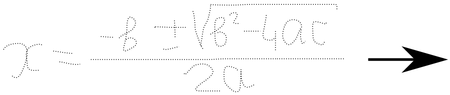
Towards Online Recognition of Handwritten Mathematics

Vadim Mazalov, joint work with Oleg Golubitsky and Stephen M. Watt

Ontario Research Centre for Computer Algebra
Department of Computer Science
Western University
London, Canada

Introduction

Our objective is to recognize handwritten mathematics in pen-based environment.



$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

Organization

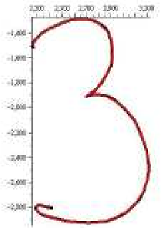
- Background of handwriting, approximation and classification
- Rotation-invariant recognition
- Shear-invariant recognition
- Optimization of recognition of isolated characters
- Optimization of recognition of distorted groups of characters

Digital handwriting

- Represented as a sequence of points

$$(x_0, y_0), (x_1, y_1), (x_2, y_2), \ldots$$

- Each point contains one value of certain channel



Decomposition of Channels

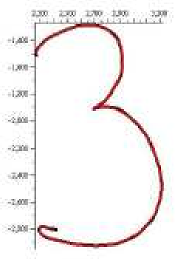
- Consider X and Y coordinates separately, as functions of arclength. Then

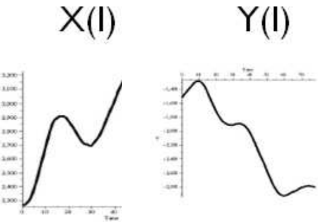
$$(x_0, y_0), (x_1, y_1), (x_2, y_2), \ldots$$

  becomes

$$(l_0, x_0), (l_1, x_1), (l_2, x_2), \ldots \text{ and } (l_0, y_0), (l_1, y_1), (l_2, y_2), \ldots$$

- And





Functional Approximation

- An inner product $\langle f, g \rangle$ gives the orthogonal basis $\{\phi_0(\lambda), ..., \phi_d(\lambda)\}$ on a subspace of polynomials by GS orthogonalization.
- We use a Legendre-Sobolev inner product so we can measure distance in the first jet space

$$\langle f, g \rangle = \int_{-1}^{1} f(\lambda) g(\lambda) \mathrm{d}\lambda + \mu \int_{-1}^{1} f'(\lambda) g'(\lambda) \mathrm{d}\lambda$$

- Can approximate

$$f(\lambda) \approx \sum_{i=0}^{d} \alpha_i \phi_i(\lambda)$$

Pre-classification

- Coefficients are computed as

$$c_i = \frac{\langle F, P_i \rangle}{\langle P_i, P_i \rangle} \quad i = 1..d$$

where $\langle \cdot, \cdot \rangle$ is the L-S inner product.

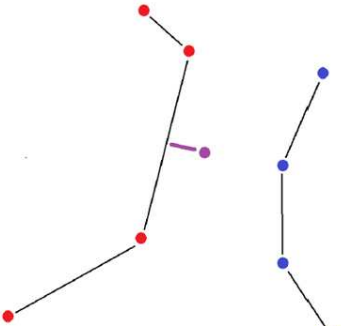
- Having the coefficients vector

$$x_0, x_1, \ldots, x_d, y_0, y_1, \ldots, y_d$$

the sample can be normalized with respect to position and size by omitting the first point (x_0, y_0) and normalizing the vector $(x_1, \ldots, x_d, y_1, \ldots, y_d)$.

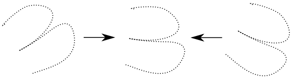
Classification

Classification is based on the distance to convex hulls of nearest neighbours

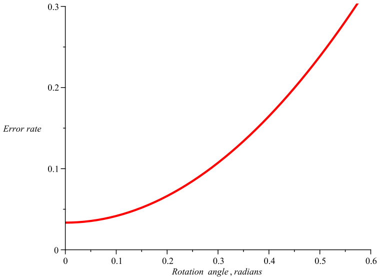


Rotation

- Commonly occurs in practice.



