Reprint: 5th Int. Workshop on Frontiers in Handwriting Recognition (IWFHR-5), Univ. of Essex, Sep. 1996, in „Progress in Handwriting Recognition (Ed.: A.C. Downton, S. Impedovo)“, World Scientific, 1997

SYMBOL SEGMENTATION AND RECOGNITION FOR UNDERSTANDING HANDWRITTEN MATHEMATICAL EXPRESSIONS

HANS-JÜRGEN WINKLER, MANFRED LANG
Institute for Human-Machine-Communication, Munich University of Technology, Arcisstr. 21, 80290 Munich, Germany

This paper is focused on the symbol segmentation and recognition problem within on-line sampled handwritten expressions, the first stage of an overall system for understanding arithmetic formulas. Within our system a statistical approach is used tolerating ambiguities within the single decision stages and resolving them either automatically by additional knowledge acquired during the following processing stages or by interaction with the user. At this state the interaction is done by displaying next to the recognition result, the most probable symbol sequence corresponding to the handwritten input, some recognition alternatives for selection by the user.

1 Introduction

We are accustomed in writing mathematical expressions containing integrals, fractions, exponents or indices by hand, but there is no human-adapted way to enter these expressions into a computer. A comfortable possibility is the analysis of the handwriting, but due to the fact that mathematical expressions contain two-dimensional information, there are two basic problems to be solved [1]: first, focused in this paper, symbol segmentation and recognition and next structure analysis for extracting the meaning of the two-dimensional symbol positioning [2][3].

2 Soft-decision approach

Our system is based on the on-line sampled handwriting data. Hence, the input data consists of a sequence of strokes $L = (L_1, L_2, \ldots, L_N)$ sampled during writing. Within the segmentation stage a symbol hypotheses net (SHN) is generated containing symbol hypotheses $G(n, g)$ consisting of the stroke sub-sequence $(L_n, \ldots, L_{n+g})$, $0 \le g \le 3$, $1 \le n \le N - g$, of the handwritten input L. Thus, soft-decision segmentation is done transforming the stroke sequence L into one or more different sequences $G^{(i)}$ of symbol hypotheses represented by the corresponding path through the SHN. Each symbol hypotheses is classified by a symbol recognition system based on Hidden Markov Models (HMMs) assigning d different symbol recognition results $S(n, g, d)$ to each symbol hypotheses $G(n, g)$ of the SHN. Hence, this classification is a soft-decision process again, transforming each symbol hypotheses sequence $G^{(i)}$ into different symbol sequences $S^{(j)}$.

Each decision within the segmentation and recognition stage is combined with a certain decision probability resulting to the corresponding sequence probabilities $P(G^{(i)}|L)$ and $P(S^{(j)}|G^{(i)})$. The final classification of the symbol sequence $S_F^{(1)}$ and its alternatives $S_F^{(k)}$, $k > 1$, is based on these probabilities.

Reprint: 5th Int. Workshop on Frontiers in Handwriting Recognition (IWFHR-5), Univ. of Essex, Sep. 1996, in „Progress in Handwriting Recognition (Ed.: A.C. Downton, S. Impedovo)“, World Scientific, 1997

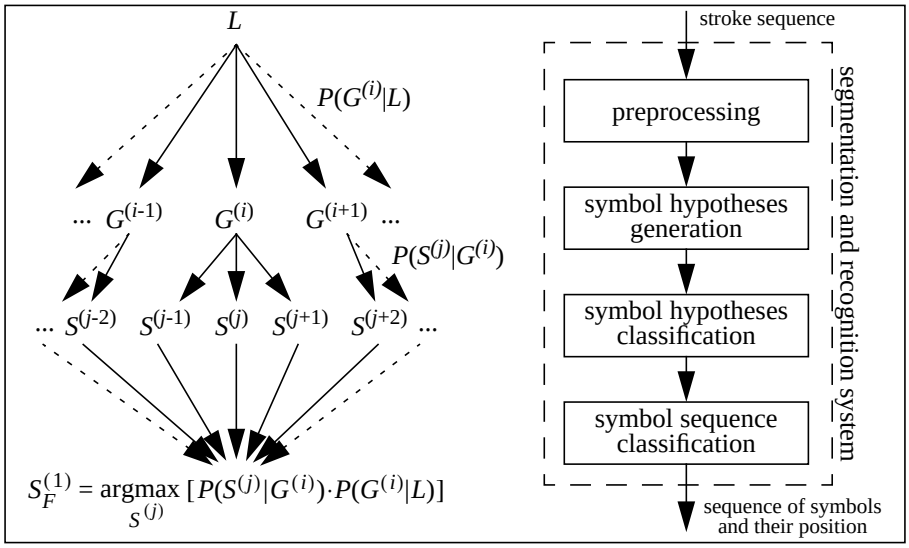

Figure 1: Soft-decision approach and the corresponding system overview.

