Reprint: 1995 Int. Conference on Acoustics, Speech, and Signal Processing (ICASSP-95), Detroit, USA, Vol. 4, pp. 2439-2442, May 1995

SEGMENTATION AND RECOGNITION OF SYMBOLS WITHIN HANDWRITTEN MATHEMATICAL EXPRESSIONS

M. Koschinski, H.-J. Winkler, M. Lang
Institute for Human-Machine-Communication
University of Technology, Munich
Arcisstr. 21, 80333 Munich, Germany

ABSTRACT

In this paper an efficient on-line recognition system for symbols within handwritten mathematical expressions is proposed. The system is based on the generation of a symbol hypotheses net and the classification of the elements within the net. The final classification is done by calculating the most probable path through the net under regard of the stroke group probabilities and the probabilities obtained by the symbol recognizer based on Hidden Markov Models.

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 would be the analysis of the handwriting, but due to the fact that a mathematical formula contains two-dimensional information, there are two problems to be solved: symbol recognition and structure analysis.
In this paper we will focus on the symbol recognition process. Knowledge resulting from this process will drive the structure analysis.

2 SYSTEM OVERVIEW

Handwriting recognition can be carried out under two different environments: off-line and on-line [1]. Off-line recognition means that the data are captured after the writing is completed. On-line recognition captures the data during the writing by using a stylus and an electronic tablet connected to a computer. The advantage is that the temporal information of the writing like the number of strokes, the order of the strokes and the direction of the writing of each stroke is available. A stroke, in this connection, is the writing from pen down to pen up.

Our recognition system is based on the on-line sampled data, therefore the input data consists of a sequence I of strokes, each stroke itself is represented by a sequence of (x,y)-coordinates corresponding to the pen positions. Due to the fact that most symbols are composed of more than one stroke, a grouping process is necessary collecting together the strokes which belong to the same symbol. Therefore, by using a soft-decision process, a symbol hypotheses net is generated, which transforms the sequence I of strokes into different sequences G_i of stroke groups.
The elements $G_i(n)$ within each stroke group sequence are classified by using a Hidden Markov Model (HMM) [2]. Again, the classification is a soft-decision process, thus each stroke group sequence G_i is transformed into one or more different sequences S_j of symbols.
The final classification of the symbols sequence S_F within the handwritten mathematical expression is done by calculating the most probable sequence S_j of symbols.
This recognition strategy and the corresponding block diagram are illustrated by fig. 1.

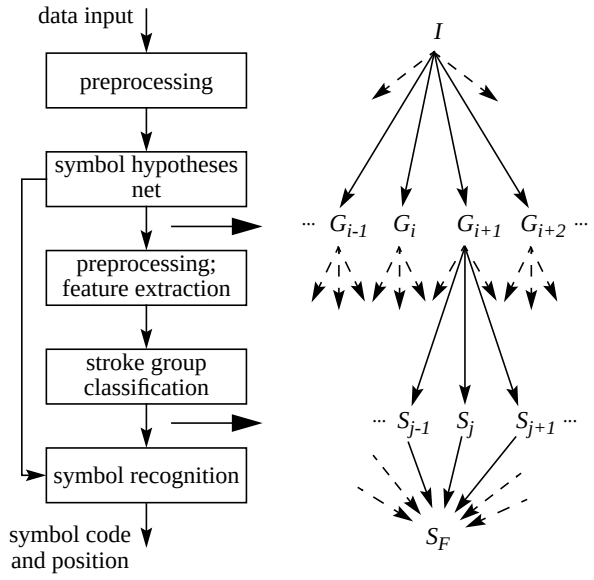

Figure 1: System overview and the corresponding recognition strategy

Reprint: 1995 Int. Conference on Acoustics, Speech, and Signal Processing (ICASSP-95), Detroit, USA, Vol. 4, pp. 2439-2442, May 1995

The output data of our recognition system consists of the symbol codes and their position. These data will drive the structure analysis process [3][4].

3 GENERATING A SYMBOL HYPOTHESES NET

3.1 Preprocessing

The preprocessing stage is subdivided into three steps:

- smoothing the input data by lowpass-filtering.
- slant correction by carrying out a shear. The shear angle is calculated by averaging the near-vertical parts of each stroke [5].
- a reference length is calculated by taking the median stroke length within the input data. The features generated in the following section are normalized to this reference length.

