Error Evaluation and Applicability of OCR Systems

Ventzislav Alexandrov

Abstract:
Using an OCR for volume document conversion, obligatory require a preliminary evaluation of its performance. This paper investigates the main recognition errors in document reading systems and the causes for their generation. Metrics for estimation of recognition rate are presented. Proposed is a methodology for error evaluation, especially for documents of large volume. An evaluation of human intervention for correction of the system output is given. Special attention is paid to the applicability of the results and it is shown that not all applications require a near zero error rate.

Key words:
Document conversion, optical character recognition (OCR), text recognition, error evaluation

INTRODUCTION

Increased power of the tools for scanning and reading paper documents in the last ten years allowed their more intensive practical implementation. Along with that, the reduced price made the use of such technologies more massive. They became a serious alternative of manual input especially for paper documents of large volume.

Converting paper documents into computer files is the main precondition for electronic management of the information contained in these documents. The purpose of the conversion may be different, for example searchable information massif, small size storage, republication, indexing, web publication and etc, but the goal is always to receive a content that is suitable for electronic processing.

Today's OCR applications are capable to distinguish automatically the different type of information (text and graphics). They can identify tables and other tab structures, different fonts and styles are also recognized. Despite all these capabilities, there is always a need of human validation and correction of the received results, because these tools are still not perfect. However, the main advantage of machine conversion is the achieved performance in speed and volume compared to the manual input. Combining the machine output and the human intervention gives beneficial results, as generally, the human errors are different form the machine ones. The amount of human corrections is always examined from the economical point of view.

Unlike the tests of commercially available OCR packages, which are often carried out in computer magazines and which in most of the cases represent a simple comparison of recognition rate, this paper will try to present an evaluation and the origin of the errors committed by the OCR systems. Here is proposed a methodology for error evaluation for large volume documents and the question about the applicability of the received OCR output is examined. Also, a general estimation of the human influence on the results after the proofing is presented.

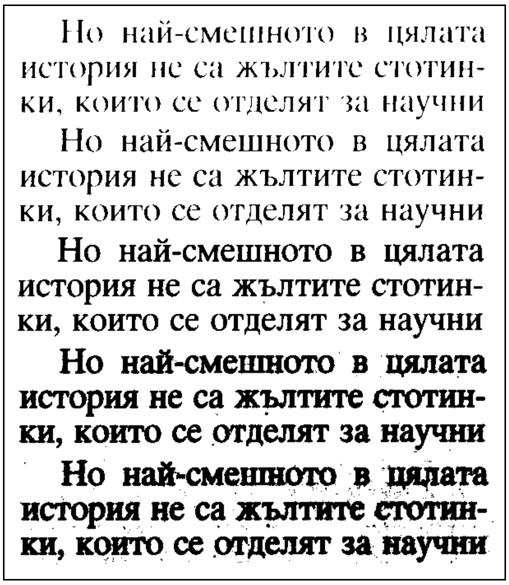
DOCUMENT QUALITY AND SCANNING SETTINGS

Recognition results depend heavily on the quality of the characters in the scanned image file. On his turn the image file depends on quality of the printed media and the correct setting of used scanning device.

Mainly black and white scanning is preferred, generally with resolution from 200dpi to 400dpi. For A4 format sheet scanned at 300dpi, the image file contains 2,520x3564 pixels, which is practically enough if the size of used fonts is not less than 8 points. If the resolution is chosen too small, a loss of important character features may happen. But if the resolution is chosen unnecessary big, then image files became bigger, while the recognition is not improving. Brightness and contrast of the scanner also play substantial role. Correct setting of all these parameters is crucial for the recognition results.

Quality of the print and paper is also important. Old yellowish or grainy paper could hamper the recognition process.

On fig. 1 is presented a typical recognition result. It is seen that when the quality of the image is getting poor the recognition is getting more incorrect and the number of errors increases.



