Feature Extraction Methods for Form Recognition Applications

JAE ADAMS, E.A. YFANTIS, DEAN CURTIS, TRENTON PACK
Digital Image Processing Lab
Department of Computer Science
University of Nevada, Las Vegas
4505 Maryland Parkway, Las Vegas, NV 89154
UNITED STATES OF AMERICA
http://web.cs.unlv.edu/digitalimageprocessinglab

Abstract - The process being used to recognize a form includes extracting a wide variety of features that allow the recognizing procedure to sufficiently characterize a form. The first feature is based on the logo of the form. The logo is a distinct picture that appears on a form. The second feature is the form identification (to be called "id") which is unique to each form. The third feature is the structural lines. The fourth feature is a check box counting routine. The last feature is a word counting routine. Together, these features are utilized to extract vital characteristics that lead to a confident classification of the form. Current results within the form classification algorithm, which uses all the aforementioned features, yields a high accuracy.

Key-Words: Feature Extraction, Image Processing, Form Recognition

1 Introduction

A recognition algorithm has been developed that autonomously processes and recognizes paper medical records. The recognition algorithm relies on the extraction of basic features within a given form.

The medical forms provided for recognition were investigated with the goal of finding features that could be used in characterization. Upon examination, a list of features was compiled that could confidently, and with high statistical probability, be utilized as part of the recognition algorithm. These features are based on components of a form that either appear in all forms or appear as a characteristic in most forms (figure 3). The features to be discussed in this paper include the form logo, the form id, the structural lines, the number of checkboxes and the number of words.

2 Logo Detection

The logo of a form is a feature that has several consistent traits. Unlike many other features in the form, the form logo has a consistent shape that does not change from form to form. For each logo type, there are structural features that can be detected which will always be present in that form logo. Thus, the search for a logo involves a search for the trait that will identify a particular logo.

Since no two logo types are alike, the initial traits are different for each kind of logo. The methods of detection will be unique, depending on what trait is being examined. However, a trait that is common among all the logo detection methods is a segmentation search. The form is separated into connected components. Then, these components are searched for those components having a trait similar to the form logo.

Fig. 1 Logo with letters isolated.

For example, the logo shown in figure 1 contains a long black, vertical bar at the base of the logo. In scanning for this logo, the connected components are searched to find a single component whose width is more than ten times the height of the component. This search produces only a few components which might be the base of the logo.

Once the logo base bar has been found, its width and height are used to determine a region of the form where the words "Nevada Corporation" should be located. The presence of these words indicates whether the logo exists in this region. So, another connected component scan is made of the designated area, searching for components whose dimensions are of the correct size to be a single letter. If there are a significant amount of characters above the logo base bar, then the letters are combined with the bar to form the complete logo.

When searching for logos, it is often the case that disconnected components may be connected due to scanning noise or other image processing errors. In our example, the base line of the logo may be connected to some of the letters within the logo. Additionally, several of the letters may be connected to each other. In these cases, special tests must be made to determine whether the logo is indeed present.

If one or more of the letters is connected to the bar, then our initial scan for the bar detects a connected region that contains most of the logo. The interior of this region is searched for additional components whose dimensions are those of letters. We do not expect to find the full seventeen letters, as we know that some of the letters are connected, but if more than half of these characters are detected, they are combined to form the logo.

In the worst case, the logo pieces are connected almost into a single component. In this case, we must identify the logo by other statistical tests. These tests include the logo having the correct width and height ratios, a minimum size test, a smooth line on the bottom of the logo, a jagged upper edge of the horizontal histogram (indicating the division between letters), and a correct ratio between black and white pixels within the logo. If all of these conditions are met, the logo has been identified.

3 ID Detection

The form ID can also be a distinguishing feature in the form classification. The location and type of the form ID can be used to determine the type of form, even if the exact form ID number is not known. Thus, it is important to determine the location of the form ID. The ID detector isolates four regions of the form where the ID may be located: the upper left corner, lower left corner, bottom center, and lower right corner. The detector then searches each region for the ID.

In order to find the ID, the region of the form is first segmented into connected components. These components are sorted to isolate any components that resemble text characters. This is done by making sure that each component is within a maximum height, minimum height, and minimum width. It is important to note that a text character is not checked for a maximum width. This is due to the fact that in examining text, even type-written characters that are directly next to each other may often be combined by connectivity to form single connected components. This is especially the case in form ID's, where the font size is typically smaller than the rest of the text. The characters are not checked for a maximum width to allow groups of connected text characters to be appropriately examined.

