Recognition a Hand Drawn Chemical Structure Diagrams Using the Discrete B-Spline Wavelet Transform

Matheel Emaduldeen Abdulmuim

Department of Computer Science- University of Technology

Received 5 September 2016 ; Accepted 16 October 2016

Abstract

Chemical structures are a suitable way to represent the chemical equations perfectly in 2D space. But sometimes a hands drawn structures have some complicated when one take them as a document image and then recognized it to its full meaning to be accepted in machine data mining techniques so far. The wavelets with Spline are very steady and commonly symmetric or anti-symmetric. B-Spline has the preferable parataxis properties over all different types of wavelets in order L-1. In this paper a unified framework was built to include the organic and inorganic expressions. A suitable way was presented to classify hand drawn chemical structures using the B-Spline wavelet transform as a tool for image classification. In empirical valuation one can show that an enforcement of this method exceed the open source system available. The proposed framework achieved in Test-5 with 84.7% data accuracy for recognition the handwritten chemical expression database. Also with 77.8% classification accuracy using discrete B-Spline wavelet transforms.

Keywords: Chemical structure, B-Spline wavelet transform, Recognition**.**

Recognition a Hand Drawn Chemical Structure Diagrams Using the Discrete B-Spline Wavelet Transform

Matheel Emaduldeen Abdulmuim

تمييز خط اليد لمخططات الهياكل الكيميائية باستخدام التحويل المويجي ب – سبلاين

مثيل عمادالدين عبدالمنعم

قسم علوم الحاسوب - الجامعة التكنولوجية

الخلاصة

الهياكل الكيميائية هي الطريقة المناسبة لتمثيل المعادلات الكيميائية بشكل مثالي في الفضاء ذو البعدين. لكن احيانا الهياكل المكتوبة بخط اليد يكون فيها تعقيدا اكثر عند اخذ صوره للوثيقة ومن ثم تمييز معناها الكامل لتكون مقبولة في تقنيات تنقيب البيانات المعروفة. الويفيليت مع السبلاين تكون ثابتة جدا وعادة متماثل او غير متماثل. ب-سبلاين تملك افضلية وتقريب للخصائص اكثر من انواع الويفيليت بالترتيب L-1. في هذا البحث تم بناء نظام متكامل ليظم التعابير الغضوية وغير العضوية. تم تقديم طريقة مناسبة لتصنيف الهياكل الكيميائية اليدوية باستخدام التحويل المويجي ب-سبلاين كاداة لتصنيف الصور. وحسب النتائج التجريبية يمكن ملاحظة قوة هذه الطريقة لتجاوز توفير النظام مفتوح المصدر. الطريقة المقترحة حققت في الفحص 5 دقة بيانات 84.7% لتمييز قواعد البيانات للتعابير الكيميائية اليدوية. كذلك مع دقة تصنيف تصل الى 77.8% باستخدام التحويل ب-سبلاين المويجي المتقطع.

الكلمات المفتاحية:الهيكل الكيميائي، التحويل المويجي ب-سبلاين، التمييز.

Introduction

Optical character recognition may be a method which will modify text that will exist in real image, to adjective text. It permits a PC to acknowledge characters by visual techniques. The method includes a method to pre-process contents of the image then conquest of vital information concerning written language [1]. Handwritten chemical term recognition is a facultative application of scientific knowledge to attain normal employee expertise in epidermal PC reciprocal action particularly within the teaching field. Whereas the favorable outcomes were reportable and some industrial computer code product were discharged on recognizing written scientific discipline expressions within the last years ago, the analysis on understand the chemical terms was abundant less energetic. Chemical terms embody a mathematical relationship and neutralization from each mineral and chemical science. Whereas the mineral terms have sturdy constitutional likeness to scientific

discipline terms, graph such as organic form are far more completely diverse and sophisticated because of numerous constraint as clarify by Figure (1). These features send affront not solely on realizing organic terms, however conjointly oncoming up with a consolidated answer for realizing each inorganic and organic term [2].

