Recognition of Manuscript Tables in Computer Processing of Technical Transport Documentation

Elena Y. Bursian
Emperor Alexander I
St. Petersburg State Transport University,
St. Petersburg, Russia
bursianeu@mail.ru

Anton M. Demin
Emperor Alexander I
St. Petersburg State Transport University,
St. Petersburg, Russia
ad2271@ya.ru

Alexander P. Glukhov
Joint Stock Company Railway Research Institute (JSC "VNIIZhT"),
Moscow, Russia

Abstract

The article discusses the process of recognizing handwritten characters presented in tables of technical railway documentation and test works of students of PGUPS. In the model under study, a skeleton graph is constructed for each recognizable region and the procedure for statistical processing of the characteristics of the branches of skeletal graphs is analyzed.Skeletal graphs are constructed for reference symbols and are dynamically replaced during recognition by skeletal graphs of recognized areas. The nature of dependencies between the components of skeletal graphs of reference symbols and dynamically added objects is investigated. In the process of automatic recognition of handwritten characters, the obtained statistical results are applied.

Introduction

The task of automatic processing and optical recognition of a pre-scanned or captured image is relevant in many fields of activity. Verification of scanned completed tables and forms, tests, questionnaires, tests during distance learning of PSU students, tables of technical railway documentation requires handwriting recognition.

Universal electronic document management systems and individual specialized programs are based on the principles of the general theory of recognition. Monographs by V. N. Vapnik [Vap74], A.Ya.Chervonenkis, Yu.I.Zhuravlyov[Zhu05], A.B. Merkov[Mer14], V.V. Ryazanov, O.V. Senko.

Researches of L.M. Mestetsky[Mes09], D. A Gavrilov[Gav19],N. A. Lomov[Lom16],Yu.V.Vizilter[Vis12], Ya.A. Furman relate to practical application and have recognized applied value. The main foreign works are presented by R. C. Gonzalez,L.Lam [Lam95], C.Suen, T. Y. Zhang [Zha84], C. Y. Suen, D. T.Lee [Lee82],H. Blum [Blu67], R. O. Duda [Dud00], P. E. Hart, R. Shapiro, L. Shapiro, G. Stockman, P. Viola and M. Jones [Vio04],S. Rosset[Ros04], L.C.Molina [Mol02].

Currently, there are computer systems for recognizing handwritten characters in almost noisy images: ABBYY FormReader, OmnPage, CuneiForm, ReadirisPro. Moreover, the use of these systems in the recognition of specialized tables is not always effective, since the recognized information in many cases has a predetermined structure. Recognition of handwritten characters in technical tables on low-quality images is an urgent scientific and technical task.

Handwritten tables and test documents usually assume that there are different density distributions for specific characters, but they are unknown. During computer processing, the characteristics of the sample can be calculated and unknown.

1 Statement of the problem

To build an automatic recognition system for individual handwritten characters, a set of basic recognizers is used with not always a high probability of object recognition. Weak recognizers are grouped, and a committee of classifiers is built using the AdaBoost algorithm.

When automatically checking test works, it is permissible to assume that the classifier divides the space of vectors of informative attributes X into two sets X1 and X2, since the test usually assumes a single answer. You can also automatically build a committee of classifiers separately for each character.

To build a committee of classifiers, you must first construct many different basic recognizers. The construction of basic recognizers by estimating the parameters of multidimensional distribution densities of the vector characteristics of handwritten characters is an urgent task.

In: A. Khomonenko, B. Sokolov, K. Ivanova (eds.): Selected Papers of the Models and Methods of Information Systems Research Workshop, St. Petersburg, Russia, 4-5 Dec. 2019, published at http://ceur-ws.org

2 Construction of basic recognizers and final classifier

After scanning the document, the image is reduced to a two-gradation view. Figure 1 shows a scanned image of a table of railway documentation.

