СЕНСОРНЫЕ СИСТЕМЫ, 2020, том 34, № 3, с. 217–225

ТЕХНИЧЕСКОЕ ЗРЕНИЕ

УДК 303.732

ANALYSIS OF A STOPPING METHOD FOR TEXT RECOGNITION IN VIDEO STREAM USING AN EXTENDED RESULT MODEL WITH PER-CHARACTER ALTERNATIVES

© 2020 г. K. B. Bulatov[1,2], B. I. Savelyev[1,2,*], V. V. Arlazarov[1,2], and N. V. Fedotova[2]

[1] Federal Research Center "Computer Science and Control" of RAS 117312 Moscow, 60-letiya Oktyabrya avenue 9, Russia

[2] Smart Engines Service LLC 121205 Moscow, Skolkovo innovation center, Nobel st. 7, 132, Russia

*E-mail: bsaveliev@smartengines.ru

Received April 7, 2020; Revised April 22, 2020; Accepted April 29, 2020

In the field of document analysis and recognition using mobile devices for capturing, and the field of object recognition in a video stream, an important problem is determining the time when the capturing process should be stopped. Efficient stopping influences not only the total time spent for performing recognition and data entry, but the expected accuracy of the result as well. This paper is directed on extending the stopping method based on the modelling of the next integrated recognition result, in order for it to be used within a string result recognition model with per-character alternatives. The stopping method and notes on its extension are described, and experimental evaluation is performed using the open datasets MIDV-500 and MIDV-2019. The method was compared with previously published methods based on input observations clustering. The obtained results indicate that the stopping method based on the next integrated result modelling allows to achieve higher accuracy, even when compared with the best achievable configuration of the competing methods, however the computations required are significant and more research should be targeted on optimizing its implementation.

Key words: recognition in video stream, mobile OCR, stopping rules, decision making, mobile document recognition, anytime algorithms

DOI: 10.31857/S0235009220030026

INTRODUCTION

Modern document entry systems allow to automatize the process of data extraction from various documents, either business, regulatory, or personal. Such systems are used for creating digital archives of historical documents (Van Phan, 2016), recognition of small-scale documents such as business cards (Dangiwa, 2018), ID documents, driving licenses, passports (Arlazarov, 2019), as well as large-scale business documents (Esser, 2013).

Increasing computational power of mobile devices and rising technical characteristics of small-scale digital cameras lead to increased interest in methods for automatic document entry using mobile devices (Арлазаров, 2017; Ravneet, 2018; Povolotskiy, 2019; Skoryukina, 2018) and forensic analysis (Chernyshova, 2019; Полевой, 2019). As a rule, regular smartphones are used for document recognition, due to relatively low cost, sufficient computational power for performing recognition tasks, and ability of capturing video (or sequence of images). The ability to capture video is one of the most important advantages over traditional scanners, as in such case more information could be retrieved in comparison with a single image, and each newly acquired document image may be used to improve the recognition result (Bulatov, 2017). Figure 1 illustrates an example of per-frame recognition results combination in a video stream. As it can be seen, the correct integrated result may be acquired even before any individual frame result is correct.

While processing the sequence of frames and combining the per-frame recognition results a single more precise one, the problem arises – when this process should terminate? The capturing process in a general case might not be naturally limited, and if a sufficiently good combination strategy is employed the increase of the number of integrated observations the expected result precision also increases (Bulatov, 2019a). However, the time required to perform recognition and output the final result is also very important and thus efficient strategies for video stream recognition stopping should be developed and further studied.

