

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



interACT



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
1.3. Limitations of Symbol Recognition	2
1.4. What is a symbol?	2
2. Data, Preprocessing and Feature extraction	3
2.1. Preprocessing	3
2.1.1. Normalization: Scaling, shifting and resampling	3
2.1.2. Noise reduction	4
2.2. Features	5
2.2.1. Local features	5
2.2.2. Global features	6
3. Baseline system	7
4. Artificial Neural Nets	9
4.1. Artificial neurons	9
4.2. Multilayer Perceptron	10
4.3. Notation	10
4.4. Evaluation	11
4.5. Supervised Training with Backpropagation	11
4.6. Parameters	13
4.7. Activation functions	13
4.7.1. Unit step function	13
4.7.2. Sigmoid function	13
4.7.3. Hyperbolic tangent	13
4.7.4. Softmax	14
4.8. Out of Vocabulary	14
4.9. Time Delay Neural Networks	14
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	25

Appendix	27
A. Algorithms	27

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. So the pen trajectory is given in on-line HWR. 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, BN72].

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 a supervised machine learning task that consists of two parts

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

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) .

Weights will be denoted with W , hyperparameters with θ .

0-indexed vectors are used.

1.3. Limitations of Symbol Recognition

The recognition capabilities of single symbols are quite limited. There are many symbols such as “.” and “.” or “0”, “O” and “o” that can only be distinguished with context and the availability of a baseline. As I chose to set up the design of write-math.com without a baseline and the associated restrictions for the user, it is impossible to distinguish these characters. The best that can be done is applying the a priori probability.

Preprocessing-operations that cannot be done with a single symbol are

- Baseline correction

Another limitation is that there is no context that can be used for calculating the hypothesis.

1.4. What is a symbol?

A symbol is an atomic semantic entity. Examples for symbols are: $\alpha, \infty, \cdot, x, \int, \sigma, \dots$

In contrast, the L^AT_EX command `\ll` which compiles to \ll are two symbols as well as `\iint` and `\iiint` which compile to \iint and \iiint are two and three symbols.

I suggest to ignore those commands in a symbol recognizer. Instead, a formula recognizer that recognizes `\int\int` should apply a post recognition method that transforms $\int\int$ to \iint . Some of those are transformations are listed in table 1.1.

Search		Replace	
L ^A T _E X	Rendered	L ^A T _E X	Rendered
<code>\int\int</code>	$\int\int$	<code>\iint</code>	\iint
<code>\int\int\int</code>	$\int\int\int$	<code>\iiint</code>	\iiint
<code><<</code>	$<<$	<code>\ll</code>	\ll
<code><<<</code>	$<<<$	<code>\lll</code>	\lll
<code>>></code>	$>>$	<code>\gg</code>	\gg
<code>>>></code>	$>>>$	<code>\ggg</code>	\ggg

Table 1.1.: Multiple symbols in one L^AT_EX command

2. Data, Preprocessing and Feature extraction

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 a time resolution of about 20 ms and are within a bounding box of a 250px \times 250px square.

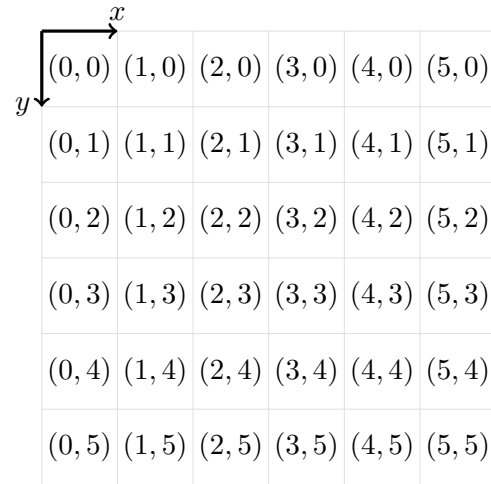


Figure 2.1.: HTML5 canvas plane. Each step is one pixel. There cannot be non-integer coordinates.