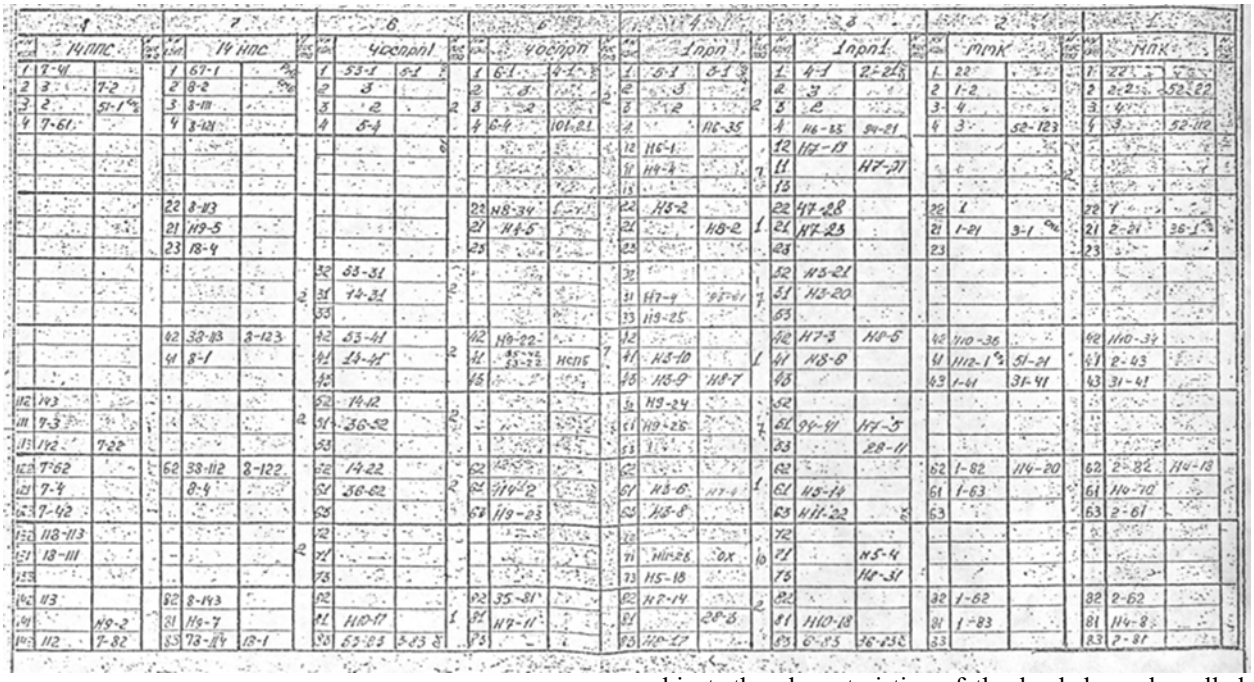

Figure 1: Railway Documentation Table

It should be noted that a modern approach to maintaining railway documentation requires the mandatory introduction of electronic document management [*].

For each recognized area, a skeletal description is constructed. The calculation of the characteristics of the skeletal representation of the region is based on the following definition.

The point P belongs to the skeletal representation of the domain D if and only if the following statement holds:

$$B_r(P) \subset D \;\&\; \nexists B_{r1}(P_1) \subset D : B_r(P) \subset B_{r1}(P_1) \& B_r(P) \neq B_{r1}(P_1),$$

where $B_r(P)$ is a circle centered at point P and radius r [Dud00].

The set of points P belonging to the skeleton representation is also called the skeleton of the region D. We can assume that the skeleton of the region is the set of centers of maximal circles lying in the region (Figure 2).

Based on the region's skeleton, for each recognized object, the characteristics of the loaded graph, called the region's skeleton graph, are calculated. The skeletal graph of a recognized object can be represented as follows (Figure 2, Figure 3).