3.2 Symbol hypotheses net

To achieve reasonable results by generating the hypotheses net, the writer has to fulfil a few prerequisites:

- the handwritten symbols must be an element of the given alphabet containing upper and lower case letters as well as digits, mathematical operators and other special symbols. Altogether, the alphabet contains currently 82 different symbols, an example is given in fig. 2.
- each symbol consists of up to four strokes.
- the writing of the actual symbol is finished before the writing of a new one is started.
- the pen is lifted after the writing of a symbol is finished. This blockletter-writing is almost usual in mathematical expressions except for functions such as „sin“, „cos“ or „log“. Currently, these functions must be written by using the single letters out of the alphabet, the reconstruction is done by the structure analysis process described in [4].

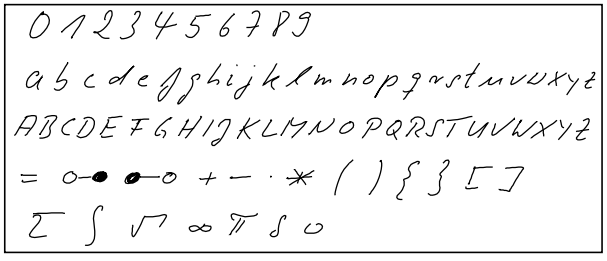

Figure 2: Symbols within the given alphabet, written by writer „wh“

The result of these prerequisites is that up to four temporal successive strokes could belong to the same symbol. Nevertheless, $(4M-6)$ different symbols can be generated if the handwritten input consists of M strokes $(M \geq 4)$.

To shrink this number of possibilities and to obtain a measurement for stroke unity, different geometrical features are calculated between stroke m and stroke $m+g$, $1 \leq g \leq 3$:

- the minimum distance between the two strokes.
- the horizontal and vertical overlapping of the surrounding rectangles of the strokes
- the distance between the starting points of the strokes
- the backward movement between the ending point of stroke m and the starting point of stroke $m+g$.

Furthermore, each stroke is classified into one out of the three categories *primitive*, *standard* or *complex* due to the complexity of writing this stroke. The classification is based on two features:

- the overall angle alteration during writing the stroke.
- the standard deviation vertical to the main axis of the stroke, calculated by the pen positions.

The use of these categories is based on the following characteristics:

- only certain combinations of these categories are possible within a symbol. For example, there is no symbol containing two *complex* strokes.
- the more strokes are belonging to a symbol, the simpler they are.

Based on the geometrical features and the information received by the stroke categories, the probabilities $P(m,g)$, that stroke m and the next g strokes belong to the same symbol, are calculated. The probability $P(m,0)$ is calculated by building the complement to the maximum probability that the stroke m belongs to any other symbol group.

Using the probabilities $P(m,g)$, $0 \leq g \leq 3$, a symbol hypotheses net (fig. 3) is generated by observing:

- $P(m,g) = 1$: Only this hypothesis is represented.
- $P(m,g) = 0$: This stroke combination is impossible and therefore not represented in the symbol hypotheses net.

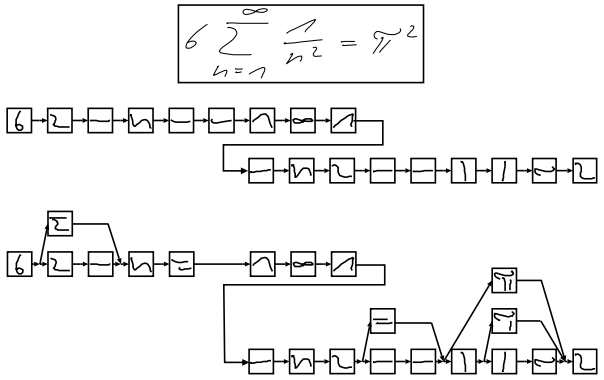

Figure 3: Handwritten expression, corresponding stroke sequence after preprocessing and the generated symbol hypotheses net

Reprint: 1995 Int. Conference on Acoustics, Speech, and Signal Processing (ICASSP-95), Detroit, USA, Vol. 4, pp. 2439-2442, May 1995

The possible sequences G_i of stroke groups are described by the different paths through the symbol hypotheses net. The probability of each sequence G_i is calculated by

$$P(G_i|I) = \prod_{Path\ i} P(m,g).$$

4 SYMBOL HYPOTHESES CLASSIFICATION

Each element within the hypotheses net represents a possible symbol and therefore has to be classified.

Before extracting any features used for classification, another slant correction is necessary because the slant of each element of the symbol hypotheses net may vary from the overall slant corrected already. After that, an image is calculated by interpolating the on-line sampled data of the symbol hypothesis. Based on this image, two kinds of feature vectors are generated.

