

On-line recognition of handwritten mathematical symbols

Bachelor's Thesis of

Martin Thoma

At the Department of Informatics
Institute for Anthropomatics and Robotics (IAR)
Karlsruhe Institute of Technology (KIT)
Karlsruhe, Germany

School of Computer Science
Interactive Systems Lab (ISL)
Carnegie Mellon University (CMU)
Pittsburgh, United States

Reviewer:	Prof. Dr. Alexander Waibel
Second reviewer:	Dr. Sebastian Stücker
Advisor:	Prof. Dr. Alexander Waibel
Second advisor:	Prof. Dr. Florian Metze

Duration: June 2014 – September 2014

I declare that I have developed and written the enclosed thesis completely by myself, and have not used sources or means without declaration in the text.

Pittsburgh, DD. MM. 2014

.....
(**Martin Thoma**)

Acknowledgement

TODO



Studienstiftung
des deutschen Volkes



Baden-Württemberg
STIPENDIUM®

Ein Programm der

**Baden-
Württemberg
Stiftung**

WIR STIFTEN ZUKUNFT



This work can be cited the following way:

```
@Misc{Thoma:2014,
  Title      = {On-line recognition of handwritten mathematical symbols},

  Author     = {Martin Thoma},
  Month      = {10},
  Year       = {2014},
  School     = Karlsruhe Institute of Technology,
  Address    = "Karlsruhe, Germany",
  Type       = "{B.S. Thesis}"

  Keywords   = {Handwriting Recognition; on-line;
                Artificial Neural Networks; Hidden Markov Models},
  Owner      = {Martin Thoma},
  Timestamp  = {2014.06.17},
  Url        = {http://write-math.com/bstheis}
}
```

Contents

1. Introduction	1
1.1. Steps in handwriting recognition	1
1.2. Mathematical notation	2
2. Data, Preprocessing and Features	3
2.1. Preprocessing	3
2.1.1. Normalization: Scaling and shifting	3
2.2. Features	3
2.2.1. Local features	4
2.2.2. Global features	4
3. Baseline system	5
4. Artificial Neural Nets	7
4.1. Artificial neurons	7
4.2. Multilayer Perceptron	8
4.3. Notation	8
4.4. Evaluation	9
4.5. Supervised Training with Backpropagation	9
4.6. Parameters	11
4.7. Activation functions	12
4.7.1. Unit step function	12
4.7.2. Sigmoid function	12
4.7.3. Hyperbolic tangent	12
4.7.4. Softmax	12
4.8. Out of Vocabulary	12
4.9. Time Delay Neural Networks	13
5. Formula Recognition	15
5.1. Nesting Structures	15
5.2. Segmentation	15
6. Evaluation	17
6.1. Baseline system: Greedy matching	17
7. Conclusion	19
Bibliography	21
Glossary	23
Appendix	25
A. Algorithms	25

B. Figures	25
----------------------	----

1. Introduction

Handwriting recognition (HWR) is the task of finding a proper textual representation given a handwritten symbol or sequence of symbols.

In off-line HWR, all algorithms have to work on pixel image information of the handwriting. On-line HWR on the other hand can use the information how symbols were written. This thesis is about on-line HWR.

On-line HWR can use techniques of off-line HWR, but studies have showed that on-line information does significantly improve recognition rates and also simplify algorithms.[GAC⁺91]

1.1. Steps in handwriting recognition

Most handwriting-recognizers perform the following steps in order to recognize characters, symbols or words:

1. **Preprocessing:** Clean the data. This step is done to get rid of information that was either generated due to errors in the hardware or is not needed at all. The details are explained in section 2.1.
2. **Segmentation:** The task of formula recognition can eventually be reduced to the task of symbol recognition combined with symbol placement. But before symbol recognition can be done the formula has to be segmented. As this thesis is only about single symbol recognition, this step will not be discussed.
3. **Feature computation:** Features is high-level information derived from the raw data after preprocessing. Some systems simply take the result of the preprocessing step, but many compute new features. This might have the advantage that less training data is needed as the developer can use a priori knowledge to compute highly discriminative features.
Various features will be explained in section 2.2.

After those steps, it's only a supervised machine learning task that consists of two parts

1. **Learning:** This is most of the time adjusting parameters.
2. **Evaluation** of new datasets.

1.2. Mathematical notation

I will use the notation $v^{(i)}$ when I want to write about the i -th element of a vector v . Vectors will always be denoted by lowercase latin letters.

Matrices will be denoted by uppercase latin letters.

The number of training examples will be denoted with m , the input vector with x and the output vector with y .

A single training example thus is given by the tuple (x, y) .

