intelligent character recognition, pattern recognition,
hospital information systems

Jerzy SAS*

HANDWRITTEN LABORATORY TEST ORDER FORM RECOGNITION MODULE FOR DISTRIBUTED CLINIC

The work describes methods used in a laboratory order form recognition module of a hospital information system. Three-level form analysis architecture is proposed. The lower alphabetical level is responsible for separate character recognition. On the intermediate level, recognised strings are verified against the lexicons of items specific for a particular form field. Probabilistic model is used to select the set of most probable items. On the upper level, the dependencies between the form data items are taken into account to further improve the recognition performance. The presented approach was implemented in the medical information system supporting clinic laboratory operation. The laboratory test orders prepared manually by the physician in the paper form, in the net of distributed outpatient clinics are processed in the central hospital laboratory. In the central laboratory the paper forms are scanned, recognised and entered into the information system. The performance tests results are discussed and some further improvements of the applied recognition method are also suggested in the paper.

1. INTRODUCTION

Automatic analysis of hand-written forms is useful in such applications where direct information insertion into the computer system is not possible or inconvenient. Such situation appears frequently in hospital medical information systems, where physicians or medical staff not always can enter the information directly at the system terminal. Form scanning is considered to be especially useful in laboratory support software, where paper forms are still frequently used as a medium for laboratory test orders representation. Hence, in many commercially available medical laboratory systems a scanning and recognition module is available.

In this paper, the problem of hand-written laboratory orders recognition is described. The concept described here has been originally elaborated for the hospital information system designed for operation in the distributed outpatient clinics net. The biological samples for laboratory tests (blood, urine, spit etc.) are taken in places, which have no connection with the central laboratory. The information flow is based on hand-written forms containing the patient data (name, date of birth, social security identifier), identifier of the institution reimbursing for the test, sample container identifier and the list of laboratory examinations to be performed. The form is partly filled by the physician, which orders the

* Computer Science Department, Wroclaw University of Technology, 50-370 Wrocław, Wyb. Wyspiańskiego 27

test, then completed by the medical staff, which take the sample of the biological material, then transferred to the central laboratory where it is scanned, and finally recognised, verified and entered into the medical system database.

The key feature determining the practical acceptance of such kind of software is the high accuracy of the form recognition. It appears that no single hand-written classification algorithm can assure sufficiently low error rate. The most accurate published methods reach the accuracy of the order of 98-99% per single numeric character and 90-95% per alphabetic character ([3], [4], [5]). If we take into account that the typical form contains on average about 50 characters then the expected number of errors per single form is about 1-2. It means that practically each form contains an error and needs manual correction. From our experiences it follows that the solution can be practically acceptable, if not more that 10-20% of forms need to be corrected manually. To achieve such level of accuracy the mixture of character recognition techniques must be applied simultaneously, both on the single character recognition level and on the complete information unit (names, dates, symbols, identifiers, whole documents) level.

2. PROBLEM STATEMENT

Let us consider the typical form recognition problem. The form contains a set of separated data item fields, where each of them is assigned a type. The following field types can be used:

- alphabetic field – containing only letters,
- numeric field – containing digits and optionally sign character and decimal separator,
- date field – containing dates in fixed format,
- check boxes,
- barcodes.

In case of forms applied to medical orders recording, the alphabetical fields contain patient name and surname, ordering physician's name and a sample container symbol. Numeric fields can be used for patient's social security identifiers, payer symbol etc. Date fields are used for birth date, date of order issue, test deadline. Checkboxes are typically used for selecting the ordered tests or marking the patient's sex. The data derived from the form – after manual verification – are entered into the medical system database.

While there is practically no problem with automatic near-perfect recognition of scanned barcodes and check boxes, automatic recognition of hand-written strings or numbers is still a difficult task. In the automatic hospital information system, where the daily throughput is in order of several hundreds forms, the very low recognition error rate is expected to minimise the risk of mistakes, even on the assumption that each automatically scanned form is manually verified by the medical staff. The method that supports automatic form processing and recognition must therefore assure high recognition accuracy to be accepted by the medical community.