Vamvakas et al. , 2010 [3] has conferred a strategy for directly written chemical expression realization. The suggested methods depends on a replacement feature extraction mechanism supported algorithmic dividing the character in the image in order that the ensuring located image at every loop have stop points numbers of foreground character pixels, as so much as the formula can be doable. Sankaran et al. , 2012 [4] has conferred a realization method for the Indian characters. The recognition accuracy of script isn't nevertheless similar to its Roman counterparts. This can be primarily because of the quality of the script, style etc. They use a type of Neural Network called Bidirectional Long- Short Term Memory (BLSTM) [1]. Another paper on the chemical term that is written to recognize was reportable antecedently. Neural net was employed in [5] to acknowledge chemical formula like rings. A work was conferred in [6] to acknowledge written organic chemical expressions. Support vector machine was illustrated for image clustering and realization, and native spatial case and a group of constraints in a used domain for interpret the chemical structure. Also, the work for automotive vehicle that are represented in three dimension views are presented in [7] to written organic structures with a main target on three dimension structure representation instead of realization. Sturdy suggest that each image written in one stroke were required for the popularity algorithmic program. [8] Conferred a system to acknowledge written image formula from a structural illustration. Yang and et al. in [9] propose a two-level algorithmic program to acknowledge written chemical expressions [10]. We propose during this paper a complete system for realizing each inorganic and organic expression. The system is composed of 3 stages – image grouping, structure analysis and linguistics verification. B-Spline wavelet transform was used to recognize the suited structure verification.

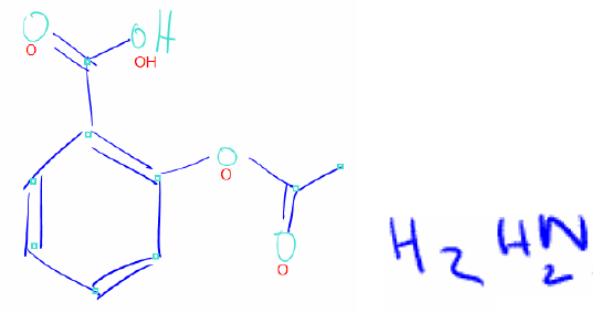
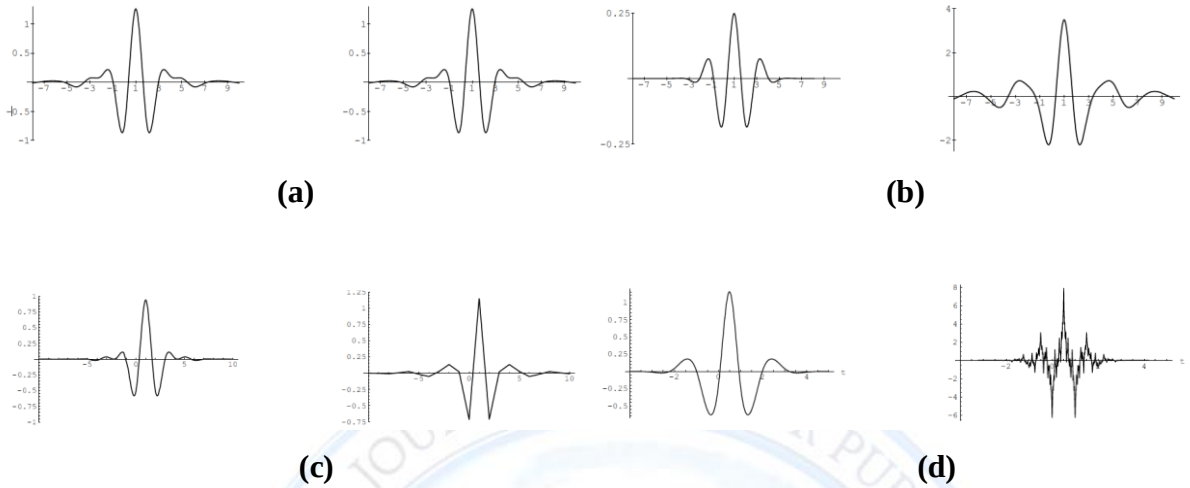


Figure 1: handwritten chemical expression examples.

