Automated Segmentation and Classification of Chemical and other Equations from Document Images

Prerana Jana, Anubhab Majumdar, Ashish Kumar Layek, Sekhar Mandal, Amit Kumar Das
Computer Science & Technology Department
IIEST Shibpur, India
Emails: (prerana.jana,anubhabmajumdar93)@gmail.com
(ashish,sekhar,amit)@cs.iiests.ac.in

***Abstract*—Segmentation of mathematical equations from document images is already a major research area for improved performance of OCR systems. Though chemical equations are also sharing similar spatial properties as that of non-chemical equations (for example, mathematical equations), efforts to segment those are still to be explored. This paper presents a novel method for segmenting and identifying chemical and any other equations in heterogeneous document images that may contain graphics, tables, text and the classifying them into two categories; chemical and non-chemical equations. This study, a first of its kind, as far our knowledge goes, not only improves the OCR performance, but also leads to creation of chemical database and formation of bond electron matrix from chemical equations or formulae. In our proposed method we extracted the equations using morphological operators and histogram analysis and the extracted equations are classified using an open source OCR engine. The effectiveness of the proposed method is demonstrated by testing it on 152 document images. Test results show an accuracy of 97.4% and 97.45% for segmentation and classification, respectively.**

***Index Terms*—Mathematical symbols, morphological operation, histogram analysis.**

I. Introduction

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, halftones, chemical and mathematical equations. Also, the existing search engines take the text-based keywords for retrieving documents overlooking chemical/mathematical equations from scientific documents.

A few studies [1], [2], [3] are directed toward math-symbol or math equation recognition assuming that the math-zones are already marked. Though, symbol recognition is a part of OCR activity, but when it is applied to the non-segmented mixed material (text with math-zone and others) computation is expensive and success far from satisfactory. We, on the other hand, contend that a better approach is to segment the mathematical/chemical equations from the mixed material thereby helping the future OCR activity to focus it's processing only on specific content. In this paper we propose fully automated segmentation technique for extracting mathematical/chemical equations exploiting spatial distribution of black pixels on a white background and subsequently classifying them using an OCR.

A number of work have been done over the past decade to detect and extract the mathematical equations present in heterogeneous document images. Fateman et al. [4] proposed a scheme which utilised character size and font information etc. to identify all connected components. Two bags, namely *text* and *math* are defined. The *text* bag is used to keep all letters and italic numbers; whereas the *math* bag collects punctuation, special symbols, Roman digits, italic letters, lines and dots. The *math* bag objects are then grouped together according to their spatial proximity. Grouping of items in text bag is done next followed by review and correction to move isolated items to their proper destinations. Math segmentation is done in [5] through physical and logical segmentation using spatial characteristics of the math zone as well as identifying some math-symbols. The document is segmented to characters, words, lines and blocks by physical segmentation. The logical segmentation process that follows consists of two steps; first the displayed math is detected by identifying their usual centered position and in the next step in-line maths is detected by identifying special symbols.

Kacem et al. [6] extracted the equations using fuzzy logic by detecting mathematical operators. Their method was tested on a dataset consisting of 300 expressions and the success rate is about 93%. Some of the operators like '+', '-', '(', and ')' do appear in chemical equations. This leads to the misclassification chemical equation as mathematical equations reducing the success rate. A similar method has been proposed in [7] to segment the mathematical expression in printed documents. The statistical approach taken by Garain [9] on the corpus of 400 pages differentiates normal text lines and lines containing equations/expressions on the basis of their white spacings which are usually larger in math-equation than the normal text. However, the chemical equations in the documents bear the same property. Jin et al.[11] proposed a similar method to extract displayed formulas using Parzen classifier. Drake and Baird [12] came up with a graphical approach; similarly Guo et al.[13] developed a Gaussian mixture model to describe spatial relationships between sub-

978-1-4799-7458-0/15/$31.00 ©2015 IEEE

components of a math expression. Another method to check text style (regular, italic, bold) at the character level has been proposed in [10]. Garain [8] proposed a method to segment the displayed and embedded mathematical formulas from the documents using a bunch of features. The method is tested on a data-set of 200 images containing 1163 embedded and 1039 displayed expressions and the success rate is 88.3% and 97.2% respectively for embedded and displayed expressions. A method proposed by Chu and Liu [14] used features based on centroid fluctuation information on non-homogeneous regions to detect displayed and embedded formulas.

In a nutshell, in all the above methods emphasis is given only on mathematical equation and in the eventual segmentation/classification chemical equations would automatically be included as a part of mathematical (or other) equations thereby reducing the success rate of the segmentation and effectiveness of the subsequent classification; if any.

Considering the possible applications of the segmentation of chemical equations we see that the recent development in the field of chemo-informatics requires precise identification of chemical equations amongst a myriad collection of chemical and non-chemical formulae/equations. This can be important for various tasks like creation of chemical database as well as to obtain bond-electron matrix from a given chemical equation, etc. The proposed work embodied in this paper is motivated by the aforementioned needs.

The paper is organized as follows. Proposed method of segmentation of displayed-equations and their classification is presented in section II. Section III presents experimental results. We conclude the paper in section IV.

II. Proposed Work

The work starts with heterogeneous binary images that may contain text and graphics. The graphics, tables and headings are extracted out from the heterogeneous documents keeping the math-zone/chemical-zone, if any, along with the segmented and skew corrected text following popular and robust methods described in [17] and [15]. We did not consider in-line expressions because, in books, chemical equations are rarely present in that form. Since, our main motivation was to classify chemical and non-chemical equations, we restrict our study to displayed-math equations only.