- Decreases recognition rate of most of algorithms as a function of the rotation angle.



Challenge

- Question

Is it possible to make an algorithm independent of rotation?

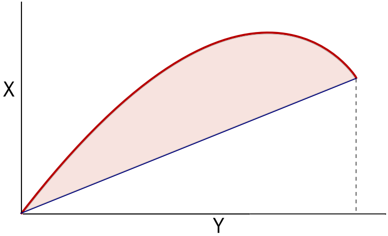
- Answer

Yes. Describe the curve in terms of rotation invariant functions:

- Integral invariants (II)
- Geometric moments (GM)

Geometric Representation of II

- Invariant $I_0(\lambda)$ is the radius to a point on a curve.
- Invariant $I_1(\lambda)$ is the area between the curve and a secant.



Functional Representation of II

- I_0 and I_1 are invariant under rotation and may be given in terms of coordinate functions $X(\lambda)$ and $Y(\lambda)$:

$$I_0(\lambda) = \sqrt{X^2(\lambda) + Y^2(\lambda)} = R(\lambda),$$
$$I_1(\lambda) = \int_0^\lambda X(\tau)dY(\tau) - \frac{1}{2}X(\lambda)Y(\lambda)$$

  where $X(\lambda)$, $Y(\lambda)$ are parameterized by Euclidean arc length.
- We calculate coefficients for $I_0(\lambda)$ and $I_1(\lambda)$ and obtain a $2d$-dimensional vector for each sample

$$(I_{0,1}, \ldots, I_{0,d}, I_{1,1}, \ldots, I_{1,d}).$$

Geometric Moments

- Of special interest for the purpose of online curve classification under pressure of computational constraints, since they are easy to calculate, while invariant under scaling, translation and rotation.
- A $(p+q)$-th order moment of f can be expressed as

$$m_{pq} = \sum_x \sum_y x^p y^q f(x, y)$$

- Moment invariants have form

$$\begin{aligned} M_0(\lambda) &= m_{00}(\lambda), \\ M_1(\lambda) &= m_{20}(\lambda) + m_{02}(\lambda), \\ M_2(\lambda) &= (m_{20}(\lambda) - m_{02}(\lambda))^2 + 4m_{11}^2(\lambda). \end{aligned}$$

Classification

The following algorithms are introduced:

- Classification with integral invariants (CII)
- Classification with coordinate functions and integral invariants (CCFII)
- Classification with coordinate functions and moment invariants (CCFMI)

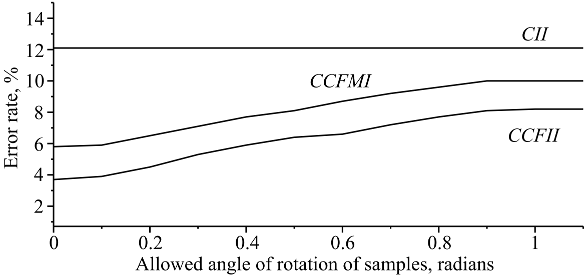
Experimental Setting

- Our dataset comprised 50,703 handwritten mathematical symbols from 242 classes.
- Testing is implemented in 10-fold cross-validation (the dataset was randomly divided into 10 parts, preserving proportions of class sizes).
- Normalized Legendre-Sobolev coefficient vectors of coordinate functions of randomly rotated symbols, as well as coefficients of integral invariants, were pre-computed for all symbols.