Figure 2: Skeletal representation of the area

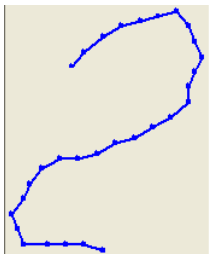

Figure 3: Skeletal graph of the area

For each branch of the skeletal graph, a vector of informative characteristics is constructed, made up of the slope coefficients of the edges of the skeleton graph or directly the slope angles of the edges of the skeleton graph. In the case when the symbols are written in one hand, between the values of the slope angles of the edges of the skeletal graph taken at equal intervals, there is a statistical dependence.

TopNumber	TopCoordinates	Neighbours
0	(384, 7)	1 4
1	(381, 8)	0 2
2	(378, 9)	1 3
3	(375, 10)	2 5
4	(386, 10)	0 6
5	(372, 12)	3 7
6	(387, 13)	4 8
7	(369, 15)	5 9
8	(388, 16)	6 10
9	(367, 18)	7
10	(387, 19)	8 11
11	(386, 22)	10 12
12	(386, 25)	11 13
13	(383, 28)	12 14
14	(380, 30)	13 15
15	(377, 32)	14 16
16	(374, 33)	15 17
17	(371, 35)	16 19
18	(365, 36)	19 20
19	(368, 36)	17 18
20	(362, 38)	18 21
21	(360, 41)	20 22
22	(359, 44)	21 23
23	(357, 47)	22 24
24	(358, 50)	23 25
25	(359, 53)	24 26
26	(363, 53)	25 27
27	(366, 53)	26 28
28	(369, 53)	27 29
29	(372, 54)	28

Figure 4: The coordinates of the vertices of the skeletal graph of the region

The regression of the slope angles of the skeletal graph of a symbol from the values of the corresponding angles taken at previous intervals can be calculated using multivariate regression analysis methods, creating a system for determining unknown regression coefficients.

$$\begin{cases} \varphi_0^{(1)} = a_0^{(1)} + a_1\varphi_1^{(1)} \dots a_k\varphi_k^{(1)} + \varepsilon^{(1)} \\ \varphi_0^{(2)} = a_0^{(2)} + a_1\varphi_1^{(2)} \dots a_k\varphi_k^{(2)} + \varepsilon^{(2)} \\ \dots \quad \dots \quad \dots \\ \varphi_0^{(n)} = a_0^{(n)} + a_1\varphi_1^{(n)} \dots a_k\varphi_k^{(n)} + \varepsilon^{(n)} \end{cases}$$

We assume that $\boldsymbol{\varphi}^T = (\varphi_1, \dots \varphi_k)$is the vector of regression factors, in the training set for the symbol with number i the regression factors take the values: $\varphi_1^{(i)}, \dots \varphi_k^{(i)}$and correspond to the angles of inclination of the edges of the skeletal graph of the symbol,$\varphi_0^{(i)}$is the response value. $\boldsymbol{a}^T = (a_0, a_1, \dots a_k)$ – estimated regression parameters,$\boldsymbol{\varepsilon}^T = (\varepsilon^{(1)}, \dots \varepsilon^{(n)})$error vector, The number of unknown regression parameters does not exceed the number characters in the training set k<n.

Considering that the angle of inclination of the edges of the skeletal graph of a symbol taken at a certain interval is a random variable, we can write its conditional expectation in the following form.

$$\boldsymbol{M}[\varphi|\boldsymbol{\varphi}] = a_0 + a_1\varphi_1 \dots a_k\varphi_k$$

An estimate of the regression parameters a can be obtained by calculating the pseudoinverse matrix.

$$\overline{\boldsymbol{a}} = (\Phi\Phi^T)^{-1}\Phi\boldsymbol{\varphi_0}$$

where

$$\Phi = \begin{pmatrix} 1 & \varphi_1^{(1)} & \dots & \varphi_k^{(1)} \\ 1 & \varphi_1^{(2)} & \dots & \varphi_k^{(2)} \\ \dots & \dots & \dots & \dots \\ 1 & \varphi_1^{(n)} & \dots & \varphi_k^{(n)} \end{pmatrix}^T,$$

$$\boldsymbol{\varphi_0}^T = (\varphi_0^{(1)}, \dots \varphi_0^{(n)}).$$

When a symbol is recognized, the vectors of the informative characteristics of the symbol are compared, in this case, the angles of inclination of the edges of the skeletal graph, with the set of angles of inclination

obtained by regression of the angles of inclination taken at previous levels.