The optimal stopping problems themselves occupy a special place in mathematical statistics and decision theory (Ferguson, 2010; Christensen, 2019). Some methods were also proposed for the video stream recognition problem (Arlazarov, 2018; Bulatov, 2019b; Bulatov, 2020b). In (Arlazarov, 2018) a method was presented which consisted of clustering the set of per-

#	Text field image	Frame result	Integrated result
1		JW	JW
2		U KW	U KW
3		IU F F LIK	U KW
4		IU, TJT LAN	U LKW
5		LIE, FFZ LKN	U LKW
6		LIU, TIL LAN	IU, T LAW
7		LIU, IIF LAN	IU, TIF LAW
8		LAU, TSF LAN	LIU, TIF LAW
9		LEE IIL LAK	LIU, TIF LAW
10		IIE II LKW	LIU, TIF LAW
11		LAU, TIL LAN	LIU, TIL LAW
12		LAU, RSZ LAN	LIU, TIL LAN
13		LAU, T SZ LAN	LIU, TSZ LAN
14		LAU, T SZ LAN	**LAU, TSZ LAN**
15		**LAU, TSZ LAN**	**LAU, TSZ LAN**
16		**LAU, TSZ LAN**	**LAU, TSZ LAN**

Fig. 1. Example of per-frame recognition results combination in a video stream. Correct recognition results are highlighted. Images are taken from MIDV-500 dataset (Arlazarov, 2019).

frame field recognition results, estimating a confidence score for each cluster, and making a stopping decision based on three parameters: cluster size, cluster confidence, and the total number of processed observations. The method can be applied in two ways: the clusters may be formed from the initial per-frame recognition results, or from the integrated results obtained on each stage. This method, however, is not fully formalized and raises the questions of tuning the clustering parameters. In (Bulatov, 2019b) a method is proposed, which considers the video stream recognition stopping as a monotone sequential decision problem. It presents a stopping strategy derived from the properties of monotone stopping problems, however it was tested only for text recognition results as simple strings, without any per-character alternatives. At the same time an extended string recognition result model with per-character alternatives is an important way of text recognition result representation: it is used for recognition results post-processing (Llobet, 2010) and was shown to be valuable for improving the integrated result precision (Bulatov, 2019a).

The goal of this paper is to investigate the applicability of stopping strategy introduced in (Bulatov, 2019b) for text string recognition with per-character alternatives and to compare it with alternative methods, which are already adapted for such recognition result model. In section 2 a brief description of the stopping method is given, and in section 3 experimental evaluation and comparison for stopping methods is presented.

METHOD DESCRIPTION

In order to provide a description of the stopping method, let us consider text string recognition in a video stream as a sequential decision problem. Let $\mathbb{X}$ represent a set of all possible text string recognition results, and the task is to recognize a text string with correct value $x^* \in \mathbb{X}$ given a sequence of images $I_1, I_2, I_3, \ldots$ which are obtained one at a time. At stage n the image I_n is recognized and the per-frame recognition result $x_n \in \mathbb{X}$ is obtained. After x_n is obtained the results $x_1, x_2, \ldots, x_n$ are combined using some combination algorithm to produce an integrated result $R_n \in \mathbb{X}$. The stopping decision is now to either stop the process and use R_n as the final recognition result, or continue the process in an effort to obtain in the future the integrated result with higher expected accuracy. If the process is stopped at stage n the penalty is paid in form of a linear combination of distance from the obtained result to the correct one (a "price for error") and the number of frames process (a "price for time"):

$$L_n = \rho(R_n, x^*) + c \cdot n, \quad (1)$$

where ρ is a metric function on the set $\mathbb{X}$, and c is a constant representing the price paid for each observation (in relation to the cost of the recognition error).

The stopping rule is formally defined as a sequence of real-valued functions, which represent the conditional probability of stopping after the stage n is reached (Ferguson, 2006). However, such conditional probability functions define a random stopping time N, which could be used to denote the stopping rule instead with more clarity (the conditional probability of stopping could also be inferred from the random stopping time N). The distribution of N depends on the obtained observations $x_1, x_2, x_3, \ldots$. The stopping problem is an optimization problem of finding a stopping rule with a goal to minimize the expected loss, which can be expressed as follows:

$$\mathrm{E}(L_N(X_1, X_2, \ldots, X_N)) \to \min_N, \tag{2}$$

where $\mathrm{E}(\cdot)$ is a mathematical expectation, and $X_1, X_2, \ldots, X_k$ are random recognition results with identical joint distribution with x^* of which $x_1, x_2, \ldots, x_k$ are realizations observed at stages $1, 2, \ldots, k$.