Once the text characters are identified, they are grouped into text lines. The ID is detected by looking for areas of grouped text that resemble an ID.

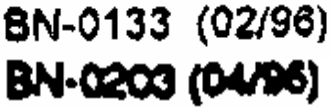


Fig 2. Two sample clips of form identification numbers. The top is not connected and the bottom one is partially connected.

The text lines are examined to find sections (or "islands") that have the dimensions of a form ID. Once these possible ID sections are found, they are examined to verify whether they contain an actual ID. This is done by recognizing whether the possible ID text begins with a valid ID prefix or not. If it has a matching prefix, then the detector has found the general ID location.

It is important to note that the detector does not recognize the ID. Meaning, if the form ID is BN-0087, for example, the detector will not be able to determine this unique ID. In order for a method to be created that uniquely identifies the ID, it must be able to separate connected characters from each other without losing their meaning.

Despite this limitation of actual recognition, the ID detector will be able to determine whether the ID exists, where it is located within the form, and what the ID prefix is. This information can then be utilized in determining the type of the form.

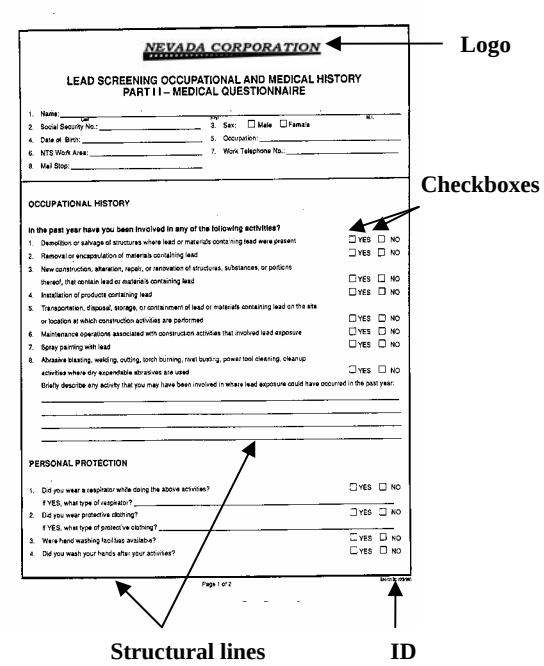


Fig. 3 Form features

4 Structural Lines

The structural lines of a form include the underlines, form boundary lines, and frame lines. Structural lines provide an additional feature that can be used in recognizing forms. Two applications were designed in terms of the structural lines: the counting of vertical and horizontal segments within designated portions of the form, and the removal of all structural lines.

In the counting of the structural lines, the form is divided into regions to establish a more concrete characteristic. In the counting process there are two steps. The first step is to determine the location of the horizontal and vertical line segments. The horizontal line finder is a top-down searching method that starts with one pixel and traces the pixels horizontally for as long as the group of pixels remains within the bounds of a horizontal line. If it is determined that the group of pixels form a line, they are added to the list. The same methodology is repeated with vertical lines except the scan is left to right. Both the horizontal and vertical line finders (which are separate functions) only create a list of pixels, and all pixels within the list are part of a line in the form. So, the second step is to group the pixels with their respective lines and count the number of lines. This is done by arranging the pixels into horizontal and vertical components which are mostly connected, and acknowledging these components as line segments.

In the removal of the structural lines, the same method to determine the horizontal and vertical lines is used, except that in the removal steps, the image used in the horizontal and vertical processing is thinned. This means that every component in the image is made to be one pixel thick. So, the pixels returned will represent the tracing through the center of all structural lines. Then, for each individual horizontal and vertical line, the average thickness is determined, and the line is erased by removing all pixels within the proximity of the average thickness of the line.

5 Checkbox Detection and Removal

The number of check boxes in a form, and the organization of the check boxes is another feature that is very characteristic to each form. The check boxes in each form contain distinct characteristics that make identifying the check boxes possible.

Before the search for a check box begins, the initial image is thinned so that all components within the image are one pixel in thickness. After thinning, an erosion algorithm is used to remove any pixels that do not form an enclosed loop. The final preprocessing step is to dilate the image slightly to improve the check box detector's accuracy. The detector can then begin scanning from the top of the image to the bottom.

The scan starts by searching for black pixels. A cursory trace is performed to see if the enclosed set of pixels resembles a check box. If, at any point, the tracing from the original pixel goes outside the rules of what a check box has been defined to be, the search is immediately abandoned, and the next pixel is inspected.