To classify the classifier, basic recognizers are also used based on algorithms for calculating the Hausdorff distances between sets of critical vertices of skeletal graphs, calculating correlation functions for the angles of inclination of edges of skeletal graphs, and comparing them with threshold values of correlation functions of coordinates of critical vertices of skeletal graphs, where the vertices of skeletal graphs are considered critical non-hanging peaks with degrees other than two[Ros04], [Mol02].

The final classifier is constructed using the AdaBoost algorithm [Vio04]. The training sample is the set of ordered pairs $A = \{(x_1, y_1), \ldots (x_i, y_i), \ldots (x_n, y_n)\}$, where x_i is the skeleton graph of the symbol, $y_i \in \{1, -1\}$ if $y_i = -1$, the skeleton graph corresponds to the symbol, $y_i = 1$ otherwise, h_t are the basic classifiers, $t \in \{1,2 \ldots T\}$.

The initial distribution of elements of the set A is initialized by the uniform distribution $P_1 = 1 / n$. In the next steps, for $t \in \{1,2 \ldots T\}$, the coefficients αt and the new distribution $P_{t+1}(i)$ are calculated using the distribution $P_t(i)$ calculated in the previous step.

$$\varepsilon_t = \min_{t \in \{1,2,\ldots T\}} (P\{h_t(x) \neq y_i\})$$

That is, ε_t is the probability of error of the classifier h_t, provided that the classifier h_t is less error than other classifiers on the distribution P_t.

$$\alpha_t = \frac{1}{2} \ln \left(\frac{1 - \varepsilon_t}{\varepsilon_t}\right)$$

The probability distribution for the next step is updated by the formula:

$$P_{t+1}(i) = \frac{P_t(i) e^{-\alpha_t h_t(x_i)}}{z_t},$$

where z_t is the normalizing factor.

The final classifier is built according to the formula:

$$H(x_i) = sign\left(\sum_{t=1}^{T} \alpha_t \, h_t(x_i)\right).$$

Conclusion

When applying the AdaBoost algorithm for recognizing individual handwritten characters in tables of railway documentation or in test works of students of PSUPS, a training set of skeletal graphs and a set of basic classifiers with a volume of at least two dozen elements are required.

It is possible to develop basic classifiers and build the final classifier, both for an arbitrary character, and for individual characters with the division into two classes: belonging to the class of the given character and relation to other sets. The development of basic classifiers based on various algorithms and heuristic procedures is a laborious task.

The use of multidimensional linear and polynomial regression methods for the angles of inclination of the edges of the skeleton graph of the symbol makes it possible to construct a set of basic recognizers, since when writing a symbol in case of deviation or curvature of the sign, a regression trend is confirmed, confirmed by experimental data.

To obtain experimental data, an experimental set of programs based on the Visual C ++ platform was developed. The construction of the skeletal representation of the symbol, the skeletal graph of the symbol, and the angles of inclination of the edges of the skeletal graph of the symbol was carried out.

The considered approach allows us to solve more complex problems of text recognition. In particular, recognition of handwritten railway documentation or recognition of handwritten text allows for a large number of mathematical symbols and its presentation in the generally accepted TEX format.

References

[Vap74] Theory of pattern recognition (statistical problemsof learning), V. N. Vapnik, A. Ya. Chervonenkis. Publishing House "Science", Main Edition of the Physics and Mathematics Literature, M., 416 pp., 1974.