2. Data, Preprocessing and Features

The data that was used for all experiments was collected with write-math.com, a website designed solely for this purpose. This website makes use of HTML5 canvas elements. Those elements can be used to track fingers or a mouse cursor touching the canvas, moving and lifting. The origin is at the upper left corner and get bigger to the right (x -coordinate) and to the bottom (y -coordinate).

The data is stored and shared in JSON format. Each handdrawing is stored as a list of lines, where each line consists of tuples $(x(t), y(t), t)$, where x and y are canvas coordinates and t is a timestamp in seconds. This timestamp gives the time in milliseconds from 1970.

The time resolution between points as well as the resolution of the image depends on the device that was used. However, most symbols have time resolution of about 20 ms and are within a bounding box of a 250px \times 250px.

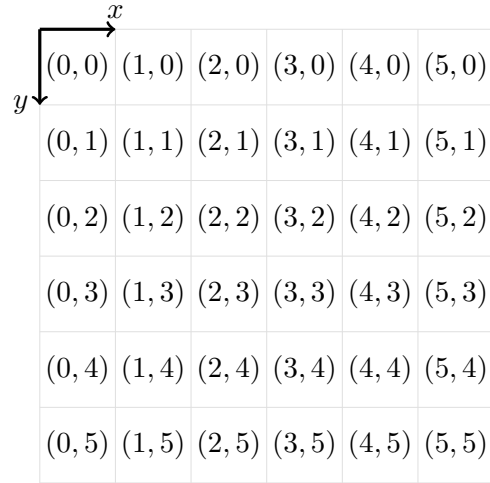


Figure 2.1.: HTML5 canvas plane. Each step is one pixel. There cannot be non-integer coordinates.

2.1. Preprocessing

2.1.1. Normalization: Scaling and shifting

In [Kir10] the datapoints were scaled to fit into a unit square while keeping their aspect ratio. Afterwards, the points were shifted to the $(0, 1) \times (0, 1)$ unit square. The algorithm is given in pseudocode on page 26. It was shown in [Kir10, HZK09] that this kind of preprocessing boosts classification accuracy significantly.

Another idea in preprocessing is resampling [GAC⁺91, SGH94].

2.2. Features

A number of different features have been suggested so far for on-line handwriting recognition. They can be grouped into local features and global features. Local features apply to

a given point on the drawing plane and sometimes even only to point on the drawn curve whereas global features apply to a complete line or even the complete image.

2.2.1. Local features

- Coordinates of the current point[GAC⁺91]
- Speed[HZK09, SGH94]
- Binary pen pressure[KR98, KRLP99, SGH94]
- Direction[MFW95, HK06]
- Curvature[MFW95, SGH94, GAC⁺91]
- Bitmap-environment[MFW95]
- Hat-Feature[SGH94, JMW00]

[KR98, KRLP99] suggest that speed is a bad feature, because it “highly inconsistent”.

The **direction** dir at the point i can be described by the vector $(\cos \theta(i), \sin \theta(i))$ as described in [GAC⁺91].

$$\cos \theta(i) = \frac{\Delta x(i)}{\Delta s(i)} \quad (2.1)$$

$$\sin \theta(i) = \frac{\Delta y(i)}{\Delta s(i)} \quad (2.2)$$

where

$$\Delta x(i) = x^{(i+1)} - x^{(i-1)} \quad (2.3)$$

$$\Delta y(i) = y^{(i+1)} - y^{(i-1)} \quad (2.4)$$

$$\Delta s(i) = \sqrt{\Delta x(i)^2 + \Delta y(i)^2} \quad (2.5)$$

The **curvature** ϕ of the point i can be described by the angle of the two neighboring points:

2.2.2. Global features

- Re-curvature[HK06] (TODO: Explain (all))
- Center point[HK06]
- Stroke length[HK06]
- Number of strokes[HZK09]
- Pseudo-Zernike Features (TODO: Really global? What is it?)[KC]
- Shadow Code Features[KC]

3. Baseline system

A system for symbol recognition was already written and is described in [Kir10]. It uses an algorithm called greedy matching which is similar to Dynamic time warping (DTW).

The Greedy Matching algorithm takes two series of points (A, B) and matches the first points (a, b) of A to B . The distance that point had to be moved is measured. Afterwards, it tries to match the next point a' of A to b as well as the next point b' of B and b' to a . It takes the matching in which the points had to be moved the shortest distance and continues like that.

Pseudocode is on page 25.

4. Artificial Neural Nets

Artificial neural networks (ANNs) are models for classification that were inspired by the brain. They consist of artificial neurons and have a lot of different subtypes like Feed Forward Neural Nets.