3. THREE-LEVEL FORM RECOGNITION METHOD

In the approach presented here we applied a combination of a few methods in order to improve the recognition reliability. All data in the form are classified into one of data types listed in the previous section. In this way we can independently recognise digits and letter characters, thus reducing the count of classes for the recognition problem. The form analysis is performed on the three levels:

- alphabetical level – where separate characters are being recognised,
- lexical level - where each character sequence consisting of the characters recognised on the alphabetic level is compared with the contents of the data items lexicon defined for the field,
- pragmatic level – where relations between form fields are being considered.

3.1. CHARACTER RECOGNITION ON THE ALPHABETIC LEVEL

The alphabetic level is fetched with the separated and normalised character field images. The classification problem on this level consists in assigning the object being recognised (normalised character image) to one of *n* classes. The count of classes is 28 for alphabetic fields and 10 to 13 for numeric fields (depending on the applied number format).

We apply the method, which is a simplification of unconstrained flexible matching concept ([7]). Let us consider two black-and-white character field images A and B, both of the equal resolution $x_{res} \times y_{res}$. A pixel in the image is active if it is covered by the strokes constituting the character. The background pixels are inactive. Let $\hat{X}$ denote the set of row and column index pairs of active pixels in the image *X*.

$$\hat{X} = \{ <i, j> : 1 \le i \le y_{res} \wedge 1 \le j \le x_{res} \wedge X[i,j] \text{ is active } \} \qquad (1)$$

Hence we have the sets of active pixels $\hat{A}$ and $\hat{B}$ for images *A* and *B* correspondingly. We will define the dissimilarity measure between two images A and B as:

$$R(A,B) = r(A,B) + r(B,A), \qquad (2)$$

where *r(A,B)* for two images A and B is defined as:

$$r(A,B) = \sum_{<i,j>\in\hat{A}} m(B,i,j) \qquad (3)$$

Here *m(B,i,j)* is a distance from the pixel <*i.j*> in the image A to the nearest active pixel in the image B:

$$m(B,i,j) = \min_{<i',j'>\in B} d(<i,j>,<i',j'>), \qquad (4)$$

where *d(<i,j>, <i',j'>)* is the Euclidean distance in the 2D pixel coordinates space.

The image dissimilarity measure defined in this way corresponds usually to our intuition: two characters seem to be similar if we can find corresponding strokes in both images and if these strokes are close each to other. The Fig.1 explains the idea behind the proposed dissimilarity measure. Black shape (A) represents the active pixels set of the reference sample. Grey shape (B) is the character being compared to the reference sample. The arrows on the left figure represent distances *d* from active pixels of reference sample to the active pixels of the character being recognised. The arrows on the right figure represent distances between corresponding pixels in A and B. The measure described above is sensitive to character image translation. This shortcoming can however be reduced by image resizing and clipping performed in the pre-processing phase.

The classic nearest neighbour algorithm is applied to recognise the character represented by character field image. For each type of field (alphabetical and numeric) there is a learning set consisting of letter or digit images. The character field image taken from the form is scaled to the standard resolution. Then the dissimilarity measures calculated according to (2) to all the elements of the learning set are calculated. This character is finally recognised, which is assigned to the most similar learning set element.