Comparison of results

Table: Error rates of CII, CCFII and CCFMI

α, rad.	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	1.0	1.1
CII	12	12	12	12	12	12	12	12	12	12
CCFII	**3.7**	**3.9**	**4.5**	**5.3**	**5.9**	**6.4**	**6.6**	**7.2**	**8.2**	**8.2**
CCFMI	5.8	5.9	6.5	7.1	7.7	8.1	8.7	9.2	10	10

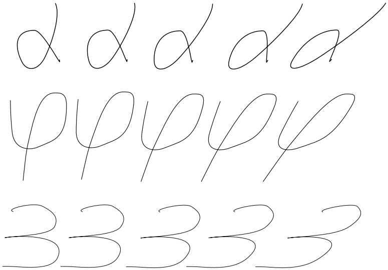


Analysis of Results

- CCFII performs noticeably better, while requiring less computation.
- As expected, we observed an increase in error rate with the rotation angle for CCFII and CCFMI. Typical misclassifications arise when distinct symbols have similar shape and are normally distinguished by their orientation, for example "1" and "/", "$+$" and "$\times$", "U" and "$\subset$".
- This result can still be improved, but in a different setting.

Shear

- Commonly occurs in practice, possibly even more often than rotation.



- Easily recognizable by a human, even for a large degree of transformation (more than 1 radian).

Shear

- Parameterization by arc length is not suitable, since the length of a curve changes under shear.
- Size normalization is not trivial.
- Shear may easily transform symbols to different characters.
- Can be treated with integral invariants similar to rotation.

Functional Representation

- In addition to I_1, introduced earlier, we take I_2

$$I_2(\lambda) = X(\lambda)\int_0^{\lambda} X(\tau)Y(\tau)dY(\tau) - \frac{1}{2}Y(\lambda)\int_0^{\lambda} X^2(\tau)dY(\tau) - \frac{1}{6}X^2(\lambda)Y^2(\lambda)$$

  where $X(\lambda)$, $Y(\lambda)$ are coordinate functions.
- Both I_1 and I_2 are invariant under special linear group $SL(2, R)$.
- I_2 can be described geometrically in terms of volume.

Size Normalization: Existing Methods that we test

- Taking the Euclidean norm of the vector of L-S coefficients of coordinate functions.
- Normalizing by the height of a sample (works for horizontal shear, but not for rotation).
- Aspect-ratio size normalization (may work for rotation, but becomes inaccurate for horizontal shear).

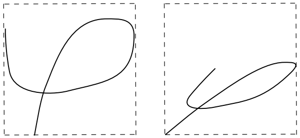


Figure: Aspect ratio size normalization.

Size Normalization: Our Approach

- For the case of shear (and affine) recognition we look at the norm $\|I_1\|$ of the coefficient vector of I_1.

- Coefficients of coordinate functions are normalized by multiplication by $1/\sqrt{\|I_1\|}$.

- Computing the norm of I_1 allows to extend the invariance of I_1 and I_2 from the special linear group, $SL(2, R)$, to the general linear group, $GL(2, R)$.

Parameterization of Coordinate Functions

We test the following parameterization approaches

- By time.
- By Euclidean arc length

$$F(\lambda) = \int_0^\lambda \sqrt{(X'(\tau))^2 + (Y'(\tau))^2} d\tau.$$

- By affine arc length

$$\hat{F}(\lambda) = \int_0^\lambda \sqrt[3]{|X'(\tau)Y''(\tau) - X''(\tau)Y'(\tau)|} d\tau.$$

Classification: Selecting the Class

- We select N classes in the space of L-S coefficients of the integral invariants.
- We consider each of the selected classes C_i to find the minimal distance to the subject sample with respect to different levels of shear in the space of L-S coefficients of the coordinate functions:

$$\min_{\phi} \mathrm{CHNN}_k(X(\phi), C_i),$$

where $X(\phi)$ is the sheared image of the test sample curve X and $\mathrm{CHNN}_k(X, C)$ is the distance from a point X (in the L-S space) to the convex hull of k nearest neighbors in class C.

Classification

Classification results for different types of parameterization of coordinate functions: by time, by Euclidean arc length (AL) and by affine arc length (AAL) for discussed size normalization approaches.

