Emergency Medicine, Disease Surveillance, and Informatics

Venu Govindaraju
Center for Unified Biometrics and Sensors, Department of Computer Science and Engineering
University at Buffalo, NY 14260
govind@buffalo.edu

ABSTRACT
Traditional handwriting recognition algorithms rely heavily on small lexicons and clean word images. Unfortunately, emergency medical documents do not satisfy either of these conditions. This paper describes a strategy whereby given an image representing a noisy handwritten word from a medical document, and a large lexicon consisting of English, medical and pharmacological words, symbols, abbreviations and acronyms, significantly reduces the size of the lexicon while keeping the unknown desired entry within the lexicon.

An approach for segmenting handwritten text is also presented. Stroke analyses are performed and image primitives are extracted for word detection. A heuristics-based approach, involving gap spacing, height transitions, and the average stroke width of the writer is used in detecting word boundaries. Experiments show perfect segmentation of 69%, outperforming the more tested and proven algorithms by as much as 15%.

Categories and Subject Descriptors
I.5 [**Pattern Recognition**]: Models –neural nets; Design Methodology – classifier design and evaluation, feature evaluation, pattern analysis; Applications- text processing. I.4 [**Image processing and Computer Vision**]: Segmentation

General Terms
Algorithms, Experimentation

Keywords
Medical Forms Processing, OCR.

INTRODUCTION
Pre-Hospital Care Reports (PCRs) are forms used by Emergency Medical Service personnel to gather vital patient information and track the patient until the handoff to the hospital. Vast amounts of PCRs are kept in State Departments of Health, for microanalysis, which often does not occur until years after the forms were received! [1]. More timely and efficient disease or bio-terrorist related analysis can be performed on the PCR data if the handwritten text on the forms can be automatically extracted, transliterated and stored in a database.

Because of the nature of the circumstances under which the PCRs are filled in (emergencies - often involving fast moving vehicles and high-stress environments), the task of digitally transliterating the PCR data is quite challenging for several reasons. (i) handwritten responses are very loosely constrained in terms of writing style, format of response, and choice of text, (ii) when a respondent is nearing the right margin of the box and has more to write, the writing gets squeezed in the space available, (iii) some input fields contain words that have been subsequently struck out, unidentifiable shorthand symbols, marks for insertion of additional text etc. (iv) availability of only carbon copies of the PCR - as shown in Figure 1.

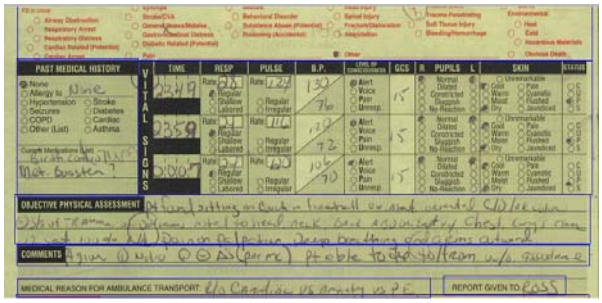

Figure 1: Sample of a portion of the PCR with the form layout regions highlighted.

MEDICAL PHRASE RECOGNITION
There are five major zones on the PCR containing the handwritten information of interest: Chief Complaint, Subject Assessment, Objective Physical Assessment, Comments, and Past Medical History. These handwriting areas contain numbers (e.g. 84), symbols (e.g. ↑ = increase), abbreviations (e.g. ABD = abdominal), acronyms (e.g. CHF = Congestive Heart Failure), anatomical descriptions (e.g. thoracic), medical conditions (e.g. pneumothorax), pharmacological words (e.g. codeine) and common English. The handwritten zones can contain data from a large heterogeneous lexicon and the text often does not fit perfectly within form boundaries. The form available for processing and data mining is a carbon copy. The carbon mesh residue in various locations on the form, and broken/unnatural handwriting due to ambulance movement and emergency environments, add further complexity to the document. The text zones therefore present a highly challenging recognition problem.

A compact lexicon is constructed before sending each segmented word from the PCR to a lexicon driven handwriting recognition algorithm. Research has shown that the smaller the lexicon the greater the odds of successful recognition, provided that the target word is in the lexicon. That is the motivation for the reduced lexicon strategy.

An enhanced binarized representation is produced for each segmented word, which is input to a lexicon free word recognizer (LFWR, [4]). The characters with the highest confidence

determined by the recognizer are then converted to a list of substring text patterns, and then to several ANN compatible input layer configurations. Using the patient's presenting problem, indicated by a checkbox on the medical form, the ANN is trained with all possible configurations against it.

When later presented with a new group of three word images, the LFWR is used to produce the substring configurations, which are used to query the ANN for the concept. The words from the lexicon associated with that concept are finally extracted. This pruned lexicon is then submitted to a lexicon driven word recognizer (LDWR, [3]) which will produce the final recognition results.

