handwriting recognition,
hospital information systems,
natural language processing

Jerzy SAS[*] Maciej PIASECKI[*]

A MULTI-LEVEL ARCHITECTURE FOR RECOGNITION OF POLISH HANDWRITTEN MEDICAL TEXTS

The paper discusses problems of handwritten text recognition in medical information systems. Aspects specific for medical notes recognition are considered. Limited set of authors, relatively small dictionary of typically used words, characteristic phrases used in particular categories of documents can be utilized in order to increase the accuracy of handwritten text recognition. The concept of five-stage recognition pipeline and its implementation in flexible text recognition system cooperating with hospital information system are described. The recognition pipeline consists of document category identification, writer recognition, isolated character and word classification and finally sentence recognition based on syntax language model. At the stage of character and word recognition, used classifiers are trained individually for each author and for each text category. Preliminary results of medical text recognition using techniques proposed in recognition pipeline are presented. The proposed recognition method can be applied to hand-printed text recognition as well as to cursive unconstrained script.

1. INTRODUCTION

The ability to import handwritten documents is a frequent functional requirement expected from modern hospital information systems (HIS). The need of automatic handwritten text processing and recognition follows usually from existence of large paper archives containing medical episodes from the past, which are necessary either for current patient treatment or for research and scientific activities. Unfortunately, physician's writing style is generally considered to be very difficult to read, even by a human. While there are still no handwriting recognition methods offering sufficiently high text recognition accuracy, recognition of texts written by doctors seems to be especially difficult task. There are however some specific features in typical problem of medical document recognition, which make recognition problem easier. Facilitations follow from the following facts:

- documents created in environment of typical hospital ward or outpatient clinic come from a few typical categories (e.g. analysis of disease after its end, observations, medical history, treatment),
- document categories can be easily recognized by finding characteristic elements of document preprinted form or text layout (e.g. logos, frames, fixed graphical elements of the form),

[*] Institute of Applied Informatics, Wroclaw University of Technology, Wyb. Wyspianskiego 27, 50-370 Wroclaw, POLAND, e-mail: maciej.piasecki@pwr.wroc.pl, jerzy.sas@pwr.wroc.pl

- the language is specific for each text category, in particular characteristic set of typical words can be approximated for each document category and set of specific phrases can be derived from the representative corpus of category-specific texts,
- texts are written by small set of authors (physicians employed in institution from which texts are analysed).

The aim of works described in this paper is to elaborate multilevel architecture for medical text recognition which makes use of mentioned above characteristic features of application domain. The architecture applies five level text recognition pipeline consisting of the following stages:

- ***document identification stage*** – where category of document is identified in order to pass this information to further stages, where it can be applied to select word and sentence recognisers which fit characteristic language features in category,
- ***author identification stage*** – where text author is identified in order to apply specific text recognisers trained individually for single author,
- ***character recognition stage*** – where text image is segmented into isolated characters which are recognized using character classifier specific for identified author,
- ***word recognition stage*** – where words are recognized using lexicons specific for the document category recognized at earlier stage; in case of cursive script recognition character and word recognition stages can be merged,
- ***sentence recognition stage*** – where complete sentences are recognized using results of word recognition and language syntactic model specific for identified document category.

There are many announcements about multilevel handwritten text recognition systems in literature ([2], [4], [5]), and generally multilevel text recognition concept obviously isn't new. The novelty of the approach presented here consists however in unique combination of author recognition with document type identification at early stages of recognition process in order to apply individual recognisers trained for identified writers and text categories. Because of limited volume of the paper, we focus our attention on proposed architecture of multilevel text recognition system. Subsequent stages are briefly described and results of preliminary experiments are presented. Details of stages implementation can be found in other referenced articles ([8], [10],[11]).