To generate the first sequence of feature vectors $\{x_v\}$, the image is divided into vertical slices of fixed size, each slice itself is subdivided into seven small windows [6][7]. The horizontal size of the slices as well as the thresholds between the subwindows are determined from the symbol size. The length of the strokes within each window in relation to the overall stroke length is calculated, the sequence of feature vector $\{x_v\}$ is calculated by averaging the results of two adjacent slices from left to right and an overlapping of one slice (fig. 4, left side).

The second set of feature vectors $\{x_h\}$ is generated analogous but this time the image is divided into horizontal slices and the calculation of the feature vectors is done from top to bottom (fig. 4, right side).

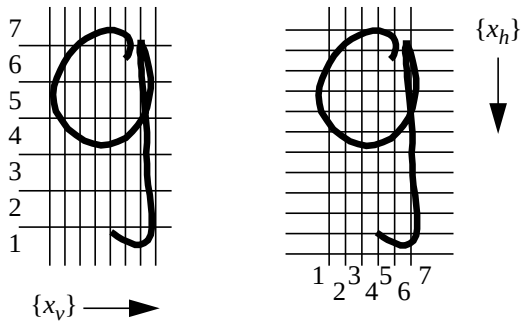

Figure 4: Generation of the sequences of feature vectors $\{x_v\}$ and $\{x_h\}$, illustrated by the number „9“

Each sequence of feature vectors is classified by the recognizer, which is based on a semi-continuous, first-order left-right Hidden Markov Model (HMM) [2]. After the classification of each feature sequence, the single recognizer results are combined.

For quality verification, writer-dependent symbol recognition experiments are carried out using single sampled symbols out of the alphabet. Each writer contributed 50 versions of the 82 symbols, 40 versions are used for training of the HMMs, the remaining are used for recognition. The results of these experiments are summarized in tab. 1.

Writer	$\{x_v\}$	$\{x_h\}$	$\{x_v\} \cap \{x_h\}$
bw	94.5%	92.7%	95.9%
fh	90.8%	88.1%	91.6%
hm	92.5%	91.1%	94.1%
kh	94.6%	92.5%	95.0%
km	91.0%	90.0%	93.2%
rp	94.6%	95.2%	96.9%
wh	90.9%	90.1%	92.4%

Table 1: Writer-dependent recognition results

Analysing these recognition results, it is realized that about 70% of the recognition errors occur due to a mix-up between almost not distinguishable symbols (fig. 2) such as „s“ and „S“, „x“ and X“ or „0“ and „O“ [7].

By classifying an element out of the symbol hypotheses net (corresponding to an element $G_i(n)$ of the stroke group sequence G_i), the best fitting symbol recognizer results $S_j(n)$ are observed by obtaining the most probable generation probabilities $P(G_i(n)|S_j(n))$ [2].

5 SYMBOL RECOGNITION

In the preceding stages the input stroke sequence I was transformed into sequences G of stroke groups including the probability $P(G_i|I)$ obtained by generating the symbol hypotheses net. Furthermore, by using the HMM recognizer, each sequence G_i of stroke groups was transformed into sequences S_j of symbols, each sequence S_j concludes a generation probability

$$P(G_i|S_j) = \prod_n P(G_i(n)|S_j(n)).$$

Assuming that $P(G_i)$ and $P(S_j)$ are constants, the Bayes theorem results in following equation:

$$P(G_i|S_j) = P(S_j|G_i).$$

Therefore, the probability $P(S_j|I)$ for transforming the input data I into a sequence S_j of symbols is calculated by

$$P(S_j|I) = P(S_j|G_i) \cdot P(G_i|I).$$

The final classification of the symbol sequence S_F within the handwritten expression is done by calculating the most probable symbol sequence out of all generated symbol sequences S_j:

$$S_F = \underset{S}{\operatorname{argmax}}\,[P(S_j|I)] = \underset{S}{\operatorname{argmax}}\,[P(S_j|G_i) \cdot P(G_i|I)].$$

Reprint: 1995 Int. Conference on Acoustics, Speech, and Signal Processing (ICASSP-95), Detroit, USA, Vol. 4, pp. 2439-2442, May 1995

Alternatives for the classified symbol sequence S_F are observed by obtaining the most probable sequences S_j.
Within these alternatives, caused by the multiplication of $P(G_i|I)$ and $P(S_j|G_i)$ within the final classification, the symbol recognizer results may change as well as the path through the symbol hypotheses net.
Furthermore, a probably preferred wrong path through the symbol hypotheses net, caused by a high path probability $P(G_i|I)$, may be devaluated by a poor symbol recognizer probability $P(S_j|G_i)$ if an element within the sequence G_i is unknown to the symbol recognizer

6 SOME RESULTS

Each of the handwritten expressions given in fig. 5 was analysed by the system. The corresponding symbol hypotheses net generated by the system is given next to each expression.