A backpropagation ANN is trained with the bi-substring sequence concepts generated from the LFWR. The ANN is used to associate a patient's presenting medical problem (i.e. the concept), with each grouping of three adjacent words, on the corresponding medical form. The word grouping is denoted as a phrase and the presenting problem is denoted as the target concept to learn. Each concept corresponds to one of these 40 unique nodes in the ANN output layer:

The five handwriting components from each PCR image are prepared for the ANN input layer as follows: The database is queried for the coordinates of each word (previously segmented by a human). The word images are passed to LFWR. All possible bi-substring patterns using the highest confidence character score combinations, are produced for each word image. Each pattern is in the format: (* C_1 * C_2 *) where C_1 and C_2 are individual characters and an asterisk (*) indicates other characters may be present in that slot. A 58 bit representation is produced for each bi-substring pattern. Three such patterns are constructed for the three adjacent words corresponding to a bi-substring phrase represented by a total of 174 bits. The 174 input layer bits are trained to the desired output layer node corresponding to the target concept. The ANN is trained with all available 174 bit bi-substring.

The performance summary from a run on 10 PCR images is as follows: Total PCR forms: 10; Total PCR Word Images: 761; Lexicon Size: 383; Training Set Size: 9,940,250; Average Lexicon Reduction: 96 %; and final recognition rate: 70%.

WORD SEGMENTATION

The motivation for our heuristics based word break detection lies in the human cues for word separation. These typically are [4] are: spatial separation between components, presence of punctuation marks, presence of uppercase character followed by string of lowercase ones, transition between numerals and alpha characters and actual recognition of words prior to word separation. Other than the actual recognition of words prior to separation, our method either explicitly or implicitly uses the human cues for word separation.

A writer's spacing style is observed to depend on the strokes located within the horizontal writing band (the computed, imaginary upper and lower half reference lines. By examining the dominant convex points along the upper contours of the text within the writing band, and extracting vertical lines from these dominant points, we can determine the initial global intra-word and inter-word spacing styles. The initial global spacings are then constantly updated with more localized spacing metrics. We refer to this process of updating the spacing values as *adaptive spacing.*

If N vertical lines are extracted from the image, let i be the index of a vertical line where $i=1,2,..N$. Also, let (x_i,h_i,C_i) be the horizontal location, the height and the connected component of the i^{th} vertical line. Also, let the interval of the i^{th} vertical line $d_i = x_i - x_{i-1}$.

If we assume there are P potential words in the text line image, and k is the index of the current word being processed, where $k=1,2…P$. Then the adaptive intra-word spacing D_k for the potential word at the k^{th} index can be calculated as:

$$D_k = \begin{cases} d' & \text{if } k = 1 \\ \frac{1}{n+1}\left(D_{k-1} + d_{local}\right) & \text{otherwise} \end{cases} \quad (1)$$

where n is the number of vertical lines in the k^{th} potential word, d' is the global average spacing, and d_{local} is the sum total of intervals in the k^{th} potential word. The inter-word spacing is also computed in a similar manner, being updated at every potential word break.

For the i^{th} vertical line, the interval-based score, s_{c_gap} is a weighted score assigned, based on the disparity between the interval d_i and the current adaptive intra-word spacing value D_{k-1}. Similarly, s_{w_gap} is the score based on the inter-word spacing. Based on the height transition between i^{th} and $(i-1)^{th}$ elements, a weighted score s_{height} is assigned. The next weighted score s_{cc} is assigned based on the whether the i^{th} and $(i-1)^{th}$ elements exist in the same connected component and if not, then on the proximity of the components.

The final total score is most influenced by s_{c_gap} and s_{cc}, which are weighted most heavily. The average stroke width of the image is used to determine the weight values. Total score for the i^{th} element is therefore given by:

$$s_{Total} = s_{c\_gap} + s_{w\_gap} + s_{cc} + s_{height} \quad (2)$$

Possible word break points are detected by thresholding, after computing the scores for all the vertical lines and their intervals. The threshold value is derived also using the average stroke width of the image text.

A total of 416 words were tested in 35 text line images and the algorithm successfully segmented 285 words correctly. Of the 416 words, 72 words were hyper-segmented while 57 words failed or were under-segmented.

REFERENCES

[1] [1] V. Govindaraju and R. Milewski, "Automated reading and mining of pre-hospital care reports," in *IEEE Symposium on Computer-Based Medical Systems*, pp. 152-157, 2001.

[2] [9] John T. Favata: Offline General Handwritten Word Recognition Using an Approximate BEAM Matching Algorithm. IEEE Trans. Pattern Anal. Mach. Intell. 23 (9): 1009-1021 (2001).

[3] [11] Gyeonghwan Kim, Venu Govindaraju: A Lexicon Driven Approach to Handwritten Word Recognition for Real-Time Applications. IEEE Trans. Pattern Anal. Mach. Intell. 19(4): 366-379 (1997).

[4] [[5] G. Kim and V. Govindaraju, "Handwritten phrase recognition as applied to street name images," *Pattern Recognition*, vol. 31, pp. 41-51, 1998.