International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. 11, No.1 (Jan-2022)

https://dx.doi.org/10.12785/ijcds/110129

Identification of Inorganic Chemical Formulas based on Support Vector Machine and SURF Key Point Descriptor

Shrikant Mapari[1] **and Navendu Chaudhary** [2]

[1]*Symbiosis Institute of Computer Studies and Research (SICSR),Symbiosis International (Deemed University) (SIU),Model Colony, Pune, Maharashtra, India.*

[2]*Symbiosis Institute of Geoinformatics (SIG) Symbiosis International (Deemed University) (SIU), Model Colony, Pune, Maharashtra, India.*

Received 25 Feb. 2021, Revised 28 Jul. 2021, Accepted 13 Nov. 2021, Published 9 Jan. 2022

Abstract: In inorganic chemistry, the chemical reactions are mostly represented by Inorganic chemical formulas of the chemical compounds. While digitizing the handwritten or printed contents related to inorganic chemistry, firstly it has to recognize handwritten inorganic chemical formulas (HICF). The HICF has represented by using alphabets and numbers, this alphabet or number is termed as Inorganic Symbols (IS). Therefore, to recognize HICF; the inorganic symbols have to be reorganized. Here in this paper, we had developed a model to recognize HICF. The developed model has based on the classification of inorganic symbols (IS). This proposed model does the classification based on geometric shapes, which are part of the IS. The classification of IS has done in three classes based on the geometric shapes used to formulate the IS. After classification, this model identifies the individual IS with the help of feature descriptor and Support Vector Machine (SVM). In this paper, each class has it's trained SVM to identify individual IS from the respective class. Once each individual IS has identified then it has put together to represent in the inorganic chemical formula. The result shows that the recognition percentage of each SVM is nearly equal to 97%. The system accepts scanned images of HICF as an input and delivers recognized HICF in text format as an output.

Keywords: Support Vector Machine (SVM); SURF descriptor; classification; feature extraction; contours based segmentation; handwritten chemical formulas

1. Introduction

The chemical reactions are made up of organic symbols, inorganic chemical formulas and operators. The organic symbols mainly content the geometrical shapes to represent bonds, rings and benzene structures, where inorganic compounds are mostly represented by alphanumerical characters. Hence the recognition of such handwritten inorganic compound formula is similar to handwritten character recognition. While recognizing Handwritten Inorganic Chemical Formulas (HICF) firstly, the individual Inorganic Symbols (IS) has to be recognized. Here in this paper, we developed a system which is capable of recognizing the HICF. This proposed system has first recognized the all IS as a part of HICF. Then complete HICF has represented in text format as an output of the system. This proposed system can be utilized in the education domain. This model can help to develop the automatized system for correction of the answer books of subject related to chemistry domain. This is one of the inspirations behind this research work.

As stated earlier that the recognition of HICF has similar to the handwritten character and numerical recognition. Though it has similarities, the recognition of HICF has constraint from the chemistry domain because to form an inorganic chemical formula it has not used all alphabets and number. Hence, to recognize HICF only the alphanumeric characters those were used as part of the inorganic chemical formula has used. Recognition of free hand-drawn character has more complexities because there has no single approach which has solved this problem in an efficient way [1]. Similarly, the recognition HICF has difficulties due to the irregularities in free handwriting. The different user has different ways to draw the same IS which was the part of HICF. This differentiation in handwriting cause errors while identifying the similar IS written differently.

Another most frequent problem in recognition of handwritten character was irregularities in representing the geometric definition of a character [2]. The handwritten IS also has geometric definitions while they have drawn on paper by the user. These hands have drawn IS has represented with some geometric shapes like lines, arcs and circles. When these shapes were drawn by different people to represent HICF, it has deformation in representation because of noise

E-mail address: shrikant.mapari@sicsr.ac.in, navendu@sig.ac.in

and variation in thickness of strokes. The awkward handwriting has put more obstacles in recognition of HICF. The awkwardness of handwriting results into misrecognition of some IS, which has similarities in representation. For example, characters like ‘C’ and ‘O’ or ‘N’ and ‘H’, which were frequently used in IS has similarities in their representation on paper. A person with clumsy handwriting can swap the representation of these above discuss characters, which results in misinterpretation of HICF. This problem has been addressed in our proposed, by classification and training the different class SVM. The SVM has used to identify the single character which represents the IS in HICF. The SVM has a good recognition rate in case of handwritten character recognition [3], [4].

We consider that each IS has contains some geometric shapes like lines, arcs and circles. Hence, while recognizing this handwritten IS we consider the presence of such basic shape inside the IS and used this information to categorize IS in different categories.

Our developed system has four main steps, namely preprocessing and segmentation, Feature extraction, Classification and Identification. The input for our system was scanned images of handwritten inorganic chemical formulas (ICF). In the first step, our system pre-processes the input image and extracts each IS from it. In the feature extraction step, the feature vector has built up by extracting the features from each IS. The classification of each IS has done in three different classes using the values of the features. Each IS identification has done using Speeded-Up Robust Feature (SURF) descriptor and Support Vector Machine (SVM). Lastly, the inorganic formula was represented by using recognized IS in text format. The rest of this paper has organized as follows: In section 2 related works and problem identification has discussed. In section 3 we have elaborated developed a system along with its components and working. The experimental work and results have explained in section 4. Section 5 brief conclusions of the work have provided. Section 6 talks about the limitations of the system with future expansions.