3 MAIN STAGES OF THE SYSTEM

3.1 Preprocessing

Based on the incoming temporally equispaced samples (pen positions) of the strokes several preprocessing steps are carried out:

- smoothing of the data by lowpass-filtering.
- slant detection by averaging the near-vertical stroke parts within the sequence L under regard to their height and its correction by carrying out a shear.
- determination of the reference size representing a measurement for the writing size. Caused by the two-dimensional symbol positioning within mathematical expressions, a normalization to a standard height or width is nonsensically. Hence, the reference size determination has to be based on the size of the strokes, the only information available at this stage. In detail, the area the surrounding rectangle of each stroke is calculated, the square root of each stroke area is determined and their median value is assigned to the reference size.
- the temporally equispaced points of each stroke are re-sampled spatially along the stroke trajectory retaining the temporal order. The re-sampling distance between two successive points is determined under regard to the reference size.

3.2 Generating the symbol hypotheses net (SHN)

Using a-priori knowledge available by prerequisites concerning the style of writing and by restricting the alphabet to currently 84 symbols, a mixed knowledge-based and

statistical technique is used for generating the symbol hypotheses net [2]. In comparison to other systems only analyzing stroke distances, the generation within our system is based on stroke-specific as well as geometrical features between strokes using additional knowledge obtained by a symbol pre-recognition stage [4].

By using this approach, a probability measurement $0 \le \tilde{P}(G(n, g)|L) \le 1$ is obtained for each possible stroke group $G(n, g)$. This probability is used for soft-decision external segmentation, i.e. only stroke groups with $\tilde{P}(G(n, g)|L) > 0$ are regarded as a potential symbol (symbol hypotheses) of the handwritten expression and, hence, only these stroke groups are included into the SHN.

Finally, probability normalization is done transferring $\tilde{P}(G(n, g)|L)$ into the probability $P(G(n, g)|L)$ influencing only the absolute value of a path through the SHN but not the value in relation to another path probability. The probability of the path i through the SHN, i.e. the symbol hypotheses sequence $G^{(i)}$, is determined by the product of the probabilities of the path elements (see eq. (1)).

3.3 Symbol hypotheses classification

Evaluating the temporal information is the most obvious kind for recognizing on-line sampled handwriting [5]. On the other hand, remarkable recognition results are also concerned with image based recognition strategies used either exclusive in recognizing off-line sampled handwriting [6] or as additional feature next to the temporal information available by on-line sampling [7].

The symbol hypotheses classification system should be able to support the already done soft-decision external pre-segmentation by an internal segmentation through symbol recognition. Focusing the symbol recognition task itself, the recognition rate should be as high as possible. Furthermore, for an automatic error correction by determining recognition alternatives, the reliability of the results should be high, if the result is correct, or low, if the result is wrong.

To meet all requirements, within our classification system three different feature vector sequences are extracted by each not pre-recognized symbol hypotheses $G_R(n, g)$ of the SHN [8]. One of these feature vector sequences is generated by analyzing the temporal information of writing [8], the two remaining sequences and are extracted by the image, i.e. the result of the writing [2].

Again, soft-decision is done by considering the d most probable recognizer results $S(n, g, d)$ to each symbol hypotheses $G_R(n, g)$. This decision is based on a combined generation probability $P(G_R(n, g)|S(n, g, d))$ determined by a weighted multiplication of normalized generation probabilities, each of them obtained by a HMM-based recognizer analyzing the different feature vector sequences. The normalization of the generation probabilities is done to the number of feature vectors within each sequence considering that each feature extraction algorithm results in a different number of feature vectors.

For homogeneity, a generation probability $P(G_P(n, g)|S(n, g, d))$ analogous to the HMM-based symbol classification is assigned to the symbol hypotheses $G_P(n, g)$ pre-recognized by the pre-recognition stage within the symbol hypotheses classification stage. The pre-recognition is done before the HMM-based classification for separating the symbol hypotheses $G_P(n, g)$ representing the symbols „Dot“, „Minus“, and „Fraction“ from the symbol hypotheses $G_R(n, g)$ representing any of the remaining symbols of the alphabet [4].

