A SOFT-DECISION APPROACH FOR SYMBOL SEGMENTATION WITHIN HANDWRITTEN MATHEMATICAL EXPRESSIONS

Stefan Lehmberg, Hans-Jürgen Winkler, Manfred Lang
Institute for Human-Machine-Communication
Munich University of Technology
Arcisstr. 21, 80290 Munich, Germany
{leh,win,lg}@mmk.e-technik.tu-muenchen.de

ABSTRACT

In this paper a soft-decision approach for symbol segmentation within on-line sampled handwritten mathematical expressions is presented. Based on stroke-specific features as well as geometrical features between the strokes a symbol hypotheses net is generated. For assistance additional knowledge obtained by a symbol prerecognition stage is used. The results achieved by the segmentation and prerecognition experiments indicate the performance of our approach.

1. INTRODUCTION

At ICASSP'95, we presented our approach for analysing on-line sampled handwritten mathematical expressions [1]. In this paper we will focus on the problem of symbol segmentation. Symbol segmentation is defined as the transformation of the incoming sequence of strokes (the on-line sampled handwriting) into a sequence of symbols, which will be classified within the following processing stage.
Based on the problems arising from handwriting such as illustrated in the following section, a soft-decision approach is used by generating a symbol hypotheses net containing possible symbols of the handwritten expression.

2. SYMBOL SEGMENTATION

Symbol segmentation based on off-line sampled data means splitting the image into subimages each containing a symbol.
Our 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. A stroke, in this connection, is the writing from pen down to pen up. Considering the prerequisites given in [1], symbol segmentation within our on-line based system means collecting together up to four temporal successive strokes.

As illustrated in fig. 1, in comparison to a line of text symbol segmentation within mathematical expressions is complicated by the fact that symbols can be placed above, below or even within other symbols. Furthermore, handwriting causes additional problems such as inaccuracies of stroke positioning resulting in:

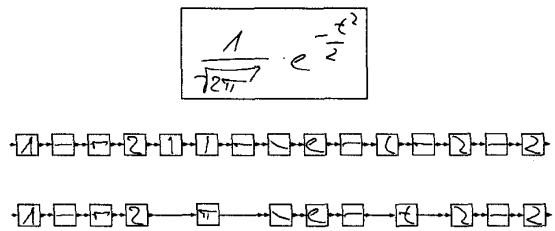

Figure 1: Image, stroke sequence and symbol sequence of a handwritten expression

- strokes belonging to the same symbol are not connected.
- the distance between strokes belonging to different symbols is very small or, in worst case, they are in touch.

Caused by these problems, hard-decision approaches may often fail in symbol segmentation. Even in the analysis of printed expressions sampled off-line using a scanner, problems arise within the segmentation process such as illustrated in [2].

Therefore, in our system a soft-decision approach is used by generating a symbol hypotheses net (SHN). The sequence of symbols within the handwritten expression is represented by one of the different paths G_i through the SHN. The final selection will be done using additional knowledge obtained by applying each element of the SHN to a symbol recognizer [1].

3. SYMBOL HYPOTHESES NET

Regarding the prerequisites given in [1], $(4M-6)$ different stroke groups can be generated if the handwritten input consists of M strokes ($M \geq 4$) . Representing all these groups within the SHN will cause problems based on

- the almost exponentially increasing number G_M of different paths through the SHN, which can be calculated by:

0-7803-3192-3/96 $5.00©1996 IEEE

$$G_M = \sum_{k=1}^{4} G_{M-k} \; ; \; G_0 = 1, \; G_k = 0 \forall (k < 0) \; .$$

For example, an expression containing $M = 30$ strokes results in $G_M \approx 2 \cdot 10^8$ different paths through the SHN containing altogether 114 different stroke groups. Applying the soft-decision symbol recognition process to the elements of the SHN will exceed the memory and, by calculating the most probable sequences of symbols, the performance of the computer.

- the large number of symbols within the alphabet consisting of other symbols. Therefore, automatic symbol segmentation by symbol recognition (such as successfully used in [3] for recognizing handwritten text) causes problems within mathematical expressions based on the lack of positional knowledge [4].