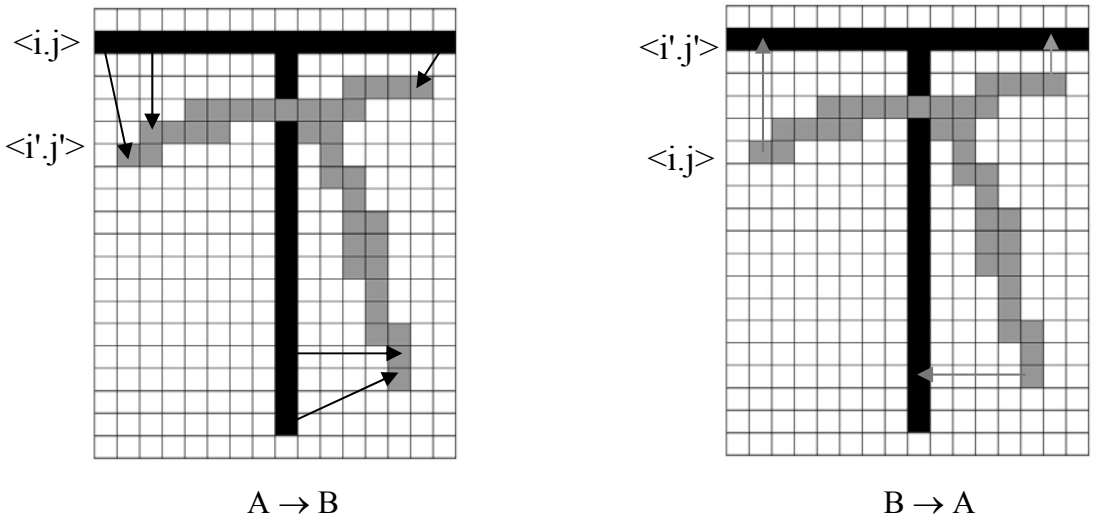

Fig. 1. Evaluation of dissimilarity measures between shapes A and B

3.2. DATA ITEMS RECOGNITION ON THE LEXICAL LEVEL

On the lexical level we use the results of classification done on the alphabetical level. Here the character sequences recognised in the consecutive character fields of the data field are used to select the data item from the lexicon defined for this field. In case of the laboratory order form recognition system being described here, two types of lexicons can be distinguished:

- general purpose lexicons defined for names and surnames,
- specialised lexicons containing data strongly related to the contents of medical information system database: identifiers of order form issuing institutions, symbols of the form variants, registered patient names etc.

The Bayes approach is applied, which minimises the risk of erroneous recognition. Let $p_{ik} = P(a_i \mid \Phi(A) = a_k)$ denote the probability of character a_i appearance, provided that the classifier Φ on the alphabetical level recognised the character a_k on the image A of the character field. The probability p_{ik} can be estimated by using verified contents of forms recognised by the system and comparing it with of automatic character classification results. Assuming that the events consisting in recognising characters on various positions of the data field are independent, we can evaluate the *aposteriori* conditional probability of each lexicon item $\alpha=a_{i1}\ a_{i2}\ ...\ a_{il}$ appearance, provided that the alphabetical classifier Φ recognised the sequence $a_{k1},a_{k2},...,a_{kl}$

$$P((a_{i1},a_{i2},...,a_{il}) \mid (a_{k1},a_{k2},...,a_{kl})) = \prod_{m=1}^{l} P(a_{im} \mid \Phi(A) = a_{km}) \qquad (5)$$

According to Bayes approach this lexicon item is finally recognised, for which the probability calculated using formula (5) is maximal. However, in the three level approaches the final recognition is deferred to the third level and the role of the lexical level consists in evaluating *aposteriori* probabilities of lexicon items. Let $p_{d,\alpha}$ denote the calculated probability of the lexicon item α for the form data field d. Having calculated the probabilities $p_{d,\alpha}$, the set D_d of the most probable lexicon elements for each field d is selected and passed to the pragmatic level. The cardinality of D_d is determined in such way, that the sum of probabilities of all selected items exceeds assumed threshold. Finally, the probabilities of each D_d elements are normalised to sum to 1.