Optimal stopping problems, such as (2), can be classified into a variety of subtypes, and for each type some theoretical results have been achieved over the years. A large class of optimal stopping problems grounded in real applications are the finite horizon problems, which could be solved using the backwards induction method (Berezovskij, 1981). The finite horizon problem is characterized by a fixed stage T such that any stopping rule calls for stopping on the stage $n = T$. For the task of text string recognition in a video stream this would correspond to a predefined "time-out": the number of processed video frames after which the procedure is always stopped (possibly with a null result). The addition of such a "time-out" T to the stopping problem (2) does not divert it from the practical relevance.

The other important subtype of stopping rule problems are the so-called monotone stopping problems (Ferguson, 2006; Chow, 1961). The definition is as follows: consider the event $A_n = \{L_n \le \mathrm{E}(L_{n+1} | X_1 = x_1, \ldots, X_n = x_n)\}$, which indicates that the loss which would be suffered if the process is stopped at stage n is not greater than the expected loss of stopping at the next stage, given that only the first n observations were obtained. The optimal stopping problem is considered monotone if the occurrence of the event A_n leads to the occurrence of A_{n+1} for all n. In other words, if the current loss is not greater than the expected loss at the next stage, then this will be true for all future stages as well. It can be proven (Ferguson, 2006) using the backwards induction method, that for monotone stopping problems with a finite horizon the optimal stopping rule for the problem (2) is the so-called "myopic" rule:

$$\begin{gathered} N_A = \min\{n \ge 0 : \\ L_n \le \mathrm{E}(L_{n+1} | X_1 = x_1, \ldots, X_n = x_n)\}, \end{gathered} \tag{3}$$

which calls the process to stop at the earliest stage when the event A_n occurs.

In terms of the loss function (1) the events A_n considered in the definition of the monotone stopping problem can be expressed as follows:

$$\begin{gathered} A_n = \{\rho(R_n, x^*) - \\ - \mathrm{E}(\rho(R_{n+1}, x^*) | X_1 = x_1, \ldots, X_n = x_n) \le c\}. \end{gathered} \tag{4}$$

Since x^* is unknown at the moment of making the stopping decision (and thus the loss function (1) cannot be computed), the occurrence of the event A_n also cannot be determined. Thus, some additional assumptions need to be made in order to be able to estimate the left-hand side of the inequality in (4). The stopping method proposed in (Bulatov, 2019b) for the problem of text string recognition in a video stream is relying on an assumption that the expected distances between two consecutive integrated results decrease over time:

$$\mathrm{E}(\rho(R_n, R_{n+1})) \ge \mathrm{E}(\rho(R_{n+1}, R_{n+2})), \quad \forall n > 0. \tag{5}$$

Let us consider the event $B_n = \{\mathrm{E}(\rho(R_n, R_{n+1}) \mid X_1 = x_1, \ldots, X_n = x_n) \le c\}$, which occurs when the current estimation of the distance from the current combined recognition result to the next one is not greater than the relative cost of the observation. Due to assumption (5) we can expect that the occurrence of B_n will lead to the occurrence of B_{n+1} for all n. At the same time, due to the triangle inequality governing the metric function ρ, the occurrence of B_n leads to the occurrence of A_n. This means, that if B_n has occurred, the events $A_n, A_{n+1}, A_{n+2}, \ldots$ will occur as well, so starting from this stage the problem is effectively monotone and it is optimal to stop (however, the true optimal stopping rule may have called for stopping the process at the earlier stages). Thus, we obtain a stopping rule which does not depend on the correct recognition result value x^*:

$$\begin{gathered} N_\Delta = \min\{n \ge 0 : \\ \mathrm{E}(\rho(R_n, R_{n+1}) | X_1 = x_1, \ldots, X_n = x_n) \le c\}. \end{gathered} \tag{6}$$