2. STAGES OF MULTILEVEL TEXT RECOGNITION

2.1. AUTHOR IDENTIFICATION

In handwritten medical texts recognition we usually deal with the situation where the texts being recognized come from relatively small set of known and permanent authors. The count of writers typically is of the order of several in typical hospital ward, where medical documentation arises. Handwriting style is very individual for each human being and there are significant differences in shapes of the same characters or words written by various writers. This fact can be misleading for automatic character or word classifier trained using common corpus of texts coming from the whole group of writers. Better results can be expected if text is recognized by the classifier trained exclusively using text samples written by actual author of the text. In very rare cases the text is explicitly marked by its writer in the way allowing for unambiguous writer recognition. The writer can be however recognized using characteristic features of writing style. Because writer identification can be also erroneous, the method of writer identification utilization at later stages of text recognition pipeline should be in some way tolerant, so as to prevent definite degradation of recognition quality in case of erroneous author identification. In our approach soft recognition concept is applied ([3]) which makes the whole recognition process more tolerant for author identification errors.

The author identification is carried out at the first stage of text recognition. Writer recognizer uses two groups of features extracted from scanned text image:

- directional features consisting of stroke direction histogram and edge hinge angle distribution extracted according to the procedure described in [7],
- additional geometric features set containing average letters width and height, variance of letters height, average distance between words and characters (in case of block handwriting).

Soft recognition paradigm is applied on writer and character recognition levels. It means that writer recognizer $\Psi^W(x)$ produces the vector of support values

$$\Psi^W(x) = (w_1, ..., w_K) \tag{1}$$

Support values w_i are non-negative and sum-up to unity. The value w_i represents the classifier confidence that the text subject to recognition is written by i-th author. Also character recognizers used at higher level apply soft recognition, providing this time the vectors of support values for letters from alphabet. For each author we have individual character classifier

$$\Psi_k^C(x) = (d_1^k, ..., d_L^k), \quad k = 1, ..., K, \tag{2}$$

where K is the number of authors, L is the number of characters in alphabet and d_l^k is support factor for l-th character from the classifier for k-th author. Final character soft classification $(c_1, ..., c_L)$ is obtained by combining support factors d_l^k where supports for authors w_k elaborated by author recogniser are used as weights:

$$c_l = \sum_{k=1}^{K} w_k d_l^k \tag{3}$$

In order to test how text recognition accuracy can be improved by author identification preliminary experiments have been performed. In the experiment original fragments of patient records concerning disease analysis extracted from HIS database were rewritten by 25 authors. Each author wrote 5 text samples. The experiment was performed in five rounds. In each round, testing set was created using 25 texts (one from each author) while remaining 100 texts were used as training set applied to train writer recognizer and individual character recognizer for each writer. In each round another sample was selected from each writer. Achieved writer identification accuracy was 78%.

Writer identification results were then applied to text recognition according to formula (3). Multi-layer perceptron was used as character classifier with directional features set extracted according to method described in [8]. Achieved character recognition accuracy of the algorithm based on writer identification was compared with the accuracy of single stage approach, where single character classifier is trained with all texts in training set, regardless of actual authorship. Character recognition error rate was reduced from 9.0% to 5.3% (41.1% of error reduction) in result of applying personalised word recognisers trained individually for single author.

2.2. DOCUMENT CATEGORY RECOGNITION

All documents categories processed in the system have 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 category has its own distinct printed form containing fixed graphical elements. These elements can be used to identify the category of document being recognized.

Correct recognition of document categories is not a problem in case, where document forms are especially designed taking automatic recognition into account. This is however not a case if old documents are to be recognized, which layout is not adapted to automatic processing. In our approach we assumed that each document template contains characteristic grid of vertical and horizontal lines, usually dividing document form into fields containing defined information. Document identification is achieved by detecting grid lines position from document image and by matching it to corresponding lines position on templates. For each template, a matching factor is calculated. This document category is finally identified for which matching factor is highest. Details of document identification procedure are discussed in [11].

Experiments have been performed with identifying medical documents belonging to four categories constituting typical patient record: disease analysis, observations, medical history and treatment. The method was tested using the set of 229 low quality document images. Documents came from paper archive of old patient records. In most cases they were significantly faded-out and deteriorated by multiply copying on xerox-machine. Despite low document image quality, only 3 documents were recognized incorrectly, what gives almost 99% of correct recognitions.