4.1. Artificial neurons

Artificial neurons are inspired by biological neurons. Signals are sent within the cell by charged particles, so called *ions*. But before a biological neuron sends a signal, a threshold charge has to be reached at the axon hillock. This threshold charge is called *action potential*. The action potential can be reached by multiple factors, but the one I want to focus on are charges sent by other neurons. Depending on where the other axon terminals are located and how long the distance to the axon hillock is, the signal contributes more or less to reaching the action potential. After that, it simply sends a signal.

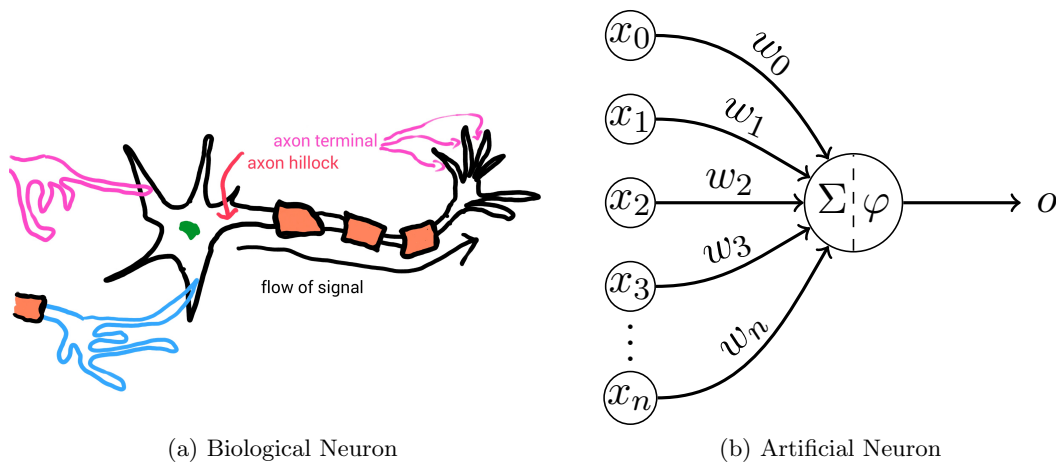


Figure 4.1.: Both neurons receive weighted input, apply a function to that and give output

Artificial neurons are similar. They receive at least one input and give at least one output. Those inputs might get weighted as well as the output.

The neurons apply a function to the sum of all weighted inputs. This function is also called *activation function*.

An artificial neuron using the unit step function (see section 4.7.1) is called a *perceptron*. The artificial neuron sums all weighted inputs $x_i \cdot w_i$ up and applies its activation function f to it.

4.2. Multilayer Perceptron

Multilayer perceptrons (MLPs) are neural nets which neurons are structured in layers. Each layer is fully connected with the next layer, but there are no other connections between neurons.

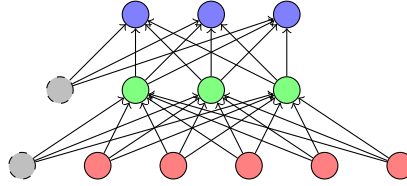


Figure 4.2.: Feedforward artificial neural network

The red neurons in figure 4.2 are input neurons, the green ones are hidden neurons and the blue one is an output node. The gray neurons are bias neurons. Bias neurons have a fixed output of 1.

Usually, you have as many output neurons as you have classes. So in the case of symbol recognition that would be about 1076 neurons.

The number of input neurons is equal to the number of features.

4.3. Notation

Let n_i be the number of neurons in the i -th layer and ℓ be the number of layers of the MLP.

Two neighboring layers of neurons are fully connected and have weights between two layers. This means you can store those weights in form of matrices. So the weights between layer i and layer $i + 1$ are

$$W_i = \begin{pmatrix} w_{1,1} & w_{1,2} & \dots & w_{1,n_{i+1}} \\ w_{2,1} & w_{2,2} & \dots & w_{2,n_{i+1}} \\ w_{3,1} & w_{3,2} & \dots & w_{3,n_{i+1}} \\ \vdots & & \ddots & \vdots \\ w_{n_i,1} & w_{n_i,2} & \dots & w_{n_i,n_{i+1}} \end{pmatrix}$$

Let $w_{ij}^{(k)}$ be the value w_{ij} in W_k . So it is the weight of the between neuron i in layer k and neuron j in layer $k + 1$.

So $W_i \in \mathbb{R}^{n_i \times n_{i+1}}$ is the matrix denoting the weights between layer i and layer $i + 1$.

The unweighted output vector of layer i is denoted by $x_i \in \mathbb{R}^{1 \times n_i}$; the weighted output vector by $\text{net}_i \in \mathbb{R}^{1 \times n_i}$. Instead of x_1 I will write x . The output of the MLP for the input x is denoted by $o_x := x_n$.