Most of the elements in the displayed equations (DE) have little difference from the normal text. Naturally, the segmentation depends on a couple of rules formed by observing the general spatial appearance of displayed-equations in common technical documents including journals. This is carried out by sampling 152 scanned pages containing mathematical/chemical equations in different possible forms. The following is the general spatial characteristics of the DE zones.

- Subscripts and superscripts are frequently present in DE zones.
- Math expressions are often written using 2-3 consecutive text lines to accommodate subscripts and superscripts leading to vertical overlap of characters; a phenomenon absent in the normal text lines.
  The characters and symbols are less dense in the mathematical expressions in comparison to the normal text lines.
- Presence of different operators in DE zones.
- Presence of a horizontal line separating the numerator and denominator portions is common.
- The DE zones are generally aligned of which central alignment is most frequent.

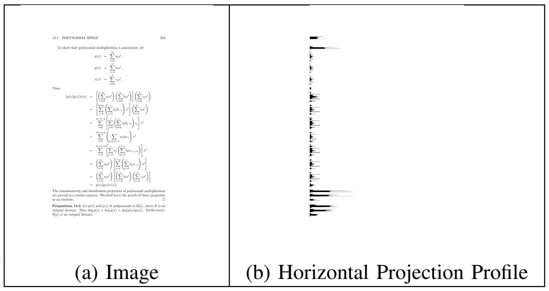
Segmentation of the DE zone starts with the removal of very small components, like dots of 'i', 'j' and small punctuation marks like comma, period, etc. by using component labelling and a suitable area threshold.

Major steps that follow for DE segmentation and classification are given below.

1) Text line segmentation
2) Blob formation using morphological tools
3) Operator identification
4) DE zone segmentation
5) Classification of DE zones into chemical and non-chemical zones.

The details of the above steps are as follows:

1. *Text line Segmentation*
   To detect DE zone, text lines have to be segmented first from which the operators are identified to determine whether a text line is a displayed equation or not. We have taken the horizontal projection profile of the document page to segment the text line. A document page and its horizontal projection profile are shown in Fig. 1



(a) Image

(b) Horizontal Projection Profile

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

2. *Creation of word blobs in Text lines*
   This is done by merging the characters in a word. 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 mathematical formulation for blob formation is as follows. Consider a binary image, I, which consists of connected components $C_k (k = 1, 2, \ldots, P)$. Let $L(C_k)$, $R(C_k)$, $T(C_k)$ and $B(C_k)$ be the four indices of the k_{th} compoent representing the leftmost, rightmost column indices as well as topmost and bottommost row indices, respectively.
   Let F be a function which ensures that the two connected components lie in the same text line. Then F may be represented as

$$F(C_a, C_b) = \begin{cases} 1 & \text{if} \quad (T(C_a) \leq B(C_b) \wedge B(C_a) \geq T(C_b)) \\ 0 & \text{otherwise} \end{cases}$$

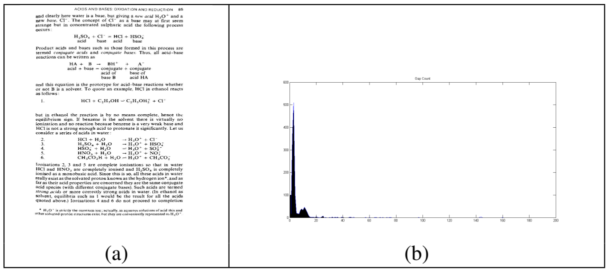


Fig. 2. Example of a page and histogram. (a) Document page; (b) Histogram of the preprocessed page.

Cluster formation requires information on inter-word gap. This may be obtained from the histogram H_1 of the distance D between two consecutive connected components C_a and C_b as follows.

The distance function (D) is defined for computing the horizontal distance between any two consecutive connected components, as below

$$D(C_a, C_b) = L(C_b) - R(C_a)$$

where $b = \min_x \{L(C_x) - R(C_a)\}$ such that $(F(C_a, C_x) = 1$ AND $L(C_x) > R(C_a))$. The histogram H_1 registers the gap between two consecutive characters. It may be noted that there may be more than two distinct humps in H_1, first one represents the character gaps in words and the second one represents the normal word gaps in text lines. An example page and corresponding histogram is shown in Fig. 2(a) and (b).

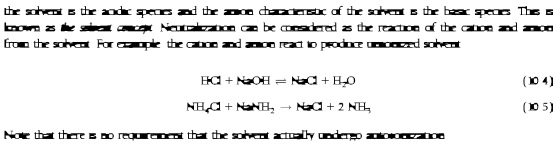


Fig. 3. Example of blob formation. Clusters formed from a portion of the image shown in Fig. 2(a).

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 (υ) of the first hump as the length of structuring element. Morphological close operation with a structuring element of area $(\upsilon \times 1)$ will form the blobs denoted as V_w (where $w = 1, 2, \ldots, Q$). The blob formation will be dictated by the following two conditions:

1) if there are two connected components C_m and C_n $(1 \leq m, n \leq P)$ obeying the relations
   - $F(C_m, C_n) = 1$
   - $D(C_m, C_n) \leq \upsilon$

   then C_m and C_n should belong to the same blob.
2) $V_a \cap V_b = \emptyset \ \forall a, b \mid (1 \leq (a, b) \leq Q \ AND \ a \neq b)$

3. Operator identification

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.