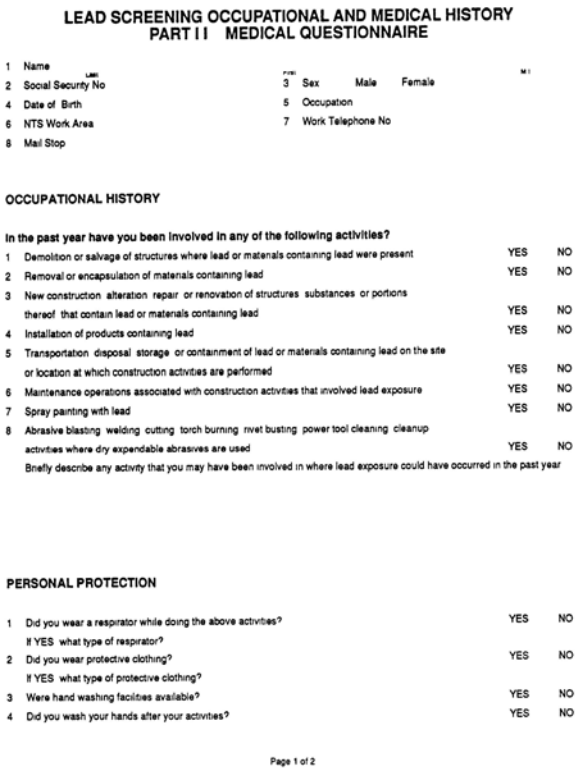
After an enclosing loop has been determined to resemble a check box, the area is compared against a more stringent set of rules. The detection rules look for the a best case of each area scanned. The following are attributes of a check box:

1) The box must be relatively square.
2) The number of steps taken to trace the enclosed pixel must closely match to a 1:1 ratio of the perimeter of the box enclosing the traced pixels.

3) The sides of the traced pixels must be relatively smooth, and not change slope too often.
4) The box must be a minimum size.

Once a check box is found to meet these specifications, it is stored in a list of check boxes. Each enclosed set of pixels will be scanned multiple times. This continues until the end of the form is reached. Next, the list of check boxes is scanned, and any duplicate rectangles are removed. Finally, a scan is performed to remove any boxes of anomalous size compared to the average.

Once the check boxes have been counted, then the check boxes are removed by removing all pixels associated with the rectangle regions stored in the list of check box rectangles.



LEAD SCREENING OCCUPATIONAL AND MEDICAL HISTORY
PART I I MEDICAL QUESTIONNAIRE

1 Name
2 Social Security No
3 Sex Male Female
4 Date of Birth
5 Occupation
6 NTS Work Area
7 Work Telephone No
8 Mail Stop

OCCUPATIONAL HISTORY

In the past year have you been involved in any of the following activities?

1 Demolition or salvage of structures where lead or materials containing lead were present YES NO
2 Removal or encapsulation of materials containing lead YES NO
3 New construction alteration repair or renovation of structures substances or portions thereof that contain lead or materials containing lead YES NO
4 Installation of products containing lead YES NO
5 Transportation disposal storage or containment of lead or materials containing lead on the site or location at which construction activities are performed YES NO
6 Maintenance operations associated with construction activities that involved lead exposure YES NO
7 Spray painting with lead YES NO
8 Abrasive blasting welding cutting torch burning rivet busting power tool cleaning cleanup activities where dry expendable abrasives are used YES NO

Briefly describe any activity that you may have been involved in where lead exposure could have occurred in the past year

PERSONAL PROTECTION

1 Did you wear a respirator while doing the above activities? YES NO
If YES what type of respirator?
2 Did you wear protective clothing? YES NO
If YES what type of protective clothing?
3 Were hand washing facilities available? YES NO
4 Did you wash your hands after your activities? YES NO

Page 1 of 2

Fig. 4 The image after the form logo, form id, checkboxes and structural lines has been removed. This done to facilitate the word count algorithm.

6 Word Count – Group Isolation

After all features have been removed, the only objects contained in the image are words, and the process of word counting can begin. The first step is isolating lines of words that are relevant. A relevant line of words consists of the following criteria:

- The line consists of letters of the same font and size
- The letters are linearly aligned so that all letters are on the same horizontal line

It is important to note that the relevancy achieved is, for the most part, based on the assumption of the inherent configuration of letters and the words they compose. It is recognized that there will be islands that contain two sets of letters of different fonts and size, but this only affects the nature of each island in a minor and negligible way.

To describe the isolation algorithm, there are a few terms that must be defined. The first is the idea of a connected component. A connected component is a group of pixels from an image that form one object through connectivity. So, the first step is to apply a component scan by applying a segmenting algorithm on the image of figure 4. The algorithm used for this routine can be found in Reference [1]. The result is a list of individual connected components. Ideally, each of these components would be single letters, but scanning problems cause letters to sometimes be joined. However, the joined characters do not cause a problem at this stage of the process.

