Pen-Based Computing in Medicine: Factorial Analysis of the Rotation-Invariant Recognition Algorithm

Vadim Mazalov[a]*, Dmitry Mazalov[b], and Anna Pauer[c]

[a]Department of Computer Science, University of Western Ontario, London, Canada
[b]Kuban State University of Technology, Krasnodar, Russia
[c]Department of Biology, University of Western Ontario, London, Canada

ABSTRACT

Classification of handwritten characters is an essential element in development of algorithms for medical text recognition. We propose an algorithm for recognition of sequences of rotated characters. The algorithm is based on finding the angle of rotation that yields the least error likelihood. There are several parameters that the recognition rate depends on. To accurately estimate the parameters, we perform $2^k r$ factorial design. Results obtained in factorial design are valid only with certain assumptions. We show that those assumptions are satisfied.

Keywords: Medical Handwriting Recognition, Medical Informatics, Rotation-Invariance, Factorial Design

1 BACKGROUND

Handwriting is one of the most natural ways for a human to record knowledge. In recent years this type of human-computer interaction has received increasing attention due to the rapid evolution of digital ink hardware. As applied to medicine and bioinformatics, pen-based input of data has a significant advantage compared to conventional techniques (with a pen and paper, or even with a mouse and a keyboard) in a way that it allows patient data acquisition in two-dimensional format (schemes, tables, graphs and formulas) along with regular text. It is useful in digital recording of symptoms, diagnosis and medical tests. Our objective is development of algorithms to allow recognition and classification of handwriting for next-generation user interfaces of clinical software. Combined with appropriate network and storage infrastructure, the patient data obtained from a tablet can be saved in a cloud in real time for analysis and sharing with eligible entities.

Some research has been done in applying pen-based technologies in clinical setting, e.g. see [1, 2, 3, 4, 5]. However, the mentioned papers do not focus on recognition primarily. In contrast, our research contributes to the art of online classification of handwritten characters. Even though considerable work has been done in the field of handwriting recognition, classification of medical symbols and text requires special attention. Among the factors that give recognition of medical texts additional challenges beyond those of normal text recognition is the high density of mathematical and chemical characters, special symbols, tables and graphs. Such variety creates limitations for syntactic verification of recognized objects. An additional challenge is that symbols are usually subjected to transformations in positioning and in form (scale, rotation and shear). An example of handwritten input of a result of blood test on iron is given in Figure 1. In general, an ability to recognize digits, special and mathematical characters is an important part of a robust medical text recognition algorithm.

In the past we have developed an algorithm for rotation-invariant classification of handwritten symbols [6], based on the ideas proposed in [7]. We represent coordinates of a sample as a parameterized curve and approximate the curve with orthogonal polynomials up to degree d. Coefficients of approximation serve as a descriptor of the sample in $2d$-dimensional space. Classification is based on the distance to convex hulls of nearest neighbours. Normalization with respect to size and position is achieved respectively by normalizing the coefficient vector and dropping the first (0-order) coefficient. To be able to recognize samples independently of rotation, we proposed to approximate the invariant functions of coordinates, rather than coordinate functions themselves. We tested 2 types of invariant functions: geometric moments [8] and integral invariants [9] and found the latter to perform significantly better while requiring less amount of resources [6]. In this paper we propose a new algorithm for rotation-invariant recognition of characters, that is based on analysis of sequences of characters at a time, rather than sample at a time. We also perform $2^k r$ factorial design to estimate which of the k parameters have the most effect on recognition rate.

*Corresponding author. E-mail: vmazalov@uwo.ca, Telephone: +1(519)661-2111 (ext. 83741).