In principle each neuron might have a different activation function, but in practice each neuron in one layer has the same activation function. However, activation functions might differ from layer to layer. This is the reason why I denote the activation function of layer i

by φ_i . Although φ_i is defined for single neurons, I will in the following apply it to vectors. In its meant to be applied pointwise.

The activation function is a function

$$\varphi_i : \mathbb{R} \rightarrow \mathbb{R}$$

but because of the short-notation it can be applied pointwise to all neurons of layer i it is also a function

$$\varphi_i : \mathbb{R}^{n_i} \rightarrow \mathbb{R}^{n_i}$$

4.4. Evaluation

Let $x_1 \in \mathbb{R}^{1 \times n_1}$ be an unweighted output of layer 1. So it's simply the input of our neural net with n_1 features.

Given x_1 one can easily compute the weighted input for layer 2:

$$\begin{array}{ccccc} x_1 & \cdot & W_1 & = & \text{net}_1 \\ \cap & & \cap & & \cap \\ \mathbb{R}^{1 \times n_1} & & \mathbb{R}^{n_1 \times n_2} & & \mathbb{R}^{1 \times n_2} \end{array}$$

After that, you can apply the activation function φ_{i+1} pointwise to net_i .

So the output vector x_3 of a 3-layer (input, hidden, output) neural net can be computed by

$$x_3 = \varphi_3(\varphi_2(x_1 \cdot W_1) \cdot W_2)$$

The output of a general MLP can be computed by

$$\begin{aligned} \Phi(x, n) &: \mathbb{R}^{n_1} \rightarrow \mathbb{R}^{n_n} \\ \Phi(x_1, 2) &:= \varphi_2(x_1 \cdot W_1) \\ \Phi(x, n) &:= \varphi_n(\Phi(x, n-1) \cdot W_{n-1}) \\ \Phi(x, n)^{(p)} &= \varphi_n \left(\sum_{i=1}^{n_{n-1}} \Phi(x, n-1)^{(i)} \cdot w_{ip}^{(n-1)} \right) \end{aligned}$$

It follows: $x_i(x_1) = \Phi(x_1, i)$.

4.5. Supervised Training with Backpropagation

The backpropagation algorithm is a supervised algorithm for training MLPs. This means the trainigset T consists of tuples (x, t_x) where x is input and t_x is the desired output.

To evaluate how good the current MLP is, an error function can be defined:

$$E : \mathbb{R}^{n_1 \times n_2} \times \mathbb{R}^{n_2 \times n_3} \times \dots \times \mathbb{R}^{n_{\ell-1}, \ell} \rightarrow \mathbb{R}_{\geq 0}$$

$$E_T(W) = \frac{1}{2} \sum_{(x, t_x) \in T} \sum_{p=1}^{n_\ell} \left(t_x^{(p)} - o_x^{(p)} \right)^2$$

This function is isomorphic to

$$E : \mathbb{R}^{\sum_{i=2}^{\ell} n_{i-1} \cdot n_i} \rightarrow \mathbb{R}_{\geq 0}$$

This error should be minimized. As the error is the sum of non-negative values, we will get a lower error by minimizing the error for a single training example. However, note that those minimizations are not independent. This means, we could get trapped in a local minimum.

The idea is to “go” into the direction in which the error E decreases most. This is the gradient and the process is thus called *gradient descent*.

At this point we need to decide how far we want to go. If we make too big steps in the direction of the gradient, we might overshoot. If we make too small steps, the algorithm will take too long to get to the minimum. As reducing the error is basically learning the number is called learning rate $\eta \in \mathbb{R}_{>0}$.

So the algorithm is

Algorithm 1 Backpropagate

```

function BACKPROPAGATE( $T, W$ )
  while True do
    for all  $(x, t_x) \in T$  do
      for all node  $i$  do
        for all nodes  $j$  following  $i$  do
           $w_{ijk} \leftarrow w_{ijk} - \eta \frac{\partial E_{\{x\}}}{\partial w_{ijk}}(W)$ 

```

Computing the partial derivatives $\frac{\partial E_{\{x\}}}{\partial w_{ijk}}$ is not a trivial task. To do that, we have to take a closer look at the error function:

Another approach:

$$E_x(W) = \frac{1}{2} \sum_{p=1}^{n_\ell} \left(t_x^{(p)} - o_x^{(p)} \right)^2 \quad (4.1)$$

$$o_x^{(p)} = \Phi(x, \ell)^{(p)} \quad (4.2)$$

$$= \varphi_\ell(\Phi(x, \ell - 1) \cdot W_{\ell-1})^{(p)} \quad (4.3)$$

$$= \varphi_\ell \left(\underbrace{\sum_{i=1}^{n_{\ell-1}} \Phi(x, \ell - 1)^{(i)} \cdot w_{ip}^{(\ell-1)}}_{\text{net}_{\ell-1}^{(p)}} \right) \quad (4.4)$$

$$\frac{\partial E_x}{\partial w_{ij}^{(k)}} = \frac{\partial E_x}{\partial \text{net}_k^{(j)}} \frac{\partial \text{net}_k^{(j)}}{\partial w_{ij}^{(k)}} \quad (4.5)$$

$$= \frac{\partial E_x}{\partial \text{net}_k^{(j)}} \frac{\sum_{i=1}^{n_k} \Phi(x, k)^{(i)} \cdot w_{ij}^{(k)}}{\partial w_{ij}^{(k)}} \quad (4.6)$$

$$= \frac{\partial E_x}{\partial \text{net}_k^{(j)}} \Phi(x, k)^{(i)} \quad (4.7)$$

Suppose $k = \ell - 1$ (weights to the output layer). Then:

$$\frac{\partial E_x}{\partial w_{ij}^{(\ell-1)}} = \frac{\partial E_x}{\partial \text{net}_{\ell-1}^{(j)}} \Phi(x, \ell - 1)^{(i)} \quad (4.8)$$

$$= \frac{\frac{1}{2} \sum_{p=1}^{n_\ell} \left(t_x^{(p)} - o_x^{(p)} \right)^2}{\partial \text{net}_{\ell-1}^{(j)}} \Phi(x, \ell - 1)^{(i)} \quad (4.9)$$

$$= \frac{\frac{1}{2} \sum_{p=1}^{n_\ell} \left(t_x^{(p)} - \varphi_\ell(\text{net}_{\ell-1})^{(p)} \right)^2}{\partial \text{net}_{\ell-1}^{(j)}} \Phi(x, \ell - 1)^{(i)} \quad (4.10)$$

$$= \left((t_x^{(j)} - \varphi_\ell(\text{net}_{\ell-1}^{(j)})) \cdot (-\varphi'_\ell(\text{net}_{\ell-1}^{(j)})) \right) \Phi(x, \ell - 1)^{(i)} \quad (4.11)$$

Suppose $k = \ell - 2$ (last hidden layer). Then:

$$\frac{\partial E_x}{\partial w_{ij}^{(\ell-2)}} = \frac{\frac{1}{2} \sum_{p=1}^{n_\ell} \left(t_x^{(p)} - \varphi_\ell(\text{net}_{\ell-1})^{(p)} \right)^2}{\partial \text{net}_{\ell-2}^{(j)}} \Phi(x, \ell - 2)^{(i)} \quad (4.12)$$

$$= \frac{\frac{1}{2} \sum_{p=1}^{n_\ell} \left(t_x^{(p)} - \varphi_\ell \left(\sum_{i=1}^{n_{\ell-1}} (\varphi_{\ell-1}(\varphi_{\ell-2}(\text{net}_{\ell-2}^{(j)})))^{(i)} \cdot w_{ip}^{(\ell-1)} \right)^{(p)} \right)^2}{\partial \text{net}_{\ell-2}^{(j)}} \Phi(x, \ell - 2)^{(i)} \quad (4.13)$$

$$(4.14)$$

4.6. Parameters

- Number of hidden layers

- Number of neurons per hidden layer
- Epochs
- Momentum
- Weight decay
- Learning rate
- Learning rate decay

4.7. Activation functions

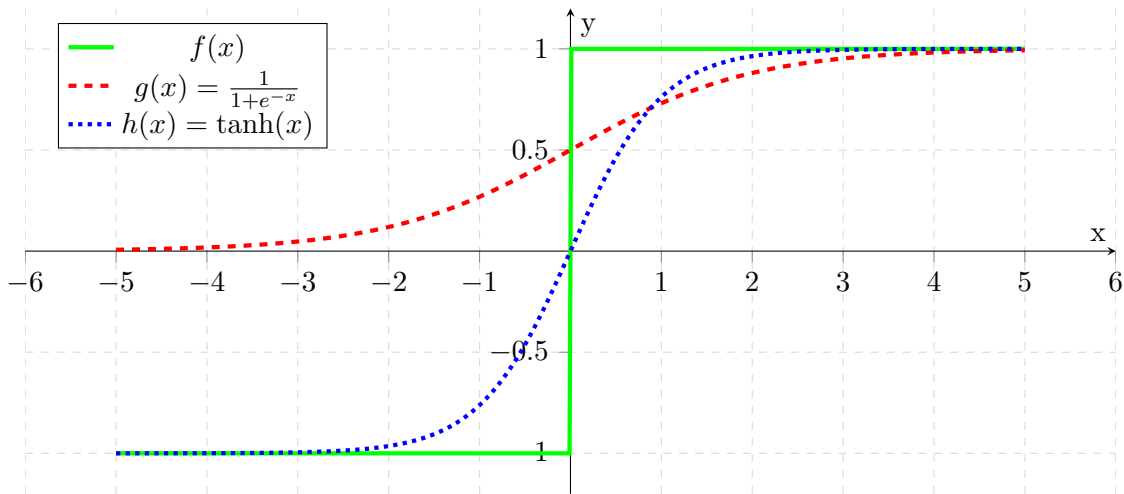


Figure 4.3.: A variation of the sign function f with $f(0) = -1$, the sigmoid function g and the hyperbolic tangent h .