The next step is to sort the letter components based on the left most coordinate of each component. Then, characters are grouped based on two conditions:

- Linearly centered proximity to account for vertical orientation
- Left and Right threshold for spans of empty space between segments

Each character that passes this test is merged into the corresponding section to which it belongs, and these groups ultimately become the islands. After the completion of this algorithm, the result is a list of islands which will then be used in the word counting.

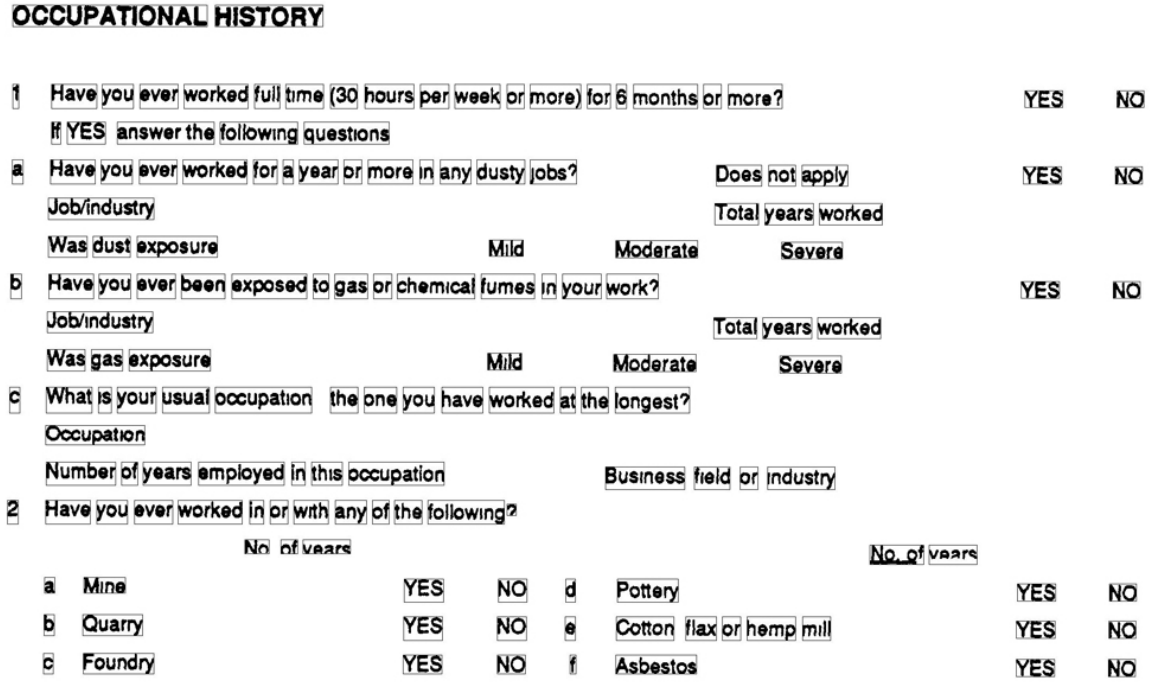


Fig. 5 Result of word count. Words are isolated in black boxes. Sample taken from a portion of a form.

7 Word Count – Algorithm and Implementation

For every form, a hand count was performed to get the precise number of words. This is the number used as part of the recognition process to compare with the result of this word count. In order to develop a system that confidently and successfully counts words, there must be a precise definition of what will be counted as a word. In the hand counting of words, a word is defined with these guidelines:

- A word is a group of letters preceded and followed by a typewritten space.
- If two words are separated by a '/', then they are counted as two words
- If two letters or two words or any combination is separated by a '-', then they are counted as two words
- Individual letters with the proper spacing before and after are counted as words
- Lettering for lists is counted as words.

To count the words, the first step is, given a group of linear aligned letters ("islands"), to scan each line of vertical pixels from left to right for the entire island. Each vertical scan line is classified as one of two possibilities: the line has no black pixels, or has at least one black pixel. The object of this organization is to find and store all the gaps between both letters and words. A gap consists of a series of consecutive vertical scan lines that have no black pixels. So, to isolate the gaps, when a vertical scan line is found that does not have a black pixel, that location is saved and each consecutive empty vertical scan line after is counted until the next pixel is found.