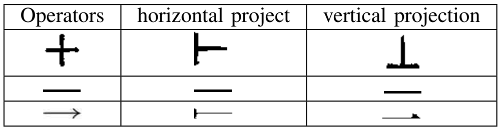

Operators
horizontal project
vertical projection

Fig. 4. Operators and their horizontal and vertical projection profiles

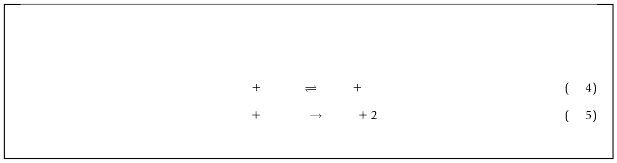


Fig. 5. Single components extracted from the word blobs shown in Fig. 3

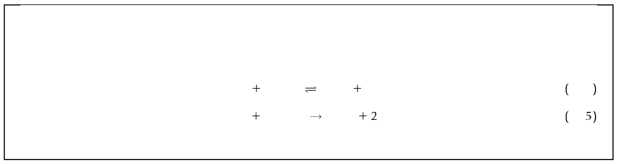


Fig. 6. After removal of alphanumerals based on Euler number from the image shown in Fig. 5

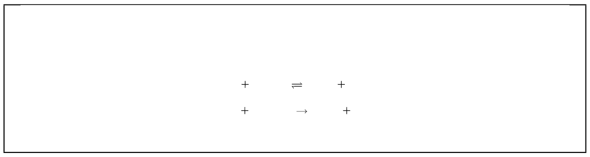


Fig. 7. Operators extracted from the image shown in Fig. 6

After blob formation, each blob is labeled. 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. 5). The remaining components in the blob image are 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 (see Fig. 6). This image is denoted by the name $I_{singleChar}$. We then extracted the following features:

- 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 (see Fig. 4) 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.

Now, $[f_a, f_d, f_{m2}^h, f_{m2}^v, f_{m3}^h, f_{m3}^v, f_l^h, f_l^v, f_m^h, f_m^v, f_p]$ serve as the feature descriptor for classification of operators from $I_{singleChar}$. A *one-class* SVM is used to classify single components into two classes; operators (+,-,$\leftharpoonup$,$\rightharpoonup$,$\rightarrow$,$\leftrightarrow$) and non-operators (all remaining single characters). In case of *multi-class* SVM total number of classes to be considered would be more than 62 (including alphanumeric, operators, etc.). Hence, *one-class* SVM is more suitable according to our requirements. 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 (-,$\leftharpoonup$,$\rightharpoonup$,$\rightarrow$,$\leftrightarrow$). 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) × (twice the median width of the text lines).

4. Segmentation of DE zones
   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 the histogram H_1 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. 8.
   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. 9) 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.

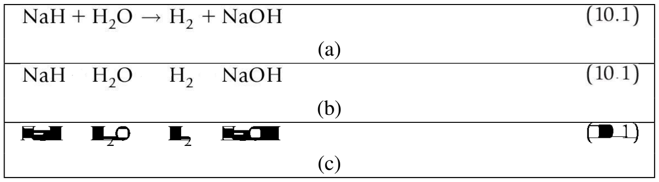


Fig. 8. The output of H-RLSA on portion of an image (a) a part of an image; (b) same part without operators; (c) result of H-RLSA operation

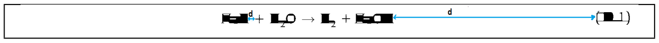


Fig. 9. Example of a equation number present in DE zone

5. Classification of segmented DE zones
   Each segment of a displayed equation is divided into three zones; namely upper zone, middle zone and lower zone (see Fig 10). 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.

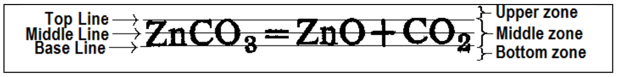


Fig. 10. Example of a equation number present in DE zone

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 removed from the DE zone. The operators are also removed from DE zone.

Now, each displayed equation is an input of an OCR engine Google Tesseract 3.02. 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:

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

For better comprehension, the working mechanism of the parser is depicted in Fig. 11. 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 733 displayed equations. The reasons, for β not being one, are

1) limitation of the OCR we used;
2) presence of broken and touching characters in the DE zone.

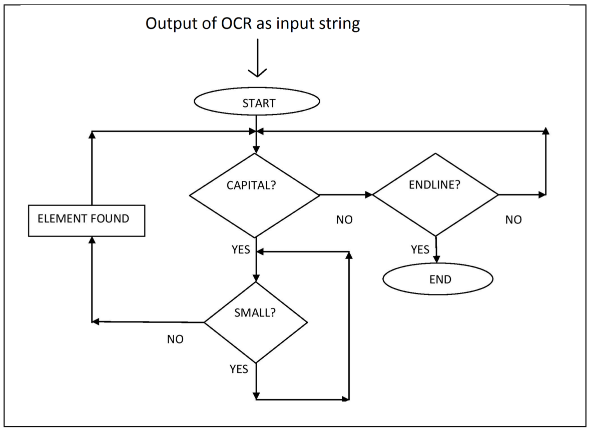


Fig. 11. Flow chart of the working mechanism of the parser

III. Experimental Result

We have implemented our algorithm in MATLAB 8.3.0.532 (R2014 a) in a PC (Intel core 2 Duo T6500 2.1 GHz CPU running Windows 8). The proposed method has been tested on a dataset consisting 152 document pages. Out of 152 pages 50 pages are taken from ICDAR 2013 Math-zone segmentation datasets and other document pages are scanned from different Chemistry books. Altogether 752 displayed equations are spotted manually from the dataset. Our method segmented 733 DE zones from the dataset which are manually verified. Our method is not applicable for segmentation of some of the displayed-math zones like conditional equations and matrices because our motivation is to classify the chemical and non-chemical equations. The summary of the experimental results is shown in Table 1. From the table we see a high degree of accuracy and low misclassification percentage for both chemical and non-chemical expressions obtained from our dataset. Other than the numerical data some of the result of the classification for few cases are shown in Fig. 12 and Fig. 13. In the figures the red and the green bounding boxes indicate chemical and non-chemical expressions respectively.