(a) Size normalization by height

	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
AAL	82.2	82.2	82.2	82.1	82.1	82.1	82.1	82.1	82.1	82
AL	96.4	96.4	96.1	95.6	95	94.1	93	91.9	90.2	88
Time	94.8	94.9	94.9	94.7	94.5	94.4	94.4	94.4	94.4	94.3

(b) Aspect ratio size normalization

	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
AAL	81.9	81.8	81.6	81.4	81.2	81	80.8	80.2	79.4	77.4
AL	96.3	96.4	96.1	95.5	94.7	93.7	92.3	90.1	85.7	77.5
Time	94.7	94.7	94.6	94.3	94.1	93.9	93.7	93.2	91.9	89.

(c) Size normalization by l_1

	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
AAL	83	83.1	83	82.9	82.9	82.8	82.8	82.8	82.8	82.7
AL	**96.3**	**96.3**	**96.1**	**95.7**	**95.1**	**94.4**	**93.3**	**91.9**	**90.2**	**87.9**
Time	94.6	94.7	94.6	94.5	94.5	94.5	94.5	94.5	94.5	94.4

Mixed Parameterization

Parameterization by time gives low recognition rate, while remains affine-invariant. It is opposite for parameterization by arc length. We, therefore, propose to unite these two parameterization approaches in the form of mixed parameterization as follows

- Divide the curve in N equal time intervals, and parameterize each interval by arc length.
- Smooth the transition from time to arc length with a mixed metric of the form $kdt^2 + dx^2 + dy^2$ inside the subintervals, where k is a parameter.
- The optimal values of N and k are found by cross-validation.

Recognition rate for different N and k

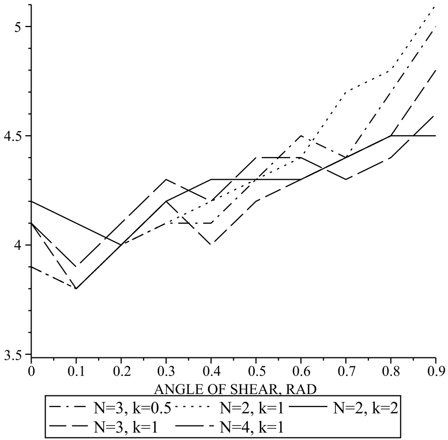


Figure: Error (%) for the mixed parameterization for different values of skew.

Further optimization

- Optimization of isolated characters recognition by adjusting the
  - "jet scale" in the LS inner product for coordinate functions and integral invariants
  - degree of approximation of coordinate functions and integral invariants.
- An in-context rotation-invariant algorithm that yields substantially better results than isolated recognition and can be extended to other transformations.

Coordinate Functions

- The optimal μ for approximation of coordinate functions was found to minimize classification error.
- The original characters in the dataset were considered without any distortions.
- The characters were approximated with a corresponding value of μ in the range from 0 to 0.2.

Integral Invariants

- The optimal μ for approximation of integral invariants was found to minimize the average relative error in coefficients of the original and rotated samples

$$\frac{\sum_{ij}\left|c_{ij}-c'_{ij}\right|}{\sum_{ij}\left|c_{ij}\right|}$$

  where c_{ij} is the j-th coefficient of the i-th original sample, and c'_{ij} is the corresponding coefficient of the sample, rotated on an angle.

- The whole collection of original samples was rotated by an angle between 0 and 2π with the step of $\pi/9$.

- The integral invariants were approximated with LS series for different values of $\mu \in (0, 0.2]$ with the step of 0.002.

Complexity of Handwritten Characters

- We considered the possibility that the optimal value of μ might depend on the nature of the characters to be recognized.