Splines and Wavelets

Researchers are currently Janus-faced with a lot of increasing kind of wave bases to decide on from. Whereas the selection of the superior wave is clearly implementation-rely, it may be helpful to insulate variety of attributes and options that are of social utility to the employee [11]. Splines have a big effect on the speculation of the wavelet family. The premature instance is that the Haar wave that may be a spline of grade zero. This structure was expanded to splines with highly order, although this proposition exist for the most part unnoticed till wavelets became what they're these days [11]. It may be the simplest familiar samples of spline wavelets equal the perpendicular "Battle-Lemarie" functions, which may be seen as precursors of Mallat's multiresolution analysis of the wavelet family. Splines have conjointly been wont to explain several of the last structure of non-orthogonal wave standard. Noticeable exemplify are the B-spline wavelets that compressed as a squared succinctly supported and win a close to best time-frequency locating [12]. Also, the foremost widespread delegate of the Cohen-Daubechies-Feauveau category of biorthogonal wavelets compacted as square splines similarly. This can be as a result of the refinement of the binomial improvement filter that may be a critical part in any wave structure assembles to the B-spline that is the generation operates for polynomial splines. Four identical samples of blockish spline wavelets and their duals are illustrated in Figure (2) [12].



(a) **(b)**

(c) **(d)**

Figure (2): Cubic Spline wavelets: (a): Orthogonal spline, (b): Semi-orthogonal B-spline, (c) Shift orthogonal spline, (d) Biorthogonal spline [11].

Chemical Structure analysis

Given a stroke sequence *(o1, o2,…,oN)* of *N* round, the objective of image gathering is to search out the best image sequence *(G1, G2,…,Gn)* with corresponding boundaries within the most probability sense [14].

The O/P of image gathering may be a sequence of round combinations every one with n nominee symbols. The objective of construction dissection is to spot the constitutional connection through the round combinations and confirm the familiar image for every stroke cluster [10]. Separation of symbols and graphics is based on the approximate value of the capital letter height. If this value is found, then all segments are classified by this height with checking some exceptional cases, such as small single bonds. But the calculation of capital letter height is not a trivial procedure, if the image contains both bonds and text [13].

Symbol recognition and Segmentation

There are different manner for chemical diagram recognition. Associate optical character recognition incentive is employed to spot letters. The attached part streams of collections distributing as letters that won't to restore their identical element collection from the image shape that successively square measure that is used as input to the optical character

recognition incentive. An effort is additionally created to spot and section pasted letters. The method begins by normalizing the taken away zones and causation them to the optical character recognition incentive. This is often pursue by analyzing the strings of letters known by the optical character recognition to diagrammatic representation any strings admire foundations into their authentic requisite construction style. Uncommented strings are neglected [14].

The confession method carries on by preliminary processing then analyzing the remainder of the wheel to represent the essential construction of the graph. Preliminary processing is completed as a result of straight lines could also be divided into two or a lot of smaller segments. This method include a cleaning method to tie any lines destroyed close to an intersection (so quite two lines halve) and to proper any elementary lines be divided into smaller pieces [15]. To valid the primary situation, any comparatively teeny lines are deleted and also the lines linked to them are joined. Another situation is treated by employing a predetermined worth for the investor at the purpose wherever two lines link. If whether lines are instituted to form an investor below the threshold chose, the purpose of gathering is extracted and also the two lines are integrated to formulate long-term line section [15].

Characters are distributed using OCR engine that supported as a candidate mechanism and anatomy of their geometric and topologic options. Characters and others of comparatively larger magnitude may be understood not withstanding their magnitude and font and also its rotation [14]. Once distinguishing symbols, they are distributed by the tow directions in step with their coordinates to create statement that are successively during seek in a database of considerably utilized feasible groups such as R-groups and other ones. An effort to dissect statement or atom vamps not institute within the information is created [16].

To locate property data, all ligament lines square measure related to free-flags at their endpoints that are two. If the two bond lines bit one another, their conformable free-flags are equal to false; else they're equal to true. Then, line-atom property is decided employing an established in advance threshold range, ensuring that the line's free-flag is set as true, the atom distance is within the orientation of the road and ensuring that this atom is that the highest to the current line. A preselected option atom is connected to any or all ligament lines that don't seem

to be linked to any atom to affix any two ligament lines (disclosure of double/triple constraints), should they need to they have to be among shut proximity of every alternative and should be equivalent [16]. An affiliation schedule clarifies the property specifics of all constraints and atoms are produced at this point and constraints collected equivalent atoms are combined to form double or triple bonds. Also, atoms connected to that finish of dowel and dotted dowel constraint also are distinct within the affiliation schedule. This is often pursuing by the translation of general forms that square measure typically pictured as structures related to variable teams (R-groups) [17].