TABLE I
Summary of experimental results

Classified As Actual	Chemical	Other
Chemical	97.8%	2.2%
Other	2.95%	97.05%

A. Error Analysis

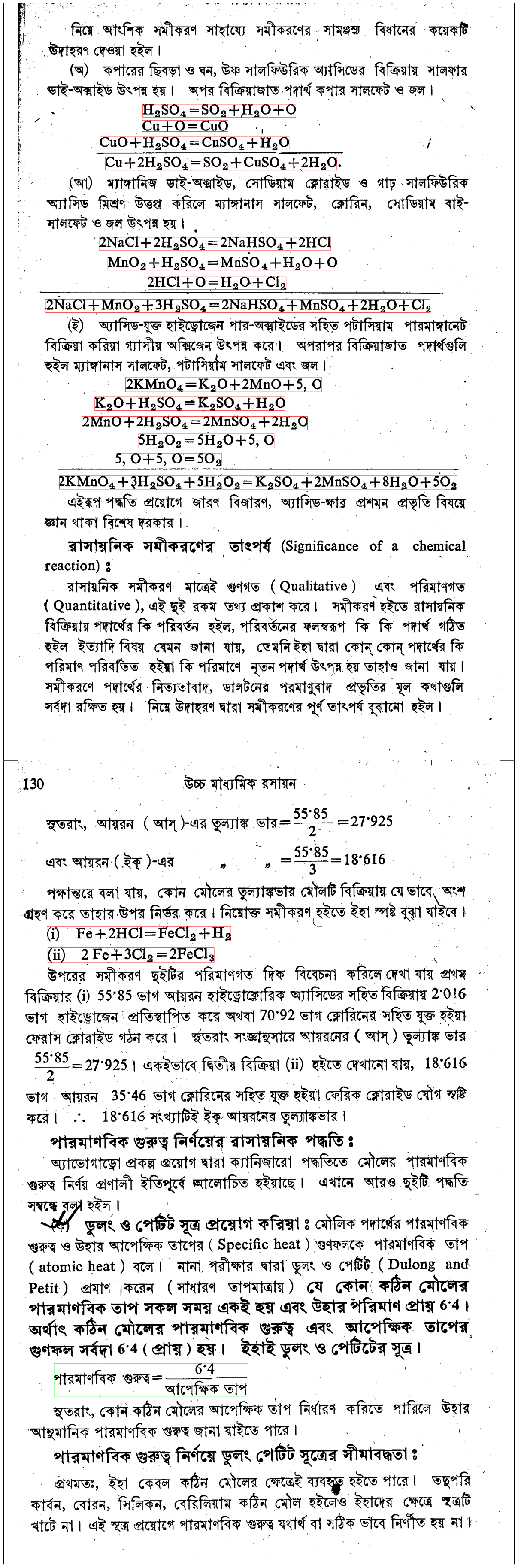
Here we try to analyse the sources of some of the errors which have negative effect on the performance figures both for segmentation and classification.

- Segmentation error:

  Some chemical equations have some reactants/symbols (such as Δ) just on top or bottom of the arrow and in some cases they enter the bounding box of the arrow. When we extract only single characters from the word blob information as mentioned in II, we do not get the arrow as it is no longer treated as a single character within its bounding box. Here the operator, arrow does not get detected and the ratio between number of operands and number of operators exceeds 2. So, this is segmented as an embedded equation. See Fig. 14(a).
  In a few cases, text lines consisting of embedded expressions are identified as displayed equations (see Fig. 14(b)), but in those cases the text lines contain more mathematics than the normal text. However, identification of mathematics intensive text lines as displayed equation is not a severe error.
- Classification error:
  In some cases if the operands of a chemical equation have Alkyl or Halide group, they are denoted as R and X respectively. But these symbols are not present in the periodic table. Hence, when the substrings from the OCR output are searched in the dictionary, it comes back negative. The equation is detected as a non-chemical one. See 4th equation bounded with a green rectangle in Fig. 15.

IV. Conclusion

We have presented an automated chemical and non-chemical equation segmentation system able to classify dis-



নিম্নে আংশিক সমীকরণ সাহায্যে সমীকরণের সামঞ্জস্য বিধানের কয়েকটি উদাহরণ দেওয়া হইল।

(অ) কপারের ছিবড়া ও ঘন, উষ্ণ সালফিউরিক অ্যাসিডের বিক্রিয়ায় সালফার ডাই-অক্সাইড উৎপন্ন হয়। অপর বিক্রিয়াজাত পদার্থ কপার সালফেট ও জল।

$$H_2SO_4=SO_2+H_2O+O$$
$$Cu+O=CuO$$
$$CuO+H_2SO_4=CuSO_4+H_2O$$
$$Cu+2H_2SO_4=SO_2+CuSO_4+2H_2O.$$

(আ) ম্যাঙ্গানিজ ডাই-অক্সাইড, সোডিয়াম ক্লোরাইড ও গাঢ় সালফিউরিক অ্যাসিড মিশ্রণ উত্তপ্ত করিলে ম্যাঙ্গানাস সালফেট, ক্লোরিন, সোডিয়াম বাই-সালফেট ও জল উৎপন্ন হয়।

$$2NaCl+2H_2SO_4=2NaHSO_4+2HCl$$
$$MnO_2+H_2SO_4=MnSO_4+H_2O+O$$
$$2HCl+O=H_2O+Cl_2$$
$$2NaCl+MnO_2+3H_2SO_4=2NaHSO_4+MnSO_4+2H_2O+Cl_2$$