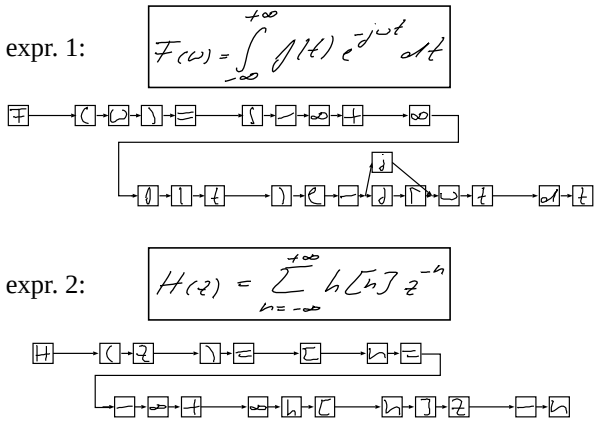

Figure 5: Two handwritten expressions (writer „wh“) and the generated symbol hypotheses net

Based on expr. 1, two different sequences G_i of stroke groups are generated (two different paths through the symbol hypotheses net), the correct path through the symbol hypotheses net was the most probable (based on the probability $P(G_i|I)$). The symbol recognizer results, obtained by the classification of the elements within the symbol hypotheses net, led to a further amplification of the probability of the correct path (in comparison to the second path).
In the second example (expr. 2) the grouping process was non-ambiguous, only one stroke group sequence G_i was generated.

Analysing the symbol recognizer results, about 88% of the elements (altogether 40 elements) within the correct path of each symbol hypotheses net are recognized correctly. Three of the recognizer errors are caused by the opening brackets „(“, which was classified as „c“ or „C“.

7 FURTHER WORK

The first results obtained by using our recognition system are very promising, the next step will be to examine the performance of our system by analysing a large number of handwritten expressions.
Furthermore, an improvement of the recognition results and their reliability will be possible by using additional feature sequences and models reproducing the making or the dynamics of writing a symbol [8][9].

8 REFERENCES

[1] C. C. Tappert, C. Y. Suen, T. Wakahara, *On-line handwriting recognition - a survey*, 9-th International Conference on Pattern Recognition, Vol. 2, pp. 1123-1131, Nov. 1988.

[2] L. R. Rabiner, *A Tutorial on Hidden Markov Models and selected Applications in Speech Recognition*, Proceedings of the IEEE Vol.77 No.2, pp. 257-286, Feb. 1989.

[3] H.-J. Lee, M.-C. Lee, *Understanding mathematical expressions using procedure-oriented transformation*, Pattern Recognition Vol.27 No.3, pp 447-457, 1994.

[4] H.-J. Winkler, H. Fahrner, M. Lang, *A Soft-Decision Approach for Structural Analysis of Handwritten Mathematical Expressions*, to be published in ICASSP 1995

[5] R. M. Bozinovic, S. N. Srihari, *Off-line cursive script recognition*, IEEE Trans. PAMI Vol.11 No.1, pp. 68-83, 1990.

[6] T.Caesar, J. Gloger, A. Kaltenmeier, E. Mandler, *Recognition of handwritten word images by statistical methods*, Proc. 3. Int. Workshop *Frontiers in Handwriting Recognition*, pp. 409-416, May 1993.

[7] H.-J. Winkler, *Symbol Recognition in Handwritten Mathematical Expressions*, Proc. Int. Workshop *Modern Modes of Man-Machine-Communication*, Maribor (Slovenia), 1994.

[8] C. C. Tappert, C. Y. Suen, T. Wakahara, *The State of the Art in On-line Handwriting Recognition*, IEEE Trans. PAMI Vol.12 No.8, pp 787-808 ,1990.

[9] K. S. Nathan, J. R. Bellegarda, D. Nahamoo, E. J. Bellegarda, *On-line Handwriting Recognition using Continuous Parameter Hidden Markov Models*, ICASSP 1993 Vol.5, pp. 121-124, 1993