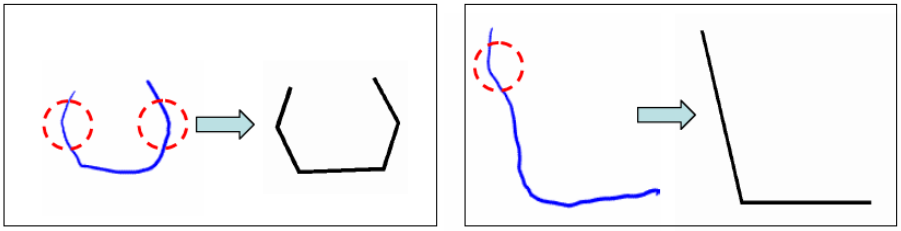


Figure 3: Stroke segmentation process.

The confidence scores were increased and labels created by the three realizers with a collection of group from the geometrical options specified source from the rounds in every filter group. The aim of those options is to assist the framework distinguishes between comparable characters and between educates and uneducated stroke congregations [16]:

1. Range of rounds: This feature occupies quality of combined design conventions: The symbol O (for oxygen), for example, is sometimes wrote with one stroke, whereas hash bonds generally include a minimum of three lines.
2. Bound-box overall dimensions (it is a vector including the dimension, elevation, and diagonal extend of the littlest axis stratify bound- box for the filter group): Bound-boxes of arrivers (e.g., differing kinds of bonds) square measure generally bigger and might have a vaster of side ratios than bound- boxes for part characters.

3. Ink intensity (the magnitude relation of the quantity of ink within the nominee cluster to the diagonal length of its bound-box). Ink intensity can facilitate indicates the kind of character: text characters and stake bonds usually coincide to zones of highly ink intensity.
4. Coat-stroke dimension (the most dimensions between personage strokes within the group): The feature described in this point will facilitate differentiate like "H" and "N" in the hash and double bonds.
5. Poly-line parataxis intermediate mean square error and intermediate phase expansion of the poly-line parataxis for the strokes within the group: This feature is helpful for distinctive bonds and hash bonds.
6. Line-segment direction (a vector of amounts that epitomizes the proportional orientations of line-segments within the filter group): This feature supported the poly-line work, contains the amount of in parallel lines, vertical lines, and intersections between line parts.

The proposed Algorithm

We suggest a consolidated system for educated each organic and inorganic expression. The system depends on three elements – image gathering, construction test and linguistics investigation. If we have a group of ink strokes, image gathering divides the strokes into combinations that represent image filters. Additionally to designing combined chemical characters, we tend to produce patterns for non-characters and bond images in character grouping so as to scale back the speed of miss-grouping. Non-character and bond designing additionally change image gathering to be behaved in a very regular method for each organic and inorganic form. The moment that the put in ink strokes are classified into possibility symbols, construction test is behave to see the constitutional relation through symbols as well as bonds. Using a tend to subedit construction test into a diagram study drawback within which supported outlined standard, the proper term construction is that the optimum diagram of the measured tendency digraph

appearing image relation. Linguistics investigation influence field information to smooth the popularity filter of construction test and manufacture the ultimate conclusion. It ought to be famed that each special and discourse data is applied mathematically combined into the suggested system and belated deciding is employed at each phase of the statistic system so as to look for the optimum discrimination outcome data from all elements. For character recognition, we use a simple classification procedure based on continues B-Spline wavelet transform descriptors of symbols' outer and inner contours. Also, because we start with the grouped symbols, some heuristics can be used to predict the next character and to improve the recognition rate. Figure 4 represent the block diagram of the recognition system. In parsing and recognition phase, each symbel or digit is determined and recognized using different stages as shown in figure 5.