2.1. Preprocessing

Preprocessing in symbol recognition is done to improve the quality and expressive power of data. It should make follow-up tasks like segmentation and feature extraction easier, more effective or faster. It does so by removing or fixing errors in the input data, reducing duplicate information and removing irrelevant information.

2.1.1. Normalization: Scaling, shifting and resampling

Size normalization is done by many handwriting recognition systems, but the way in which size normalization is done varies.

Single symbol recognizers such as the one presented in [Kir10] scale the datapoints 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 29. It was shown in [Kir10, HZK09] that this kind of preprocessing boosts classification accuracy significantly.

[GAC⁺91] shifts the symbol to $[-1, 1] \times [-1, 1]$.

Multiple symbol recognizers such as NPen++ as presented in [JMRW01] scale words in respect to a previously calculated baseline and a corpus line.

Another method to normalize data is *resampling*, sometimes also called stroke length normalization. [GAC⁺91] resampled characters and digits to 81 points each, where different lines were also connected by “pen-up” segments. They resampled to get points regularly spaced in arc length, not in time. [JMRW01] also resampled the points to be equidistant in space, but they used a distance of $\frac{\text{corpus height}}{13}$. They found an improvement of 5% with this preprocessing step. [SGH94] also resampled data to get points regularly spaced in arc length, but they encoded speed as an extra feature.

2.1.2. Noise reduction

The following list of noise reduction techniques was created by [TSW90] and is still up-to-date.

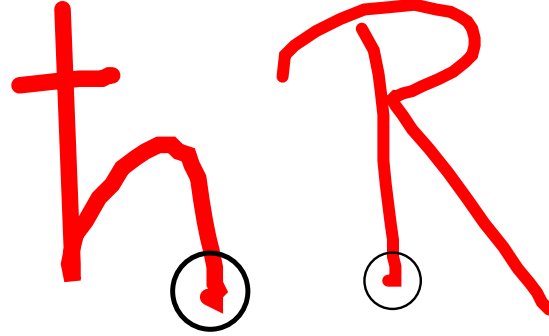
- **Dot reduction** reduces dots to single points. Sometimes multiple points get recorded although the user wanted to make only a single point, e.g. for one of the following symbols: \cdot , \cdot , \dots , $\dot{\cdot}$, $\ddot{\cdot}$, i , ä , ö , ü . This can be detected by calculating the maximum distance d two points in a stroke have. If d is smaller than a threshold, then it is a single point. In that case all points of the line get reduced to a single dot. This dot could be the center of mass of all points in the stroke.
- **Dehooking** is the removal of hooks which the author did not want to write. Hooks appear sometimes at the end of strokes. Examples can be seen in figure 2.2. An algorithm in pseudocode can be found on page 27. A more sophisticated method was applied in [HZK09].
- **Filtering** is the process of removing points by some criteria. Those criteria include:
 - Duplicate points as applied in [HZK09, GP93],
 - Enforcing a minimal distance between consecutive points[TSW90].
 - Maximum velocity / acceleration[Tap87]
 - Enforcing a minimal change in direction[TSW90].

A special reason for the application of filtering methods are occasional spurious points which are also called *wild points*.

- **Smoothing** can be done in multiple ways. An approach that was used quite often is applying a weighted average [Gro66, Tap87, Ara83]. Algorithm 6 describes in pseudocode how weighted average smoothing can be implemented.
- **Stroke connection** might be used if the distance between pen-up and pen-down is below a threshold. [GP93] describes that such maliciously disconnected components can get detected by observing angular continuity and the shortness of distance between two strokes.

- **Deskewing** corrects character slant. Although this technique was applied by some authors [BS89, GP93, HBT94], it seems not to be applicable to the domain of mathematical handwriting, because on the one hand symbols might occur in variations with slant, like \rightarrow and \nearrow . On the other hand it is questionable if slant is as consistent with symbols as it is with cursive handwriting.
- **Baseline drift correction** moves words to be on a baseline.

