HMM-based Online Recognition of Handwritten Chemical Symbols

Yang Zhang, Guangshun Shi, Jufeng Yang
Institute of Machine Intelligence, Nankai University, Tianjin China
{yangzhang, gsshi, jfyang}@imi.nankai.edu.cn

Abstract

In this paper, we present an online handwritten recognition method for Chemical Symbols, a widely used symbol in education and academic interactions. This method is based on Hidden Markov Models (HMMs), which are increasingly being used to model characters. We built an HMM for each symbol and used 11-dimensional local features which are suitable for online handwritten recognition, and obtained top-1 accuracy of 89.5% and top-3 accuracy of 98.7% on a dataset containing 5,670 train samples and 2,016 test samples. These initial results are promising and warrant further research in this direction.

1. Introduction

With the growing popularity of pen-based input devices, precise and robust handwritten recognition technologies have become the key to extend the application of pen-based device. Among them, the handwritten symbols recognition is a very active area. Haton[1] used the structure-based methods; Chou[2] applied the traditional template matching method to character recognition; still other methods based on neural networks [3], and statistical methods [4]. These technologies are widely used in western characters, CJK, mathematical symbols or other symbols and have achieved good results.

However, in the field of handwritten character recognition, how to recognize the chemical symbols in chemical expression is still a challenging problem. Due to the special structure of chemical expressions, first of all, pen-based input often results in broken and conglutinate characters; In addition, cursive and freely handwriting with random noise is unavoidable because of various human factors and different writing styles; again, the online characteristic of handwritten recognition also requires a certain degree of recognition speed; Finally, as the basic component of the chemical expressions, correct recognition of the chemical symbols will have an significant effect on the following understanding of the chemical expressions.

To our best knowledge, related and published works are still scarce, and more on offline recognition of the chemical structure and expressions. Casey and others [5] proposed a prototype encoding handwritten chemical expression in printed documents, the system used spatial features to locate and computed a convex polygon to describe the expression. Ramel [6] proposed a method recognizing the graphic entities in handwritten chemical expressions and acquired accuracy of 93% in a small 300dpi offline document set. Until recently, a small number of works concerning online recognition of handwritten chemical symbols occurred. Jiang and others [7] proposed a recognition algorithm based on online segmentation. This algorithm used the local optimization rather than global optimization. Jufeng Yang [8] proposed an online handwritten chemical expression recognition algorithm, this algorithm recognized in substance level and symbols level, meantime, was assisted in HCI to modify the results.

In this paper, we mainly focus on online recognition of handwritten chemical symbols. In our experiment, we use 11-dimensional local features which are suitable for online recognition and build an HMM for each chemical symbol to train. The reminder of this paper is organized as follows: Section 2 simply introduces the method including our thoughts, the features used and HMM classifier. Section 3 presents the application including preprocessing, feature extraction and the strategy of parameter training and symbol recognition based on the method in Section 2. Section 4 gives the dataset, analysis for experimental results and some possible improvements in our method. Section 5 concludes the paper and puts forward the prospect of our future research.

2. HMM-based recognition method

Handwritten chemical symbol is a temporal sequence of X-Y points captured using a digitizer that senses the pen-tip position while writing, it has online

978-0-7695-3725-2/09 $25.00 © 2009 IEEE
DOI 10.1109/ICDAR.2009.99

properties. In addition, as a result of arbitrary handwriting and different writing styles, there are some noises, broken and conglutinate strokes, which are inevitable and make the recognition difficult. However, the stochastic model can effectively deal with the noise and variation caused by handwriting. As a stochastic model, hidden markov models are suitable for handwritten recognition for a number of reasons [9]. Hence, in this work, HMM, which are shown to be successful for western cursive recognition, and CJK script recognition, are applied to model chemical symbols.

There is no doubt that recognition performance depends largely on the quality of the features, therefore, we must select the appropriate features according to different recognition. Different from the offline handwritten recognition, online recognition are more dependent on local features which can properly describe the inherent nature of handwritten symbols. For each preprocessed sample, we extract 11-dimensional local features at each point of a stroke. The features selected are normalized vertical position, normalized first-order, two-order derivative, curvature which describes the extent of bending, writing direction characterizing the change of direction, aspect ratio, and linearity. These features are computed in Section 3.

HMM is a doubly embedded stochastic process. The stochastic process which describes the state transition is hidden, and the transitions are observed through another stochastic process that generates the output observations. According to the type of the output observations, HMMs are classified into discrete and continuous HMMs. In our experiment, continuous HMMs are chosen to model the chemical symbols, and the chosen HMMs have a left-to-right topology known as Bakis model [10] which takes into account the temporal order of the signal.