To shrink the number of stroke groups and to obtain a measurement for stroke unity, symbol hypotheses are generated and represented within the SHN after preprocessing the online sampled handwriting [1].

The generation of the symbol hypotheses is based on stroke-specific features as well as geometrical features between the strokes supported by results obtained by symbol prerecognition.

The determination of using these kind of features was done by conducting experiments analogous to [5]. Within these experiments different writers were asked to write single symbols out of the alphabet, their style of writing was analysed. An illustration of the alphabet is given in [1] and [4].

3.1 Stroke-specific features

Due to the complexity of writing a stroke, each stroke is classified into one out of the categories *primitive* (p), *standard* (s) or *complex* (c). The classification is based on:

- 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 length of the stroke in relation to the reference length calculated within the preprocessing stage [1].

The use of these categories is based on the following characteristics, an illustration for a few symbols is given in [5]:

- only certain combinations of these categories are possible within a symbol.
- the more strokes are belonging to a symbol, in most cases the simpler they are.

By using this knowledge, permitted combinations of stroke categories for each stroke group size are extracted and stored. Thus, for example, the $3^3 = 27$ different combinations of a stroke group containing $g + 1 = 3$ strokes can be reduced to 15 combinations, almost half of them only caused by the two symbols indicating the Fourier Transform and its inversion.

This combinational knowledge is transformed into a binary probability $P_c(m, g)$ which is set to 1 if the stroke category combination of stroke m and the g successive strokes is permitted, otherwise $P_c(m, g)$ is set to 0.

3.2 Symbol prerecognition

Each stroke is applied to a soft-decision prerecognition stage. This prerecognition stage is used twice within the system, once at this stage applying each stroke of the sequence I, a second time after generating the SHN applying its elements. A description of this stage is given later in chap. 3.5.

3.3 Geometrical features between the strokes

A unity matrix U of dimension $(M,3)$ is used for representing the geometrical relations between stroke m and stroke $m + g$, $1 \le g \le 3$, by the matrix element $u_{m,g}$.

For each pair of strokes different geometrical features f_k are extracted by analysing:

- the minimum distance between the strokes.
- the horizontal overlapping of the surrounding rectangles of the strokes.
- the distance as well as the horizontal offset between the starting positions of the strokes, the analogous calculation is done by the ending positions of the strokes.

For temporal successive strokes ($g = 1$) additional features are calculated by analysing:

- the backward movement between the ending position of stroke m and the starting position of the successive stroke $m + 1$.
- the parallelity of the two strokes.

Each calculated feature makes a contribution to $u_{m,g}$ by $u_{m,g} = \sum_k w_k \cdot f_k$ using feature specific weights w_k.

If one of the focused strokes m or $m + g$ is, unequivocal or not, prerecognized as the symbol „Dot" and the second one is positioned below within a certain angle, the corresponding matrix element $u_{m,g}$ is set to a minimum value by

$$u_{m,g} = \max [u_{m,g}, (z_1 + z_0)/2] \; .$$

By applying special search patterns (necessary for $g > 1$) to the unity matrix U, the measurement for stroke unity $z(m,g)$ between stroke m and the next g successive strokes is calculated by the relations of the stroke pairs within this group. For example, the calculation of $z(m,2)$ is done by

$$\max[\min[u_{m,1}, u_{m+1,1}], \min[u_{m,1}, u_{m,2}], \min[u_{m,2}, u_{m+1,1}]].$$

Finally, by using an upper and a lower threshold z_1 and z_0, the probability $P_z(m, g)$ is calculated by:

- $z(m,g) \ge z_1$: $P_z(m, g) = 1$.
- $z_0 < z(m,g) < z_1$: $P_z(m, g) = \sqrt[g]{(z(m, g) - z_0)/(z_1 - z_0)}$.
- $z(m,g) \le z_0$: $P_z(m, g) = 0$.

