handwriting recognition,
hospital information systems,
automatic document processing

Jerzy SAS[*] Jerzy PEJCZ[**]

APPLICATION OF DOCUMENT TYPE IDENTIFICATION IN MEDICAL HANDWRITTEN TEXTS RECOGNITION

In the paper a method of handwritten document type identification and its application in handwritten medical texts recognition are discussed. Document type identification procedure utilizes fixed graphical elements of the forms on which documents are written. It is assumed that the form contains pre-printed frames consisting of distinct vertical and horizontal line segments. Line segments are detected using Hough transform. Next, the document image is rotated in order to make detected lines exactly vertical or horizontal. The grid of lines is matched by translation to best fit the set of lines on the document template. Finally this document is recognized for which the matching factor is highest. Results of document type recognition are used by handwritten text recogniser, where document identification is used to select appropriate probabilistic lexicon prepared for recognized text type. The method was elaborates to be applied at the first stage of multilevel system for handwritten medical documents recognition. Results of experiments with authentic medical texts extracted from hospital information system described in the article show that significant text recognition accuracy improvement can be obtained by applying document identification results.

1. INTRODUCTION

Modern healthcare cannot be imagined without Hospital Information Systems (HIS), where complete documentation about patient treatment and medical episode accounting is gathered and made accessible. There are however huge volumes of medical documentation written by hand, either prepared in the past or created in places where terminals of hospital information system are not available. In case of complete information flow in electronic form, handwritten documents have to be finally entered into HIS. Manual retyping is costly and error prone process. For this reason natural expectation from HIS user is to have a functionality for automatic handwritten documents recognition or at least for effective computer-supported handwritten documents translation to electronic form based on handwriting recognition techniques.

Medical documentation consists of texts describing certain aspects of medical episode related to patient. It seems to be general tendency to use specific vocabulary and language phrases in each type of document. This fact can greatly simplify the problem of automatic recognition of text written on document of known type by applying text recogniser trained for the type of document currently being recognized. Unfortunately, handwritten documents are rarely intentionally marked in

[*] Institute of Applied Informatics, Wroclaw University of Technology, Wyb. Wyspianskiego 27, 50-370 Wroclaw, POLAND, e-mail: jerzy.sas@pwr.wroc.pl
[**] Department of Pediatrics, St Hedwig Hospital, 55-100 Trzebnica, POLAND

the way allowing for reliable automatic recognition of document type, unless they were prepared on forms specially designed for automatic text processing. This is typically not the case as far as old documentation archives are considered. Some features of the document layout or pre-printed elements on the printed form specific for given handwritten document type can be however used to identify document type.

The aim of work described in this article is to elaborate the simple and fast method of handwritten document identification and to test how the knowledge about document type can contribute to improvement in handwritten documents recognition in case of authentic documents appearing in patient records archive.

Problem of automatic document recognition in not new and was considered in many earlier publications ([1], [3], [4]). The novelty of approach described here consists in application of results of document type recognition at further stages of handwritten text recognition process. Additionally, proposed document identification method provides as a by-product the parameters of afine transform, which precisely matches document image to the template of its type. It may be useful to determine the image area where handwritten text to be recognized is located or to eliminate pre-printed background graphic elements intersecting handwritten text.

2. FRAME-BASED DOCUMENT TYPE IDENTIFICATION ALGORITHM

Handwritten documents subject to text recognition have usually characteristic graphical elements. Typically the most characteristic element is the frame or grid printed in the document form, possibly containing some guidance texts. It is assumed that each document type has its own distinct printed form containing fixed graphical elements. These elements can be used to identify the document type. The method proposed here is based on line segment identification and matching. The method is simpler than others reported in literature ([1], [4]), yet sufficiently robust for recognition of typical documents used in medical documentation.

Let D be the number of document types processed in the system. For each document type there exists its binary template image f_k, $k = 1,...D$. The problem of document type identification can be stated as follows. For given document image o find such document template index k^*, which minimizes certain dissimilarity measure $d(f_k, o)$:

$$k^* = \arg\min_{k\in\{1,\ldots,D\}} d(f_k, o) \tag{4}$$

The similarity measure is based on relation between fixed graphical elements extracted from the image and fixed elements of the document template. Let us assume that each document template has characteristic layout of vertical and horizontal grid lines, what is typical to documents with pre-printed frames. Fig. 1 shows document templates typically used in hospital for medical episode summary, observations, medical history and treatment. All of them contain characteristic frames consisting of vertical and horizontal lines.