2.3. APPLICATION OF DOCUMENT IDENTIFICATION IN WORD RECOGNITION

Each document category contains handwritten text concerning different aspect of medical episode related to patient. Language features such as typically used words and phrases appear to be significantly different in different document categories. Word classifier applied at the next stage of text recognition can make effective use of known text category, resulting in improvement of word recognition accuracy. In our approach we applied a concept of probabilistic lexicons supporting the process of word recognition. Probabilistic lexicon contains the set of words most frequently appearing in texts being recognized and relative frequencies of words appearance. For each text category we have category-specific probabilistic lexicon. The lexicon specific for given text category can be used to support word recognition, provided that the category is identified.

The experiment was carried out in order to evaluate how distinguishing document categories can improve accuracy of word recognition. For four document types recognizable by document classifier described in previous section, individual probabilistic lexicons were created using text corpus extracted from authentic HIS database. Each text processed in the system was first classified by document category and next the word classifier using category-specific lexicon was applied. This approach was compared with another one, where document identification was not performed and words were recognized using single common probabilistic lexicon. Word recognition error rate was reduced from 5.50% to 4.55% (17% of error rate reduction) in result of category specific probabilistic lexicons application. Details of experiment are described in [11]

2.4. FINAL TEXT RECOGNITION ON SYNTAX LEVEL

The errors produced at earlier stages of recognition pipeline can be corrected by application of a *language model* describing somehow proper natural language expressions. The language model is commonly a stochastic model of possible sequences of words and their *grammatical classes* ([6]).

However, Polish possesses huge number of word forms due to inflection and almost free word order. Moreover, medical texts include often mistakes, abbreviations, foreign words etc. These makes creation of a stochastic language model very difficult task. Thus, we want to investigate an application of more complex morpho-syntactic constraints expressed in symbolic rules, less dependent on statistic features. As the first step, a *rule-based tagger* called TaKIPI ([9]) was integrated with the architecture. The tagger is a program that chooses for each word one morpho-syntactic analysis from the set of possible ones, i.e. it chooses between verb and noun, or between masculine and feminine gender. TaKIPI combines hand-written rules and C4.5 Decision Trees ([1]) (DTs). It was trained on the IPI PAN Corpus (IPIC) [6] joined with the corpus of medical texts extracted form HIS database.

TaKIPI returns for a word the probability of its choice. We transformed it to a measure of *confidence of decision* (CD), which is a difference between the probabilities of the least possible choice and the most possible one for a given word.

TaKIPI is used to asses which candidate from the a list of the k=10 most probable ones returned by soft word classifier fits best a given language expression. For each candidate, we pass to the tagger each possible permutation of candidates from the context of $\langle -n, +m \rangle$ plus the best candidates from outside of the context up to the sentence boundaries. From outside of the window the best candidates for that moment are taken: before the window — as evaluated by our algorithm, after the window — according to the support factors from soft word classifier. The CDs of processed words are collected. However, CD of unknown word, number, punctuation mark etc. is just 0. We decided to omit zero values of CD and to calculate the average from non-zero CDs. This average is called a *context consistency measure* (CCM). This way of calculation proved its superiority against maximum or minimum over CDs that were also tested experimentally. CCM alone allows to achieve the results comparable with the word classifier. In order to use both CCM and probabilities of word classifier, a simple multiplication of both has been applied. The candidate with the highest combined measure is chosen as the best one. After testing on the same test sample, error rate of combined algorithm decreased by 61% as the effect of TaKIPI application.

3. SYSTEM ARCHITECTURE

Five-stage handwritten document recognition concept presented in previous sections has been elaborated taking into account specific needs of application to medical documents. For the sake of practical implementation useful in daily clinical practice, corresponding architecture of flexible document recognition system must be designed and implemented. To be practically useable, the implementation of handwritten document recognition in medical information system must have the following basic functionalities:

- processing of form-like structured documents as well as unconstrained script handwriting,
- ability to define recognizable document types and their logical and graphical structure,

- automatic recognition of document category,
- flexible manual text verification supported by soft words classification results,
- ability to call text recognition module from within user interface of HIS,
- batch processing of scanned documents packages,
- gathering of recognized and manually verified text images in tagged text sample database used periodically to re-train classifiers,
- ability to perform initial training procedure which trains writer classifier and writer-specific text recogniser using the set of correctly classified sample text images,
- periodic re-training of writer and text classifiers using gathered set of text images,
- automatic update of probabilistic lexicons and syntax language models using actual contents of HIS database.