An idea for filtering that seemingly nobody has tried before is applying the Douglas-Peucker algorithm.



(a) Hook at the end of h (b) Hook in the letter R

Figure 2.2.: Examples for hooks

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 are used by [GAC⁺91].
- **Speed** has been used by [SGH94], but [KR98, KRLP99] suggest that speed is a bad feature, because they think that speed is “highly inconsistent”.
- **Binary pen pressure** has been used by [KR98, KRLP99, SGH94, MFW94, GAC⁺91].
- **Direction** has been used by [MFW95, HK06]. The **direction** 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

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

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

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

- **Curvature** has been used by [Gro66, MFW95, SGH94, GAC⁺91]. It is calculated in [GAC⁺91] by the angle of two neighboring lines like this:

$$\varphi(i) = \theta(i+1) - \theta(i-1) \quad (2.6)$$

$$\cos \varphi(i) = \cos \theta(i-1) \cdot \cos \theta(i+1) \quad (2.7)$$

$$+ \sin \theta(i-1) \cdot \sin \theta(i+1) \quad (2.8)$$

$$\cos \varphi(i) = \cos \theta(i-1) \cdot \cos \theta(i+1) \quad (2.9)$$

$$- \sin \theta(i-1) \cdot \sin \theta(i+1) \quad (2.10)$$

- **Bitmap-environment** has been used by [MFW94]make. This feature is a 3×3 pixel environment around the current point. It allows the recognizer to determine points that cross or touch strokes.
- **Hat-Feature** has been used by [SGH94, JMW00].

2.2.2. Global features

- **Re-curvature** is defined in [HK06, HZK09] as the ratio between the height of a stroke and the distance between its start and end points.
- **Center point** was used in [HK06].
- **Stroke length** was used in [HK06]. It can be calculated by using a linear interpolation.
- **Number of strokes** was used in [HZK09].
- **Sequence features**
 - *Pentip sequence*: [Kir10] used the raw pentip sequence combined with dynamic time warping (DTW) to recognize mathematical symbols. Other authors like [KWL95] used pentip sequences, too, but made use of hidden Markov models (HMMs) or artificial neural networks (ANNs) to recognize symbols
 - *Zone sequences* are used by [Bro64, iHY80]. The idea is to recognize symbols by dividing the box in which the character is written into zones. By examining the position of the pen-tip a sequence of zones can be generated for a written symbol.
 - *Direction sequences* were used in [IMP76, Pow73].

There are other global features used for off-line handwriting recognition which I will not examine. Examples are Pseudo-Zernike moments and Shadow Code features which were used in [KC98].

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 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 28.

4. Artificial Neural Nets

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.

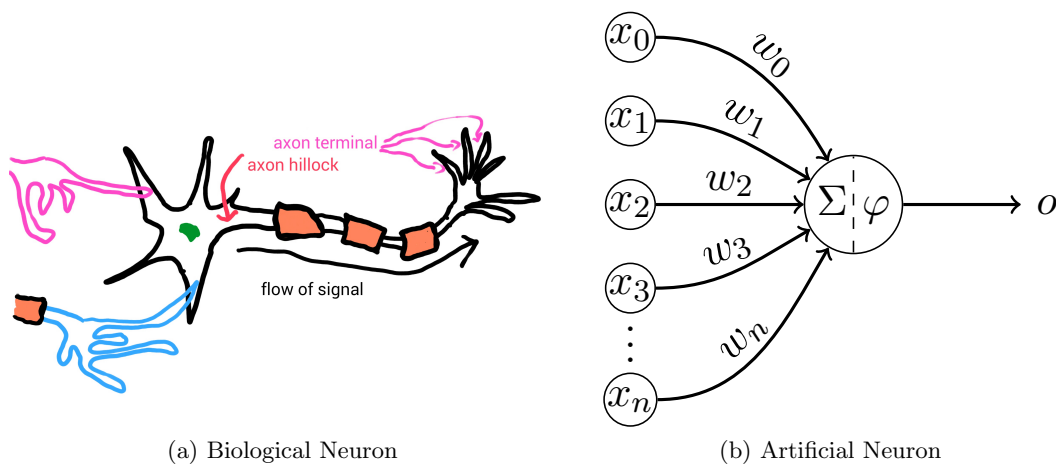


Figure 4.1.: Both neurons receive weighted input, apply a function to that and give output

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.