(ই) অ্যাসিড-যুক্ত হাইড্রোজেন পার-অক্সাইডের সহিত পটাসিয়াম পারমাঙ্গানেট বিক্রিয়া করিয়া গ্যাসীয় অক্সিজেন উৎপন্ন করে। অপরাপর বিক্রিয়াজাত পদার্থগুলি হইল ম্যাঙ্গানাস সালফেট, পটাসিয়াম সালফেট এবং জল।

$$2KMnO_4=K_2O+2MnO+5,\ O$$
$$K_2O+H_2SO_4=K_2SO_4+H_2O$$
$$2MnO+2H_2SO_4=2MnSO_4+2H_2O$$
$$5H_2O_2=5H_2O+5,\ O$$
$$5,\ O+5,\ O=5O_2$$
$$2KMnO_4+3H_2SO_4+5H_2O_2=K_2SO_4+2MnSO_4+8H_2O+5O_2$$

এইরূপ পদ্ধতি প্রয়োগে জারণ বিজারণ, অ্যাসিড-ক্ষার প্রশমন প্রভৃতি বিষয়ে জ্ঞান থাকা বিশেষ দরকার।

রাসায়নিক সমীকরণের তাৎপর্য (Significance of a chemical reaction) ঃ

রাসায়নিক সমীকরণ মাত্রেই গুণগত (Qualitative) এবং পরিমাণগত (Quantitative), এই দুই রকম তথ্য প্রকাশ করে। সমীকরণ হইতে রাসায়নিক বিক্রিয়ায় পদার্থের কি পরিবর্তন হইল, পরিবর্তনের ফলস্বরূপ কি কি পদার্থ গঠিত হইল ইত্যাদি বিষয় যেমন জানা যায়, তেমনি ইহা দ্বারা কোন্ কোন্ পদার্থের কি পরিমাণ পরিবর্তিত হইয়া কি পরিমাণে নূতন পদার্থ উৎপন্ন হয় তাহাও জানা যায়। সমীকরণে পদার্থের নিত্যতাবাদ, ডালটনের পরমাণুবাদ প্রভৃতির মূল কথাগুলি সর্বদা রক্ষিত হয়। নিম্নে উদাহরণ দ্বারা সমীকরণের পূর্ণ তাৎপর্য বুঝানো হইল।

130 উচ্চ মাধ্যমিক রসায়ন

সুতরাং, আয়রন (আস্)-এর তুল্যাঙ্ক ভার $=\frac{55{\cdot}85}{2}=27{\cdot}925$

এবং আয়রন (ইক্)-এর „ „ $=\frac{55{\cdot}85}{3}=18{\cdot}616$

পক্ষান্তরে বলা যায়, কোন মৌলের তুল্যাঙ্কভার মৌলটি বিক্রিয়ায় যে ভাবে অংশ গ্রহণ করে তাহার উপর নির্ভর করে। নিম্নোক্ত সমীকরণ হইতে ইহা স্পষ্ট বুঝা যাইবে।

(i) $Fe+2HCl=FeCl_2+H_2$

(ii) $2\,Fe+3Cl_2=2FeCl_3$

উপরের সমীকরণ দুইটির পরিমাণগত দিক বিবেচনা করিলে দেখা যায় প্রথম বিক্রিয়ার (i) 55·85 ভাগ আয়রন হাইড্রোক্লোরিক অ্যাসিডের সহিত বিক্রিয়ায় 2·016 ভাগ হাইড্রোজেন প্রতিস্থাপিত করে অথবা 70·92 ভাগ ক্লোরিনের সহিত যুক্ত হইয়া ফেরাস ক্লোরাইড গঠন করে। সুতরাং সংজ্ঞানুসারে আয়রনের (আস্) তুল্যাঙ্ক ভার $\frac{55{\cdot}85}{2}=27{\cdot}925$। একইভাবে দ্বিতীয় বিক্রিয়া (ii) হইতে দেখানো যায়, 18·616 ভাগ আয়রন 35·46 ভাগ ক্লোরিনের সহিত যুক্ত হইয়া ফেরিক ক্লোরাইড যৌগ সৃষ্টি করে। ∴ 18·616 সংখ্যাটিই ইক্ আয়রনের তুল্যাঙ্কভার।

পারমাণবিক গুরুত্ব নির্ণয়ের রাসায়নিক পদ্ধতি ঃ

অ্যাভোগাড্রো প্রকল্প প্রয়োগ দ্বারা ক্যানিজারো পদ্ধতিতে মৌলের পারমাণবিক গুরুত্ব নির্ণয় প্রণালী ইতিপূর্বে আলোচিত হইয়াছে। এখানে আরও দুইটি পদ্ধতি সম্বন্ধে বলা হইল।

(ক) **ডুলং ও পেটিট সূত্র প্রয়োগ করিয়া ঃ** মৌলিক পদার্থের পারমাণবিক গুরুত্ব ও উহার আপেক্ষিক তাপের (Specific heat) গুণফলকে পারমাণবিক তাপ (atomic heat) বলে। নানা পরীক্ষার দ্বারা ডুলং ও পেটিট (Dulong and Petit) প্রমাণ করেন (সাধারণ তাপমাত্রায়) **যে কোন কঠিন মৌলের পারমাণবিক তাপ সকল সময় একই হয় এবং উহার পরিমাণ প্রায় 6·4। অর্থাৎ কঠিন মৌলের পারমাণবিক গুরুত্ব এবং আপেক্ষিক তাপের গুণফল সর্বদা 6·4 (প্রায়) হয়। ইহাই ডুলং ও পেটিটের সূত্র।**