In this paper, an HMM is described as $\lambda = \{\pi, A, B\}$, where π denotes the initial distribution of the states, A denotes the transition matrix of the states, and B is used to describe the probability of the generated observations in a state.

3. Recognition application

3.1. Preprocessing

Before features are derived from the handwritten symbol, the raw data recorded by the Tablet-PC should undergo a series of processing steps. The process is presented as in Figure 1.

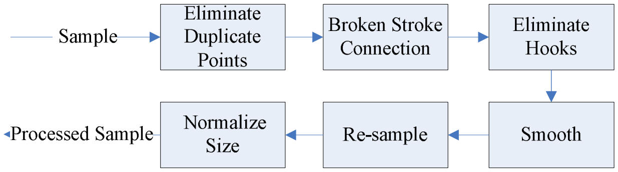

Figure 1. Preprocessing

Duplicate points Elimination: So-called duplicate points are that those who have identical coordinates in a stroke and can not provide any useful information. Therefore, they should be eliminated before feature extraction.
Connect broken strokes: Due to various writing style, emotion, writing speed of writers, and the sampling rate of machine, broken strokes that occurred when writing would affect the characteristics of symbols. So, we use Bezie curve [11] to interpolate missing points.
Eliminate hooks: Hooks often appear in the beginning or the end of a stroke. They are characterized by short length and great change in perspective which results in "*false*" characteristics, they must be removed.
Smoothing: Smoothing is used to remove the noises caused by erratic movements of the pen.
Re-sampling: In digital ink capture, some points appear equidistant in time but not in space. In order to remove this variation, we re-sample each sample keeping the original trajectory.
Size normalization: This step is necessary to remove the variation of different writing size.

3.2. Feature extraction

In order to obtain the sequence of feature vectors, for each sample, we choose a small area of five-point which contains the current point (x_t, y_t), and two points to the left and two points to the right of the current point. Our features [12][13] are computed according to Figure 2 as below.

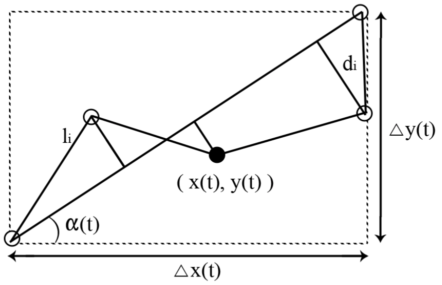

Figure 2. Five-point area

Normalized vertical position: Each sample sequence, $(x(t), y(t))$, where $t = 1, ..., T$, are scaled and translated to get the new vertical position which is mapped into the range [0, 1], preserving the original aspect-ratio of the sample.

Normalized first derivatives: x', y' are calculated as:

$$x_t' = \frac{\Sigma_{i=1}^{2} i \cdot (x_{t+i} - x_{t-i})}{2 \cdot \Sigma_{i=1}^{2} i^2} \qquad y_t' = \frac{\Sigma_{i=1}^{2} i \cdot (y_{t+i} - y_{t-i})}{2 \cdot \Sigma_{i=1}^{2} i^2}$$

$$\hat{x}_t' = \frac{x_t'}{\sqrt{x_t'^2 + y_t'^2}} \qquad \hat{y}_t' = \frac{y_t'}{\sqrt{x_t'^2 + y_t'^2}}$$

The range of i corresponds to the size of the small area of five-point.

Normalized second derivatives: The second normalized derivatives are computed as the first normalized derivatives by using x' and y' other than x and y.

Curvature: Curvature describes the degree of bending of the curve and it is calculated as below:

$$k_t = \frac{x' \cdot y'' - x'' \cdot y'}{(x'^2 + y'^2)^{3/2}}$$

Writing direction: The local writing direction at the current point (x_t, y_t) is calculated using the cosine and sine given below:

$$\cos \alpha(t) = \frac{x(t-1) - x(t+1)}{\sqrt{\sum_{k=x,y} (k(t-1) - k(t+1))^2}}$$

Sine is calculated as cosine by replacing the numerator $x(t-1) - x(t+1)$ with $y(t-1) - y(t+1)$.

Aspect: The aspect of a point characterizes the ratio of the height to the width of the small five-point area. It is computed as:

$$A(t) = \frac{\Delta y(t) - \Delta x(t)}{\Delta y(t) + \Delta x(t)}$$

Where Δx and Δy respectively represent the width and the height of the small five-point area in Figure 2.

Curliness: The curliness is a feature that characterizes the derivation from the straight line connecting the first point and the last point in the small area in Figure 2. It is calculated as:

$$C(t) = \frac{\sum_i l_i(t)}{\max(\Delta x, \Delta y)} - 2$$

where $l_i(t)$ denotes the length of the ith segment between the neighboring points.