To implement the stopping rule (6) we need to estimate the expected distance from the current combined result to the next one. To achieve this, a modelling of the next integrated result is proposed in (Bulatov, 2019b), defining the following stopping method:

$$\begin{gathered} \text{Stop if} \quad \hat{\Delta}_n \le c, \\ \hat{\Delta}_n = \frac{1}{n+1}\left(\delta + \sum_{i=1}^{n} \rho(R_n, R(x_1, x_2, \ldots, x_n, x_i))\right), \end{gathered} \tag{7}$$

where c is an observation cost (essentially, a threshold parameter of the stopping method), δ is an external parameter, and $R(x_1, x_2, \ldots, x_n, x_i)$ is a modeled integration result of all consecutive observations obtained by the stage n concatenated with i-th observation.

It is noted in (Bulatov, 2019b) that the concrete method of modelling the next integrated result might depend on the nature of the combination algorithm and other specifics of the problem, however the proposed method could still be used in a quite general case, by replacing the recognition results combination method and the metric function ρ. In the original paper the experiments were conducted using Tesseract (Smith, 2007) as the recognition algorithm, simple string of characters as a recognized string representation, a normalized Levenshtein distance (Yujian, 2007) as a metric function ρ, and ROVER (Fiscus, 1997) as a combination algorithm. It was not clear whether this stopping method would be effective for an extended string recognition result model, containing per-character classification alternatives. In the extended model, the string recognition result can be represented as a matrix of alternatives:

$$x = \begin{pmatrix} (q_{11}, c_{11}) & \cdots & (q_{M1}, c_{M1}) \\ \vdots & \ddots & \vdots \\ (q_{1K}, c_{1K}) & \cdots & (q_{MK}, c_{MK}) \end{pmatrix}, \quad \forall i \quad \sum_{j=1}^{K} q_{ij} = 1, \quad (8)$$

where c_{ij} are character labels, $q_{ij} \in [0,1]$ – class membership estimations for each character, K – the size of the alphabet, and M – the length of the string. The combination algorithms for string recognition result in this extended model can be viewed as a generalization of the ROVER approach, and to define the metric function ρ a generalized Levenshtein distance may be used after defining the metric on the individual character classification results (Bulatov, 2019a).

To compare different stopping methods the expected performance profiles can be used – a methodology from the field of anytime algorithms (Zilberstein, 1996). Expected performance profiles are graphical plots which show dependence of the expected accuracy on the expected time required to obtain it.

EVALUATION

In order to evaluate the stopping method described in section 2 we used an open dataset MIDV-500 (Arlazarov, 2019) which contains 500 video sequences of 50 types of identity documents with ground truth. Each original clip contained 30 frames. The frames on which the document was not fully visible were removed from the consideration, and the resulting clip was repeated in a loop until the original size of 30 frames was reached. In addition we used an extension of this dataset called MIDV-2019 (Bulatov, 2020a), which has the same ground truth format, but features two additional capturing conditions.

The ground truth in the MIDV-500 contains both ideal values for text field recognition and the ideal geometric coordinates, i.e. for each field its geometric position in the document boundaries is known, making it possible to crop the field from any frame of the dataset. Text fields were cropped with margins with width equal to 30% of the smallest text field bounding box side. Since physical dimensions of each document type in the MIDV-500 dataset is also known, it is possible to crop each field in a uniform resolution. For recognition, all text fields were cropped with the resolution of 300 DPI. After cropping each text field was recognized using a text string recognition subsystem of Smart IDReader document recognition software (Bulatov, 2017), obtaining the recognized value as a sequence of character classification results with alternatives. For combination of per-frame recognition result a method from (Bulatov, 2019a) was used, which could be regarded as a generalization of the ROVER (Fiscus, 1997) approach for string recognition results with per-character alternatives. As a distance metric ρ a normalized version of the generalized Levenshtein distance was used, with a taxicab metric for individual character classification results.