The main reason for this separation is that on the hand no writing is done for the making of a „Dot“ and on the other hand the distinction of the symbols „Minus“ and „Fraction“, both mostly represented by a horizontal line, requires knowledge about the positions of the remaining SHN-elements. The pre-recognition process itself is almost identical the stage used during the symbol hypotheses generation, in contrast to its first use this time symbol hypotheses (up to four strokes) instead of single strokes are applied and ambiguous recognition results are considered.

The determination of the „generation probability“ $P(G_P(n, g)|S(n, g, 1))$ of the most probable pre-recognition result $S(n, g, 1)$ is done by using the median value of the most probable generation probabilities $P(G_R(n, g)|S(n, g, 1))$ obtained by the HMM-based classification of the symbol hypotheses $G_R(n, g)$. In case of an ambiguous classification, the generation probability $P(G_P(n, g)|S(n, g, 2))$ of the recognition alternative $S(n, g, 2)$ is fixed to $P(G_P(n, g)|S(n, g, 1)) - \Delta P$, $\Delta P \to +0$.

3.4 Symbol sequence classification

The information available at this final decision stage consists of the probabilities $P(G(n, g)|L) > 0$ obtained by generating the symbol hypotheses $G(n, g)$ and the probabilities $P(G(n, g)|S(n, g, d))$ obtained by their recognition.

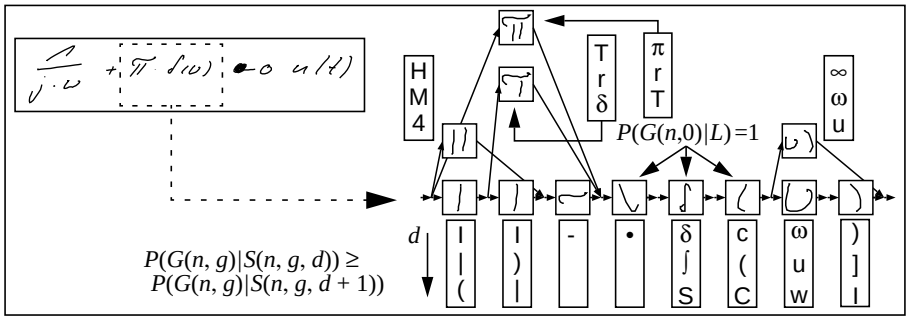

Figure 2: Image and the SHN generated by analyzing the on-line sampled data, the top-3 recognition results are shown next to each symbol hypotheses.

Using the SHN-elements, the decision criterion given in fig. 1 can be transformed to:

$$S_F^{(1)} = \underset{S^{(j)}}{\operatorname{argmax}}\left[\prod_{S(n, g, d) \in \text{path } j} P(S(n, g, d)|G(n, g)) \cdot P(G(n, g)|L)\right]. \qquad (1)$$

By using the Bayes theorem and assuming, that all „a-priori“-probabilities are constants ($P(G(n, g)) = P_G$ and $P(S(n, g, d)) = P_S$), the decision criterion results in:

$$S_F^{(1)} = \underset{S^{(j)}}{\operatorname{argmax}}\left[\left(\frac{P_S}{P_G}\right)^{N(S(n,g,d) \in \text{ path } j)} \cdot \prod_{S(n,g,d) \in \text{ path } j} P(G(n,g)|S(n,g,d)) \cdot P(G(n,g)|L) \right]. \quad (2)$$

Focusing the left term within this decision criterion, either long or short paths through the SHN will be preferred depending on the relation of the „a-priories“. Within the right part, always short paths will be preferred caused by the normalization done during the symbol hypotheses generation and classification. Hence, normalization is necessary to the number of elements $N(S(n, g, d))$ within the path.

The final decision criterion for the most probable symbol sequence based on the handwritten input and its recognition alternatives results by this normalization in:

$$\tilde{S}_F^{(k)} = \underset{S^{(j)} \notin \{\tilde{S}_F^1, \ldots, \tilde{S}_F^{(k-1)}\}}{\operatorname{argmax}} \left[\sqrt[N(S(n,g,d) \in \text{ path } j)]{\prod_{S(n,g,d) \in \text{ path } j} P(G(n,g)|S(n,g,d)) \cdot P(G(n,g)|L)} \right]. \quad (3)$$

Finally, a verification concerning the mathematical syntax based on the number of parentheses, brackets, braces and absolute value symbols is carried out. Recognition results failing this verification are considered as invalid.