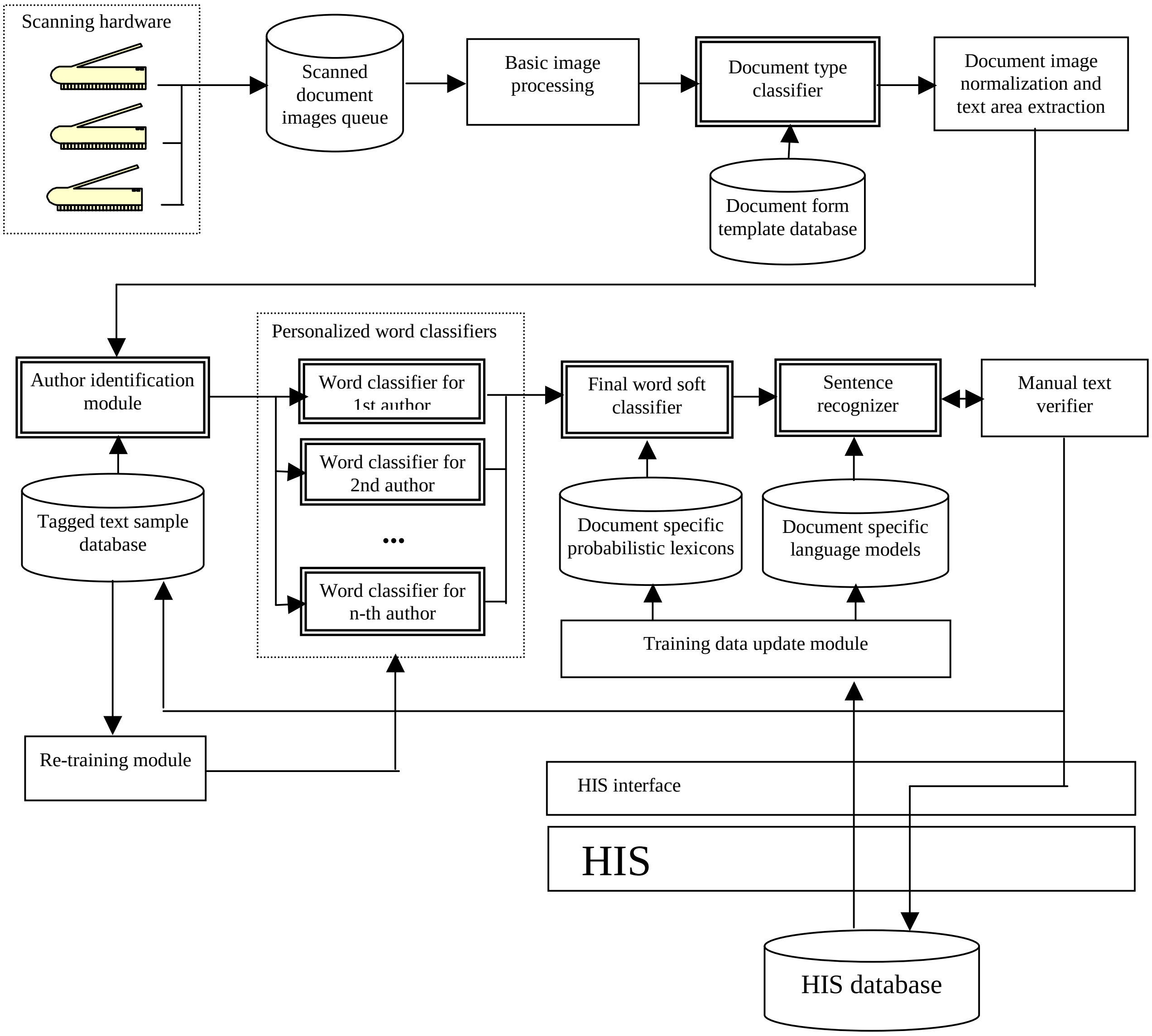
Scanning hardware
Scanned document images queue
Basic image processing
Document type classifier
Document image normalization and text area extraction
Document form template database
Personalized word classifiers
Author identification module
Word classifier for 1st author
Word classifier for 2nd author
...
Word classifier for n-th author
Final word soft classifier
Sentence recognizer
Manual text verifier
Tagged text sample database
Document specific probabilistic lexicons
Document specific language models
Training data update module
Re-training module
HIS interface
HIS
HIS database

