any lack of correspondence due to word-length differences. This allows for the association of an abbreviated word with its full-word counterpart (as in the topmost example in Fig. 8). However, this can also lead to an increase in misclassification (as in the bottom-most example in Fig. 9). In this case, the use of a penalty term would be advantageous.

In the second set of experiments we augment the database to 4 keywords: *el rey, governador, peru, provincia*, which were contained in 50 documents. Using profile information alone percent correct is **54.5%** (see Table 4), for cavity information alone the classification accuracy is **64.0%** (see Table 5). Combining profiles with cavities in the Bayesian framework increased the classification rate to

keyword	correct	error
el rey	45	5
governador	11	39
peru	21	29
provincia	32	18

Table 4. Classification results on four keywords, using DCT encoded profile information. Overall correct classification is 54.5%.

72% (see Table 6). Examples of correctly classified samples and incorrectly classified samples are presented in Fig. 10 and Fig. 11 respectively. The lower classification percentage can be attributed to the addition of the keyword *peru*, which introduces confusions since it shares the same shape with the abbreviated forms of *governador* and *provincia*. In order to improve the discrimination ability of keyword signatures, we intend to extend our feature set to include directional line segments and infliction points.

keyword	correct	error
el rey	45	5
governador	27	23
peru	27	23
provincia	29	21

Table 5. Classification results on four keywords, using coarsely encoded cavity information. Overall correct classification is 64.0%.

keyword	correct	error
el rey	48	2
governador	27	23
peru	27	23
provincia	42	8

Table 6. Classification results on four keywords, using keyword signatures and Bayesian inferencing. Overall correct classification is 72.0%.

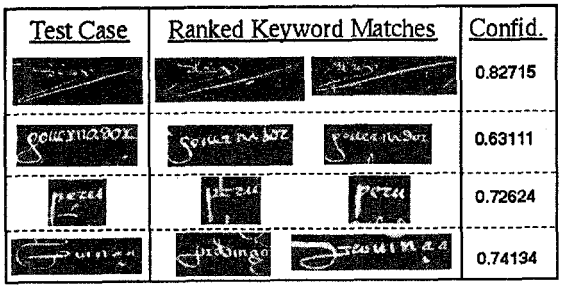

Figure 10. Correctly identified words (4 keyword experiment).

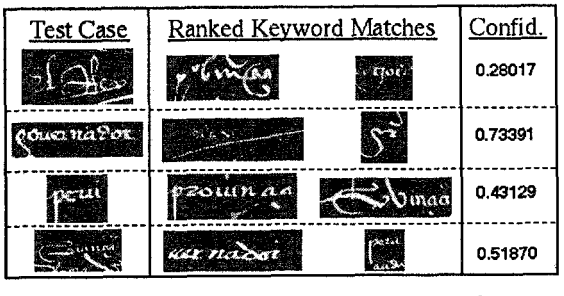

Figure 11. Incorrectly identified words (4 keyword experiment).

6. Conclusion

In this work we are presenting one of the first attempts in the literature to handle documents in the noisy environment of unconstrained, multiple author, cursive handwritten forms. We introduce a framework which is motivated from human visual cognition, and which utilizes tools from pattern recognition, learning and information fusion. Learning techniques are utilized in the encoded-signature domain, to learn characteristic signatures of keywords for storage in a database. Our goal is to demonstrate that bringing in these new perspectives will allow for new methods to be developed in document indexing and retrieval.

Appendix

A. Normalized Cross-Correlation Computation For Focus-Of-Attention

A keyword template, T, is compared the document image, I, using normalized cross-correlation. The computed similarity measure is insensitive to linear transformations of gray scale:

$$p = \frac{\sum_{x,y}(I(x,y) - \mu_I)(T(x,y) - \mu_T)}{\sigma_I \sigma_T} \tag{7}$$

$$\mu_T = \frac{1}{N_x N_y} \sum_{x,y} T(x,y);$$

$$\sigma_T^2 = \frac{1}{N_x N_y} \sum_{x,y} (T(x,y) - \mu_T)^2.$$

In eqn. 7, p is a scalar which varies between -1 (anti-correlated) and $+1$ (perfectly correlated). The most likely locations of a keyword in an document image are found using eqn. 7. Direct calculation of the normalized cross-correlation equation is very time consuming, therefore we perform the computation in the frequency domain. A normalized keyword template T' is computed by:

$$T'(x,y) = \frac{(T(x,y) - \mu_T)}{\sigma_T} \tag{8}$$

Next, the μ_I term of eqn. 8 is approximated by convolving the image with a Gaussian kernel, G:

$$\mu_I(x,y) = (G * *I)(x,y) \tag{9}$$

Once the local mean image is computed, we can then subtract it from the original image and convolve the squared result with a Gaussian. This approximates the local variance:

$$\sigma_I^2(x,y) = (G * *(I - \mu_I)^2)(x,y) \tag{10}$$

The normalized the document image is then computed as:

$$I'(x,y) = \frac{(I(x,y) - \mu_I)}{\sigma_I} \quad (11)$$

Finally, an approximation of the correlation coefficient at each image location (x,y) is given by correlating T' and I':

$$\rho(x,y) \approx (T' \circ I')(x,y). \quad (12)$$

This technique has several advantages. Once the document image is normalized in eqn. 11, any number of keyword templates can be efficiently matched against the document using eqn. 12.The locations of the top correlation peaks are then passed along to the feature extraction stage (see Fig. 2).

B. The Zhang-Suen Method For Skeletonization

The Zhang-Suen method determines whether a pixel can be eroded by looking only at its eight neighbors. There are two rules used to decide whether or not a pixel may be removed. The first rule is that a pixel can be deleted only if it has more than one and fewer than seven neighbors. By neighbors we mean 8-adjacent object pixels. This rule ensures that the end points of the skeleton are not eroded away and that pixels are stripped away from the boundary of the region, not from the inside. The second rule states that a pixel can be deleted only if it is connected to only one other region. This ensures that the skeletal pixels form the same number of regions as those of the original binary image.

To thin a region these rules must be applied to all of the pixels that belong to the region, and those pixels satisfying the previous conditions can be removed. This is done again and again until no more pixels can be deleted, at which point the remaining pixels should be a skeleton. When a pass through the image results in no pixel deletions, the thinning procedure is finished.

C. Cavity Extraction

Let N, S ,E, and W denote structuring elements which are rays in the directions north, south, east, and west. The cavity feature images are then computed according to the following morphological algorithm:

$$\begin{aligned}
NF &= I \oplus N \cap (I \oplus S)^C \cap I \oplus E \cap I \oplus W \cap B \\
SF &= (I \oplus N)^C \cap I \oplus S \cap I \oplus E \cap I \oplus W \cap B \\
EF &= I \oplus N \cap I \oplus S \cap I \oplus E \cap (I \oplus W)^C \cap B \\
WF &= I \oplus N \cap I \oplus S \cap (I \oplus E)^C \cap I \oplus W \cap B \\
CF &= I \oplus N \cap I \oplus S \cap I \oplus E \cap I \oplus W \cap B \\
HF &= (span - until(BORDER, B, T) \cup I)^C
\end{aligned}$$

The cavity feature images are denoted as HF, CF, NF, SF, EF, and WF, which are the hole, center, north, south, east and west feature maps respectively. I denotes the original binary image of the word, and B the background, or complement, of the image I.

A hole is any region of background that is completely surrounded by the foreground in the word image. In the expression for the hole feature image (HF), BORDER denotes the image that consists of the one-pixel-wide border around the edge of the image, which is assumed to be completely contained in the background. T represents the 3×3 binary structuring element (N,S,W,E). The function span-until represents the iteration of the conditional dilation operation.

Acknowledgments

We would like to thank Victoria Carmona Vergara of Seville, Spain, and Kanna Shimizu for their help in generating our database of documents. This work was supported in part by ARPA and ONR grant no. N00014093-1-0990. Patricia Keaton is supported by a Hughes Doctoral Fellowship, and Hayit Greenspan is supported in part by and Intel research grant.

References

[1] Mori, S., Suen, C., Yamamoto, K., "Historical review of OCR research and development", *Proceedings of the IEEE*, vol. 80, no. 7, pp. 1029-1058, 1992.

[2] Manmatha, R., Han, C., Riseman, E. M., "Word spotting: A new approach to indexing handwriting", *IEEE Proceedings CVPR-96*, pp. 631-637, June 1996.

[3] Parker, J. R., " Practical Computer Vision using C," Wiley, 1994.

[4] Zhang,T.Y. and Suen,C.Y., "A Fast Parallel Algorithm for Thinning Digital Patterns", *Communications of the ACM*, Vol. 27, No. 3, pp.236-239, 1984.

[5] Gader, P., Gillies. A., Hepp, D., "Handwritten Character Recognition", Digital Image Processing Methods, Marcel Dekker, pp. 223-260, 1994.

[6] Humphreys, G. W. and Bruce, V., "Visual Cognition - computational, experimental and neuropsychological perspectives", Chapter 7, Lawrence Erlbaum Associates Ltd., Publishers, U.K., 1989.

[7] Jensen, Finn V., "An Introduction to Bayesian Networks", Springer-Verlag, New York, New York, 1996.