3.3. FORM OBJECTS RECOGNITION ON THE PRAGMATIC LEVEL

Final laboratory test order form contents recognition is performed on the pragmatic level. We assume that the form fields can be grouped into clusters, which describe the compound objects. The objects that appeared on forms in the past are recorded in the medical information system database. In most cases the object described by the form is already recorded in the database, so the database can be used as a kind of lexicon for the whole form – similarly as the lexicons for data fields were used on the lexical level. The aim is to identify the object described by the fields' cluster and to find it in the database, or to decide that the form describes the new object, which is not represented in the database yet. In the second case the recognition algorithm is expected to fetch the field contents.

In case of laboratory test order form, the three form fields describe objects: a patient subject to test, a reimbursing institution and a form type. We do not consider here the list of the ordered tests, because they are identified by easy recognisable check boxes and sufficiently reliable recognition of such form element is not a problem. The most crucial and difficult element of the order form recognition is identification of the patient. The patient is defined on the form by five data fields: name, surname, sex, date of birth and social security identifier.

Let us consider a general case of object type described by the cluster consisting of N form fields $<d_1, d_2, ..., d_N>$. As a set of candidate objects to be recognised we consider the

Cartesian product C of the sets D_{di} $i = 1, 2, ..., N$ fetched by the lexical level recogniser. For each candidate $c \in C$ the "rating" $v(c)$ is calculated which is the product of the probabilities $p_{d,\alpha}$ evaluated on the lexical level for each candidate element. The set C is then ordered by rating values and compared with the database contents. If there are candidates with rating above the given rejection threshold T represented in the database then the represented candidate with the highest rating is selected. Otherwise, it is assumed that the form describes a new patient and in that case the candidate with the highest rating among all elements of C is selected.

The rejection threshold T depends on the database contents. At the early stages of the medical information system usage, when it contains only small amount of data, it cannot be used as the reliable lexicon of patients – hence the threshold T should be high. As the database grows, it becomes more and more probable that the form being recognised concerns the patient who is already registered in the database – therefore the threshold T should be gradually reduced. We applied the solution consisting in taking as T the estimated probability of the event, that the form being processed concerns a new patient, not yet registered in the database. This probability can be easily estimated using the contents of the recognised forms database.

The experiment has been performed to test how the value of T depends on the number of processed forms in the environment of actual big hospital laboratory. In the experiment the medical information system database resulting from over three years of the system operation was used. The database contained 42525 laboratory order form records and 8290 patient records. The estimates of T for various moments in the past could be calculated due to patient record and form record insertion dates stored in the database. The dependency of the threshold T on the length of the system operation period and on the stored forms count is presented in Tab.1.

Table 1. Dependence of rejection threshold and entered forms count on the system operation period.

System operation period [months]	6	12	18	24	30	36
Entered forms count	5230	12873	23455	29312	36129	42525
T estimated after operation period	0.72	0.46	0.38	0.22	0.14	0.12

4. IMPLEMENTATION

The concept presented in the previous chapters was utilised and implemented in the hospital information system, which support wide range of hospital activities including biochemical laboratory operation. One of methods of entering laboratory test orders into the system database is scanning of a paper test order forms. The method is especially useful for registering the orders coming from outer units in the distributed outpatient clinic. The form used in the system is shown on Fig.2. Laboratory tests are selected by filling the appropriate checkboxes. Forms can appear in many variants. A variant determines mapping between the

checkbox positions on the form and the associated tests. The form variant is identified by the contents of variant field. The form contains positioning markers in the form corners. The markers are used to transform the scanned form image into the standard position, in which data fields can be precisely localised.

Before the form is recognised it goes through the series of image processing operations. First, it is scanned and stored in an ordinary greyscale image file. The image is then converted to black and white binary form. Next, the positioning markers are localised and the image is transformed into the standard position. At the next stage the character fields are extracted and median filtering is applied in order to remove the printed frame lines enclosing character fields. Then the characters are thinned and slant correction is performed. Finally, field images are clipped to the minimal rectangular region containing the character contour and the clipped rectangles are scaled to standard resolution 60x70. The character images are packed into the data structures grouping them into data fields and the structures are passed to the three-level form recogniser.