Но най-смешното в цялата
история не са жълтите стотин-
ки, които се отделят за научни

Но най-смешното в цялата
история не са жълтите стотин-
ки, които се отделят за научни

Но най-смешното в цялата
история не са жълтите стотин-
ки, които се отделят за научни

Но най-смешното в цялата
история не са жълтите стотин-
ки, които се отделят за научни

Но най-смешното в цялата
история не са жълтите стотин-
ки, които се отделят за научни

)1п н^^й-^м^п^нп^п п п^л^^тп
И^1'(п)ИЯ ((^ С^1 Ж1^71'И'('С ^'(^)1'ИГ(-
КИ, К()И'1'() ^^ ()'(')(^Л^')' '^:1 ^):)У'1Г(И

Но най-смешното в пялата
история нс са жълтите стотин-
ки, които се отделят ^а научни

Но най-смешното в цялата
история не са жълтите стотин-
ки, които се отделят за научни

Но най-смешното в цялата
история не са якълтите стотин-
ки, които се отделят за науч^^и

Но на^смевмо^о ^^дя^е,та
ястория не са акълтяте е^отии-
ки, ковто се отделят за наушп^

Fig.1. Results depend heavily on image quality

It is quite difficult to determine the optimal setting for a given set of documents. Very often the only possibility of finding the optimum is the method of "trials and errors". To find the best setting, a careful examination of the quality of the scanned image must be performed. The easiest way to do that is by zooming the image on the screen. Chosen setting is always a compromise because for volume tasks the documents are seldom uniform. Generally, during the process of conversion some change of the settings may be required, but this slows down the whole process.

RECOGNITION ERRORS AND ETALON TEXT COMPARISON

Every text is a gathering of characters, which belong to a given alphabet, combined in words, which on their turn pertain to a given language vocabulary. The task of every OCR system is to recognize individually every character and then validate it according to the surround context in the row and page. This is not an easy job because characters may have different shape for every font and language even in one and the same document. This multilingual and omnifont generalization is difficult to achieve and it is one of major error generators in OCR conversion along with the bad quality of the images.

OCR systems can make four types of errors in recognition process:

- substitution – recognizing of one character as another. This often happens for structurally close characters
- deletion – ignoring a character because it is regarded as noise
- rejection – either the system can't recognize a symbol or it is not sure in its recognition.

- addition – one symbol is recognized as two symbols or a noise is recognized as character.

Automatic search of these errors is not a straightforward process. For doing it, an etalon text for comparison is needed. Receiving such etalon text is very responsible task and could be done in three ways:

- typing the document – it is better to use double input for assuring correct texts
- receiving it with another OCR and proofing all errors – better to be reviewed by more than one reviewer
- using ready computer text files from which are generated test image pages

For automatic comparison of the sample text and the etalon text string comparison techniques are used. The editing distance between the OCR result and the etalon document should be calculated. The weight of substitutions, suppressions and additions must be defined, which will allow computing the minimal number of edit operations required to correct the results. This minimal number could be used as measure of the correction cost.

MEASURES FOR RECOGNITION PERFORMANCE

The main measure is the global error rate

$$G_{err} = 100\frac{n_e}{n_c} \qquad (1)$$

where n_e is the number of committed errors and n_c is the number of all characters in the text.

An error is registered if recognized character and the character in the etalon text are different. In this parameter are included the deletions, substitutions and additions.

The other important measure is the rejection rate

$$G_{rej} = 100\frac{n_r}{n_c} \qquad (2)$$

where n_r is the number of rejections.

Recognition rate, which is mostly used for describing the efficiency of a given OCR system, is given by the equation

$$G_{rec} = 100\frac{n_c - n_r - n_e}{n_c} \qquad (3)$$

It is clear that for these three parameters the following is true:

$$G_{rec} + G_{rej} + G_{err} = 100\% \qquad (4)$$

Nevertheless that the above measures are widely used, they do not give a direct evaluation of the reliability of the OCR systems. Often another measure is needed, which could show the reliability of the results. This measure is the level of reliability

$$G_{rel} = 100\frac{n_c - n_r - n_e}{n_c - n_r} = \frac{G_{rec}}{G_{rec} + G_{err}} \qquad (5)$$

Obviously, the goal is to maximize G_{rec} and to minimize G_{err} but this is not easy in practice. Often G_{rec}+G_{err} is called total percentage of recognition. Sometimes it is preferable to increase the rejection level (if the system allows it) in order to diminish the confusion and increase the reliability of the recognition results [4].