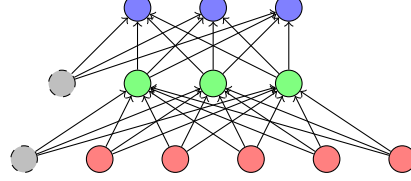


Figure 4.2.: Feedforward artificial neural network

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:

$$\begin{aligned} E &: \mathbb{R}^{n_1 \times n_2} \times \mathbb{R}^{n_2 \times n_3} \times \dots \times \mathbb{R}^{n_{\ell-1}, \ell} \rightarrow \mathbb{R}_{\geq 0} \\ E_T(W) &= \frac{1}{2} \sum_{(x, t_x) \in T} \sum_{p=1}^{n_\ell} \left(t_x^{(p)} - o_x^{(p)} \right)^2 \end{aligned}$$

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

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?)

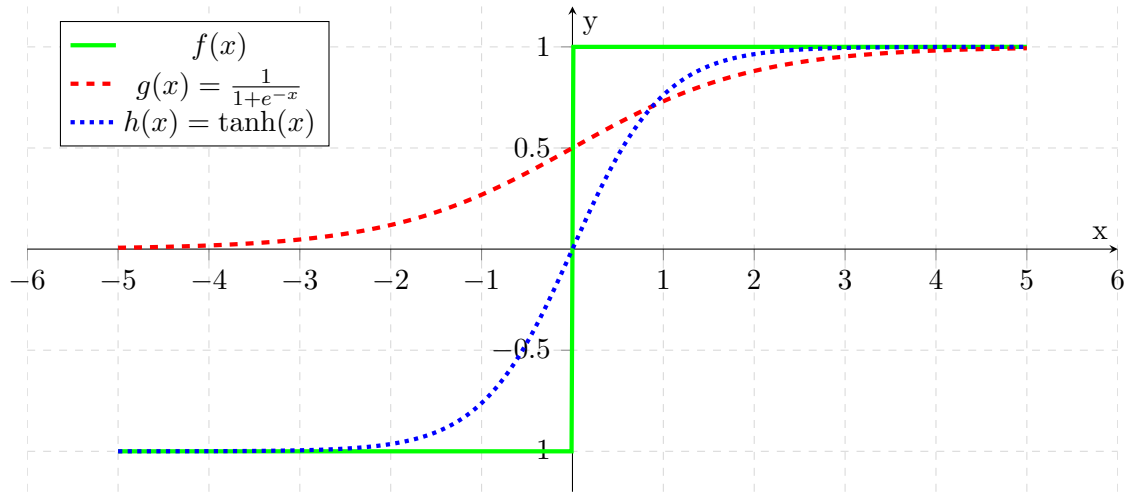


Figure 4.3.: A variation of the sign function f with $f(0) = -1$, the sigmoid function g and the hyperbolic tangend h .

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 `\Omega`.

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	<code>\Mundus</code> - 	7 lines	– <code>\fax</code> - 
11 lines	<code>\FAX</code> - 		– <code>\Emailct</code> - 
9 lines	<code>\sun</code>		– <code>\Letter</code> - 
8 lines	<code>\idotsint</code> - $\int \cdots \int$		– <code>\EyesDollar</code> - $\$$

	- \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 29) 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