- We took the notion of a sample's complexity as

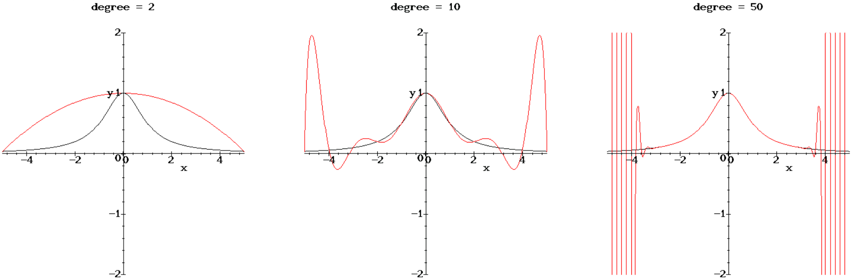
$$\eta = \sum_{i=1}^{d} (X_i^{1/i} + Y_i^{1/i}), |X_i| \leq 1 \text{ and } |Y_i| \leq 1$$

  where X_i and Y_i are normalized coefficients of approximation of the sample with orthogonal polynomials

- The sample complexity function is derived from the fact that coefficients of higher degree are typically greater for "complex" characters – characters that contain large number of loops and/or amount of curvature.

Degree of Approximation

- Small degree leads to high approximation error.
- High degree may cause extreme oscillation at the edges of an interval.



- The optimal degree of the truncated series was determined to minimize the
  - recognition error for coordinate functions
  - approximation error for integral invariants.

The images are from http://www.cs.laurentian.ca/badams/numeric/runge/runge.html

In-context transformation-invariant recognition

- Recognition of distorted math symbols without context is sometimes impossible

$$\text{“}<\text{”} \text{ vs } \angle, \text{ “}|\text{” vs “}/\text{”}, \text{ “}\cup\text{” vs “}\subset$$

- We use context to improve distortion-independent classification, taking advantage of the fact that samples written by a person typically exhibit similar degree of transformation.

- We consider the case of rotation (shear and other transformations may be handled similarly).

The algorithm

Input: A set of n rotated test samples and an angle β of the maximum possible rotation angle of the samples.

Output: A set of n recognized samples and the angle γ of rotation of the samples.

for $i = 1$ to n

Approximate coordinate functions with LS polynomials.

Normalize the sample with respect to position and size.

Approximate I_0 and I_1 with LS polynomials.

With LS coefs of I_0 and I_1, find T closest CHNN_k. These T classes serve as candidates for the i-th sample in the sequence.

end for.

// continued on next slide

The algorithm, cont.

for $\alpha = -\beta$ to β by step of 1 degree

Compute

$$\epsilon_\alpha = \prod_{i=1}^{n} \frac{D_{i\alpha}^1}{\sum_{j=1}^{p} D_{i\alpha}^j}$$

where $D_{i\alpha}^j$ is the Euclidean distance to the j-th *closest* CHNN_k among the candidate classes T for the sample i in the sequence, rotated by angle α, and p is a parameter to be evaluated. Distance D is computed in the space of coefficients of LS polynomials of coordinate functions.

end for

Find $\epsilon_\gamma = \min_{-\beta \leq \alpha \leq \beta} \epsilon_\alpha$

return n and γ.

Coordinate functions

- The figure shows the error of recognition using coefficients of X and Y.

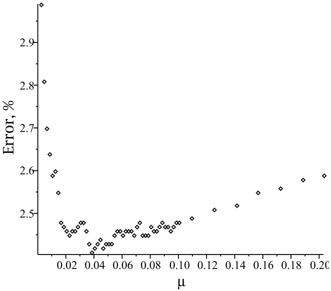
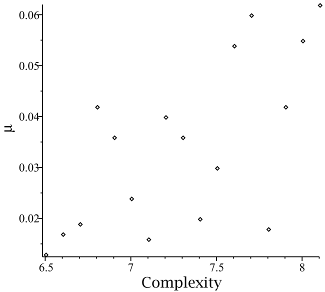


Figure: Recognition error of non-transformed characters for different values of μ

- $\approx 2.4\%$ error is reached for $\mu = 0.04$.

Optimal μ for characters with different complexities

- We found that the optimal μ is not strongly correlated with the complexity of characters.