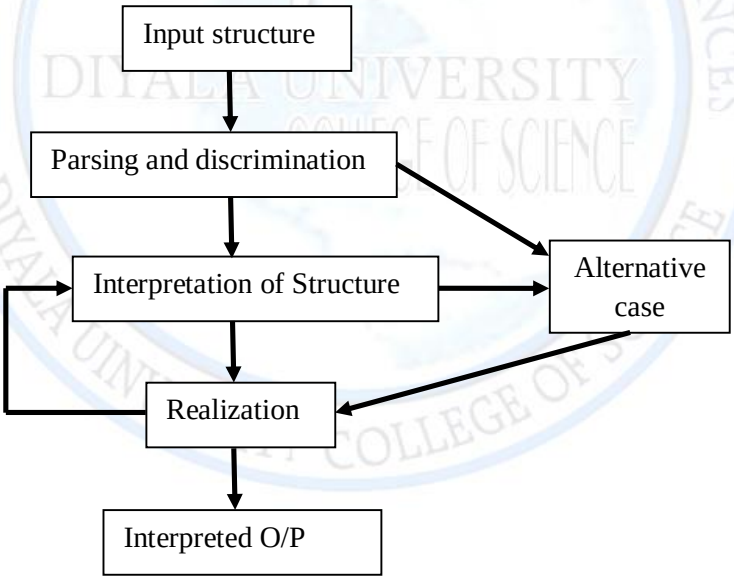


Figure 4: Recognition system.

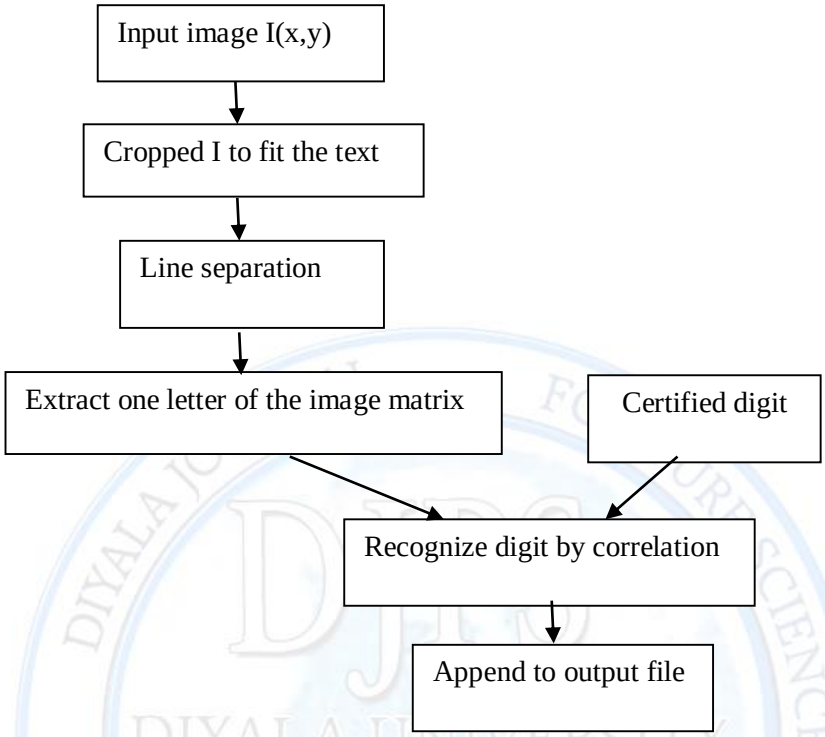


Figure 5: Parsing and recognition phase.

Results and Discussion

We are planning to discuss the experimentation and resulting analysis of the designed rules during this section. We have a tendency to conducted experiments exploitation the projected framework on our own information that consists of different databases with two hundred written chemical equations drawn by thirty completely various individuals. The information includes twenty five distinctive chemical equation derived from the chemistry books of schools and university. The whole variety of chemical characters lined within the information is thirty as well as chemical components, digits, response situation codes, e.g., '<>', '=' and ' ', factors, e.g., '-', '+', and different used codes reminiscent of '↑', '↓', '℃', '%', etc.

In our experiments, we have a tendency to live discrimination accuracy by equation that is far rigid than activity by image. An equation is taken into account to be properly realized when its codes are properly segmental and also the implicit construct of these symbols is properly known. The typical extend of equations in our information is regarding twenty one symbols. Every part in our realized system was estimated and also the outcome square measure concluded in Table (1).