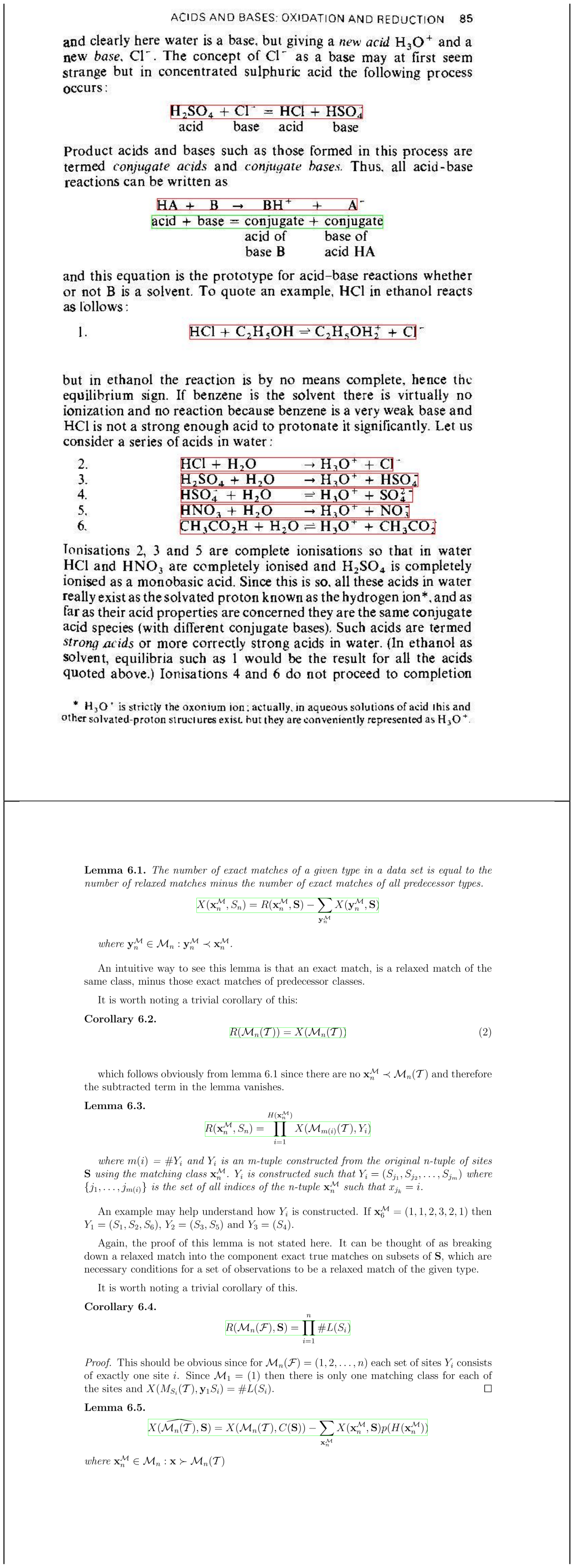
$$\text{পারমাণবিক গুরুত্ব}=\frac{6{\cdot}4}{\text{আপেক্ষিক তাপ}}$$

সুতরাং, কোন কঠিন মৌলের আপেক্ষিক তাপ নির্ধারণ করিতে পারিলে উহার আনুমানিক পারমাণবিক গুরুত্ব জানা যাইতে পারে।

পারমাণবিক গুরুত্ব নির্ণয়ে ডুলং পেটিট সূত্রের সীমাবদ্ধতা ঃ

প্রথমতঃ, ইহা কেবল কঠিন মৌলের ক্ষেত্রেই ব্যবহৃত হইতে পারে। তদুপরি কার্বন, বোরন, সিলিকন, বেরিলিয়াম কঠিন মৌল হইলেও ইহাদের ক্ষেত্রে সূত্রটি খাটে না। এই সূত্র প্রয়োগে পারমাণবিক গুরুত্ব যথার্থ বা সঠিক ভাবে নির্ণীত হয় না।

Fig. 12. Samples (1 and 2) of classification result



ACIDS AND BASES: OXIDATION AND REDUCTION 85

and clearly here water is a base, but giving a *new acid* H_3O^+ and a new *base*, Cl^-. The concept of Cl^- as a base may at first seem strange but in concentrated sulphuric acid the following process occurs:

$$\underset{\text{acid}}{H_2SO_4} + \underset{\text{base}}{Cl^-} = \underset{\text{acid}}{HCl} + \underset{\text{base}}{HSO_4^-}$$

Product acids and bases such as those formed in this process are termed *conjugate acids* and *conjugate bases*. Thus, all acid-base reactions can be written as

$$\begin{array}{ccccccc} HA & + & B & \rightarrow & BH^+ & + & A^- \\ \text{acid} & + & \text{base} & = & \text{conjugate} & + & \text{conjugate} \\ & & & & \text{acid of} & & \text{base of} \\ & & & & \text{base B} & & \text{acid HA} \end{array}$$

and this equation is the prototype for acid–base reactions whether or not B is a solvent. To quote an example, HCl in ethanol reacts as follows:

1. $HCl + C_2H_5OH \rightleftharpoons C_2H_5OH_2^+ + Cl^-$

but in ethanol the reaction is by no means complete, hence the equilibrium sign. If benzene is the solvent there is virtually no ionization and no reaction because benzene is a very weak base and HCl is not a strong enough acid to protonate it significantly. Let us consider a series of acids in water:

2. $HCl + H_2O \rightarrow H_3O^+ + Cl^-$
3. $H_2SO_4 + H_2O \rightarrow H_3O^+ + HSO_4^-$
4. $HSO_4^- + H_2O \rightleftharpoons H_3O^+ + SO_4^{2-}$
5. $HNO_3 + H_2O \rightarrow H_3O^+ + NO_3^-$
6. $CH_3CO_2H + H_2O \rightleftharpoons H_3O^+ + CH_3CO_2^-$

Ionisations 2, 3 and 5 are complete ionisations so that in water HCl and HNO_3 are completely ionised and H_2SO_4 is completely ionised as a monobasic acid. Since this is so, all these acids in water really exist as the solvated proton known as the hydrogen ion*, and as far as their acid properties are concerned they are the same conjugate acid species (with different conjugate bases). Such acids are termed *strong acids* or more correctly strong acids in water. (In ethanol as solvent, equilibria such as 1 would be the result for all the acids quoted above.) Ionisations 4 and 6 do not proceed to completion

* H_3O^+ is strictly the oxonium ion; actually, in aqueous solutions of acid this and other solvated-proton structures exist but they are conveniently represented as H_3O^+.