- [Ara83] H. Arakawa, “On-line recognition of handwritten characters – Alphanumerics, Hiragana, Katakana, Kanji,” *Pattern Recognition*, vol. 16, no. 1, pp. 9 – 21, 1983. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/0031320383900031>
- [BN72] P. W. Becker and K. A. Nielsen, “Pattern recognition using dynamic pictorial information,” *Systems, Man and Cybernetics, IEEE Transactions on*, vol. SMC-2, no. 3, pp. 434–437, July 1972. [Online]. Available: <http://ieeexplore.ieee.org/xpl/abstractKeywords.jsp?arnumber=4309141>
- [Boa12] E. Board, *Concise Dictionary of Mathematics*, unknown, Ed. V&S Publishers, 2012. [Online]. Available: <http://books.google.de/books?id=7OqVdc2LGSUC>
- [Bro64] R. M. Brown, “On-line computer recognition of handprinted characters,” *Electronic Computers, IEEE Transactions on*, vol. EC-13, no. 6, pp. 750–752, Dec 1964. [Online]. Available: <http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=4038313>
- [BS89] R. Bozinovic and S. Srihari, “Off-line cursive script word recognition,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 11, no. 1, pp. 68–83, Jan 1989. [Online]. Available: <http://ieeexplore.ieee.org/xpl/articleDetails.jsp?tp=&arnumber=23114>
- [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>
- [GP93] W. Guerfali and R. Plamondon, “Normalizing and restoring on-line handwriting,” *Pattern Recognition*, vol. 26, no. 3, pp. 419–431, 1993. [Online]. Available: [http://dx.doi.org/10.1016/0031-3203\(93\)90169-W](http://dx.doi.org/10.1016/0031-3203(93)90169-W)
- [Gro66] G. F. Groner, “Real-time recognition of handprinted text,” in *Proceedings of the November 7-10, 1966, Fall Joint Computer Conference*, ser. AFIPS ’66 (Fall). New York, NY, USA: ACM, 1966, pp. 591–601. [Online]. Available: <http://doi.acm.org/10.1145/1464291.1464355>
- [HBT94] J. Hu, M. K. Brown, and W. Turin, “Handwriting recognition with hidden markov models and grammatical constraints,” in *In Proceedings of the Fourth International Workshop on Frontiers in Handwriting Recognition*, 1994. [Online]. Available: <http://www.bell-labs.com/user/jianhu/papers/iwfh94.ps>
- [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
- [iHY80] S. ichi Hanaki and T. Yamazaki, "On-line recognition of handprinted kanji characters," *Pattern Recognition*, vol. 12, no. 6, pp. 421 – 429, 1980. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/0031320380900187>
- [IMP76] S. Impedovo, B. Marangelli, and V. L. Plantamura, "Real-time recognition of handwritten numerals," *Systems, Man and Cybernetics, IEEE Transactions on*, vol. SMC-6, no. 2, pp. 145–148, Feb 1976. [Online]. Available: <http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=5409186>
- [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
- [KC98] A. Khotanzad and C. Chung, "Hand written digit recognition using combination of neural network classifiers," in *Image Analysis and Interpretation, 1998 IEEE Southwest Symposium on*, 4 1998, pp. 168–173. [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. Laviotte, 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-etal1999.pdf>
- [KWL95] M. Koschinski, H.-J. Winkler, and M. Lang, "Segmentation and recognition of symbols within handwritten mathematical expressions," in *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, vol. 4, May 1995, pp. 2439–2442 vol.4.
- [MFW94] S. Manke, M. Finke, and A. Waibel, "Combining bitmaps with dynamic writing information for on-line handwriting recognition," in *Proceedings of the ICPR-94*, 1994, pp. 596–598.