Proposed document identification procedure recognizes frame lines of the image, matches them to lines on the template, calculates the matching factor. and uses it as dissimilarity measure $d(f_k, o)$. The method is based on the following assumptions:

- each document template contains a set of distinct horizontal and vertical lines,
- layout of lines is significantly different on all templates,

- recognized document does not contain handwritten lines nor sketches which may confuse line extraction algorithm,
- scanning process preserves scanned image scale, i.e. the document can be matched to its template merely by translation and rotation.

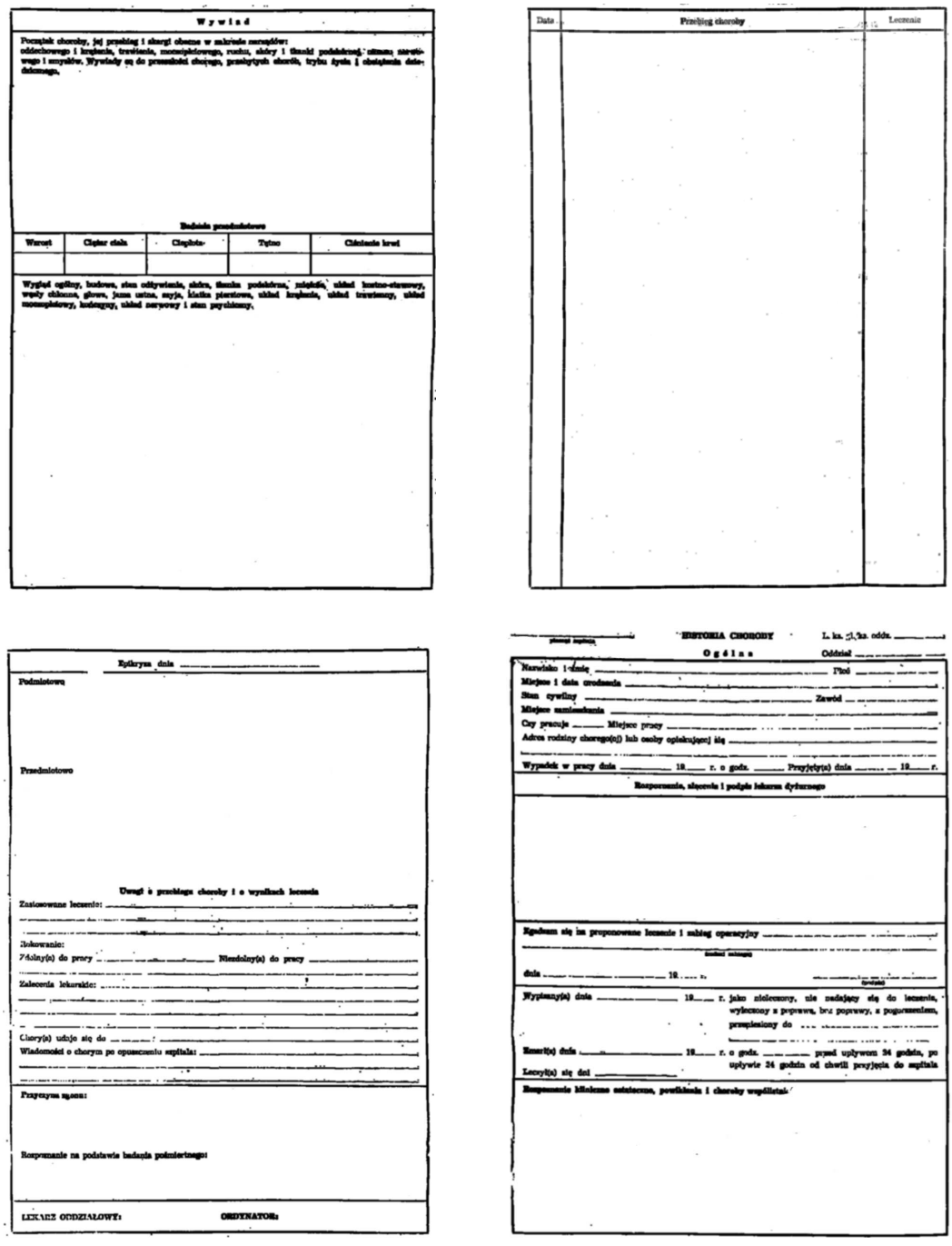

Fig.1. Templates of medical documents used in document type identification experiment

The document type recognition procedure is as follows:

- identify dominant lines on the document image using classical Hough transform for lines detection ([5]),

- correct detected lines position to obtain two sets of parallel lines – one set contains lines corresponding to vertical lines on the template, the second one contains lines which are horizontal,
- rotate the grid of recognized lines so as to obtain perfectly vertical and horizontal lines,
- for each template *k* find such translation (t_x, t_y), which minimizes certain dissimilarity measure between *k*-th template and processed document image,
- finally recognize the template k^* having minimal dissimilarity measure.