2. Problem identification and related work

A novel approach to, understanding and analysing a chemical formula has been proposed by Wang et al. [5] in 2009 based on structural characteristics, semantic rules, and more importantly grammatical rules. A formal description of the chemical formula based-on the grammatical rules is summed up and applied to the analysing process which generates grammar spanning graphs from the analysed result step-by-step, and that is used for the further structure representation and data retrieval. This proposed model was used to recognize the inorganic formula from online handwritten samples, where a digital ink is used to draw the formula. The model uses the concept of trees to store individual IS which was more complex in terms of memory and time. Sun et al. [6] have proposed an algorithm for ranking chemical formulae and tagging chemical names in digital documents. This algorithm uses Conditional Random Fields (CRFs) and Support Vector Machines (SVMs). The proposed algorithm accepts and chemical formula from the user in digital format and searches that formula in printed digital documents. The model based on a weighted directed graph has been proposed by Chang et al. [7] to recognize online handwritten inorganic chemical expressions. This model presented a set of novel statistical algorithms in two key components of this framework: symbol grouping and structure analysis. A novel two-level algorithm model to recognize online handwritten inorganic chemical expression was proposed by Yang et al. [8]. In the first level, structural information is used to distinguish different parts and recognize chemical compounds. Then the algorithm segments expressions and recognizes isolated symbols. For recognition of the handwritten chemical operators used in chemical expressions, Yang et al. [9] proposed a novel three-layer recognition model. They summarized and categorized chemical operators into different types. The three layers of the model consist of preprocessing, segmentation and substance recognizing.

The above discussion summarizes that there were lots of work has done for recognizing inorganic formulas for online handwritten. In this online method, the formulas had drawn using a digital pen or ink and with the help of some software. In this paper, we have proposed a system to recognize inorganic formula wrote in offline mode. It means the user has to write an inorganic formula on paper using a pen or pencil and scanned image has used for recognition.

3. Proposed system

The developed proposed system can recognize a HICF. This system accepts the input as an image of HICF. The proposed system has four phases labelled as segmentation, feature extraction, classification and identification. It delivers a recognized HICF in text format as an output. The flow of these stages along with input and output as shown in Figure 1

A. Input

The proposed system has accepted a scanned image of HICF as an input. The Image database of HICF images has built by collecting the samples of HICF from different age group people. The different inorganic chemical formulas are identified from the text books of *10* and *12* standards. This Identified formulas are provided to different people and collected handwritten inorganic chemical formulas from them on paper. This HICF on paper has scanned as image and the image database has built for experiment. The user has to write an inorganic chemical formula on paper with pen or pencil and scanned it using a standard scanner and with standard DPI image resolution. This scanned image should be in PNG file format. At the time of scanning, the image has to store in PNG file format or after scanning it has to convert in PNG format.

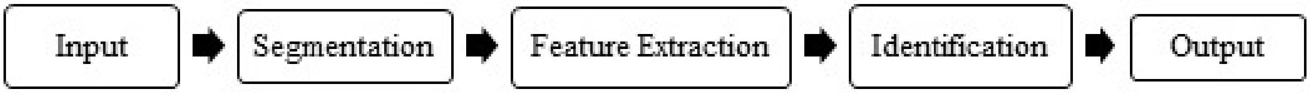


Figure 1. The architecture of the proposed system

B. Pre-Processing and Segmentation

This system pre-processes the scanned image of HICF which is the input for this system. The preprocessing has done to remove the image noise and convert the colored image to grayscale. The noise removal has done using a Laplacian filter and the input image got sharper. This sharpens image has used for the segmentation process. In segmentation, we have segmented individual IS used in the inorganic chemical formula.

In this segmentation process first, we had separated the image object from its background using distance transform [10]. The image objects are termed as continues group of pixels in the foreground. Then contour-based method [11], [12] has used for segmentation of image of HICF. In this method, the different contours had detected from inputted images of HICF. The contour of an image has termed as continues group of pixels which represented in the foreground of the image. Here for HICF, the contours were representing the individual IS. Once we have detected the contours, they were used for segmenting the individual IS. The image segment, which has represented the IS has separated from HICF by using a bounded rectangle of the respective contour. The bounded rectangle of a contour is the rectangle occupied by that contour within the image. The area of a bounded rectangle of each contour has to be calculated. Then by using this bounded rectangle area, the occupied portion of the image has to be separated as an image segment. This separated image segment represents the individual IS from HICF. The contours were detected by start reading the grayscale image from the top-left position until the last pixel of the image.