3.4 Generating the symbol hypotheses net

Based on the two probabilities $P_z(m, g)$ and $P_c(m, g)$ obtained in the previous sections, the final probability $P(m, g)$, $1 \le g \le 3$, is calculated by:

$$P(m, g) = P_z(m, g) \cdot P_c(m, g).$$

The probability $P(m, 0)$ is calculated by building the complement to the maximum probability that stroke m belongs to any other symbol group. Finally, normalisation is done by

$$\sum_{g=0}^{3} P(m, g) = 1.$$

Using the probabilities $P(m, g)$, $0 \le g \le 3$, a symbol hypotheses net (fig. 2) is generated starting with the maximum group size $g+1 = 4$ by observing:

- $P(m, g) = 1$: The hypothesis is represented exclusively, subgroups of this hypotheses are not tolerated.
- $0 < P(m, g) < 1$: The hypothesis is represented, subgroups of this hypotheses are tolerated.
- $P(m, g) = 0$: This stroke group is no symbol and therefore not represented within the SHN

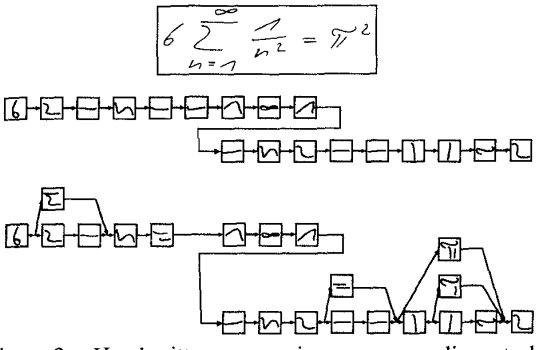

Figure 2: Handwritten expression, corresponding stroke sequence and the generated SHN

3.5 Prerecognition stage

Using prerecognition results for generating the SHN is necessary for avoiding errors caused by the symbols „i" and „j", both containing small dots placed in a considerable distance above their main body. Otherwise, the probability $P_z(m, g)$ based on the geometry of these strokes will be zero resulting in no representation within the SHN.
Additionally, prerecognition is done after generating the SHN applying its elements. A reliable recognition of the symbol „Dot" by the system presented in [4] is dubious caused by size normalization resulting in analysing the noise of pen positioning.

However, prerecognition is not limited to the symbol „Dot", additionally the two symbols „Minus" and „Fraction" are prerecognized. The selection of these three symbols is done for tolerating ambiguous recognition results between „Dot" and „Minus" and „Minus" and „Fraction".

A stroke or a symbol hypotheses within the SHN respectively is prerecognized if:

- it contains only stroke(s) of complexity p and
- its height is small or the ratio between width and height is large.

The differentiation between the three symbols is done by:

- the width of the symbol hypotheses and
- by analysing the position of the remaining elements of the SHN.

4. RESULTS AND DISCUSSION

4.1 Data sets

To train the parameters used for generating the SHN, six writers contributed 17 different mathematical expressions each written five times within several weeks. These data are additionally used for fixing the parameters within the prerecognition stage.

For the segmentation and prerecognition experiments the same writers contributed five new versions of the expression set and additional five versions of a second expression set containing 10 new expressions.
Furthermore, six unknown writers contributed mathematical expressions written up to 10 times out of the first and/or second expressions set.
Altogether, 1538 handwritten expressions are sampled for the experiments consisting of about 55700 strokes representing more than 42700 symbols, about 6200 of them representing „Dot", „Minus" and „Fraction".

4.2 Symbol segmentation

The training for generating the SHN was performed to minimize the error rate as well as the number of symbol hypotheses within the SHN representing no symbol of the expressions. Regarding the prerequisites, about 213700 symbol hypotheses can be generated resulting in a symbol hypotheses overhead of 400% but no segmentation errors.

For each writer category (known/unknown), the segmentation results are summarized in tab. 1.

Writers	Number of			
	expressions	symbols	symbol hypotheses	symbols $\notin$ SHN
known	810	22595	32783	19
unknown	728	20144	28111	8
Σ	1538	42739	60894	27

Table 1: Symbol segmentation results by SHN generation

As illustrated by the results given in tab. 1, the symbol hypotheses overhead is reduced to 42.5%, 27 symbols (0.063%) are not represented within the generated SHN. The segmentation errors as well as the symbol hypotheses

overhead depends on the style of writing the expressions. Just one of the known writers caused 15 of the 19 errors, therefore the error rate of the known writers is higher than the error rate of the unknown writers.