The dissimilarity measure $d(f_k, o)$ is in fact dissimilarity measure between original frame in the template and set of extracted lines on appropriately rotated document image. The dissimilarity measure is calculated in the following way. Let $(x_1^o, x_2^o, ..., x_{M^o}^o)$ and $(y_1^o, y_2^o, ..., y_{N^o}^o)$ denote *x* and *y* coordinates of vertical and horizontal lines on the image being recognized. Let $(x_1^k, x_2^k, ..., x_{M^k}^k)$ and $(y_1^k, y_2^k, ..., y_{N^k}^k)$ denote the sets of vertical and horizontal line coordinates on the *k*-th template. The extracted lines set is translated by (t_x, t_y) to maximize matching (or equivalently - minimize dissimilarity) between $((x_1^o, x_2^o, ..., x_{M^o}^o, y_1^o, y_2^o, ..., y_{N^o}^o)$ and $((x_1^k, x_2^k, ..., x_{M^k}^k, y_1^k, y_2^k, ..., y_{N^k}^k)$. For given translation vector (t_x, t_y) dissimilarity is defined as:

$$\begin{aligned} d(f_k, o, t_x, t_y) &= d((x_1^k, x_2^k, ..., x_{M^k}, y_1^k, y_2^k, ..., y_{N^k}), (x_1^o, x_2^o, ..., x_{M^o}^o, y_1^o, y_2^o, ..., y_{N^o}^o), t_x, t_y) = \\ &= \sum_{i=1}^{M^K} g_i^V \min_{j \in \{1, M^o\}} |x_i^k - (x_j^o + t_x)| + \sum_{i=1}^{N^K} g_i^H \min_{j \in \{1, N^o\}} |y_i^k - (y_j^o + t_y)| \end{aligned} \quad (1)$$

where g_i^V and g_i^H denote the importance of vertical and horizontal lines on the template. The dissimilarity measure between document image *o* and the template f_k is obtained by minimizing $d(f_k, o, t_x, t_y)$ over t_x and t_y.

$$d(f_k, o) = \min_{t_x \in (-x_{res}^0, x_{res}^k), t_y \in (-y_{res}^o, y_{res}^k)} d(f_k, o, t_x, t_y) \quad (2)$$

The method has been applied to distinguishing four types of documents being the part of typical patient record: analysis of disease, observations, medical history, treatment. All of them are in clinical practice written on specific black and white forms shown on Fig.1. Test set consisted of 229 document images. Document images were scanned in 300 dpi resolution and stored in lossless compression greyscale image files. Each document type was represented by approximately the same number of images. The method proved to be very robust. Only 3 document images were recognized incorrectly. Recognition errors appeared in case of document images where most of the frame image were screened in order to hide some confidential information.

3. APPLICATION OF DOCUMENT CLASSIFICATION IN HANDWRITTEN WORDS RECOGNITION

Document type recognition makes it possible to apply specific text recognisers for each document type taking into account language properties specific for recognized type of document. In case of handwritten parts of medical patient documentation for each text type there is a set of specific words and phrases typically used. These specific language features are significantly

different for various text types. Language features specific for document type can be utilized to improve the text recognition quality, provided the text type is reliably recognized.

In particular, probabilistic lexicons build separately for all document types can be used to boost handwritten words recognition accuracy. Probabilistic lexicon is the set of words w_i, $i=1,...N$ that can appear in the text with associated probability p_i of each word appearance.

$$\Lambda = \{(w_1, p_1),...,(w_N, p_N)) \tag{3}$$

It can be built by simple analysis of text corpora representative for each document type.

Probabilistic lexicon can significantly improve accuracy of words recognition, when combined with results of word classification based merely on features extracted from text image. Combination can be easily obtained if soft recognition paradigm is applied ([2]). Soft recogniser provides the vector of support values. Support value represents classifier confidence that the object being recognized belongs to given class. Let $\Phi^I(o) = (d_1^I(o),...,d_N^I(o))$ denotes the soft word classifier based merely on features extracted from the image. The soft word recogniser assigns support value $d_i(o)$ to each word w_i from probabilistic lexicon. Combined classifier $\Phi^L(o) = (d_1^L(o),...,d_N^L(o))$ takes into account prior words probabilities p_n from the lexicon. Support factors d_n^L can be calculated as follows:

$$d_n^L = d_n^I(1+\alpha(\exp(p_n / p_{\max}) - 1), \tag{4}$$

where p_{max} is prior probability of most probable word in the lexicon.