[Zhu05] Yu.I. Zhuravlev, V.V. Ryazanov, O.V. Senko. RECOGNITION Mathematical methods.Software system.Practical applications. - M .: FIZMATLIT -- 159 p., 2005 .

[Zhu78] Yu. I. Zhuravlev.On the algebraic approach to recognition and classification problems//

Problems in Cybernetics, Nauka, Moscow, Vol. 33, pp. 5–68,1978.

[Zhu78] Yu.I. Zhuravlev.Correct algebras over sets of incorrect (heuristic) algorithms, Parts I-III," Kibernetika, No. 2,pp. 35−43,1978.

[Mer11] A.B. Merkov. Pattern Recognition: An Introduction to Statistical Learning Methods. URSS, Moscow, 2011.

[Mer14] A.B. Merkov. Pattern Recognition: Building and training probabilistic models. Lenand, Moscow, 2014.

[Mes09] L.M.Mestetsky Continuous morphology of binary images: figures, skeletons, circulars. - M .: FIZMATLIT .-- 288 p. - ISBN 978-5-922-1050-1, 2009.

[Gav19] D. A.Gavrilov ,L. M.Mestetskiy, Semenov A. *B.* A method for aircraft labeling in aerial and satellite images based on continuous morphological models /Programming and Computer Software. —. — Vol. 45, no. 6. — P. 303–310, 2019.

[Lom16] N. A. Lomov, L. M. Mestetskiy. Area of the disk cover as an image shape descriptor. Computer Optics, 40(4):516–525, 2016.

[Vis12]VisilterYu.V., Sidyakin S.V. Construction of morphological spectrahalftone image // Bulletin of computer and informationTechnologies, N4, pp. 8-17, 2012.

[Vis08] Yu. V. Vizilter, S. Yu. Zheltov, V. A. Prince, A. N. Khodarev, A. V. Morzhin. Processing andanalysis of digital images with examples on LabVIEW and IMAQ Vision.M .: DMK Press. 464 p, 2008.

[Lam95] L.Lam,C.Suen. An Evaluation of Parallel Thinning Algorithms for Character Recognition // IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 17, № 9 p. 914-919, 1995.

[Zha84] T. Y.Zhang, C. Y.Suen. A fast parallel algorithm for thinning digital patterns. // IEEE, Communications of the ACM. – Vol. 27. – № 3 – 236 p. 1984.

[Lee82] D. T.Lee Medial axes transform of planar shape // IEEE Transactions on Pattern Analysis and Machine Intelligence. PAMI-4, 363-369p., 1982.

[Blu67] H. Blum. A transformation for extracting new descriptors of shape // Models for the Perception of Speech and Visual Form, MIT Press, 362-380p, 1967.

[Bul12] P.E. Bulavsky, M.N. Vasilenko, A.A. Kornienko, A.D. Khomonenko. Automation of information support for railway managers on the basis of the introduction of electronic document management.News of St. Petersburg University of Railways Messages, no. 2 (31),p. 116-118, 2012.

[Dud00] R.O. Duda, P.E. Hart. D G. Stork.Pattern Classification. -Wiley-Interscience New York, NY, USA, ISBN:047105669 688p. 2000 .

[Dud76] R.O. Duda, P.E. Hart. Pattern recognition and scene analysis. - M .: World. -- 507 p. 1976 .

[Vio04] P. Viola, M. Jones, Robust real-time face detection, International Journal of Computer Vision, vol. 57, no. 2,137–154 p. , 2004

[Ros04] S. Rosset,J. Zhu, T. Hastie. Boosting as a Regularized Path to a Maximum Margin Classifier.Journal of Machine Learning Research no.5. 941-973 p. , 2004.

[Mol02] L.C.Molina, L.Belanche, A.Nebot. Feature Selection Algorithms: A Survey And Experimental Evaluation. Proceedings of the 2002 IEEE International Conference on Data Mining, IEEE Computer Society, 306–313 p. , 2002.