Generally, a good commercial OCR product achieves about 99% (or even more) of total percentage of recognition (G_{rec} + G_{err}). This measure is received form the statistics of the OCR system and the question is how reliable it is, because the percentage of errors G_{err} can't be calculated by the system itself and could be significant.

Error rate and rejection arte are inversely proportional [1], which means that the only way to diminish the percentage of confusion and to increase the level of reliability is to increase the level of rejections. In this way the number of characters for correction will increase, but at least the rejections are shown by the OCR system itself and are easier to correct.

Sometimes, it is necessary to calculate recognition or error rate on word basis instead on character basis. Nevertheless that $G_{err(w)}$ can be calculate directly from the OCR output, it is interesting to know the relation between this parameter and $G_{err(ch).}$ Their relation can be expressed by

$$G_{err(w)} = G_{err(ch)} * \beta \tag{9}$$

where β is the average number of characters per word

This formula is an approximation and is correct only if all the character errors are in different words, which in fact is the worst case and is not always true. It should be noted that $G_{err(w)}$ is generally much bigger than $G_{err(ch)}$ and shouldn't be confounded with it. If for example $G_{err(ch)}$ = 0.7% and the average number of character per word is 9, then the maximum $G_{err(w)}$ = 6.3%.

HUMAN CORRECTIONS

The most costly process in document conversion is the manual correction after the OCR recognition. What is important from industrial point of view is not the error itself, but the difficulty of correcting this error. The correction can be made either only on rejections or on the all errors (which in practice means on the whole text). The second option is seldom preferred for the documents of large volume because the price of the whole process increases too much. If the human correction is taken only on the rejections, then

$$G_{rej} = G_{cor} + G_{hum} \tag{6}$$

where G_{cor} is the percentage of really corrected errors while G_{hum} is the percentage of human errors. Human errors are intrinsic to the operators and they shouldn't be neglected in the evaluation of the whole document conversion process. The strange thing is that the human errors are more tolerated, while we expect too much from machines.

It should be noted that it is difficult to achieve less than 3 human errors to one thousand entered characters [3].

Taking into account (4) and (6) it could be written that

$$G_{rec} + G_{err} + G_{cor} + G_{hum} = 100\% \tag{7}$$

or

$$(G_{rec} + G_{cor}) + (G_{err} + G_{hum}) = 100\% \tag{8}$$

where $(G_{err} + G_{hum})$ represent the value for minimization in order to achieve important economical gain when doing document conversion.

The percentage of errors G_{err} can't be received directly from OCR software because its statistics gives only the rejections G_{rej} and total percentage of recognition (G_{rec}+G_{err}). In other to calculate G_{err} there have to be an etalon text, which is must free from errors, to compare the result with. In this way the number of additions, suppressions and substitutions can be found.

PERFORMANCE EVALUATION

For the passed years of OCR investigation it is turned out that viable recognition is not possible when every character is recognized individually, independently of all the other symbols in page images. All characters are in relation with their neighbours, at least with the neighbours in the row. That is why more and more rules based OCR algorithms are used for improving recognition results. OCR systems have become like expert systems and on great extent they are based on rules, which integrate different relations and help recognition when ambiguity and uncertainty appear. Systems of such a type have behaviour, which can change in an unexpected manner. Sometimes, small variations of character image my lead to completely different result. As the modern OCR systems use more than ever rules, their performance is not theoretically foreseeable. That is why the evaluation of their recognition performance could be made only on the basis of their output results, either on all converted documents or on a sample set of pages. The last option is preferred, especially for large volume conversion.