Lemma 6.1. *The number of exact matches of a given type in a data set is equal to the number of relaxed matches minus the number of exact matches of all predecessor types.*

$$X(\mathbf{x}_n^{\mathcal{M}}, S_n) = R(\mathbf{x}_n^{\mathcal{M}}, \mathbf{S}) - \sum_{\mathbf{y}_n^{\mathcal{M}}} X(\mathbf{y}_n^{\mathcal{M}}, \mathbf{S})$$

where $\mathbf{y}_n^{\mathcal{M}} \in \mathcal{M}_n : \mathbf{y}_n^{\mathcal{M}} \prec \mathbf{x}_n^{\mathcal{M}}$.

An intuitive way to see this lemma is that an exact match, is a relaxed match of the same class, minus those exact matches of predecessor classes.

It is worth noting a trivial corollary of this:

Corollary 6.2.

$$R(\mathcal{M}_n(\mathcal{T})) = X(\mathcal{M}_n(\mathcal{T})) \tag{2}$$

which follows obviously from lemma 6.1 since there are no $\mathbf{x}_n^{\mathcal{M}} \prec \mathcal{M}_n(\mathcal{T})$ and therefore the subtracted term in the lemma vanishes.

Lemma 6.3.

$$R(\mathbf{x}_n^{\mathcal{M}}, S_n) = \prod_{i=1}^{H(\mathbf{x}_n^{\mathcal{M}})} X(\mathcal{M}_{m(i)}(\mathcal{T}), Y_i)$$

where $m(i) = \#Y_i$ *and* Y_i *is an* m*-tuple constructed from the original* n*-tuple of sites* $\mathbf{S}$ *using the matching class* $\mathbf{x}_n^{\mathcal{M}}$. Y_i *is constructed such that* $Y_i = (S_{j_1}, S_{j_2}, \ldots, S_{j_m})$ *where* $\{j_1, \ldots, j_{m(i)}\}$ *is the set of all indices of the* n*-tuple* $\mathbf{x}_n^{\mathcal{M}}$ *such that* $x_{j_k} = i$.

An example may help understand how Y_i is constructed. If $\mathbf{x}_6^{\mathcal{M}} = (1,1,2,3,2,1)$ then $Y_1 = (S_1, S_2, S_6)$, $Y_2 = (S_3, S_5)$ and $Y_3 = (S_4)$.

Again, the proof of this lemma is not stated here. It can be thought of as breaking down a relaxed match into the component exact true matches on subsets of $\mathbf{S}$, which are necessary conditions for a set of observations to be a relaxed match of the given type.

It is worth noting a trivial corollary of this.

Corollary 6.4.

$$R(\mathcal{M}_n(\mathcal{F}), \mathbf{S}) = \prod_{i=1}^{n} \#L(S_i)$$

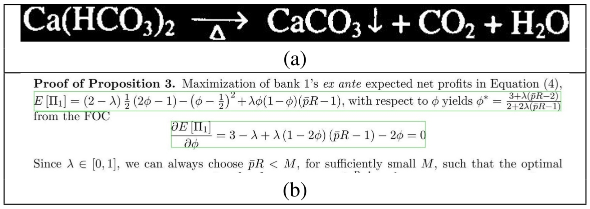
Proof. This should be obvious since for $\mathcal{M}_n(\mathcal{F}) = (1, 2, \ldots, n)$ each set of sites Y_i consists of exactly one site i. Since $\mathcal{M}_1 = (1)$ then there is only one matching class for each of the sites and $X(M_{S_i}(\mathcal{T}), \mathbf{y}_1 S_i) = \#L(S_i)$. □

Lemma 6.5.

$$X(\widehat{\mathcal{M}_n(\mathcal{T})}, \mathbf{S}) = X(\mathcal{M}_n(\mathcal{T}), C(\mathbf{S})) - \sum_{\mathbf{x}_n^{\mathcal{M}}} X(\mathbf{x}_n^{\mathcal{M}}, \mathbf{S}) p(H(\mathbf{x}_n^{\mathcal{M}}))$$

where $\mathbf{x}_n^{\mathcal{M}} \in \mathcal{M}_n : \mathbf{x} \succ \mathcal{M}_n(\mathcal{T})$

Fig. 13. Samples (3 and 4) of classification result



$Ca(HCO_3)_2 \xrightarrow[\Delta]{} CaCO_3\downarrow + CO_2 + H_2O$

(a)