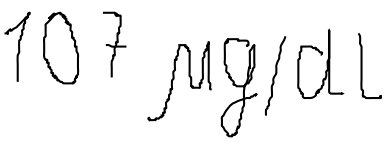

Figure 1. An example of handwritten input of a medical test result

The rest of the paper is organized as follows. In section 2 we describe the proposed algorithm and the parameters on which it depends. Section 3 contains results of the algorithm on five different datasets for corresponding values of the parameters. Factorial analysis is presented in Section 3. Section 4 concludes the paper.

2 RECOGNITION OF SEQUENCES OF ROTATED CHARACTERS

To improve classification of rotated characters, we propose to study sequences of samples with an assumption that characters in the sequence are rotated on the same angle with a minor difference. The difference in rotation between every pair of characters in a sequence is $\leq 2\beta$, where β is a noise angle. For each sequence we look for an angle that allows to minimize likelihood of an error in recognition of the samples in the sequence. We define likelihood of an error of a test sample as the relation

$$\gamma_\alpha = \frac{d_\alpha}{\sum_{i=1}^{p} d^i_{min}} \quad \text{and} \quad \gamma_{[\alpha_1;\alpha_2]} = \min_\alpha \{\gamma_\alpha, \alpha \in [\alpha_1;\alpha_2]\}$$

where d_α is the minimal distance of the test sample (rotated on angle α) among the distances to convex hulls of nearest neighbours of all training classes, and $\sum_{i=1}^{p} d^i_{min}$ is the sum of p minimal distances for all angles, e.g.

$$d^1_{min} = \min_\alpha \{d_\alpha, \alpha \in [\alpha_{\min};\alpha_{\max}]\} \quad \text{and} \quad d^i_{min} = \min_\alpha \left\{d_\alpha, d_\alpha > d^{i-1}_{\min} \;\&\; \alpha \in [\alpha_{\min};\alpha_{\max}]\right\}$$

where $\alpha_{\min} = -\alpha_{\max}$ and angle $\alpha_{\max}$ is one of the parameters in the factorial design.

Total error likelihood of samples in a sequence is computed as $\gamma'_\alpha = \sum_{i=1}^{n} \gamma^i_{[\alpha-\beta;\alpha+\beta]}$, where $\gamma^i_{[\alpha-\beta;\alpha+\beta]}$ is the minimal likelihood of the i-th sample in the sequence on $[\alpha-\beta;\alpha+\beta]$.

We then find the rotation angle θ that yields the minimal error likelihood $\gamma_\theta = \min \gamma'_\alpha$ among all the angles on the interval with the step of 1 degree. Having found the transformation angle, we rotate the test samples in the sequence on the angle and recognize the samples as proposed in [6].

3 2^4 Factorial Design with 5 Replications

A 2^k factorial design allows to evaluate performance of a system depending on k factors and each factor takes 2 values. This type of analysis received significant attention due to its simplicity and sufficient power in sorting out factors depending on their impact on performance. The 2^k factorial design can be applied in a setting, when effect of factors is unidirectional – performance decreases or increases continuously while a factor is being changed. Therefore, selecting two significantly different values of a factor and measuring difference in performance is a good starting point in performance evaluation. Later on, if the difference in performance is significant enough, a detailed examination may take place [10]. We deploy the 2^k analysis to evaluate which factors have the most effect on the recognition rate of the algorithm proposed in section 2.

A $2^k r$ factorial design is used to isolate experimental errors. In this design each of the experiments is repeated r times and error term is added to the model [10]. We implement $2^4 5$ factorial design with 4 factors (listed in the following subsection) and 5 repetitions.

3.1 Recognition Results

We select the following factors for analysis (two levels are given for each of the factors): *rotation angle*, α (0.3 and 0.6 rad.); *noise angle*, β (0.0 and 0.1 rad.); *size of the sequence*, n (3 and 5) and the *number of distances* to consider in $\sum_{i=1}^{p} d^i_{min}$, p (3 and 5).

