Automated Generation of Search-able PDF Format of Chemical Equations from Document Images

Prerana Jana, Anubhab Majumdar, Sekhar Mandal
Department of Computer Science and Technology
Indian Institute of Engineer Science and Technology
Shibpur, India
Email: (prerana.jana, anubhabmajumdar93)@gmail.com
sekhar@cs.iiests.ac.in

Bhabatosh Chanda
Electronics and Communication Sciences Unit
Indian Statistical Institute
Kolkata, India
Email: chanda@isical.ac.in

***Abstract*—PDF format of scanned document images is not searchable. OCR tries to remedy this adversity by converting images or PDF files into editable and searchable data, but it has its own limitations in presence of equations - both mathematical and chemical. OCR system for mathematical document images is already a major research area and has provided successful result. However, chemical equation segmentation has been a less ventured road. In this paper we present a novel method for automated generation of searchable PDF format of segmented chemical equations from scanned document images by performing chemical symbol recognition and auto-correction of chemical reactants. We use existing OCR systems, pattern recognition, contextual data analysis and a standard LaTeX package to generate the chemical equation in searchable PDF format. The effectiveness of the proposed method is demonstrated by testing it on 240 document images.**

***Keywords*—Chemical equations, mathematical symbols, morphological operation.**

I. Introduction

We use search engines like Google to find location of documents in WWW. Usually, text-based keywords are used for retrieving of documents. A large number of documents are being digitized today for the purpose of archival analysis, transmission and browsing. The existing OCR systems show high accuracy in interpreting text portions, but fail to properly process other components like graphics, half-tones, chemical and mathematical equations.

A few studies [4], [5], [6] are directed toward math-symbol or math equation recognition assuming that the math-zones are already marked. We, on the other hand, contend that a better approach is to segment the chemical equations from the mixed material thereby helping the future OCR activity to focus its processing only on specific content. In this paper we propose fully automated segmentation and detection technique of chemical equations present in heterogeneous document images followed by generation of LaTeX file of the segmented equations.

II. Proposed Method

The proposed method consists of four distinct steps and they are as follows: (i) Segmentation of displayed equations; (ii) Identification of chemical equations; (iii) Auto correction of reactants and the chemical equation itself; and (iv) Generation of chemical equation in search-able PDF format.

A. Segmentation of displayed equations

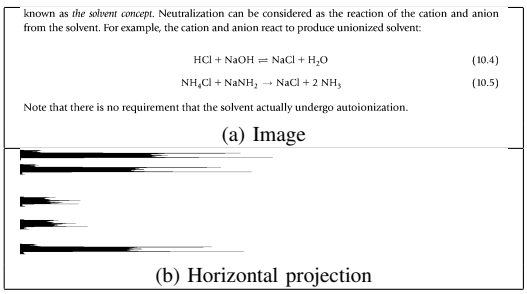
A skew free heterogeneous binary image is the input of the proposed algorithm. The tables and graphics are extracted out from the heterogeneous documents using the methods described in [3] and [2]. The reset of the document contains normal text lines and displayed equations if any. Our main motivation was to classify chemical and non-chemical equations, we restrict our study to displayed equations only.

Major steps of segmentation of displayed equations (DE) are given below.

- Text line segmentation
- Blob formation using morphological tools
- DE zone extraction

The details of the above steps are given in the following subsections.

1) Text line segmentation: As the document image is skew free, the horizontal projection profile of the document page is taken to segment the text line. A part of a document page and its horizontal projection profile are shown in Fig. 1.



(a) Image

(b) Horizontal projection

Fig. 1. A document page and its horizontal projection profile.

2) Creation of word blobs: This is done by coalescing the characters in a word using morphological closing operation. Such character coalescing process depends on the accuracy in detecting the normal character gap and the gap between the consecutive connected components in that text line. The component analysis is done first. Let C_a and C_b be the two consecutive connected components in a text line. L_b is the left most x coordinate of C_b and R_a is the right most x coordinate of C_a. A distance function D is defined as follows:

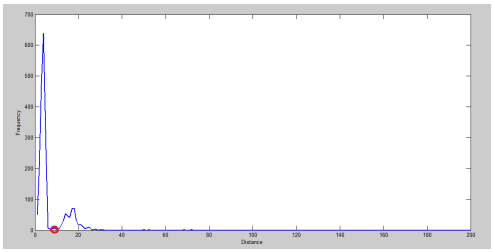


Fig. 2. Distance histogram of the image shown in Fig. 1 (a)

$D = L_b - R_a$. The histogram of D is obtained (see Fig. 2). The blob formation requires information on inter-word gap. The distance histogram is a multi-modal histogram. The first peak is corresponding to the character gaps. Our intention is to find out character gaps in running texts of a document page so that we could combine the consecutive characters into a single blob. Hence, we consider the upper boundary (l) of the first hump as the length of structuring element. Morphological close operation with a structuring element of size ($l \times 1$) will form the blobs. The result of blob formation is shown in Fig. 3.