The procedure for contour detection and segmentation named as *Procedure1*is shown below:

```
Procedure: Procedure1.
Input:
Scanned image of HICF
Output:
Seg :- An array of image segment, which contains the individual Inorganic symbols
Assumption:
Img : Matrix of size R, C to store image data.
R:- Number of rows in image
C:- Number of a column in image
Counts:- An array of image contours
brect :- individual bounded rectangle of single contour.
Img = Read scanned image of HICF
ImgGray = GrayScale(Img)
Img = Laplacian_Filter(ImgGray)
Img = Distance_transform(Img)
Counts = Contours(Img)
foreach  Cnt ∈ Counts  do
set I = 0
brect = Bounded_Rect(Cnt)
Seg( I ) = Img(bract)
I = I + 0
end for
```

The *Procedure1* is not only capable to segment the IS from HICF but also it can segment the inorganic formula which has embedded in the text for expression. The above algorithm yields the array of image segments of handwritten IS. This array further used for extracting the features of individual IS.

C. Feature extraction

In this step, teach image segment, which has generated in the segmentation phase has been used as input. We had extracted the features to construct the feature vector, which has further used for classification and identification.

Here we had identified some geometrical features which described the presence of basic shapes in each inorganic symbol. We had found an Euler number of each segment as a feature, which describes the presence of a hole in the image segment. These geometric features and Euler number features had further used to classify an image segment in some class based on the geometrical shape of IS. This geometrical shape-based classification as explained in section 3.4. Also, we had used a SURF descriptor [13] as a feature for the identification of individual IS. The features we had used in our system are described in the following sections.

1) Geometrical Features

The geometrical properties of the shape have been used by Les, Z. and Les, M [14] to describe the different aspects of the shape. The IS also consists of some shape in their formation. Hence the classification of IS was done based on the geometric feature, where we detected the presence of an arc or circle and line in the IS. These inputs are used to classify the image segment into different classes as an IS contents line or IS contents an arc or circle. We have detected lines and an arc or circles using a well-known method called as Hough transform [15].

Hough Transform: The Hough Transform (HT) [15] was introduced by P.V.C. Hough [16] to detect the line from photographs. After some time, HT gets more generalized to detect and recognize more shapes like parameterized curves [15], arbitrary 2D shapes [17], the limb shape of hands [18], and 3D shapes [19]. Here in this paper, we had used HT to detect the line and arc or circle presents in image segments of IS. The HT for a line is denoted by using the equation 1.

$$y = -\left(\frac{\cos\theta}{\sin\theta}\right)\left(\frac{x+r}{\sin\theta}\right) \tag{1}$$

Where θ is the angle of the line and r is the distance of a line from the origin.
The HT for a circle with radius R and center point (a, b) can be described by a parametric equation as shown in equation 2 and equation 3.

$$x = a + R\cos\theta \tag{2}$$

$$y = b + R\sin\theta \tag{3}$$

2) Euler Number

This This feature described the relation between the number of the continues part and number of holes presents in an inorganic symbol. The Euler Number [20] was one of the shape descriptors which has mostly used to describe the shape of images. The Euler number has calculated using the following equation 4.

$$EluN = S - N \tag{4}$$

Where EluN is Euler number, S be the number of the continues part and N is the number of holes.
This feature is used to classify the IS based on its value as follows:
If EluN =1 then classify symbol with no hole.
If EluN = 0 or EluN ¡ 0 then classify symbol with one or more hole.

3) SURF Descriptor

Speeded-Up Robust Feature (SURF) descriptor had the best performance over another descriptor. SURF is used as the detector and descriptor for object recognition in computer vision area. This SURF detector and descriptor algorithm proposed by Bay et al. [13] yields a fast and robust descriptor. The SURF algorithm works in two steps. The first step is based on Hessian Matrix [21] and uses basic Laplacian-based detector to find an interesting point which is also called key points. In the second step, it describes the distribution of Haar-wavelet responses within the interest point neighborhoods. The SURF algorithm extracts a feature descriptor of 64 floating-point values.

We had consumed SURF descriptor to identify the individual character along with an artificial neural network. We have calculated SURF descriptor for each IS and special character which is used to formulate an inorganic chemical formula (ICF). This calculated feature vector is further used as input to a Support Vector Machine (SVM) neural network for identification of the individual character.

D. Classification

To identify an ICF we have to first identify an individual IS. As per the periodic table of chemistry, a total of 46 alphabets were used to form any ICF. Hence, we have to consider these all *46* alphabets along with some numbers and special characters used in ICF as IS for recognition. Here we had categorized these all IS into three categories. This categorization has done by classifying, those inorganic Symbols (IS) in three classes on the basics of the presence of geometric feature and a hole in that IS. This process of classification helped us to identify individual IS.

We had used the calculated values of HT and Euler number of each image segment form feature extraction phase as input for classification. We have labelled these three classes as Class H, Class NHC and Class NHL. We had used an Euler number (EluN) feature and results of the Hough Transform to do this classification.

These classes are described as follows:

Class H: The IS has labelled with this class H if that IS has holes and contents an arc or circle shape in their representation. For example, inorganic symbols (IS) like A, B, O, etc. belong to this class. To detect the presence of a hole and arc or circle in IS we had observed the values of HT for circle detection and Euler number. If the value of the Euler number (EluN) is equal to zero or negative, then it has either one or more hole in IS. Then we observed values of HT for circle detection to find out the presence of the circle, ellipse or an arc in IS. If this value indicates presence, then we had labelled such image segments with a class label as class H.