4.7.1. Unit step function

Not so good, because it's not differentiable. Therefore, the backpropagation algorithm cannot be used.

4.7.2. Sigmoid function

Is great because it is infinitely often differentiable.

4.7.3. Hyperbolic tangent

Also differentiable, but gradient descent converges faster (sometimes?)

4.7.4. Softmax

$$\varphi(a_j) = \frac{e^{a_j}}{\sum_k e^{a_k}}$$

4.8. Out of Vocabulary

It might be desirable to have a possibility to reject drawings that were not recognized. Such a rejection could be realized in two ways:

- A threshold for the answer,
- a minimum delta between the highest rated answer and the second highest rated answer or
- explicitly training for out of vocabulary (OOV).

4.9. Time Delay Neural Networks

Time Delay Neural Networks were successfully applied to symbol recognition task in the past[GAC⁺91, JMRW01].

TODO!

5. Formula Recognition

5.1. Nesting Structures

One major issue of formula recognition are nesting structures:

- `\frac{[arbitrary math]}{[arbitrary math]}`
- `\begin{pmatrix}[arbitrary math; structure with &]\end{pmatrix}`
- `\begin{align}[arbitrary math; structure with &]\end{align}`
- `\stackrel{[arbitrary math]}{[arbitrary math]}`

Another point that is special about handwritten math recognition compared to natural language text are *decorators*:


- `[symbol]_{[arbitrary math]}`
- `[symbol]^{[arbitrary math]}`

A third problem that comes with formulas that consist of multiple symbols are semantics: For single symbols, it is not possible to distinguish `\Sigma` from `\sum`, so we don't expect an algorithm to be able to do so. However, if we have a complete formula we might have enough context and thus we want the recognition algorithm to be able to distinguish `\Sigma` from `\sum` and other pairs of similar symbols like `\Pi` and `\prod` or `\Omega` and `\Ohm`.

5.2. Segmentation


I assume that writers finish one handwritten symbol before they start the next symbol.

Except for the following symbols, every symbol is composed of 4 strokes:

12 lines `\Mundus` - 

7 lines — `\fax` - 

11 lines `\FAX` - 

— `\Emailct` - 

9 lines `\sun`

— `\Letter` - 

8 lines `\idotsint` - $\int \cdots \int$

— `\EyesDollar` - $\$$

	- \textreferencemark - ※	- \uranus -
	- \neptune -	- \Uranus - ⛢
	- \ataribox -	- \neptune -
6 lines	- \dots - ... (wild points appear more often with single points)	- \textcurrency - ₭
	- \vdots - ⋮	- \Lleftarrow - ⇐
	- \upuparrows - ↑↑	- \Pi - Π
	- \Neptune - ♆	- \nexists - ∄
5 lines	- \Xi - Ξ	- \updownarrow - ⇕
	- \dotsint - ∫⋯∫	- \divideontimes - ⋈
	- \idotsint - ∫⋯∫	- \permil - ‰ (wasysym)
	- \sqiint - ∫⋈∫	- \textpertenthousand - ‰■
	- \nVDash - ⋈	- \textdiscount - ‰
	- \nLeftrightarrow - ⇔	- \mathds{E} - E
	- \boxtimes - ⊠	- \mathds{F} - F
	- \Smiley, \smiley, \Frowny - ☺, ☹	- \textsca -

6. Evaluation

A recognition system has two important characteristics: Its recognition accuracy and the time it needs to recognize a new symbol. Recognition accuracy and time are measured with new data which was not seen before.

Tests can be divided into two groups: Tests where some examples of the handwriting of the writer were known at training time and tests where that's not the case.

Known-writer tests are created this way:

Algorithm 2 Creation of k bins of datasets

```
data  $\leftarrow$   $k$ -dimensional array of Lists  
 $i \leftarrow 0$   
Group labeled datasets by symbol  
Filter all symbols that have less than  $k$  datasets  
for all Group  $g$  in datasets do  
  for all Dataset  $(x, t)$  in  $g$  do  
    data[ $i$ ].APPEND( $(x, t)$ )  
     $i \leftarrow (i + 1) \bmod k$ 
```

After that, a k -fold cross validation is run:

I will call result of such a 10-fold cross-validation *classification accuracy*. The first part of the tuple is called *Top-1 accuracy* and the second one is called *Top-10 accuracy*.