In (Arlazarov, 2018) a stopping method was proposed, which was based on clusterization of the set of text field recognition results to n clusters, and making a stopping decision based on some properties of the most populous cluster. The method proposed in (Bulatov, 2019b) and described in section 2 was compared with this method in the original paper, however since the paper was focused on a simplified string recognition result model, not all features of the stopping method presented in (Arlazarov, 2018) were used, as the per-character alternatives were not available when using Tesseract as the text string recognition algorithm.

The clusterization of the observations is performed by their lengths (i.e. by the number of characters in the obtained string recognition results). For each cluster its confidence value is computed according to the following formula:

$$Q(C) = 1 - \prod_{x \in C} \left(1 - \min_{i=1}^{M(C)} \left\{ \max_{j=1}^{K} q_{ij}(x) \right\} \right), \quad (9)$$

where C is a cluster of observations with the same length $M(C)$. The stopping decision is made by three thresholding: the size of the largest cluster, the confidence of the largest cluster, and, if there is more than one cluster, the difference between confidences of the two largest clusters. Such thresholding meant that there are three stopping method parameters (three thresholds).

Two variations of the stopping method proposed in (Arlazarov, 2018) can be realized – the first, denoted hereinafter as N_{CX} which treats input observations $x_1, x_2, \ldots, x_n$ as strings to compose clusters with, and the second – N_{CR} – treats the integrated results $R_1, R_2, \ldots, R_n$ as observations and components of the clusters. Figure 2 and 3 illustrate the quality maps of the both approaches with variation of all three thresholds: each data point represents the mean number of

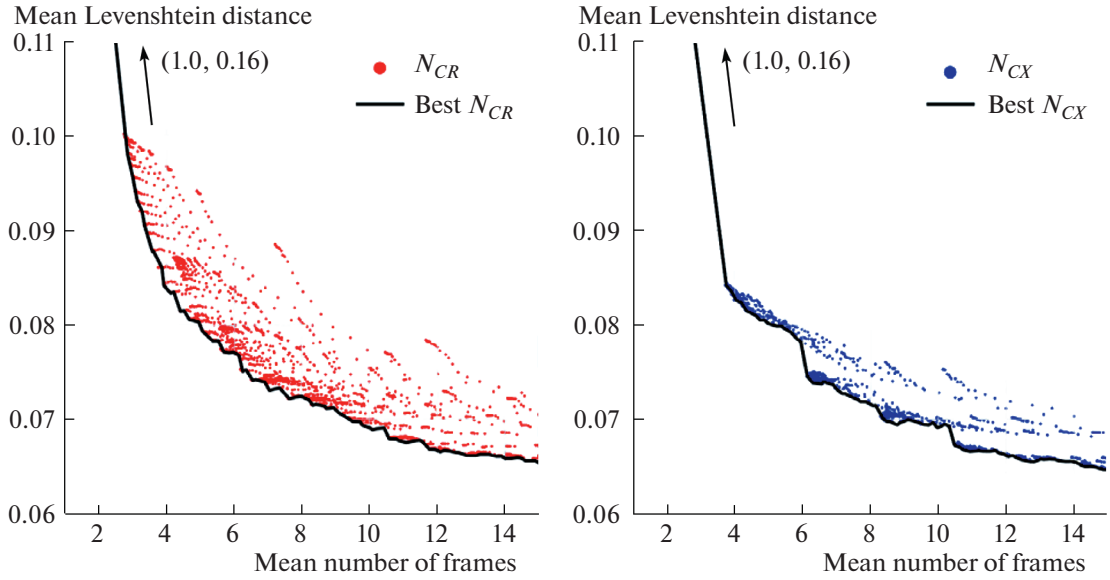

Fig. 2. Quality maps for stopping method described in (Arlazarov, 2018) in two implementation variations: clusterization of integrate results (left) and of the per-frame results (right). Black line designates the best achievable result. MIDV-500.