The model has the form, as discussed in [10]

$$y = q_0 + q_\alpha x_\alpha + q_\beta x_\beta + q_n x_n + q_p x_p + q_{\alpha\beta} x_\alpha x_\beta + \ldots + q_{\alpha\beta np} x_\alpha x_\beta x_n x_p + e$$

where q's are effects and e is experimental error.

y_1	94.9	95.0	94.4	94.8	96.4	96.4	95.9	96.2	95.0	95.1	94.6	95.0	96.4	96.5	96.2	96.3
y_2	94.4	94.6	93.5	94.0	96.0	96.1	95.8	96.0	94.6	94.6	94.1	94.3	96.0	96.1	95.9	96.0
y_3	94.9	95.0	94.0	94.4	96.1	96.0	95.8	95.9	95.0	95.1	94.5	94.8	96.1	96.1	95.9	96.1
y_4	94.6	95.0	93.9	94.5	96.0	95.9	95.7	95.8	95.0	95.1	94.6	94.8	96.1	96.1	96.0	96.0
y_5	94.7	94.8	94.3	94.5	96.2	96.3	96.0	96.1	94.8	94.9	94.4	94.6	96.2	96.3	96.1	96.2

Table 1. Recognition rate for 2^4 runs for corresponding sign table

α	β	n	p	α , β	α , n	β,n	α,p	n,p	β,p	α, β,n	α, β,p	α, n,p	β, n,p	α, β, n,p	Error
1.06%	86.08%	3.99%	1.00%	0.16%	0.35%	0.95%	0.03%	0.27%	0.23%	0.03%	0.03%	0.02%	0.04%	0.00%	5.76%

Table 2. Percentage variation of factors

3.2 Allocation of Variation

Having computed effects from the experimental data shown in Table 1, we evaluate response $\hat{y}_i$ for each combination of factors $(x_{\alpha_i}, x_{\beta_i}, x_{n_i}, x_{p_i})$ as

$$\hat{y}_i = q_0 + q_\alpha x_{\alpha_i} + q_\beta x_{\beta_i} + q_n x_{n_i} + q_p x_{p_i} + q_{\alpha\beta} x_{\alpha_i} x_{\beta_i} + \ldots + q_{\alpha\beta np} x_{\alpha_i} x_{\beta_i} x_{n_i} x_{p_i}.$$

Experimental errors are computed as the difference between the experimental and estimated values $e_{ij} = y_{ij} - \hat{y}_i$. We then compute the total variation or the total sum of squares as

$$SST = \sum_{ij} (y_{ij} - \bar{y}_{..})^2$$

where $\bar{y}_{..}$ is the average response for all replications of all combinations of factors.

The value of SST can be divided into parts as

$$SST = SS_\alpha + SS_\beta + SS_n + SS_p + \ldots + SS_{\alpha\beta np} + SSE$$

where $SSE = \sum_{ij} e_{ij}^2$ and $SS_\alpha = 2^k r q_\alpha^2$, etc.

Values for percentage variation are presented in Table 2

3.3 Confidence Intervals for Effects

The standard deviation of errors, σ_e

$$\sigma_e = \sqrt{\frac{SSE}{2^k(r-1)}}.$$

Then, standard deviation of terms is

$$\sigma_{terms} = \frac{\sigma_e}{\sqrt{2^k r}}.$$

Confidence intervals, computed as $q_i \mp t \cdot \sigma_{terms}$ for 95% confidence level, are presented in Table 3. From the table we see with 95% confidence that all of the factors are significant, but some of the interactions are insignificant.