Fig. 1. Handwritten document recognition system architecture

Proposed architecture of medical document recognition system is presented on Fig.1. The system is intended to work in batch processing mode. It automatically scans a bunch of documents inserted into scanner, performs automatic document category classification as well as writer identification and finally recognizes texts contained in documents. The result is data structure that can be read by *manual text verifier*. The manual text verifier is a module which supports manual correction of automatically recognized text. It is a specialised text editor which:

- makes possible to select actual word for each word position by selection from n-best words fetched by soft word classifier and modified by sentence recogniser,
- guides the text correction process by showing the word area on the complete document image for currently selected word in recognized text,
- updates the soft classification of words in the remaining part of the sentence as the left-most part of sentence is fixed by verifier proceeding in left-to-right order,
- prepares tagged text image where word areas in whole text image are appropriately extracted and for each word area its actual word is determined.

To be practically useful in daily clinical practice, the system must be almost maintenance-free. In particular, it must be highly integrated with existing HIS. The integration must allow at least for:

- automatic updating of training data used to periodically re-train classifiers in the system,
- insertion of recognized and verified texts into HIS database.

Update of training data is performed by *training data update module*. It extracts texts form HIS database and builds probabilistic lexicons and language models for distinguished and supported document categories. The module is executed periodically to keep data important for word and sentence recognition up-to-date. In similar way *re-training module* is used. The role of this module consists in periodically re-training of document category classifier and personalized word recognisers using correctly classified text images. It is assumed that each document recognized in system undergoes manual verification. After verification procedure the segmented text is obtained, where for each word image actual word is known and additionally actual authorship of document is determined. These data are stored in *tagged text sample database*. The database contents is then used as the training set for author recogniser and for personalised word recognisers. The re-training procedure is called when the personalized text sample database contents grows by determined amount of new texts.

4. CONCLUSIONS

In the paper, multilevel medical document recognition architecture has been discussed. The architecture takes into account specificity of medical text recognition. Preliminary experiments with author and document identification carried out using authentic medical text corpus showed that accuracy of text recognition can be observably improved by applying text recognisers trained individually for identified authors and for identified document category. Application of language syntax model resulted in further significant reduction of word recognition error rate.

The idea of multistage text recognition can be useful in practical clinical application only if the handwritten document recognition system is integrated with existing HIS. Integration of proposed system architecture with HIS consists in automatic extraction of probabilistic lexicons and language models from corpus of texts collected in HIS database and in automatic attachment of recognized documents to data objects stored in HIS.

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] QUINLAN, J. C4.5: Programs for Machine Learning. Morgan Kaufmann, 1993.

[2] PFISTER M., BEHNKE S.ROJAS R., Recognition of handwritten ZIP codes in real – word non-standard-letter sorting system, Journ. of Appl. Intelligenc, Vol 12, No 1, pp. 95-114, 2000

[3] 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

[4] GORSKI N., ANISIMOV V., AUGUSTIN E., Industrial bank check processing the A2iA CheckReader™, Int. Journ. on Document Analysis and Recognition No 3, pp. 196-206, 2001

[5] 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

[6] PRZEPIÓRKOWSKI, A., The IPI PAN Corpus Preliminary Version. Institute of Computer Science PAS, 2004

[7] SCHOMAKER L., BULACU M., FRANKE K., Automatic writer identification using fragmented connected-component contours. Proc. of 9th IWFHR, Japan, Los Alami-tos: IEEE Computer Society, pp. 185-190, 2004

[8] 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

[9] PIASECKI, M., GODLEWSKI, G., Reductionistic, tree and rule based tagger for Polish. In Kłopotek, M.A., Wierzchoń, S.T., Trojanowski, K., eds. Proceedings of Intelligent Information Processing and Web Mining 2006, Springer Verlag, 2006.

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

[11] SAS.J., PEJCZ J., Application of document type identification in medical handwritten text recognition, Proc of XI Int. Conf. MIT-2006 (submited to the same conference, not published yet, eventually will be removed in final paper)