Once the size of each gap has been computed, then the decision of whether or not that gap is a space between two words or a space between two letters is determined. This is done through a statistical analysis of the nature of the organization of typewritten words. For each island of words, the maximum height of a character within that line is determined. Then, from that number, a threshold is created that separates the gaps that are between letters and the gaps that are between words (Fig. 6).

For the forms being studied, it was found that when the maximum height of a letter within a line was less than 28 pixels, the boundary for which the two different types of spaces are separated is 18% of the maximum height (rounded down). So, if the

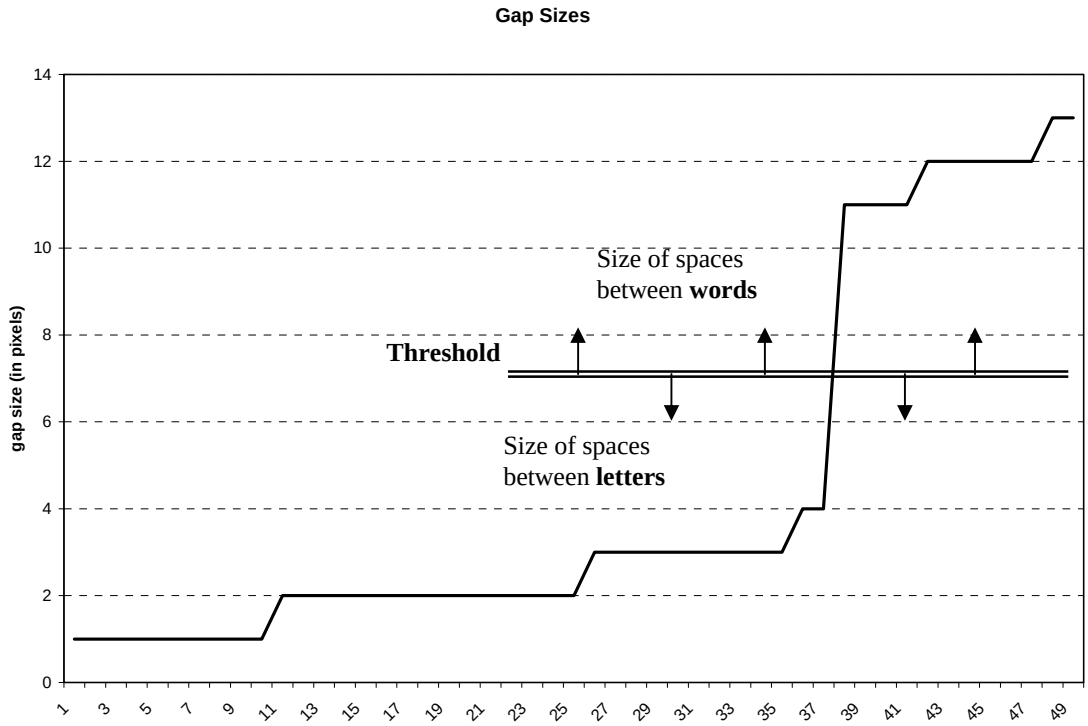


Fig. 6 Statistical Gap Analysis Gaps bigger than 8 pixels for this line, are spaces.

maximum height of a letter in the line is 9, then the largest gap between letters is no greater than 1 pixel, and if the maximum height of the letter is 15, the largest a gap can be between two letters is 2 pixels. The next range of heights is for 28 pixels or more, of which the boundary is 25% of the maximum height.

So, once a gap is found in the left to right movement within the island, if the size of the gap is greater than the boundary, then that gap is classified as a space between words, and if the gap size is smaller, then the scan simply moves forward. Then, based on the assumption that words occur in between spaces (or gaps greater than the boundary), the process of isolating words is merely creating a box between the end of one space and the start of the next. Within that range, a single word is contained.

8 Conclusion

To aid in the recognition capabilities of medical forms, it was necessary to develop methods that could extract individual features of a given form. The features that have been found to be both characteristic and consistent from record to record are the form identification, the form logo, the number of checkboxes, the number of structural lines, and the number of words.

Methods were developed that could extract these features individually, and these methods facilitate the algorithms used in actual form recognition. Feature extraction methods have proven to be highly accurate and quantitative.

Acknowledgements

"Medical Records" (E-records) is research being conducted with the collaboration with Quest Technologies Inc. and the support of the Department of Energy. Special thanks extended to Dr. Stephen Rice, Vice President of Research at UNLV.

References:

[1] Bunke, H. and Wang, P.S.P. "Handbook of Character Recognition and Document Image Analysis" World Scientific (1997)

[2] Mori, S., Nishida, H., Yamada, H. "Optical Character Recognition" Wiley-Interscience (1999)