Table 1: Recognition average (%).

	Test 1	Test 2	Test 3	Test 4	Test 5
Symbol Grouping (SG)	82.2	93.2	95.4	96.1	96.8
Structure Analysis (SA)	80.9	86.4	86.7	86.9	87.0
SG+SA	77.3	78.6	82.7	83.6	84.1
SG+SA+ Semantic verification	76.1	81.5	83.5	83.9	84.7

It ought to be noted that once appreciate the accuracy of construction test, valid splitting results were utilized in order to decouple the construction test part from the image grouping part. The results show that linguistics verification participate 1.9% proportional accuracy improvements. Overall, our projected framework achieved (Test-5) eighty 4.7% recognition accuracy on our giant written chemical expression information.It should be noted that when evaluating the accuracy of structure analysis, correct segmentation results were used in order to decouple the structure analysis component from the symbol grouping component. The thresholding used are depending on the accuracy rate that concluded from the accuracy equation. Also, another test is done in different runs to determine the recognition and miss-recognition of the chemical expressions with their accuracy. Table (2) gives the results about those runs.

Table 2: Recognition comparison.

Run	#Recognition	#Mis-Recognition	Accuracy %
1	44	51	46.32
2	56	39	58.95
3	44	51	46.32
4	54	41	56.84

To compare between the discrete B-Spline Wavelet Transform (BSWT), wavelet and Fourier transforms in results, Table (3) gives average results between those transforms. The classification accuracy are dampened by combination of the experiential operator here ***r*** is that the practice collection magnitude and ***w*** is that the practice collection structure. The mean and variance of the computer of the unusual knowledge sets properly categorized are notified.

Table 3: The classification accuracies.

		Filling			Truncating		
R	***w***	**Fourier**	**Wavelet**	**BSWT**	**Fourier**	**Wavelet**	**BSWT**
2	**0**	60.4	66.7	72.5	63.1	67.1	73.9
2	**1**	60.2	64.8	72.8	61.7	65.2	74.0
4	**0**	60.1	66.6	72.8	60.5	66.1	72.7
4	**1**	60.5	64.9	72.4	60.3	64.9	73.2
4	**2**	60.8	64.0	73.0	60.9	64.4	74.8
8	**0**	61.0	64.2	71.5	61.7	62.7	72.9
8	**1**	60.9	64.3	72.2	60.2	65.8	73.2
8	**2**	61.2	64.6	72.3	63.2	66.0	73.7
8	**4**	61.8	65.0	72.7	63.7	66.8	75.0
16	**0**	62.4	67.7	73.5	64.4	67.7	72.9
16	**1**	62.2	62.8	73.4	65.2	67.8	73.8
16	**2**	62.5	64.7	73.5	65.8	67.8	74.6
16	**4**	62.7	64.9	73.8	65.9	67.9	76.3
16	**8**	62.9	64.4	73.8	64.2	68.8	77.8

When we compare our proposed system with other recognition systems in the same datasets used, the results appear the recognition rate with high degree compared with other systems. Table (4) gives this comparison.

Table 4: Recognition rate comparison

Set/Version	Structures	v. 1.2.2	v. 1.3.6
CLiDe small set	46	26 (56%)	27 (58%)
USPTO CWU	5735	3984 (69%)	4341 (75%)
Proposed system	200	172 (86%)	179 (89.5%)

Conclusions

Constitutional Pattern Recognition (PR) is a robust test appliance inside fields wherever an outline collected of morphologic sub patterns and their mutual relations is predominant to correct rating choices. A constitutional PR framework generally contains feature extractions to spot example of morphologic property of the info that, in turn, are used because the rule for compilation mistreatment grammar. The field data substantial to instructor feature extraction and descriptive linguistics evolution is collected mistreatment data conquest mechanism. Besides that, such mechanisms are time exhaustion, inaccurate, and don't forever manufacture a whole content of the field.Thus, constitutional accession to PR are troublesome to use to unknown or shaky-comprehend fields, therefore limiting them to domains wherever the quality sorts and also the grammars have either become decided within the urbanity or are plain upon examination of the info. Completely remove the hassle indispensable to perform feature extraction and rating for constitutional PR framework can expand the relevance of constitutional accession to complicated, poorly-comprehend fields. We conferred a completely unique unified framework for written chemical expression recognition. A mix image grouping algorithmic rule that handles common symbols and bonds during a consistent manner is proposed. A graph-based illustration is outlined for each inorganic and organic expressions and structure analysis is developed as an exploration drawback for this illustration over a weighted direction graph created mistreatment applied mathematical modeling that leverages in numerous frequency domains. Linguistics verification is employed to re-rank take a look

at N best illustration generated by structure analysis. The experiment results on an out sized knowledge set show the effectiveness of our framework.