The missing of a symbol within the SHN can be caused by two different reasons: strokes belonging to the same symbol are not grouped together (1) or (parts of) different symbols are unified (2). Some examples are given in fig. 3.

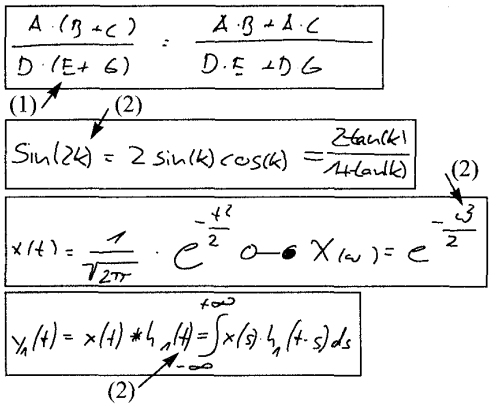

Figure 3: Some segmentation errors; the kind of error as well as the position is indicated

The significant reduction of symbol hypotheses within the generated SHNs results in an even more significant reduction of the number of paths through the net. This fact is based on the linear relationship between the number of strokes M and the number of possible stroke groups ($4M$-6) regarding the prerequisites on the one hand and on the almost exponential relationship between M and the number of paths G_M through the SHN on the other hand.

4.3 Prerecognition

Concerning the prerecognition experiment, two categories of symbols are used within the expressions, which have to be separated by the prerecognition stage:

- symbols S_p representing „Dot", „Minus" and „Fraction", which have to be prerecognized.
- symbols S_r which have to be rejected. Their recognition is done by the system presented in [4].

The results obtained by the separation experiment are given in tab. 2.

Writers	prerecognized		rejected	
	S_p	S_r	S_p	S_r
known	3245	8	13	19329
unknown	2934	0	7	17203
Σ	6179	8	20	36532

Table 2: Prerecognition and rejection of S_p and S_r within the prerecognition stage

In tab. 3 the detailed prerecognition results are given obtained by applying the symbols S_p. The differentiation into unequivocal and equivocal correct recognition results is based on the toleration of ambiguities within the prerecognition stage.

Writers	correct		wrong	re-jected
	unequivocal	equivocal		
known	3214	21	10	13
unknown	2905	24	5	7
Σ	6119	45	15	20

Table 3: Results obtained by applying the symbols S_p to the prerecognition stage

Summarizing the results given in tab. 2 and tab. 3, the average error rate in rejecting the symbols S_r results in 0.02%, the rate of prerecognizing the symbols S_p correctly (unequivocal or not) results in 99.4%.

5. CONCLUSIONS

In this paper a soft-decision approach for symbol segmentation within on-line sampled handwritten mathematical expressions is presented. Based on stroke-specific features as well as geometrical features between the strokes supported by prerecognition results a SHN is generated. Within the SHN symbols of the handwritten input are represented by the elements of the net, the symbol sequence of the handwritten input is represented by the corresponding path. The results achieved by the segmentation and prerecognition experiments indicate the performance of our system.

6. REFERENCES

[1] M. Koschinski, H.-J. Winkler, M. Lang, *Segmentation and Recognition of Symbols within Handwritten Mathematical Expressions*, ICASSP 1995 Vol.4, pp. 2439-2442, 1995.

[2] H.-J. Lee, M.C. Lee, *Understanding Mathematical Expressions using Procedure-Oriented Transformation*, Pattern Recognition Vol. 27 No. 3, pp. 447-457, 1994.

[3] K.S. Nathan, H.S.M. Beigi, J. Subrahmonia, G.C. Clary, H. Maruyama, *Real-Time On-line Unconstrained Handwriting Recognition using Statistical Methods*, ICASSP 1995 Vol.4, pp. 2619-2622, 1995.

[4] H.-J. Winkler, *HMM-Based Handwritten Symbol Recognition using On-line and Off-line Features*, to be published in ICASSP 1996.

[5] C.Y. Suen, *Handwriting Generation, Perception and Recognition*, Acta Psychologica 54, pp. 295-312, 1983.