Linearity: The linearity characterizes the linear property. It is calculated as:

$$LN(t) = \frac{1}{N} \sum_i d_i^2$$

where d_i denotes the distance of each point in the small area to the straight line connecting the first and the last point.

Finally, each sample is transformed into a matrix $V = [v_1, ..., v_T]$, where $v_t (t = 1, ..., T)$ represents the 11-dimensional real-valued feature vector. Then we perform the final signal normalization, which can provide the zero mean and unit standard deviation values. It is calculated as:

$$o_t = \Sigma^{-1/2}(v_t - \mu), (t = 1, ..., T)$$

where μ and Σ are the sample mean and sample diagonal covariance matrix . As a result, each sample is represented by a feature matrix .

3.3. Symbol model and training

For each symbol of the chemical symbols set in Section 4, we build an HMM and use the feature matrix calculated in Section 3.2 to train the HMM. Let $\lambda = \{\pi, A, B\}$ represents an HMM. Then, we can use K feature matrix $\{O^{(1)}, ..., O^{(K)}\}$ corresponding to K training samples, where $O^{(k)} = [o_1^k, ..., o_{N_k}^k], k = 1, ..., K$, we use the following training strategy:

i Initialize HMM

π : Let the initial distribution $\pi = \{\pi_1, \pi_2, ..., \pi_N\}$ be $\{1, 0, ..., 0\}$, where N denotes the number of states.

A : We use the uniform distribution to obtain the initial parameter of transition matrix A with considering the restrictions in the left-to-right HMM model [10].

B : We use the segmental k-means method [14] to obtain the initial parameter of B. A good initial B will improve the speed of the convergence.

ii Re-estimate parameter

In the re-estimation step, we used the remarked Baum-Welch re-estimation formula [10] which can warrant that:

$$\Sigma_{k=1}^{K} \log(P(o^{(k)} \mid \bar{\lambda})) \geq \Sigma_{k=1}^{K} \log(P(o^{(k)} \mid \lambda))$$

where λ denotes the original HMM and $\bar{\lambda}$ the re-estimated HMM.

iii Iteration

We replace the original HMM λ with the re-estimated $\bar{\lambda}$ until the iteration satisfies the maximum iterations or the condition as below:

$$\frac{\sum_{k=1}^{K}\log(P(o^{(k)}|\bar{\lambda}))-\sum_{k=1}^{K}\log(P(o^{(k)}|\lambda))}{\sum_{k=1}^{K}\log(P(o^{(k)}|\lambda))} \le \Theta$$

where $O^{(k)} = [o_1^k, ..., o_{N_k}^k], k = 1, ..., K$ denotes the feature matrix of kth training sample, Θ denotes the threshold selected empirically.

After the training, we use the maximum likelihood law to make decision:

$$\lambda^* = \arg\max_i P(O \mid \lambda_i)$$

For a test sample, the probability associated with each of the HMMs is calculated and the symbol whose HMM has the maximum probability is the recognition result.

4. Dataset and experiment results

Our symbol set contains the basic chemical symbols. They are usually used to constitute the chemical substance, chemical structure, and to describe the chemical reaction process. They are listed below:

Table 1. Chemical symbols set

0	1	2	3	4	5	6
7	8	9	A	B	C	D
E	F	G	H	I	K	L
M	N	O	P	R	S	T
U	V	W	X	Y	Z	a
b	c	d	e	f	g	h
i	k	l	m	n	o	p
r	s	t	u	y	=	+
-	→	↑	↓	□	(	)

In the laboratory environment, we have invited 20 college teachers and students coming from chemical and non-chemical field to write samples. These samples are tagged in character level and marked with the hardware environment, samples specification and other information. For each symbol, we collected 122 valid samples, a total of 7808 samples. The train set contains 5670 samples (90 for each symbol) and test set contains 2016 samples(32 for each symbol). For various combinations of state and Gaussians, the experiment results on the test set are given in Table 2.

Table 2. Recognition performance of the method forvarious combinations of state and Gaussians.