- Results of Spearman and Kendall tau-a correlation tests are respectively: $\rho_{\mu,\eta}(13) = 0.52, p = 0.047$ and $\tau_{\mu,\eta}(13) = 0.38, p = 0.053$..

I_0

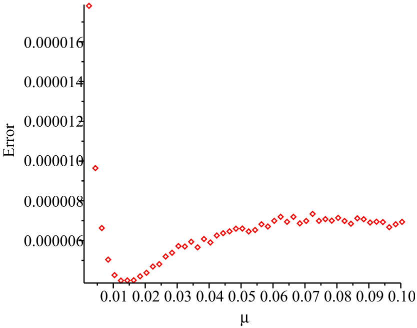
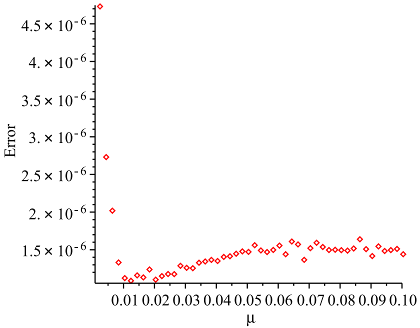


Figure: Average relative error in coefficients of I_0 depending on μ

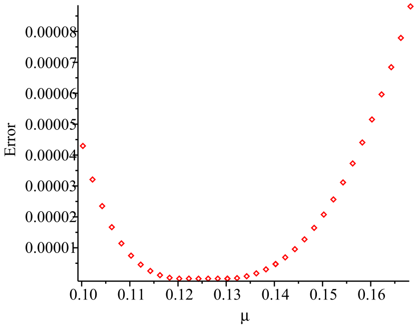
I_1



(a)

Figure: Average relative error in coefficients of I_1 depending on μ

I_2



(a)

Figure: Average relative error in coefficients of I_2 depending on μ

Degree of Approximation

Table: Recognition error (Rec.Err.), maximum approximation error (Max.Err.) and average relative error (Avg.Err.) for different degrees of approximation d, $\mu = 0.04$

d	9	10	11	12	13	14	15
Rec.Err. %	2.57	2.49	2.46	2.43	2.44	2.45	2.46
Max.Err.	707	539	539	484	475	494	500
Avg.Err.($\times 10^{-3}$)	1.9	1.6	1.4	1.2	1.1	1.0	1.2

We find degree 12 to be the optimum for recognition of symbols in our collection.

Parameters

There are 3 parameters that in-context recognition rate can depend on:

- Number p of closest classes in computation of error likelihood.
- Rotation angle.
- Size n of the set of characters.

Evaluatoin of p

- We fixed the parameter $n = 3$.
- Performed classification for values p of 2, 3 and 4.
- We found that p has almost no effect on recognition error.
- We took $p = 3$ and continued the experiments.

Evaluation of *n* and the rotation angle

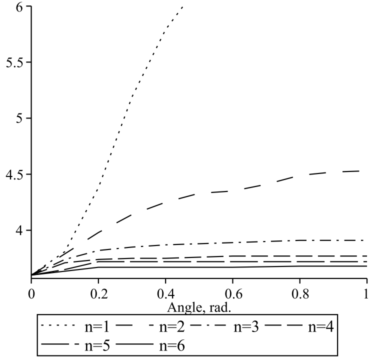


Figure: Recognition error (%) for different size of context *n* and different angles of rotation (in radians)

Conclusion

We have developed

- Orientation and shear-independent algorithm of recognition of handwritten characters.
- Size normalization of samples, when shear (and more generally, affine) transformations take place.
- Mixed parameterization of coordinate functions that allows to obtain high recognition rate.
- Shear-invariant algorithm that in conjunction with rotation-invariance (introduced in the previous work) allows to model recognition of samples, subjected to the most common affine transformations.
- Optimization of recognition of characters by themselves and in groups.