Class NHC: This class NHC has classified the inputted IS as a member of this class when this IS containing no hole and has the presence of the circle, ellipse or an arc. The IS belong to this class are C, S, 3, G, etc. To identify this class IS we have observed the values of Euler number and HT for circle detection. If the value of EulN is equal to one, then the IS does not contain a hole in it. Then the value of HT for circle detection has to find out. If this value indicates the presence of an arc, circle or ellipse, then we had labelled those image segments with this class label as class NHC.

Class NHL: This class NHL classifies the IS, which contents no hole and has the presence of line in it. The IS like M, N, W, Y, etc. belong to this class. The values of the extracted feature, an Euler number (EluN) and HT for line detection has observed for each image segment. If the value of EluN is equal to one *(1)*, then there were no holes in that IS. Then the system had checked the values of HT for line detection for the presence of the line in IS. If this value indicates the presence, then such image segments were labelled with this class label as class NHL.

The procedure to label an image segment to a particular class label is named as *Procedure2* and as shown below:

Procedure: Procedure2
Input:
Seg:- An array of the image segments, which is an output of Proceure1. Output:
LblSeg: An array of class labels for the respective image segment.
Assumptions: EluN:- An Eulre number of the single image segment.
NHTC:- Number of circle, ellipse or an arc found in the single image segment by applying HT for circle detection.
NHTL:- Number of lines in the single image segment by HT for line detection.

```
$foreach \ \ seg \in Seg \ \ do$
    $setI = 0$
    $EluN = Euler\_Number(Seg)$
    $NHTC = HoughTransform\_Circle(Seg)$
    $NHTL = HoughTransform\_Line(Seg)$
    $if \ \ ((EluN \leq 0) and NHTC > 0) \ \ then$
        $LblSeg[I] = "H"$
    end if
    $if \ \ ((EluN = 1) and (HTC > 0)) \ \ then$
        $LblSeg[I] = "NHC"$
    end if
    $if \ \ ((EluN = 1) and (HTL > 0)) \ \ then$
        $LblSeg[I] = "NHL"$
    end if
    $I = I + 1$
end for
```

This above *Procedure2* has classified each image segment of ICF into its particular class and assign a class label to it.

The Table I has shown the alphabet and numbers that had identified as IS as per the periodic table of chemistry. These identified IS are the part of ICF. This shows the each IS with its respective class

Once the image segment has got a proper label of the respective class, then it has to enter into the identification phase to identify exact IS and then proper ICF has identified.

E. Identification

In this step of the experiment, the class labelled image segments along with a label array from the classification step have been used as input. For this each image segment a SURF descriptor has calculated. This descriptor has represented by a feature vector of 64 floating-point values. This feature vector has used to recognize an individual IS by training a Support Vector Machine (SVM). This all possible IS are categorized into three different classes. Hence three different SVM for each class has to develop. To develop such SVM a feature vector has built by calculating SURF descriptor for samples of each IS belonging to the respective class. This calculated SURF descriptor feature vector has been used as input for respective class SVM. The target vector of integer value has built by assigning an integer number to each individual IS from that respective class. For example, for the SVM of Class H, we had assigned value 1 to symbol A, value 2 to B and so on. This crated target vector has used as a target for respective SVM while training. We had trained and test all three SVM by using samples of SURF descriptors of respective IS and its respective target vector.

1) Support Vector Machine (SVM)

Support Vector Machine was one of the best learning machines for the problem in the machine learning domain. SVM has introduced by Vapnik [22] and Cortes et al. to minimize the empirical training errors [23]. SVM has been used by many researchers to recognize handwritten character [3], cursive handwriting [4] and handwritten digit recognition [24]. Most researchers have also coated a good recognition rate for SVM. The SVM has worked on kernel-based learning [25]; it has three types of the kernel, namely linear, polynomial and Radial Basic Function (RBF). The formula (5) describes the linear kernel function, formula (6) denotes a function for the polynomial kernel, and formula (7) shows the kernel function for RBF.

$$f(x) = w \cdot x + b \tag{5}$$

Here, f(x) is linear kernel function, x is input feature vector, w is weight vector and b is offset.

$$K(x, x') = (x \cdot x' + 1)^d \tag{6}$$

Here, K (x, x‘) is polynomial function, x, x‘ is input vector, and d is the degree of polynomial.

$$K(x, x') = exp\left(-\gamma \left\| x - x' \right\|^2\right) \tag{7}$$

Here K (x, x‘) is RBF kernel function, x, x‘ is input vector, and ℷ is a free parameter.

Here in this paper, we had developed a classifier based on SVM to identify an individual IS from the respective class, which has labelled to the image segment. We had created separate multiclass SVM for each Class H, Class NHC and Class NHL to identify the IS from that class. These SVM has trained with input as SURF descriptor of all IS belonging to the respective class and RBF kernel parameter.