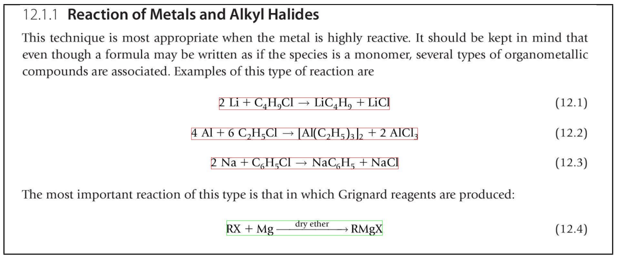
Proof of Proposition 3. Maximization of bank 1's *ex ante* expected net profits in Equation (4), $E[\Pi_1] = (2-\lambda)\frac{1}{2}(2\phi-1) - \left(\phi-\frac{1}{2}\right)^2 + \lambda\phi(1-\phi)(\bar{p}R-1)$, with respect to ϕ yields $\phi^* = \frac{3+\lambda(\bar{p}R-2)}{2+2\lambda(\bar{p}R-1)}$ from the FOC

$$\frac{\partial E[\Pi_1]}{\partial\phi} = 3-\lambda+\lambda(1-2\phi)(\bar{p}R-1)-2\phi = 0$$

Since $\lambda \in [0,1]$, we can always choose $\bar{p}R < M$, for sufficiently small M, such that the optimal

(b)

Fig. 14. Examples of segmentation error



12.1.1 **Reaction of Metals and Alkyl Halides**

This technique is most appropriate when the metal is highly reactive. It should be kept in mind that even though a formula may be written as if the species is a monomer, several types of organometallic compounds are associated. Examples of this type of reaction are

$$2\ Li + C_4H_9Cl \rightarrow LiC_4H_9 + LiCl \quad (12.1)$$

$$4\ Al + 6\ C_2H_5Cl \rightarrow [Al(C_2H_5)_3]_2 + 2\ AlCl_3 \quad (12.2)$$

$$2\ Na + C_6H_5Cl \rightarrow NaC_6H_5 + NaCl \quad (12.3)$$

The most important reaction of this type is that in which Grignard reagents are produced:

$$RX + Mg \xrightarrow{\text{dry ether}} RMgX \quad (12.4)$$

Fig. 15. An Example of classification error.

played equations of both categories from heterogeneous document images. The method is based on detecting operators as +, - and → which are common in both chemical and non-chemical equations, segmenting the displayed equations and then running them through an open source OCR to classify. A publicly available database for mathematical documents and scanned images of various chemistry books, both in English and Bengla scripts, are used in this study. The experimental results demonstrate the efficiency of our proposed method. The present work points to several new research avenues to be explored further. The over-all performance itself demands further research in this area. Excessive degradation due to aging and improper digitization of the document images result in broken and merged characters which give conflicting output from the OCR. In the future, design of a better integrated OCR specifically for chemical equations would be ventured in. In addition, formation of "electron bond matrix" of a chemical compound in a reaction mentioned in the introduction section would be a high-level goal in the future.

References

[1] D. Blostein and A. Grabavec, *Recognition of Mathematical Notation*, Handbook of Character Recognition and document Image Analysis, 557–582, 1997.

[2] K-F. Chan and D-Y. Yeung, *Mathematical Expression Recognition: A Survey*, IJDAR, Vol. 3, no: 1, 3–15, 2000.

[3] U. Garain and B. B. Chaudhuri, *On OCR of Printed mathematical Expressions*, Digital Document Processing, Ed: B. B. Chaudhuri, Advances in pattern Recognition, 235–259 , 2007.

[4] R. Fateman, T. Tokuyasu, B. Berman, and N. Mitchell, *Optical character recognition and parsing of typeset mathematics*, Visual Commun. And Image Representation, Vol 7, no. 1, 2–15, 1996.

[5] J. Y. Toumit, S. Garcia-Salicetti, and H. Emptoz, *A hierarchical and recursive model of mathematical expressions for automatic reading of mathematical documents*, In Proc. of ICDAR, 116–122, 1999.

[6] A. Kacem, A. Beliad and M. Ben Ahmed, *Automated Extraction of printed mathematical formulas using fuzzy logic and propagation of context*, IJDAR, vol. 4, no. 2, 97–108, 2001.

[7] M. Suzuki, F. Tamari, R. Fukuda, S. Uchida and T. Kanahori, *INFTY - An Integrated OCR system for Mathematical Documents'*, Proc. of ACM Symposium on Document Engineering, 95–104, 2003.

[8] Utpal Garain, *Identification of Mathematical Expressions in Document Images*, Proc. of ICDAR, 1340–1344, 2009.

[9] Utpal Garain. *Recognition of Printed Handwritten Mathematical Expressions*, Ph.D Thesis, ISI, Kolkata, India, 2005.

[10] B. B. Chaudhuri and U. Garain, *Extraction of type atyle based meta-information from Imaged documents*, IJDAR, vol. 3 no. 3, 138–149, 2001.

[11] J. Jin, X. Han and Q. Wang, *Mathematical formulas extraction*, Proc. of ICDAR, 1138–1141, 2003.

[12] D. M. Drake and H. S. Baird, *Distinguishing mathematical notation from english text using computational geometry'*, Proc. of ICDAR, 1270–1274, 2005.

[13] Y.-S. Guo, L. Huang and C.-P. Liu, *A new approach for understanding of structure of printed mathematical expressions*, Proc. of ICMLC, 2633–2638, 2007.

[14] We-Te Chu and Fan liu, *Mathematical formula detection from heterogeneous document Images*, Proc. of CTAAI, 140–146, 2013.

[15] S. P. Chowdhury, S. Mandal, A. K. Das, and B. Chanda, *Segmentation of Text and Graphics from Document Images*, Proc. of ICDAR, 619–623, 2007.

[16] R. C. Gonzalez and R. Wood, *Digital Image Processing*, Addision-Wesley, 1992.

[17] S. P. Chowdhury, S. Mandal, A. K. Das, and B. Chanda, *A simple and effective table detection system from Document Images*, Proc. of IJDAR, Vol. 8(2), 172–182, 2006.