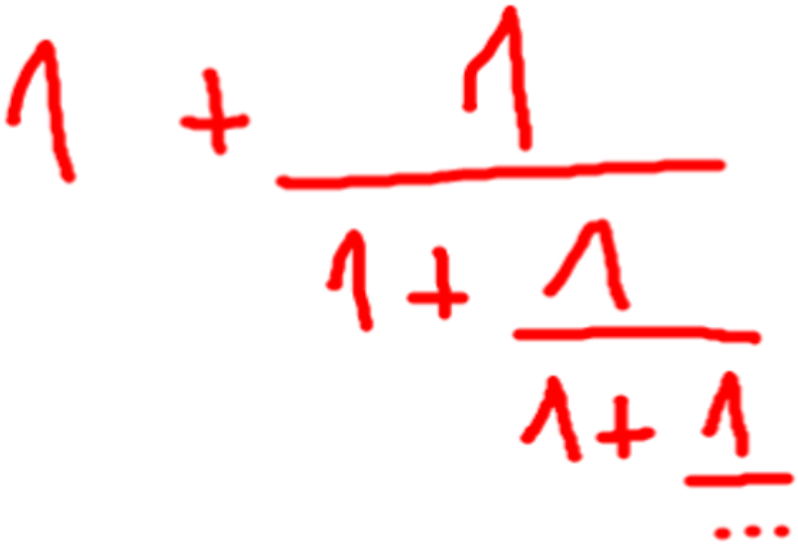
Karlsruhe Institute of Technology

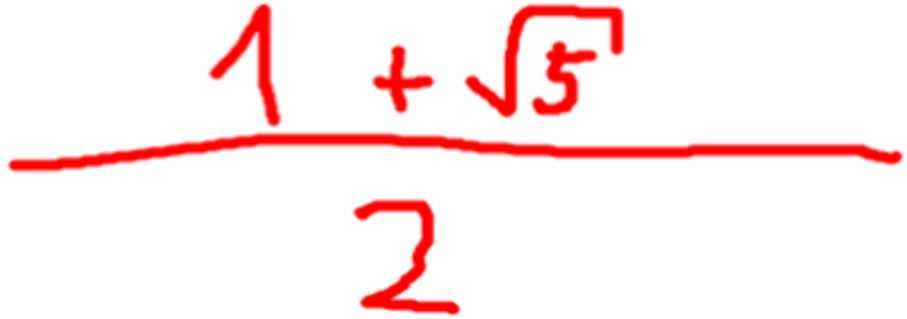
Carnegie Mellon University

On-line recognition of handwritten mathematical symbols

Bachelor's thesis of Martin Thoma

Martin Thoma | 5th of June, 2014





Contents

What is my Bachelor's thesis about?

- Recognition of handwritten mathematical symbols
- On-line recognition, not OCR!
- Given a series of points $(x(t), y(t), b(t))$
  I want to get the proper LaTeX code.

Why do I work on this topic?

- LaTeX is easy as soon as you know the \codes.
- It's hard to find the LaTeX code of single symbols.
- It's much harder to find complete formulas.

For now: recognition of isolated symbols. That means:
single symbol "formulae" rather than multi symbol formulae

write-math.com

- a website where users can add labeled training data and unlabeled data which they want to classify. I call this data "recording"