6.1. Baseline system: Greedy matching

The greedy matching algorithm got with scaling and shifting (see page 26) a Top-1 accuracy of 83.11% and a Top-10 accuracy of 97.66%.

Algorithm 3 k -fold cross-validation

```

function CROSSVALIDATION( $k$ , grouped dataset  $d$ , classifier  $c$ )
  correct, wrong  $\leftarrow 0, 0$ 
  c10, w10  $\leftarrow 0, 0$ 
  for  $i \in 0, \dots, k-1$  do
    for  $j \in 0, \dots, k-1$  do
      if  $i \neq j$  then
         $c$ .TRAIN( $d[i]$ )
    for all  $(x, t) \in d[i]$  do  $\triangleright$  List of possible classifications descending by probability
       $L \leftarrow c$ .CLASSIFY( $x$ )
      if  $L[0] == t$  then
        correct  $\leftarrow$  correct + 1
        c10  $\leftarrow$  c10 + 1
      else if  $t \in L$  then
        c10  $\leftarrow$  c10 + 1
        wrong  $\leftarrow$  wrong + 1
      else
        w10  $\leftarrow$  wrong + 1
        wrong  $\leftarrow$  wrong + 1
  return ( $\frac{\text{correct}}{\text{correct}+\text{wrong}}, \frac{\text{c10}}{\text{c10}+\text{w10}}$ )

```

7. Conclusion

...

Bibliography

- [Boa12] E. Board, *Concise Dictionary of Mathematics*, unknown, Ed. V&S Publishers, 2012. [Online]. Available: <http://books.google.de/books?id=7OqVdc2LGSUC>
- [GAC⁺91] I. Guyon, P. Albrecht, Y. L. Cun, J. Denker, and W. Hubbard, “Design of a neural network character recognizer for a touch terminal,” *Pattern Recognition*, vol. 24, no. 2, pp. 105 – 119, 1991. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/003132039190081F>
- [HK06] B. Huang and M.-T. Kechadi, “An HMM-SNN method for online handwriting symbol recognition,” in *Image Analysis and Recognition*, ser. Lecture Notes in Computer Science, A. Campilho and M. Kamel, Eds. Springer Berlin Heidelberg, 2006, vol. 4142, pp. 897–905. [Online]. Available: http://dx.doi.org/10.1007/11867661_81
- [HZK09] B. Q. Huang, Y. Zhang, and M.-T. Kechadi, “Preprocessing techniques for online handwriting recognition,” in *Intelligent Text Categorization and Clustering*, ser. Studies in Computational Intelligence, N. Nedjah, L. de Macedo Mourelle, J. Kacprzyk, F. França, and A. de De Souza, Eds. Springer Berlin Heidelberg, 2009, vol. 164, ch. Preprocessing Techniques for Online Handwriting Recognition, pp. 25–45. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-85644-3_2
- [JMRW01] S. Jaeger, S. Manke, J. Reichert, and A. Waibel, “Online handwriting recognition: the npen++ recognizer,” in *International Journal on Document Analysis and Recognition*, 2001, pp. 169–180.
- [JMW00] S. Jaeger, S. Manke, and A. Waibel, “Npen++: An on-line handwriting recognition system,” in *7th International Workshop on Frontiers in Handwriting Recognition*, 2000, pp. 249–260. [Online]. Available: http://isl.anthropomatik.kit.edu/cmu-kit/IWFHR_stephen1.pdf
- [KC] A. Khotanzad and C. Chung, “Hand written digit recognition using combination of neural network classifiers.” [Online]. Available: <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=666880>
- [Kir10] D. Kirsch, “Detexify: Erkennung handgemalter LaTeX-symbole,” Diploma thesis, Westfälische Wilhelms-Universität Münster, 10 2010. [Online]. Available: <http://danielkirs.ch/thesis.pdf>
- [KR98] A. Kosmala and G. Rigoll, “Recognition of on-line handwritten formulas,” in *In Proceedings of the Sixth International Workshop on Frontiers in Handwriting Recognition*, 1998, pp. 219–228. [Online]. Available: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.41.9056>
- [KRLP99] A. Kosmala, G. Rigoll, S. Lavirotte, and L. Pottier, “On-line handwritten formula recognition using hidden markov models and context dependent graph

