# C. Optical Character Recognition of each reactant:

Each segment of a displayed chemical equation is divided into three zones; namely upper zone, middle zone and lower zone (see Fig 1). To identify the three zones of a DCE zone, uppermost and lowermost co-ordinates of each connected component below the same DCE zone are also obtained. The median of uppermost coordinate, and median of lowermost co-ordinate of such components in DCE 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 DCE zone. Similarly, the median of uppermost co-ordinate of the components in the DCE zone generates a horizontal line. This horizontal line, called top line, separates the middle and upper zones of the DCE zone. The subscripts in a DCE 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 DCE zone we have detected the subscripts and superscripts.



Fig 1. Presence of different zones in DCE

Each reactant form one word blob. Each character forming one word blob is run through an OCR engine, Matlab OCR. Each character image is returned by OCR engine in corresponding text format and saved in a string, S Reactant for further processing. For each superscript and subscript, ‘^’ and ‘\_’ are inserted before them respectively in S Reactant for denoting purposes.

# D. Auto Correction of Reactants:

Matlab OCR does not give perfect result for chemical equations which has subscripts and superscripts as well. Hence, to generate the exact equation from the image, auto correction is required in chemical database. First an error map is created based upon our observation on multiple databases. The following table contains the error list.

g=8 3

n=l1 1'1 I1 11 X1 Il

O=0

3=8 'E s w

a=3. 3 21 8 El 8.

q=Cl Q

H=1-l 1-1 l-l l-1

2=7 4 Z z

l=1 I

I=l

u=11 U l1 ll 1l

i=1 l I

5='S

4=A

Z=2

r=1‘

e=C

s=S

The first column contains the correct input and the rest of the columns contain all outputs given by OCR. This table is generated based on our observation on x number of documents. Next this table is kept into a hash map where the key is the OCR output and its corresponding value is possible input set.

Auto correction is performed based on this error hash map, H. Each chemical reactant in any equation has the following format-

[Co-efficient]\*[Chemical compound][state]\*

Coefficient has numerical values, its regular expression is taken as [2-9]+[0-9]\*. The first digit will not be 0 as coefficient cannot start with that. There are 4 states of a chemical compound which are represented by (s), (g), (l) and (aq). To detect if the compound has any state representation, the following regular expression [(][a-z]+[)] is checked in the substring of length 4 extracted from the end of S Reactant. If a match is found and …

After extracting the state,

For each ooec

Possible input set = H.Get(ooec)

If length of ooec = 1

Possible input set = [ooec, possible input set]

End

End

All combinations = Cartesian product between possible input sets of each OOEC

Find match with the exhaustive compound list.

If perfect match found, that compound replaces S Reactant

Else,

Match= Longest common substring(s) found between All combinations and compound list

End

If number (match)>1

Consider each match as a possibility

LHS = periodic elements in the left hand side of the equal sign following the regular expression [A-Z][a-z]\*

RHS = periodic elements in the right hand side of the equal sign following the regular expression [A-Z][a-z]\*

For every possibility

If LHS=RHS

Match = possibility (i)

End

End

/\* if there are more than one matches, it’s a failure case \*/

End