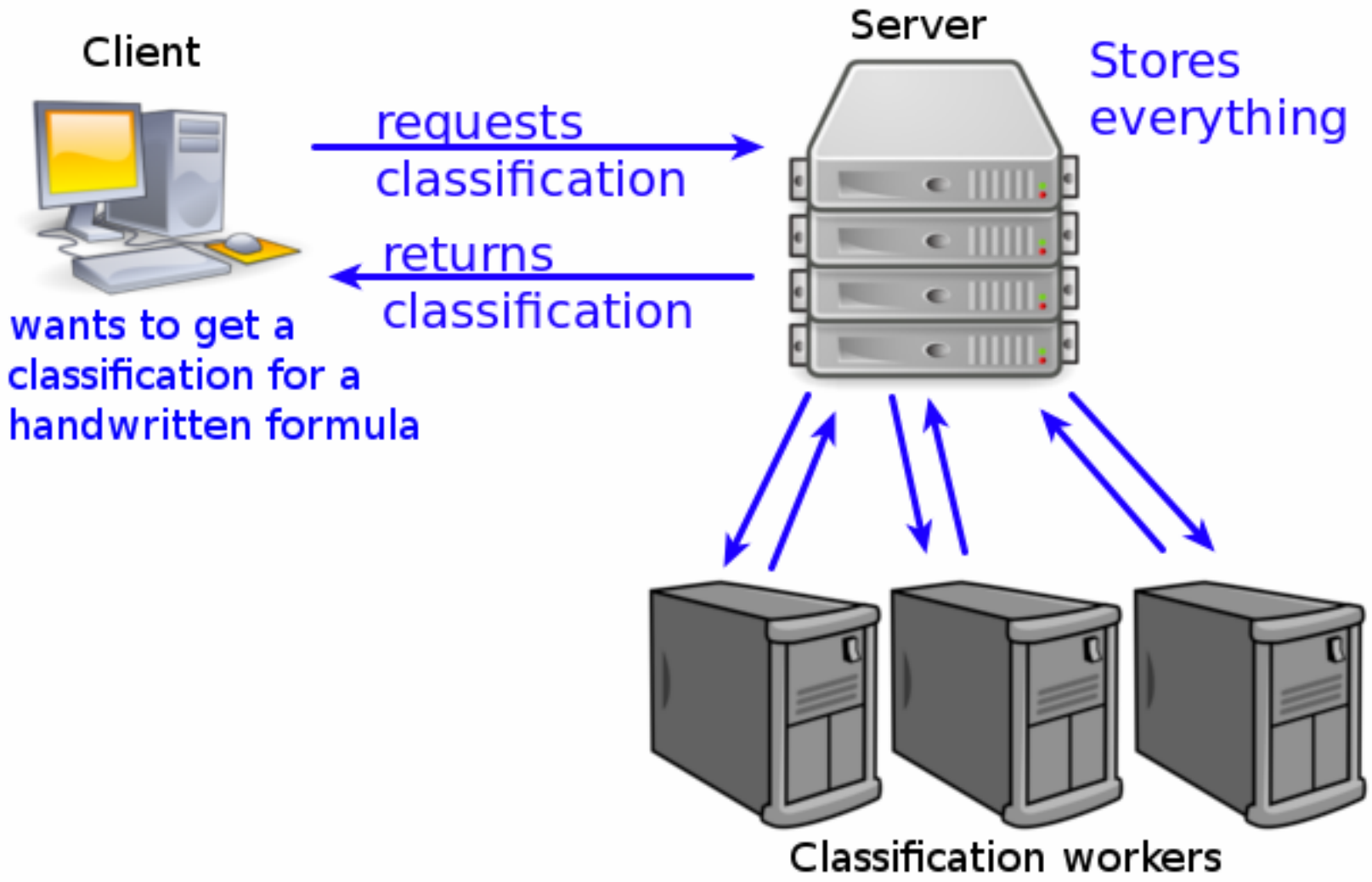
4 recordings

- works with desktop computers and touch devices
- symbol recognition can be done by multiple classifiers
- users can contribute formulas as recordings and as LaTeX answers for recordings
- users can vote for LaTeX answers: $\leq$, $\leqq$, $\leqslant$, ...
- user who entered the recording can accept one answer

Classify


Write Math
Classify
Train
Gallery
Ranking
About
Martin Thoma
Classify
Submit
Clear

Workflow


Client
Server
Stores everything
requests classification
returns classification
wants to get a classification for a handwritten formula
Classification workers

Ranking

Write Math | Classify | Train | Gallery | Ranking | About | Martin Thoma

Ranking

Only users with at least 5 written formulas will be listed below.

#	User	Written formulas	Distinct symbols
1	Detexify	217684	1125
2	Martin Thoma	4382	523
3	user_639125948	3071	430
4	Eva	1134	566
5	John	781	722
6	TorbjornT	572	253
7	user_1904016610	510	124
8	Marienkaefer	458	260
9	percusse	411	317
10	Brent	374	196

Statistics

- 127 users with at least 5 recordings
- 1111 symbols, but only 369 used for experiments
- 235 831 recordings (e.g. 3489 times `\int`, but only 50 times `X`)

First classification worker

- preprocessing: Scale to fit into unit square while keeping the aspect ratio
- applies greedy time warping
- compares a new recording with every recording in the database

$\Rightarrow$ Classification time is in $\mathcal{O}(\text{recordings})$, but we rather would like $\mathcal{O}(\text{symbols})$

- the current server / workflow can only handle about 4000 recordings

$\Rightarrow$ Another way to classify is necessary

Preprocessing

- Normalizing
  - Scaling
  - Shifting
  - Resampling
- Noise reduction
  - Smoothing (e.g. moving average)
  - Dot reduction
  - Filtering (by distance, speed or angle)
  - Stroke connection

Features

- Local
  - Coordinates
  - Speed
  - Binary pen pressure
  - Direction
  - Curvature
  - Bitmap-environment
  - Hat-Feature
- Global
  - # of points
  - # of strokes
  - Center point
  - Bitmap
  - Bounding box (width, height, time)

Experiments

Preprocessing: Scaling, shifting and linear interpolation
Features: Coordinates of 80 points (4 strokes with 20 points each)
Learning: MLP, 300 epochs, LR of 0.1, Momentum 0.1

Topology	Error	Training time
160:500:369	30.62 %	9min 08s
160:500:500:369	27.73 %	11min 49s
160:500:500:500:369	34.79 %	14min 09s
160:500:500:500:500:369	33.61 %	14min 06s

Examples of confusable symbols

LaTeX	Rendered	LaTeX	Rendered
\sum	$\sum$	Σ	Σ
\coprod	$\coprod$	$\amalg$	$\amalg$
\perp	$\perp$	$\bot$	$\bot$
\models	$\models$	$\vDash$	$\vDash$
\emptyset	$\emptyset$	$\diameter$	$\diameter$
		$\o$	$\o$
		$\varnothing$	$\varnothing$
\Delta	Δ	$\triangle$	$\triangle$
\varepsilon	ε	$\mathcal{E}$	$\mathcal{E}$

When those confusions are not counted as errors, the current best system has an classification error rate of 12.7% (otherwise 22.2%).

What will I do next?

- Include the currently best model in write-math.com
- Evaluate preprocessing steps
- Try other features
- Try other topologies / trainings (e.g. pretraining, newbob)
- Eventually try convolutional neural nets

Image Sources

- Server by RRZEicons
- Desktop Computer by Ed g2s, Ironbrother, Kierancassel and Msgj
- Server by Mimooh

The presentation can be found at
http://tinyurl.com/write-math-short-presentation

Thanks for Your Attention!