Character recognition on the alphabetical level is based on the learning set. Because each of the form fields is purely alphabetical or numeric, then the learning set is divided into two subsets containing letters and digits. For a particular field of known type the appropriate set is used. The initial contents of the learning set was gathered by imitating typical character writing styles observed on the sample set of test sheets filled by over 300 students. Initial learning sets consist of 279 letter samples and 142 digit samples. The learning set adapts to changing writing styles appearing most frequently on the processed sheets. The adaptation algorithm gradually replaces samples in the learning set by the new ones coming from verified forms. The algorithm prevents uncontrolled growth of the learning sets by limiting its size to predefined bounds.

General-purpose lexicons for Polish names and surnames used in the system contain about 1400 and 35000 items correspondingly.

The forms are scanned by the mid-range scanner equipped with the form feeder. The images are scanned in 300 dpi resolution and stored by the scanner software in grey-scale image files. The forms are scanned, processed and recognised automatically as soon as they are put into the feeder.

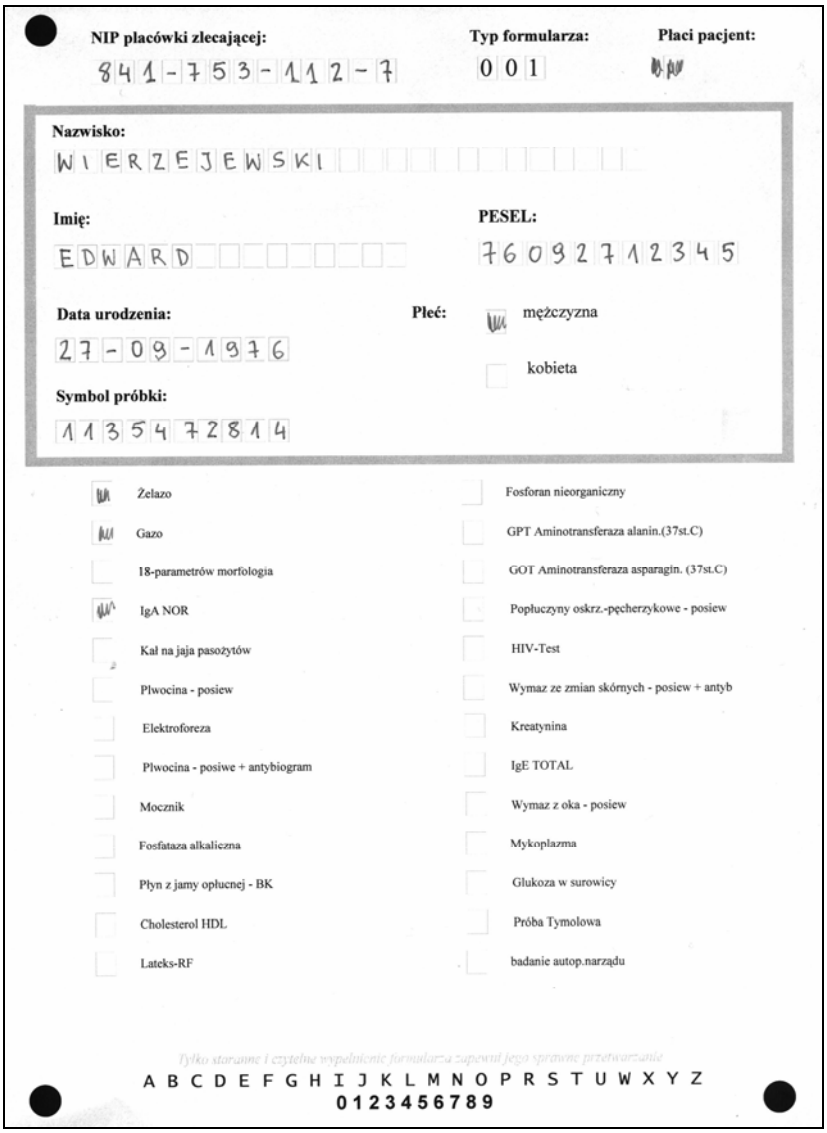

Fig. 2. Laboratory test order form layout (field labels in Polish)