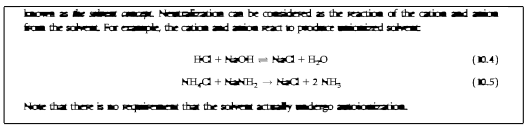


Fig. 3. Blob formation of the image shown in Fig. 1 (a)

3) DE zone extraction: We have considered the set of operators that is commonly used both in chemical equations as well as mathematical equations to fulfil our aim to classify displayed zones containing chemical and non-chemical equations.

After blob formation, small component like dots of *i and j* are eliminated on the basis area. The region, corresponding to each blob is considered from the original image and the number of connected component(s) present in that region is counted. If the number of components is more than one that blob is not an operator and is removed from the blob image (see Fig. 4). The remaining components in the blob image are

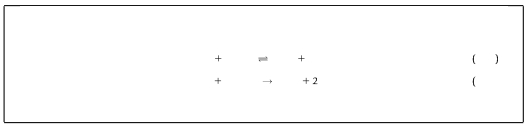


Fig. 4. Single components extracted from the blob image shown in Fig. 3

operators along with some alphanumerics like a, A, (, etc. The logical AND operation is performed between the blob image and the original image. The Euler number of the operators that we have considered is 1 (one) and based on this feature some of the alphanumerics are discarded. This image is denoted by I_s. The components present in I_s are divided into two classes; (i) one is operator class, and other is non-operator class. We have considered the following operators (+,-,$\leftharpoondown$,$\rightharpoonup$,$\rightarrow$,$\leftrightarrow$) which are normally present in chemical equations. *One class* Support Vector Machine (SVM) classifier is trained to identify the operators from I_s. The features that we have used are as follows.

- Aspect ratio (f_a) of each component
- Density
$$f_d = \frac{\#pixels_o}{\#pixels_b},$$
where $\#pixels_o$ denotes the number of object pixels and $\#pixels_b$ denotes area of the bounding box.
- Each component is resized and horizontal and vertical projection profiles are obtained.
(i) For each profile the second (f_{m2}) and third moments (f_{m3}) are calculated.
(ii) Again for each profile the location (f_l) and the magnitude (f_m) of the global maximum are determined.
- Perimeter (f_p) of each component is also obtained.

The accuracy of the classifier is 98.4%.

To detect '=' or '$\rightleftharpoons$' one extra step is required. The operators having $f_a \leq 0.6$ are considered thin symbols (-,$\leftharpoondown$,$\rightharpoonup$,$\rightarrow$,$\leftrightarrow$). For each symbol denoting thin operator, a rectangular mask is placed below the symbol to check if there is another one within the mask. If the two thin operators are present within the mask, they are considered to form either an '=' or '$\rightleftharpoons$' sign. Let the length of the thin operator be l. The area of the mask is ($l \times l/2$).

The horizontal line separating the numerator and the denominator is identified as its length is greater than the median length of the operators. Two windows are placed above and below the separating line to merge all the components within the windows with the separating line to form a single logical line. Otherwise, they would be treated as three consecutive text lines and we will not be able to associate the intermediate math-symbols (+, -, =) to a single expression. The area of the window is (length of the separating line)$\times$ (twice the median width of the text lines).

Initially, all the text lines consisting at least one operator are considered candidate displayed equations (CDE). The operators are eliminated from CDE. The upper boundary (u_v) of the second hump of distance the histogram is obtained which represents the word gaps in the text line. For each CDE zone Run Length Smoothing Algorithm in horizontal direction (H-RLSA) is carried out. If the distance between two neighbouring components is less than u_v, it means they belong to a same word and are merged by H-RLSA. H-RLSA has a similar effect as of dilation of black areas in horizontal direction. The characters in a word are dilated and coalesced to the other characters of the same word. The output of H-RLSA is shown in Fig. **??**.