2) Procedure for Identification

The identification process accepts image segments and their respective class label as input and then identifies an individual IS. The output of this process was identified handwritten ICF has shown in or drawn in text format. The procedure for identification has named as Procedure3. and shown as follows:
Procedure: - Procedure3
Input:
Seg: - An array of image segment, which is an output of Procedure1.
LblSeg:- An array of class label for respective image segments from Seg, which is output of Procedure2
Output:

TABLE I. INORGANIC SYMBOLS WITH RESPECTIVE ITS CLASS.

Class Label	Inorganic Symbols (IS) belong to the Class
Class H	A, B, D, O, P, R, a, b, e, g, o, 8, 9, 6.
Class NHC	C, G, S, U, c, h, n, r, s, t, u, 2, 3, 5.
Class NHL	F, H, I, K, L, M, N, V, W, X, Y, Z, f, i, l, 1, 4, 7

An identified image of HICF with a inorganic chemical formula shown as text on it.
Assumptions:
Sfd:- A vector of 64 floating point values represents the SURF descriptor of single image segments.
HLbl:- An array of string indicating the class labels of Inorganic Symbol belongs to class H
NHCLbl:- An array of string indicating the class labels of Inorganic Symbol belongs to class NHC
NHLLbl:- An array of string indicating class labels of Inorganic Symbol belongs to class NHL
ICF:- A string represents predicted inorganic formula.

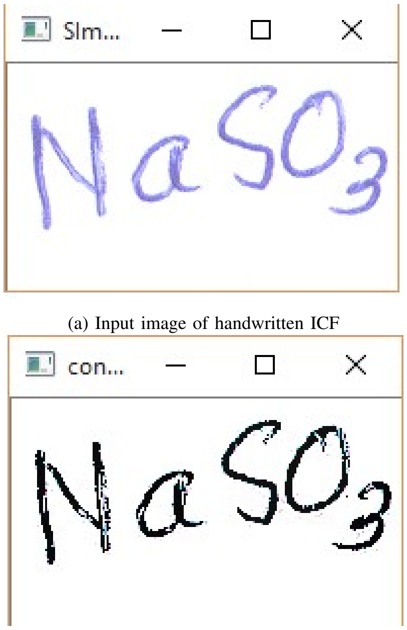
```
$foreach\ \ lblseg \in LblSeg\ \ and\ \ seg \in Seg\ \ do$
$Sfd = SURF\_Descriptor(seg)$
$if\ \ (lblseg = "H")\ \ then$
$Load_SVM(\backslash HClass")$
$intresult = SVMPredict(Sfd)$
$StringpredictedIS\ = HLbl[result]$
end if
$if\ \ (lblseg = "NHC")\ \ then$
$Load\_SVM(\backslash NHCClass")$
$intresult = SVMPredict(Sfd)$
$StringpredictedIS\ = NHCLbl[result]$
end if
$if\ \ (LblSeg[I] = "NHL")\ \ then$
$Load\_SVM(\backslash NHLClass")$
$intresult = SVMPredict(Sfd)$
$StringpredictedIS\ = NHLLbl[result]$
end if
$ICF = ICF + predictedIS$
end for
```

Show or draw ICF to show the identified Handwritten Inorganic Chemical Formula.

At the end of this identification phase, the recognized ICF has drawn as a result.

4. EXPERIMENT AND RESULTS

We had set up an experiment to identify a handwritten inorganic chemical formula (ICF). Input for this experiment was scanned images of handwritten ICF. We had collected samples of handwritten ICF from people of different age group. We had identified approx. 50 different isolated ICF from the various textbook of inorganic chemistry used by academicians. After this, we had collected approx. 50 to 60 samples of each ICF. It makes the total sample size of all handwritten ICF up to 2700. Considering that each ICF has averaged 3 IS present in it, we had approx. 5100 samples of most of used IS in inorganic chemistry. We had scanned this all handwritten ICF and converted them into PNG format. This PNG formatted images are used as input for our experiment.



(a) Input image of handwritten ICF

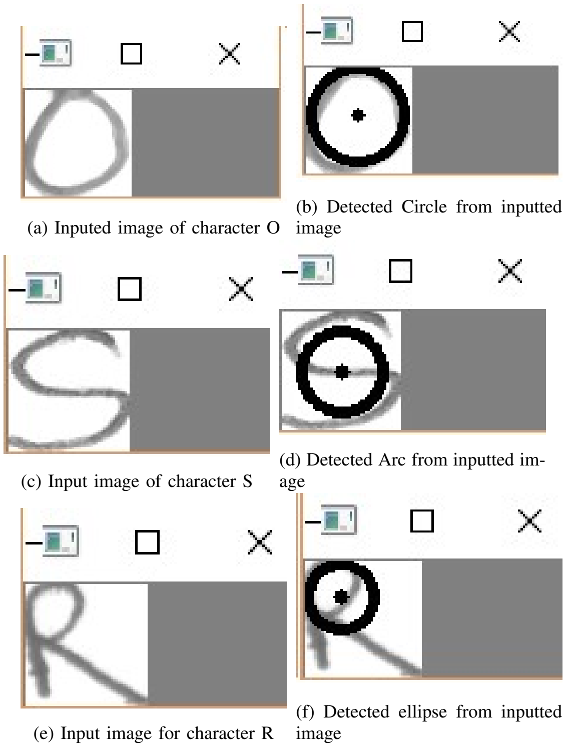
(b) Output image with detected contours

Figure 2. Input and Output Image

Then in the second step, these sample images are preprocessed and get segmented as per the Procedure1. discussed in section 3-B This segmentation has done based on contour detection technique. The experimental result of segmentation has shown in Figure 2 The Figure 2a Shows the inputted image of handwritten ICF for segmentation. Figure 2b shows the detected contours are drawn, which are used for segmentation.