- [MFW95] —, “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)
- [Pow73] V. M. Powers, “Pen direction sequences in character recognition,” *Pattern Recognition*, vol. 5, no. 4, pp. 291 – 302, 1973. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/0031320373900228>
- [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
- [Tap87] C. C. Tappert, *Speed, Accuracy, Flexibility Trade-offs in On-line Character Recognition*, ser. Research report. IBM T.J. Watson Research Center, 1987. [Online]. Available: http://books.google.com/books?id=5br_HAAACAAJ
- [TSW90] C. C. Tappert, C. Y. Suen, and T. Wakahara, “The state of the art in online handwriting recognition,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 12, no. 8, pp. 787–808, 8 1990. [Online]. Available: <http://dx.doi.org/10.1109/34.57669>

Glossary

ANN artificial neural network. 6, 9

DTW dynamic time warping. 6, 7

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]. 13

HMM hidden Markov model. 6

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. 13

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 $\eta \leftarrow \eta \times \alpha$. 13

MLP multilayer perceptron. 10, 11

OOV out of vocabulary. 14

Appendix

A. Algorithms

Algorithm 4 Dehooking

```
function DEHOOK_LINE( $line, \theta \in \mathbb{R}_{\geq 0}$ )  
  if COUNT( $line$ ) < 3 then  
    return  $line$   
  else  
     $new\_line \leftarrow line[0 : \text{COUNT}(line) - 1]$   $\triangleright$  Get everything but the last point  
     $line \leftarrow line[\text{COUNT}(line) - 3 : ]$   $\triangleright$  get the last 3 points  
     $p \leftarrow line[-1]$   $\triangleright$  last point  
    if CALCULATE_ANGLE( $line$ ) <  $\theta$  then  
       $new\_line.APPEND(p)$   
    else  
       $new\_line \leftarrow \text{DEHOOK\_LINE}(new\_line, \theta)$   
  return  $new\_pointlist$ 
```

Algorithm 5 Dot reduction

```
function DOT_REDUCTION( $pointlist, \theta \in \mathbb{R}_{\geq 0}$ )  
   $new\_pointlist \leftarrow []$   
   $current\_line \leftarrow 0$   
  for all  $line$  in  $pointlist$  do  
     $new\_line \leftarrow line$   
     $max\_distance = \text{GET\_MAX\_DISTANCE}(line)$   
    if  $max\_distance < \theta$  then  
       $new\_line \leftarrow [\text{GET\_AVERAGE\_POINT}(line)]$   
     $new\_pointlist.APPEND(new\_line)$   
  return  $new\_pointlist$ 
```

Algorithm 6 Weighted average smoothing

```

function WEIGHTED_AVERAGE_SMOOTHING(pointlist,  $\theta = [\frac{1}{6}, \frac{4}{6}, \frac{1}{6}]$ )
   $\theta \leftarrow \frac{1}{\text{SUM}(\theta)} \cdot \theta$  ▷ Normalize parameters to a sum of 1
  new_pointlist  $\leftarrow []$ 
  current_line  $\leftarrow 0$ 
  for all line in pointlist do
    new_pointlist.APPEND([line[0]])
    if LENGTH(line) > 1 then
      for  $i \leftarrow 1; i < \text{LENGTH}(\text{line}) - 1; i \leftarrow i + 1$  do
         $p \leftarrow \theta_0 \cdot \text{line}[i - 1] + \theta_1 \cdot \text{line}[i] + \theta_2 \cdot \text{line}[i + 1]$ 
        new_pointlist[current_line].APPEND(p)
      new_pointlist[current_line].APPEND(line[-1])
    current_line  $\leftarrow \text{current\_line} + 1$ 
  return new_pointlist

```

Algorithm 7 Greedy matching as described in [Kir10]

```

a  $\leftarrow$  next from A
b  $\leftarrow$  next from B
d  $\leftarrow \delta(a, b)$ 
a'  $\leftarrow$  next from A
b'  $\leftarrow$  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'  $\leftarrow$  next from A
  else if  $r = \mu$  then
    b  $\leftarrow b'$ 
    b'  $\leftarrow$  next from B
  else
    a  $\leftarrow a'$ 
    b  $\leftarrow b'$ 
    a'  $\leftarrow$  next from A
    b'  $\leftarrow$  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  $\leftarrow d + \delta(b', p)$ 

```

Algorithm 8 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

```