Equation numbers are common in the displayed equation zones. These numbers have to be removed because for each CDE we have counted the number of operators and corresponding other components in the output of H-RLSA. If the number of components $\leq 2\times$ number of operators, then the CDE is considered displayed equation; otherwise some embedded formulae/equations may exist in the line. To eliminate the equation number from the output of H-RLSA the operators are moved to the output of H-RLSA and the component analysis is done. From both ends distance (d) (see Fig. **??**) between the first two consecutive components

is measured and if $d > 5\times\ u_v$, then the first component from the end is considered the equation number and is removed. The accuracy of the DE zone segmentation is 97.4%.

B. Identification of chemical equations

Each segment of a displayed equation is divided into three zones; namely upper zone, middle zone and lower zone (see Fig 5). To identify the three zones of a DE zone, uppermost and lowermost co-ordinates of each connected component below the same DE zone are also obtained. The median of uppermost coordinate, and median of lowermost co-ordinate of such components in DE zone are computed. A horizontal line, called the baseline, is drawn through the median of lowermost coordinates of components and this baseline separates the middle zone and lower zone of DE zone. Similarly, the median of uppermost co-ordinate of the components in the DE zone generates a horizontal line. This horizontal line, called top line, separates the middle and upper zones of the DE zone.

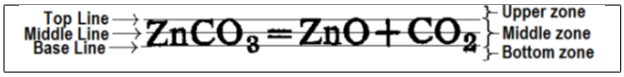


Fig. 5. Different zones of DE equation

The subscripts in a DE zone belong to lower-half of the middle zone and lower zone whereas the superscripts belong to upper zone and upper-half of the middle zone. Based on the location of the components in a DE zone we have detected the subscripts and superscripts and are separated from the DE zone. The operators are also separated from DE zone.

Now, each displayed equation is an input to an OCR of MATLAB R2014a. The OCR returns each DE zone as a text string. We made a dictionary out of all the elements in the periodic table. An important observation is that an element always starts with a capital letter. Using this property, an element can be expressed by a regular expression [A-Z][a-z]*. It means an element's symbol starts with an upper case letter and may or may not have lower case letters. We have designed a parser to extract the sub-string matching the regular expression mentioned above with the following grammar:

$$\begin{gathered} start \rightarrow capital.follow \\ follow \rightarrow small.follow| \in \\ capital \rightarrow A|B| \ldots |X|Y|Z \\ small \rightarrow a|b| \ldots |x|y|z \end{gathered}$$

Each of the substring returned by the parser is matched against the aforesaid dictionary and if it is a positive match then that substring is considered a symbol of the chemical element. Let us consider, the number of substrings extracted from the OCR output by the parser is *n* and the number of positive matches the aforementioned dictionary is *m*. If m:n ratio is more than a threshold value β then this DE is considered a Chemical Equation. This threshold (β) is set to 0.7 by running our algorithm on our dataset containing 985 displayed equations. The reason for the ratio not being 1 are (i) Limitations of OCR; (ii) Touching and broken characters.

III. Conclusions

This paper presents a simple, but efficient binarization technique that can handle different types of degradation such as faint characters, bleeding-through, large background ink stains and the noise introduced in the scanning process (block noise) which are normally present at the border of images. We contend that the proposed method may be used as 'de facto' standard due to its simplicity and efficiency for a variety of degraded images including degraded handwritten historical documents.

References

[1] N. Otsu, *A threshold selection method from gray-level histograms*, Systems, Man and Cybernetics, IEEE Transactions, vol.9, no.1, pp. 62–66, 1979.

[2] S. P. Chowdhury, S. Mandal, A. K. Das, and B. Chanda, *Segmentation of Text and Graphics from Document Images*, In Proc. of ICDAR, pp. 619623,2007.

[3] S. Mandal, S. P. Chowdhury, A. K. Das, and B. Chanda, *A simple and effective table detection system from Document Images*, IJDAR, Vol. 8(2), 172182, 2006.

[4] D. Blostein and A. Grabavec, *Recognition of Mathematical Notation*, Handbook of Character Recognition and document Image Analysis, 577–582, 1997.

[5] K-F. Chan and D-Y. Yeung, *Mathematical Expression Recognition: A Survey*, IJDAR, Vol. 3, no: 1, 315, 2000.

[6] U. Garain and B. B. Chaudhuri *On OCR of Printed mathematical Expressions, Digital Document Processing*, Ed: B. B. Chaudhuri, Advances in pattern Recognition, 235259 , 2007.