In the next phase of feature extraction for each image segment, which had segmented in the previous phase, an Euler Number (EluN) and HT for circle and line Detection (HTC and HTL) have calculated and used as features. The Figure 3b, 3d and 3f shows the sampled images of IS which have detected arc or circle in its representation using HT for circle detection. The Figure 4b, 4d and 4f shows the sampled images of IS which have detected line in its representation using HT for line detection.

These extracted features were used to classify and labelled these image segments with the appropriate class label in the classification phase. The classification has done



(a) Inputed image of character O

(b) Detected Circle from inputted image

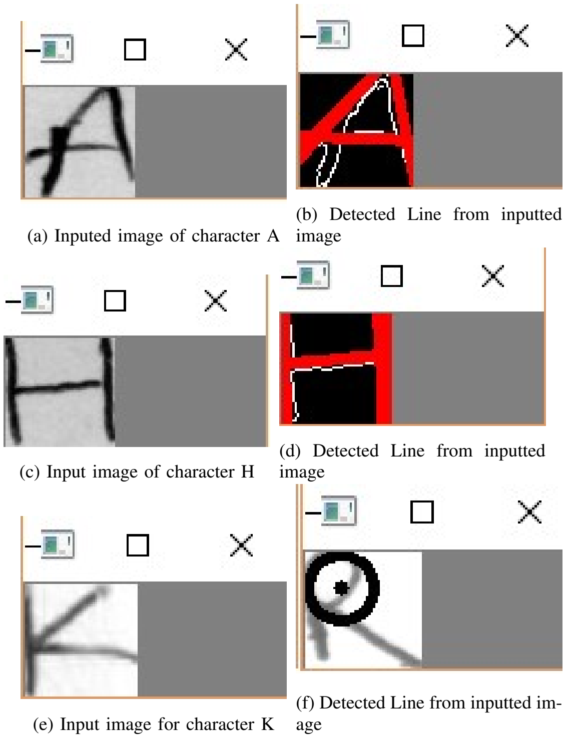
(c) Input image of character S

(d) Detected Arc from inputted image

(e) Input image for character R

(f) Detected ellipse from inputted image

Figure 3. Figures showing Hough Transform Circle



(a) Inputed image of character A

(b) Detected Line from inputted image

(c) Input image of character H

(d) Detected Line from inputted image

(e) Input image for character K

(f) Detected Line from inputted image

Figure 4. Figures showing Hough Transform Line

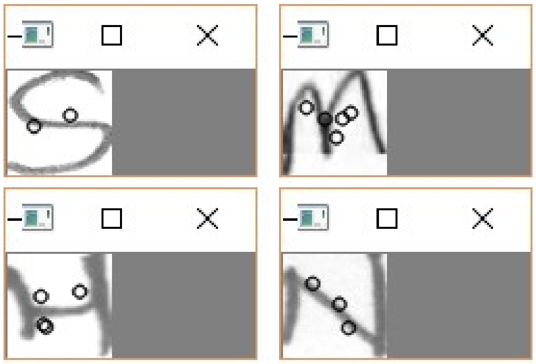


Figure 5. SURF Key points drawn for Inorganic Symbol: S, M, H and N

according to the Procedure2. as discussed above. Once the classification for all image segments had finished the system enters into the identification phase.

During the identification procedure first, a SURF descriptor has to be calculated for each IS for each class. This SURF descriptor method first detects the key points from the image segment and then computes descriptor for those key points. The detected key points are shown in Figure 5 for some sample images of IS.

Then this calculated SURF descriptor has used to find out individual IS. The SVM has trained for each class using handwritten image samples of IS belonging to that class. The feature vector of size 64 values has constructed for training SVM by calculating values of SURF descriptor for these all IS for each class. Then this feature vector has inputted to train the SVM of each class.
The SVM parameters that had set for training are as follows:
Input: Vector of 64 floating-point value
Kernel: Radial Basic Function (RBF)
Target: Vector of size total number of IS belonging to the respective class of integer values, which represent the individual Inorganic Symbol.
The training for each class SVM has done using feature vectors of samples of IS of that class. Following Table II summarized the training result of the SVM of each class.

The above Table II shows the recognition rate for each class SVM has 97 % accuracy approximately. The first column shows the number of sample images, which contains the handwritten inorganic symbols used for training and testing. The second column shows the number of samples corrected recognized for its respective class. The last column shows the recognition percentage for an individual class. The rows of this table describe the individual class samples.