References

1. S. Singh and Y. T. Sabo, "Improve Optical Character Recognition Using Templates & Correlation", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), Volume 3 Issue 9, September 2014.
2. I. V. Filippov, M. C. Nicklaus and J. Kinney, "Improvements in Optical Structure Recognition Application", DAS '10, Boston, MA, USA , June 9-11, 2010.
3. Georgios Vamvakas, Basilis Gatos, Stavros J. Perantonis, "Handwritten character recognition through two-stage foreground sub-sampling" ,Pattern Recognition, Volume 43, Issue 8, August 2010.
4. Naveen Sankaran and C.V Jawahar, "Recognition of Printed Devanagari Text Using BLSTM Neural Network", IEEE, 2012.
5. N. Hewahi, M.Al Nono, M. Nasar, M. Hamed, H. Hamed, "Chemical Ring Handwritten Recognition Based on Neural Networks", Ubiquitous Computing And Communication Journal 3, no. 3, 2008.
6. Ouyang T.Y., Davis R., "Recognition of Hand Drawn Chemical Diagrams", In Proc. of the National Conf. on Artificial Intelligence, 2007, pp.846-851.
7. Tenneson D., Becker S. "ChemPad: Generating 3D Molecules From 2D Sketches", In Proceedings of SIGGRAPH'05: ACM SIGGRAPH 2005 Posters. 2005. 8.
8. J. Ramel, G. Boissier, H. Emptoz, "Automatic Reading of Handwritten Chemical Formulas from a Structural Representation of the Image", In Proc. of 5th Intl. Conf. on Document Analysis and Recognition, 1999, pp. 83-86.
9. J.F. Yang, G.S. Shi, K. Wang, Q. Geng, Q.R. Wang, "A Study of On-line Handwritten Chemical Expressions Recognition", In Proc. of 19th Intl. Conf. on Patten Recognition, 2008.

10. Ming Chang, Shi Han, Dongmei Zhang, " A Unified Framework for Recognizing Handwritten Chemical Expressions", 2009 10th International Conference on Document Analysis and Recognition, 2009 IEEE DOI 10.1109/ICDAR.2009.64 1345.

11. A. Aldroubi, M. Unser and M. Eden, "Cardinal spline filters : stability and convergence to the ideal sinc interpolator", Signal Processing, Vol. 28, No. 2, pp. 127-138, August 1992.

12. C.K. Chui and J.Z. Wang, "On compactly supported spline wavelets and a duality principle", Trans. Amer. Math. Soc., Vol. 330, No. 2, pp. 903-915, 1992.

13. Viktor Smolov , Fedor Zentsev and Mikhail Rybalkin, " Imago: open-source toolkit for 2D chemical structure image recognition ", GGA Software Services LLC.

14. M. E. Algorri, M. Zimmermann, C. M. Friedrich, S. Akle, and M. Hofmann-Apitius, " Reconstruction of Chemical Molecules from Images", In Proc. of EMBS 2007, pages 4609–4612, 2007. 4.3.

15. M. E. Algorri, M. Zimmermann, and M. Hofmann-Apitius, "Automatic Recognition of Chemical Images", In Proc. of ENC 2007, pages 41–46, 2007. 4.3.

16. Tetsuo Asano, Danny Z. Chen, Naoki Katoh, and Takeshi Tokuyama, "Polynomial-time Solutions to Image Segmentation", In Proceedings of the seventh annual ACM-SIAM symposium on Discrete algorithms, SODA '96, pages 104–113, Philadelphia, PA, USA, 1996. Society for Industrial and Applied Mathematics.

17. Christopher M. Bishop, "Pattern Recognition and Machine Learning (Information Science and Statistics)", Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.