4 WRITER-DEPENDENT RECOGNITION EXPERIMENT

For the recognition experiment nine writer contributed five versions of 17 different expressions. The number of symbols within the expressions are ranging from at least 13 up to 45 symbols, on average an expressions consists of 27 symbols (in comparison: english words contain 5 characters on average).

Within this experiment the 10 most probable symbol sequences are generated by the decision criterion given in eq. (3) without using any language model but using the knowledge obtained by the (in this state quite simple) syntax verification. In tab. 1, the averaged expression-normalized recognition result R_E determined by

$$R_E = \frac{N(\text{ error-free segmented and recognized expression by } \tilde{S}_F^{(k)})}{N(\text{expressions} \in \text{ test data base})} \quad (4)$$

and its ranges depending on the writer as well as on the expression are given. Some samples of well and poor analyzed expressions are given in [9].

Table 1: Symbol sequence classification: average results and the writer- and expression-depending ranges.

Recognition rate R_E	Correct symbol sequence within $\tilde{S}_F^{(k)}$,				
	$k = 1$	$k \leq 2$	$k \leq 3$	$k \leq 4$	$k \leq 10$
Average:	44%	56%	60%	63%	72%
Writer-dep. range:	28% - 68%	39% - 79%	40% - 84%	45% - 84%	54% - 87%
Expr.-dep. range:	2% - 91%	4% - 96%	11% - 96%	11% - 96%	16% - 100%

Normalizing the recognition results to the number of symbols for achieving independency of the complexity of the expressions within the test data base, a symbol-normalized recognition rate R_S based on the most probable symbol sequence $\tilde{S}_F^{(1)}$ can be determined analogous to [1] by:

$$R_S = \frac{\sum_{\text{expressions} \in \text{ test data base}} N(\text{correct segmented and recognized symbols} \in \tilde{S}_F^{(1)})}{N(\text{symbols} \in \text{ test data base})}. \qquad (5)$$

In this case, the average recognition rate R_S results in more than 95% on average ranging between 93% and 98% for the different writers and between 91% and 99% for the different expressions of the test data set. If an error occurs, in 80% of all cases the error is only based on wrong symbol recognizer decisions which are, furthermore, mainly caused by a mix-up of upper and lower case letters having the same shape [8].

5 REFERENCES

[1] H.-J. Lee, M.-C. Lee: *Understanding Mathematical Expressions using Procedure-Oriented Transformations*, Pattern Recognition 27(3), pp. 447-457, 1994.

[2] M. Koschinski, H.-J. Winkler, M. Lang: *Segmentation and Recognition of Symbols within Handwritten Mathematical Expressions*, Int. Conf. on Acoustics, Speech, and Signal Processing ICASSP-95, Detroit, pp. 2439-2442, May1995.

[3] H.-J. Winkler, H. Fahrner, M. Lang: *A Soft-Decision Approach for Structural Analysis of Handwritten Mathematical Expressions*, ICASSP-95, Detroit, pp. 2459-2462, May 1995.

[4] S. Lehmberg, H.-J. Winkler, M. Lang: *A Soft-Decision Approach for Symbol Segmentation within Handwritten Mathematical Expressions*, ICASSP-96, Atlanta, pp. 3434-3437, May 1996.

[5] E.J. Bellegarda, J.R. Bellegarda, D. Nahamoo, K.S. Nathan: *A Probabilistic Framework for On-line Handwriting Recognition*, 3rd Int. Workshop on Frontiers in Handwriting Recognition IWFHR-3, Buffalo, pp. 225-234, May 1993.

[6] T. Caesar, J. Gloger, A. Kaltenmaier, E. Mandler: *Recognition of Handwritten Word Images by Statistical Methods*, IWFHR-3, Buffalo, pp. 409-416, May 1993.

[7] R.F. Lyon, L.S. Yaeger: *On-line Hand-Printing Recognition with Neural Networks*, 5th Int. Conf. on Microelectronics for Neural Networks and Fuzzy Systems, Lausanne, Switzerland, pp. 201-212, Feb. 1996.

[8] H.-J. Winkler: *HMM-based Handwritten Symbol Recognition using On-line and Off-line Features*, ICASSP-96, Atlanta, pp. 3438-3441, May 1996.

[9] H.-J. Winkler, M. Lang: *On-line Symbol Segmentation and Recognition in Handwritten Mathematical Expressions*, to be published at ICASSP-97, Munich, Germany, April 1997.