Text type	Error rate with global lexicon e_g	Error rate with lexicon specific for document type e_t	Relative error reduction $(e_g-e_t)/e_g$
disease analysis	4.96%	4.01%	0.19
observations,	5.57%	4.84%	0.13
medical history,	5.28%	4.13%	0.22
treatment	6.19%	5.22%	0.16
AVERAGE:	5.50%	4.55%	0.17

Tab.2. Comparison of word recognition error rates with global and type-specific lexicons

Common probabilistic lexicon can be built for all text types or individual lexicon can be created for each type of document. To test how document type identification can improve word recognition quality in case of medical texts we have compared both approaches. Four text types described in previous section have been used in experiment. First, global probabilistic lexicon was created and applied to each text recognition independently of its actual type. Then individual lexicons were created for all text types, text type was identified using method described in section 2. and words were recognized using soft classifier applying the formula (4). Corpora of texts representative for distinguished document types used to build global and type-specific probabilistic lexicons were extracted from original contents of hospital information system database. Due to lack of sufficiently large set of scanned handwritten documents simulated experiment was carried out where word images were constructed from isolated character images. Details of simulated

experiment are presented in [8]. As a character recogniser multi layer perceptron was used. Directional features vector was applied. Feature extraction technique described in [7] was applied. Achieved word recognition error rates are compared in Tab. 2.

4. CONCLUSIONS

In the paper a method of document type identification and its application to handwritten medical document text recognition was presented. Proposed simple, yet robust method of document type recognition gives accuracy of document type recognition close to 99%. Obtained document type identification can be then used in handwritten text recognition. Applying specific language features characteristic for given text type can improve handwritten text recognition accuracy. Prior text type recognition makes it possible to apply word classifiers based on domain-specific probabilistic lexicons prepared for each text type. Experiments carried out with the set of authentic medical texts confirmed that statistical properties of lexicons extracted from different text types are significantly different. These differences can be effectively utilized to improved text recognition if the type of document is known. Experiments showed that application of domain-specific probabilistic lexicons can reduce word recognition error rate by about 17% in average.

The work described here is a part of wider project related to multilevel algorithms of handwritten medical texts recognition, where the results of soft word recogniser are input data to final recognition stage applying natural language syntactic model. Also on syntactic level there are significant differences in language models built for different text types. Knowledge about actual text type can be used on this level as well in order to select appropriate language model and in result to further improve text recognition accuracy.

Acknowledgement. This work was financed from the State Committee for Scientific Research (KBN) resources in 2005 - 2007 years as a research project No 3 T11E 005 28.

BIBLIOGRAPHY

[1] CESARINI F., GORI M., MARINAI S., SODA G., INFORMys: a flexible invoice-like form reader system, IEEE Trans. on PAMI, Vol 20, No 7, pp. 730-745, 1998

[2] KUNCHEVA L., Combining Classifiers: Soft Computing Solutions, in: Pal S., Pal A. [eds.]: Pattern Recognition: from Classical to Modern Approaches, , World Scientific pp. 427-451, 2001

[3] PENG H., LONG F., CHI Z., SIU W., Document image template matching based on component block list. Pattern Recognition Letters, No 22, pp 1033-1042, 2001

[4] MARTI V.M., BUNKE H., Using a statistical language model to improve the performance of an HMM-based cursive handwriting recognition system. Int. Journ. of Pattern Recognition and Artificial Intelligence, Vol 15, No 1, pp. 65-90, 2001

[5] NIXON M., AGUADO A., Feature extraction & image processing. Newnes Press. Oxford, 2002

[6] LI X.Y., TAN C. L., DING X., A hybrid post-processing system for offline handwritten Chinese script recognition Pattern Alan. Applic, No 8, pp. 272-286, 2005

[7] KURZYNSKI M., SAS J., Application of three-level handprinted documents recognition in medical information systems, In: J.L. Oliveira [ed]: Biological and Medical Data analysis, Proc. of 6th Int. Symposium ISBMDA, Springer Verlag (LNBI) , pp. 1-12, 2005

[8] SAS J., Handwriting Recognition Accuracy Improvement by Author Identification, Proc of VIII Int. Conf. on Artificial Intelligence and Soft Computing, LNAI, Springer Verlag, 2006