- grammars,” in *Proceedings of the Fifth International Conference on Document Analysis and Recognition (ICDAR)*, 1999, pp. 107–110. [Online]. Available: http://hal.inria.fr/docs/00/56/46/45/PDF/kosmala-rigoll-etal_1999.pdf
- [MFW95] S. Manke, M. Finke, and A. Waibel, “The use of dynamic writing information in a connectionist on-line cursive handwriting recognition system,” in *Advances in Neural Information Processing Systems 7*, G. Tesauero, D. Touretzky, and T. Leen, Eds. MIT Press, 1995, pp. 1093–1100. [Online]. Available: [http://isl.anthropomatik.kit.edu/cmu-kit/downloads/The_Use_of_Dynamic_Writing_Information_in_a_Connectionist_On-Line_Cursive_Handwriting_Recognition_System\(3\).pdf](http://isl.anthropomatik.kit.edu/cmu-kit/downloads/The_Use_of_Dynamic_Writing_Information_in_a_Connectionist_On-Line_Cursive_Handwriting_Recognition_System(3).pdf)
- [SGH94] M. Schenkely, I. Guyonz, and D. Hendersonz, “On-line cursive script recognition using time delay neural networks and hidden markov models,” in *1994 IEEE International Conference on Acoustics, Speech, and Signal Processing.*, vol. ii, 4 1994, pp. 637–640. [Online]. Available: http://pdf.aminer.org/003/076/160/on_line_cursive_script_recognition_using_time_delay_neural_networks.pdf

Glossary

ANN artificial neural network. 7

DTW dynamic time warping. 5

epoch During iterative training of a neural network, an *epoch* is a single pass through the entire training set, followed by testing of the verification set.[Boa12]. 12

HWR handwriting recognition. 1

learning rate A factor $0 \leq \eta \in \mathbb{R}$ that affects how fast new weights are learned. $\eta = 0$ means that no new data is learned. 12

learning rate decay The learning rate decay $0 < \alpha \leq 1$ is used to adjust the learning rate. After each epoch the learning rate η is updated to $\eta \leftarrow \eta \times \alpha$. 12

MLP multilayer perceptron. 8, 9

OOV out of vocabulary. 12

Appendix

A. Algorithms

Algorithm 4 Greedy matching as described in [Kir10]

```
 $a \leftarrow \text{next from } A$   
 $b \leftarrow \text{next from } B$   
 $d \leftarrow \delta(a, b)$   
 $a' \leftarrow \text{next from } A$   
 $b' \leftarrow \text{next from } B$   
while points left in  $A \wedge$  points left in  $B$  do  
   $l, m, r \leftarrow \delta(a', b), \delta(a', b'), \delta(a, b')$   
   $\mu \leftarrow \min \{l, m, r\}$   
   $d \leftarrow d + \mu$   
  if  $l = \mu$  then  
     $a \leftarrow a'$   
     $a' \leftarrow \text{next from } A$   
  else if  $r = \mu$  then  
     $b \leftarrow b'$   
     $b' \leftarrow \text{next from } B$   
  else  
     $a \leftarrow a'$   
     $b \leftarrow b'$   
     $a' \leftarrow \text{next from } A$   
     $b' \leftarrow \text{next from } B$   
if no points left in  $A$  then  
  for all points  $p$  in  $B$  do  
     $d \leftarrow d + \delta(a', p)$   
else if no points left in  $B$  then  
  for all points  $p$  in  $A$  do  
     $d \leftarrow d + \delta(b', p)$ 
```

B. Figures

ein Bild

Figure B.1.: A figure

Algorithm 5 Scale and shift a list of lines to the $(0, 1) \times (0, 1)$ unit square

```

function SCALE_AND_SHIFT(pointlist)
   $min_x, min_y = pointlist[0][x'], pointlist[0][y']$ 
   $max_x, max_y = pointlist[0][x'], pointlist[0][y']$ 
  for all lines in pointlist do
    for all p in lines do
      if  $p[x'] < min_x$  then
         $min_x \leftarrow p[x']$ 
      else if  $p[x'] > max_x$  then
         $max_x \leftarrow p[x']$ 
      if  $p[y'] < min_y$  then
         $min_y \leftarrow p[y']$ 
      else if  $p[y'] > max_y$  then
         $max_y \leftarrow p[y']$ 
   $width, height \leftarrow max_x - min_x, max_y - min_y$ 
   $factor = 1$ 
  if  $width \neq 0$  then
     $factor_x = \frac{1}{width}$ 
  if  $height \neq 0$  then
     $factor_y = \frac{1}{height}$ 
   $factor = \min(factor_x, factor_y)$ 
   $add_x, add_y = 0, 0$ 
  for all lines in pointlist do
    for all p in lines do
       $p[x'] \leftarrow (p[x''] - min_x) \cdot factor$ 
       $p[y'] \leftarrow (p[y''] - min_y) \cdot factor$ 
  return pointlist

```