For making reliable evaluation, the sample part must be as much representative as possible for all the documents, which will be subject to conversion. A random choice of pages or parts of pages from all the types of printed information should be done. It is very important to make a random choice because choosing the most difficult or the easiest parts will not allow correct evaluation of the OCR software. Along of being random the choice must be proportional to the present types of the printed information. Let's assume that the documents are printed for example 90% with Arial font and 10% with Times font and let's assume that the OCR recognize 100% the texts with Arial and 50% the texts with Times (this percentage is given only as an example, but really the Arial font is better recognized by OCR systems than Times font, because it is much more noise proof [5]). If we take an equal amount of samples for every font, 50% for Arial and 50% for Times, the recognition rate for this sample will be 75% but for all the documents the recognition will be 95%. As we see the estimation will be not quite correct. If we convert the same sample with another OCR programme, which for example recognize Arial and Times at 80%, the sample text will be recognized at 80%, better recognition than the first OCR, and all the documents will be recognized at the same rate. As you can see, nevertheless the evaluation rate of the second OCR is better, the recognition rate for all the documents will be worse. This show how important is the selection of the sample part.

Table1. Evaluation of OCR output

	Arial rec. rate	Times rec. rate	Sample rec. rate (A50%, T50%)	All documents rec. rate (A90%, T10%)
OCR1	100%	50%	75%	95%
OCR2	80%	80%	80%	80%
OCR1+OCR2	100%	80%	90%	98%

Some authors recommend the usage of more than one OCR software and combining after that the results. With the two OCR systems in our example, in theory we would achieve a recognition rate of 90% for the sample and 98% for the all documents but in practice it is very difficult to cooperate the work of two or more OCR systems, especially for low quality documents. The decision for every symbol is made according to vote results and a specially designed SDK of the OCR packages are needed. These SDKs must give an extended output, not only the code for every symbol image but all the information for performing the voting process. In literature is showed that the error could be reduced considerably when combining different OCR engines [2].

APPLICABILITY OF RESULTS

One of the main problems in volume document conversion is the evaluation of the applicability of the results. To say it in other words, tolerable percentage of errors depends on the area of application. Not always a recognition rate of near 100% is needed. For some tasks recognition of 90% may be acceptable, while for other it will be catastrophic. For example, conversion of text information containing digits is very sensitive to errors, especially when no special features, as of checking sum, are used. Today's most used conversion is the mixed conversion, which combine the automatic recognition with human intervention. In this way verification, correction and completion of the documents is performed at the same time.

Bellow, in Table1 is given an example of various OCR applications.

Table2. OCR applications

Applications	**Volume**	**Manual work**	**Acceptable error rate**
Office document reentry	Low	Zoning and error check	Custom decision
Book re/publication	Medium-high	Zoning and error check	Near zero
Bank checks/Mail sorting	High	Rejected pages	Near zero
Information Retrieval Systems (Indexing Systems)	High	Almost none	Not very strict, some characters/words have no importance at all

CONCLUSIONS

In our days OCR systems achieved maturity, which allows their real use in practice, especially for volume document conversion. They substantially increase the productivity but it is unrealistic to believe that now they can completely replace humans. As their errors are generally different from the errors committed by humans, it is a fruitful approach to combine the efforts of the software and operators. Achieved economical advantages are big when the received error rate is not higher than the appropriate for a given application. If the desired percentage of errors should be lower than the achieved, then instead of OCR conversion other options must be envisaged, as of double operator input with comparison or a single operator input with systematic proofing.

REFERENCES

[1] Chow C.K., "On Optimum Recognition Error and Reject Tradeoff", IEEE Transactions on Information Theory, V IT-16, No. 1, Jan. 1970, 41-46.

[2] Ho T. K., Hull J. J., and Srihary S. N., "Decision Combination in Multiple Classifier Systems", IEEE Trans. on PAMI, 16(1), pp66-75, 1994

[3] Nagy G., "At the frontiers of OCR", Proceedings of IEEE, pp. 1093-1100, July 1992

[4] Wilson C. L., "Evaluation of Character Recognition Systems", In Newral Networks for Signal Processing III, IEEE, New York, pp. 485-496, 1993.

[5] Alexandrov, V.I. "Фактори, влияещи върху работата на тексторазпознаващите програми", Computer N9, 1996, pp 69-71

ABOUT THE AUTHOR

Ventzislav Alexandrov, MSc, Institute of Computer and Communication Systems, Bulgarian Academy of Sciences, Phone (+352 2) 732 951 (ext.119), e-mail: ventzi@bas.bg