Gaussian	state	Accuracy (%)		
		Top-1	Top-2	Top-3
3	4	85.9	94.5	96.7
	6	87.1	95.3	97.6
	8	88.2	96.1	98.0
6	4	87.9	95.9	98.4
	6	88.8	96.6	98.3
	8	89.0	96.7	98.5
9	4	87.9	95.9	98.3
	6	**89.5**	97.0	98.4
	8	89.3	97.0	98.7
12	4	88.9	96.7	98.5
	6	89.3	96.9	98.3
	8	89.0	97.2	98.6

The figures in Table 2 show that: In our experiment, the highest top-1 accuracy is 89.5% corresponding for the combination of 6 state and 9 Gaussians per state. In addition, the more the state and Gaussians per state, the better the recognition result. However, the cost of the machine will be greater and the efficiency will be reduced, therefore, the parameter of HMMs must be selected between recognition accuracy and efficiency according to different online requirements. Through our data analysis, classification errors appear in such cases where the symbols are similar to each other, such as the number “0”, the capital letter “O” and the lowercase letter “o”; capital letter “C”, lowercase letter “c” and the operator “(”. Fortunately, these symbols appear in only a small amount. To solve this problem, on the one hand, we can improve the recognition rate by providing candidate results. For the same combination of HMM parameters, the difference between top-1 accuracy and top-3 accuracy in Table 2 also illustrates this point. On the other hand, we can improve the recognition results by establishing discriminatory classifiers targeting those confused symbols or constituting combination of multiple classifiers. In addition, We can also train HMMs with different combination of states and Gaussians for different symbols according to their difference in shape and other features.

5. Conclusions and outlook

This paper proposed a method for online recognition of handwritten chemical symbols. Our method is based on HMM and we used the 11-dimensional normalized local features which are more suitable for online recognition. For various combinations of states and Gaussians per state, we obtained various accuracy of recognition, these initial results are promising and warrant our further research in this direction. Although this method is designed for chemical symbols, it is independent from chemical symbols and can be applied to other symbols recognition.

In the future, we will extend our research on recognition of chemical symbols to the recognition of chemical expressions and analysis of complex chemical structures.

6. Acknowledgements

This work is supported in part by Microsoft Research Asia (MSRA) under Mobile Computing in Education Theme.

7. References

[1] A. Belάid and J. P. Haton, "A syntactic approach for handwritten mathematical formula recognition", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(1):105-111, Jan. 1984.

[2] P.A. Chou, "Recognition of equations using a two-dimensional stochastic context-free grammar", Proceedings of the SPIE Visual Communications and Image Processing IV, volumn 1199, pp 852-863, Philadelphic, PA, Nov. 1989.

[3] Y. A. Dimitriadis, J. L. Coronado, and C. de la Maza, "A new interactive mathematical editor, using on-line handwritten symbol recognition, and error detection-correction with an attribute grammar", *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*, pp 344-351. 1991.

[4] L. H. Chen and P. Y. Yin, "A system for on-line recognition of handwritten mathematical expression", *Computer Processing of Chinese and Oriental Languages*, 6(1):19-39, June 1992.

[5] R. Casey, S. Boyer, P.Healy, A. Miller, B. Oudot, K. Zilles, "Optical Recognition of Chemical Graphics", *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*, pp 627-631. 1993.

[6] J. Ramel, G. Boissier, and H. Emptoz, "Automatic reading of handwritten chemical formulas from a structural representation of the image", *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*, Bangalore India, pp 83-86. 1999.

[7] Y. Jiang, X, Wang, X. Ao, G. Dai, "Online Recognition of Handwritten Chemical Formula", *Proceedings of the 2nd Joint Conference on Harmonious Human Machine Environment*, Hangzhou, China, 2006.

[8] Jufeng. Yang, Guangshun. Shi, Qingren. Wang, "Recognition of On-line Handwritten Chemical Expression", *Proceedings of International Joint Conference on neural networks*, pp 2360-2365, 2008.

[9] Junko TOKUNO, "Context-dependent Substroke Model for HMM-based On-line Handwriting Recognition", *8th International Workshop on Frontiers in Handwriting Recognition (IWFHR-8)*, pp. 78-83, 2002.

[10] L. Rabiner, A Tutorial of Hidden Markov Models and Selected Application in Speech Recognition. *Proc. IEEE*, 77:257-286, 1989.

[11] S. Abramowski, H. Müller, : Geometrisches Modellieren (in German), *Reihe Informatik*. Vol. 75, 1991.

[12] M. Pastor, A. Toselli, and E. Vidal, "Writing Speed Normalization for On-line Handwritten Text Recognition", *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*, 2005.

[13] S. Jaeger, S. Manke, J. Reichert, A. Waibel, "Online handwriting recognition: the NPen++ recognizer", International Journal on Document Analysis and Recognition. March, 2001, vol.3(3), pp. 169-180.

[14] L.R. Rabiner, B.H. Juang, S.E. Levinson, and M.M. Sondhi, "Recognition of isolated digits using hidden markov models with continuous mixture densities", AT&T Tech. J., vol. 64, no. 6, pp. 1211-1222, July-Aug. 1986.