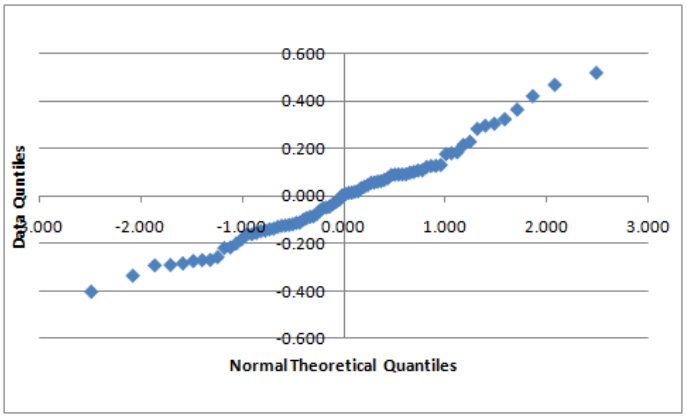

Figure 2. Normal quantile-quantile plot for errors

Confidence Intervals	lower	upper	Significance
α	95.30	95.40	Significant
β	-0.13	-0.03	Significant
n	-0.77	-0.68	Significant
p	0.11	0.20	Significant
α and β	-0.13	-0.03	Significant
α and n	-0.08	0.02	Insignificant
β and n	0.00	0.09	Insignificant
α and p	0.03	0.12	Significant
n and p	-0.06	0.03	Insignificant
β and p	-0.01	0.09	Insignificant
α, β and n	-0.08	0.01	Insignificant
α, β and p	-0.03	0.06	Insignificant
α, n and p	-0.06	0.03	Insignificant
β, n and p	-0.04	0.06	Insignificant
α, β, n and p	-0.03	0.06	Insignificant

Table 3. Confidence Intervals for the factors and interactions

Exper.	$\hat{y}$	Low(1)	High (1)	Low (5)	High (5)	Low (∞)	High (∞)
1	94.73	94.26	95.19	94.46	95.00	94.53	94.92
2	94.88	94.42	95.34	94.61	95.15	94.69	95.08
3	94.01	93.55	94.48	93.74	94.28	93.82	94.21
4	94.43	93.96	94.89	94.16	94.70	94.23	94.62
5	96.14	95.68	96.60	95.87	96.41	95.95	96.33
6	96.14	95.68	96.61	95.87	96.41	95.95	96.34
7	95.84	95.37	96.30	95.57	96.11	95.64	96.03
8	95.99	95.53	96.46	95.72	96.26	95.80	96.19
9	94.87	94.41	95.33	94.60	95.14	94.68	95.06
10	94.95	94.49	95.41	94.68	95.22	94.76	95.14
11	94.42	93.96	94.89	94.15	94.69	94.23	94.62
12	94.70	94.24	95.17	94.43	94.97	94.51	94.90
13	96.14	95.68	96.61	95.87	96.41	95.95	96.34
14	96.20	95.74	96.67	95.93	96.47	96.01	96.40
15	96.03	95.57	96.49	95.76	96.30	95.83	96.22
16	96.13	95.67	96.59	95.86	96.40	95.94	96.33

Table 4. Confidence intervals for $m = 1, m = 5$ and $m = \infty$

3.4 Confidence Intervals for Predicted Responses

We compute confidence intervals for responses for combinations of factors. The standard deviation of the mean response, depending on the number of replications, m is

$$s_{\hat{y}_m} = s_e\left(\frac{1}{n_{eff}} + \frac{1}{m}\right)^{1/2},$$

where n_{eff} is the effective number of degrees of freedom (DFs) computed as

$$n_{eff} = \frac{\text{total number of runs}}{1 + \text{sum of DFs of parameters used in } \hat{y}}.$$

We find the 95% confidence intervals for the mean response as $\hat{y} \mp t_{[1-\alpha/2;2^k(r-1)]} s_{\hat{y}_m} = \hat{y} \mp t_{[0.975;64]} s_{\hat{y}_m} \approx \hat{y} \mp 2 s_{\hat{y}_m}$ and present confidence intervals for $m = 1, m = 5$ and $m = \infty$ in Table 4.