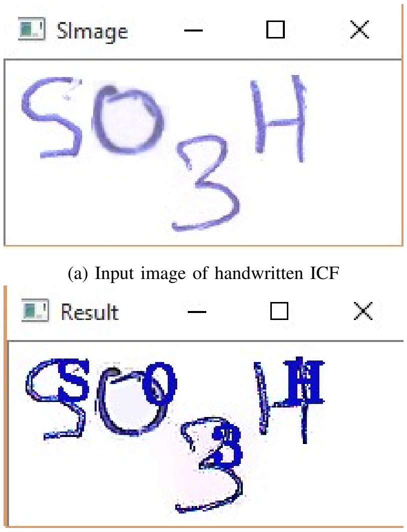
The Table III has shown the sensitivity means predictive percentage of true positive and specificity means predictive percentage of true negative for SVM algorithm.

TABLE II. TRAINING SUMMARY OF SVM

Class	Sampled used		Sampled recognized		Recognition %	
	Training Samples	Testing Samples	Training Samples	Testing Samples	Training Samples	Testing Samples
Class H	1050	200	1035	195	98.57	97.5
Class NHC	1300	250	1277	242	98.23	96.8
Class NHL	1950	350	1899	340	97.38	97.14

TABLE III. SENSITIVITY AND SPECIFICITY OF SVM

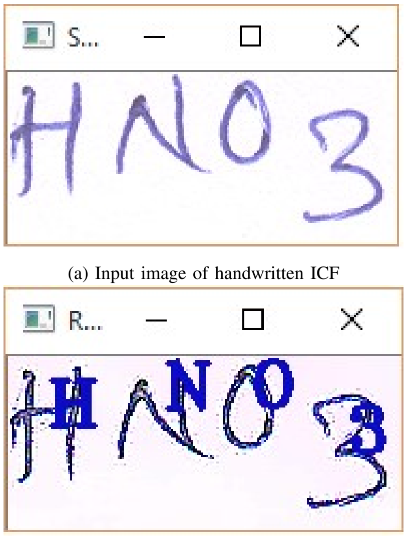
Class	Sensitivity of SVM in %	Specificity of SVM in %
Class H	96.45	96.87
Class NHC	95.85	96.2
Class NHL	96	95.5



(a) Input image of handwritten ICF

(b) Recognized ICF drawn on original image

Figure 6. Input and output images of experiment



(a) Input image of handwritten ICF

(b) Recognized ICF drawn on original image

Figure 7. Input and output images of experiment

Once these three SVM have trained and tested, they had used in this experiment to identify the individual IS. The identification has carried out as per the Procedure3. discussed as above. This trained SVM returns an integer value which represents the IS for that class. This obtained integer value from the trained SVM had mapped to its associated IS and formulate the ICF by placing these IS one after another using the values of bounded rectangle. The following Figure 6 to Figure 8 shows the result of the experiment. Figure 6a to Figure 8a shows inputted handwritten ICF and Figure 6b to Figure 8b shows recognized ICF, which has either drawn on the original image with different colors or ICF drawn on a black background with blue color.

5. Conclusion

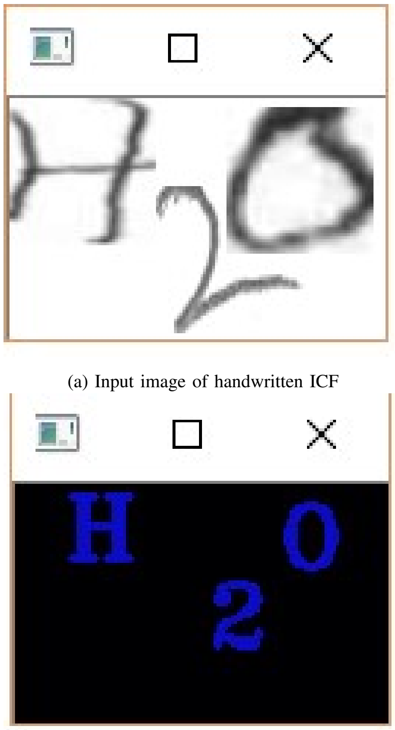
There were most of the researcher has proposed a framework or model to recognize the online handwritten inorganic chemical symbols. It is observed that no one has done the recognition work for recognizing the offline handwritten chemical symbols. There were lots of work has done for recognition of offline handwritten character. This work is near to the recognition of ICF In this paper, we had discussed the developed system to recognize offline handwritten ICF. The result that we had discussed shows that this system can recognize each inorganic symbol from the ICF and finally it constructs the same. The utilization of geometric features for the classification of IS reduces the total number of inputs and outputs for SVM avoid the complexity of the system. The result of this system also proves the accuracy of SVM to identify the IS.

6. Limitation and Future Work

The developed approach has recognized any alphanumeric character as an inorganic symbol if it has a presence in an inorganic formula. The system does not implement any mechanism for finding out that the entered formula is really a chemical formula or just a combination of an alphanumeric symbol. Hence this system can be improved in future to do the semantic verification of chemical formula using domain knowledge. It also has scope to improve the recognition rate of SVM.

References

[1] W. Kacalak, K. D. Stuart, and M. Majewski, "Selected problems of intelligent handwriting recognition," in *Analysis and design of*



(a) Input image of handwritten ICF

(b) Recognized ICF has drawn on a black background

Figure 8. Input and output images of experiment

intelligent systems using soft computing techniques. Springer, 2007, pp. 298–305.