5. EXPERIMENTS

Series of experiments have been performed to evaluate the recognition quality of the whole three-level recognition method as well as the performance of its components. The experiments were performed using the initial learning set described in the previous chapter in the environment of real medical information system database as described in the section 2.3.

In the first experiment the character level recogniser accuracy was assessed. During the experiment 86 filled forms were processed and recognised. They contained total 2465 digits in numeric fields and 1086 letters in alphabetical fields. The forms came from

9 writers. Achieved accuracy recognition on the character level was 95,7% for digits and 85.9% for letters.

In the second experiment the gain from lexicon usage on the lexical level was assessed. The lexical recogniser was used to select single (most probable) lexicon item for name and surname field. The recognition was considered as correct on this level if all characters of the recognised string matched their counterparts in the data field. The field recognition accuracy achieved by simply concatenating characters fetched by character recogniser was compared to the accuracy achieved with lexicon support. The results are shown in the Tab.2.

Table 2. Accuracy achieved on the lexical level

	names	surnames
without lexicon	51,1%	33.7%
with lexicon	79,1%	52.3%

Finally, the quality of the whole form recognition was tested. The content of medical information system database was used to improve the quality of patients' identification. The form was considered as correctly recognised in this experiment if all fields in the patient cluster were recognised correctly. 5 of 86 forms used in the experiment described patients not registered in the database. Remaining 81 forms concerned already registered patients. 2 of 5 forms concerning not registered patients were recognised incorrectly (one error in surname, one in social security identifier). Only 8 of remaining 81 forms were not recognised correctly.

6. CONCLUSIONS, FURTHER WORKS

The three-level form recogniser described in this article, despite of its conceptual simplicity, seems to give quite promising results. The overall correctness of the complete laboratory test order form recognition reaches 88% when the recogniser makes use of the database containing records of previously recognised and verified forms. This means that on average only 1 of 10 scanned orders needs to be corrected manually. In clinical practice it significantly reduces the staff efforts necessary to manually enter the complete laboratory order data and makes it less error-prone.

There are many possible ways of the system performance improvements. The improvements of character level algorithm seem to be the most promising. In particular, applying combined character classification algorithms based on various features, as well as better adapting to writers style will be investigated and implemented in future. Also utilising the information about frequencies of various lexicon items occurrence may boost the performance of the lexical level recogniser. Similar improvement can be done on the pragmatic level.

BIBLIOGRAPHY

[1] KULIKOWSK J.L., Cybernetic Recognition Systems, PWN, Warsaw, 1972
[2] KURZYŃSKI M., Pattern Recognition. Statistical Methods. Techn. Univ. of Wroclaw Press, Wrocław, 1997
[3] LIU Z.Q., CAI J., BUSE R., Handwriting Recognition. Soft Computing and Probabilistic Approaches, Springer, Berlin, Heidelberg, New York 2003
[4] LIU C., NAKASHIMA K., SAKO H., FUJISAWA H., Handwritten Digit Recognition: benchmarking of state-of-the-art techniques, Pattern Recognition, No. 36, 2003
[5] PLAMONDON R., SRIHARI S., On-line and Off-line Handwriting Recognition: A Comprehensive Survey. IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 22, No. 1, Jan. 2000
[6] TRĄBKA W., Hospital Information Systems, University Press "Vesalius", Kraków 1999
[7] UCHIDA S., SAKOE H., Eigen-deformations for elastic matching based handwritten character recognition. Pattern Recognition No 36, 2003