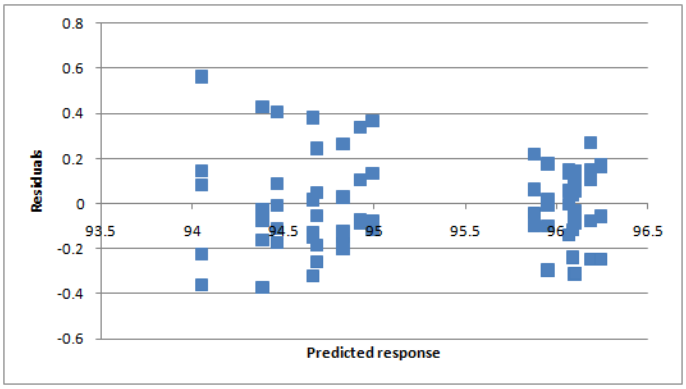

Figure 3. Scatter plot of errors versus the predicted response

3.5 Verification of Assumptions

The expressions for effects shown above are based on the following assumptions: errors are statistically independent, normally distributed and have constant standard deviation. We use visual tests to verify these assumptions.

Independent Errors To show that errors are independently and identically distributed in the model, we plot errors versus the predicted response. The scaterlpot is given in Figure 3. The scaterplot shows that the assumption is correct, since there is no visible trend in the points.

Normally distributed errors To validate this assumption, we build normal quantile-quantile plot of errors, shown in Figure 2 to verify that the graph is approximately linear.

Constant standard deviation of errors (homoscedasticity) We study the spread of points in Figure 3. The points have similar distribution and, therefore, errors have constant standard deviation.

4 CONCLUSION

We presented an algorithm for classification of rotated sequences of characters that can serve as a basis for recognition of handwritten medical text. The algorithm depends on several parameters and is based on computation of error likelihood of samples in the sequence. The method finds the rotation angle that yeilds the least error likelihood and significantly simplifies further recognition. To evaluate effect of the selected parameters, we implement $2^4 5$ factorial design. We found that the noise angle (β) has the most effect on recognition rate. All of the factors in the model are significant with 95% confidence, but some of the interactions are insignificant. We also presented confidence intervals for effects and graphically validated assumptions that form in the kernel of factorial design.

References

[1] B. B.P., "Pen-based computing: applications in clinical medicine," *The Journal of Medical Practice Management* **16(3)**, pp. 148–150, Nov-Dec 2000.

[2] A. R.D., F. L.M., R. T.C., C. R.W., T. M.S., and S. D.D., "Integration of pen-based computer technology in clinical settings," in *Proc. of the Annual Symposium on Computer Application in Medical Care*, 1994.

[3] C. K. and B. H.B., "Applicability of handheld computers in clinical information systems: comparison of evaluation methods," in *Proc. Mobile Computing in Medicine, Second Conference on Mobile Computing in Medicine, Workshop of the Project Group MoCoMed*, 2002.

[4] E. M.H., "Test advisor: a pen-based computer program for bayesian decision-making in the clinical setting," in *Proc. of the Annual Symposium on Computer Application in Medical Care*, 1994.

[5] S. A., G. R., and F. M., "A modelled classification approach to the automatic analysis of hand drawn neuropsychological tests,"

[6] V. Mazalov, "Geometric techniques for digital ink," Master's thesis, Department of Computer Science, University of Western Ontario, London, Canada, August 2010.

[7] O. Golubitsky and S. M. Watt, "Distance-based classification of handwritten symbols," *International J. Document Analysis and Recognition* **13**(2), pp. 113–146, 2010.

[8] M.-K. Hu, "Visual pattern recognition by moment invariants," *Information Theory, IRE Transactions on* **8**, pp. 179–187, February 1962.

[9] S. Feng, I. Kogan, and H. Krim, "Classification of curves in 2d and 3d via affine integral signatures," *Acta Applicandae Mathematicae* **109**, pp. 903–937, 2010. 10.1007/s10440-008-9353-9.

[10] R. Jain, *The art of computer systems performance analysis*, John Wiley and Sons, Inc, 1991.