[2] S. Uchida and H. Sakoe, "A survey of elastic matching techniques for handwritten character recognition," *IEICE transactions on information and systems*, vol. 88, no. 8, pp. 1781–1790, 2005.

[3] D. Nasien, H. Haron, and S. S. Yuhaniz, "Support vector machine (svm) for english handwritten character recognition," in *2010 Second International Conference on Computer Engineering and Applications*, vol. 1. IEEE, 2010, pp. 249–252.

[4] F. Camastra, "A svm-based cursive character recognizer," *Pattern Recognition*, vol. 40, no. 12, pp. 3721–3727, 2007.

[5] X. Wang, G. Shi, and J. Yang, "The understanding and structure analyzing for online handwritten chemical formulas," in *2009 10th International Conference on Document Analysis and Recognition*. IEEE, 2009, pp. 1056–1060.

[6] B. Sun, P. Mitra, C. Lee Giles, and K. T. Mueller, "Identifying, indexing, and ranking chemical formulae and chemical names in digital documents," *ACM Transactions on Information Systems (TOIS)*, vol. 29, no. 2, pp. 1–38, 2011.

[7] M. Chang, S. Han, and D. Zhang, "A unified framework for recognizing handwritten chemical expressions," in *2009 10th International Conference on Document Analysis and Recognition*. IEEE, 2009, pp. 1345–1349.

[8] J. Yang, G. Shi, K. Wang, Q. Geng, and Q. Wang, "A study of online handwritten chemical expressions recognition," in *2008 19th International Conference on Pattern Recognition*. IEEE, 2008, pp. 1–4.

[9] J. Yang, K. Wang, and G. Shi, "Structure-based recognition of handwritten chemical operators," in *2010 8th World Congress on Intelligent Control and Automation*. IEEE, 2010, pp. 6371–6374.

[10] F. Y. Shih and Y.-T. Wu, "Fast euclidean distance transformation in two scans using a 3× 3 neighborhood," *Computer Vision and Image Understanding*, vol. 93, no. 2, pp. 195–205, 2004.

[11] Y.-T. Hsiao, C.-L. Chuang, J.-A. Jiang, and C.-C. Chien, "A contour based image segmentation algorithm using morphological edge detection," in *2005 IEEE International Conference on systems, man and cybernetics*, vol. 3. IEEE, 2005, pp. 2962–2967.

[12] J. Malik, S. Belongie, T. Leung, and J. Shi, "Contour and texture analysis for image segmentation," *International journal of computer vision*, vol. 43, no. 1, pp. 7–27, 2001.

[13] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (surf)," *Computer vision and image understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[14] Z. Les and M. Les, "Shape-understanding system: A system of experts," *International journal of intelligent systems*, vol. 19, no. 10, pp. 949–978, 2004.

[15] R. O. Duda and P. E. Hart, "Use of the hough transformation to detect lines and curves in pictures," *Communications of the ACM*, vol. 15, no. 1, pp. 11–15, 1972.

[16] P. V. Hough, "Method and means for recognizing complex patterns," Dec. 18 1962, uS Patent 3,069,654.

[17] D. H. Ballard, "Generalizing the hough transform to detect arbitrary shapes," *Pattern recognition*, vol. 13, no. 2, pp. 111–122, 1981.

[18] R. Okada, "Discriminative generalized hough transform for object dectection," in *2009 IEEE 12th International Conference on Computer Vision*. IEEE, 2009, pp. 2000–2005.

[19] M.-T. Pham, O. J. Woodford, F. Perbet, A. Maki, B. Stenger, and R. Cipolla, "A new distance for scale-invariant 3d shape recognition and registration," in *2011 International Conference on Computer Vision*. IEEE, 2011, pp. 145–152.

[20] D. Zhang and G. Lu, "Review of shape representation and description techniques," *Pattern recognition*, vol. 37, no. 1, pp. 1–19, 2004.

[21] K. Mikolajczyk and C. Schmid, "Indexing based on scale invariant interest points," in *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, vol. 1. IEEE, 2001, pp. 525–531.

[22] V. Vapnik, *The nature of statistical learning theory*. Springer science & business media, 1999.

[23] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[24] A. Bellili, M. Gilloux, and P. Gallinari, "An hybrid mlp-svm handwritten digit recognizer," in *Proceedings of sixth international conference on document analysis and recognition*. IEEE, 2001, pp. 28–32.

[25] K.-R. Muller, S. Mika, G. Ratsch, K. Tsuda, and B. Scholkopf, "An introduction to kernel-based learning algorithms," *IEEE transactions on neural networks*, vol. 12, no. 2, pp. 181–201, 2001.

Dr. Shrikant Mapari working with Symbiosis International (Deemed University), Pune, India as Asst. Prof. and Completed Ph.D in offline handwritten recognition . His broad area of research is on digital image processing, pattern recognition

Dr. Navendu Chaudhary working with Symbiosis International (Deemed University), Pune, India as Assco. Prof. He has completed his Ph.D. from University of Cincinnati, USA. His broad research areas are satellite image processing,