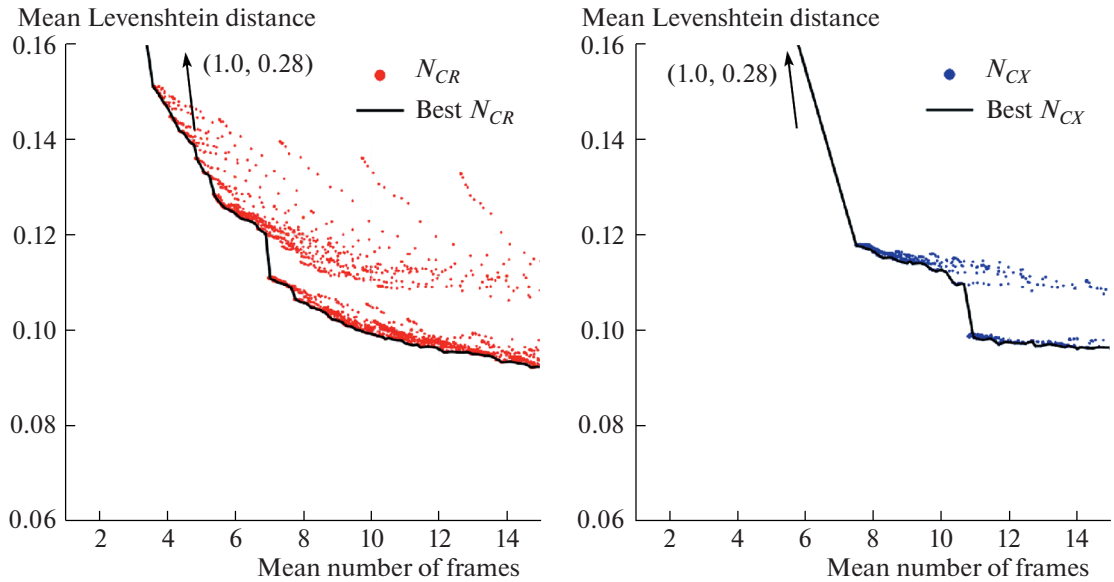

Fig. 3. Quality maps for stopping method described in (Arlazarov, 2018) in two implementation variations: clusterization of integrate results (left) and of the per-frame results (right). Black line designates the best achievable result. MIDV-2019.

observations processed before stopping and the mean distance of the integrated result to the correct value.

One of the main disadvantages of this stopping methods is that it is unclear how to jointly select the values for all thresholds to achieve the highest efficiency. In Figure 2 the black line represents the best option constructed a posteriori, which will be used for comparison with the method described in section 2.

Figures 4 and 5 illustrate the expected performance profiles comparison for the best achievable versions of the stopping methods N_{CX} and N_{CR}, the stopping method based on the modelling of the next integrated result N_{Δ}, described in section 2, and, as a baseline, a simple stopping method N_K which stops after observing K-th per-frame result. It can be seen that even though the best versions of the clustering stopping methods were evaluated, without clear understanding of how to obtain these jointly optimal threshold values, the method N_{Δ} still outperforms them.

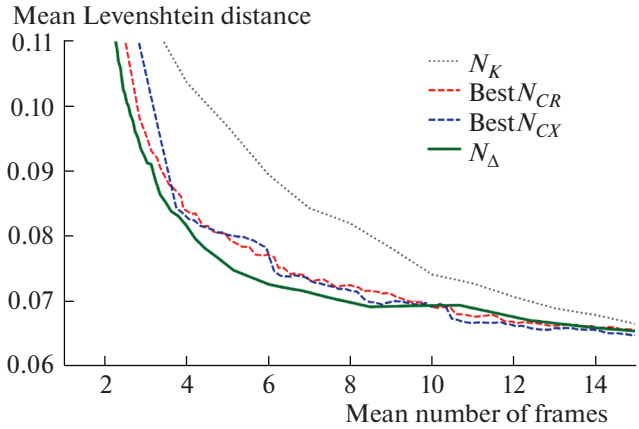

Fig. 4. Expected performance profiles for the baseline stopping method (simple integration, N_K), best versions of the clustering stopping methods, and the stopping method N_Δ, described in section 2. MIDV-500.

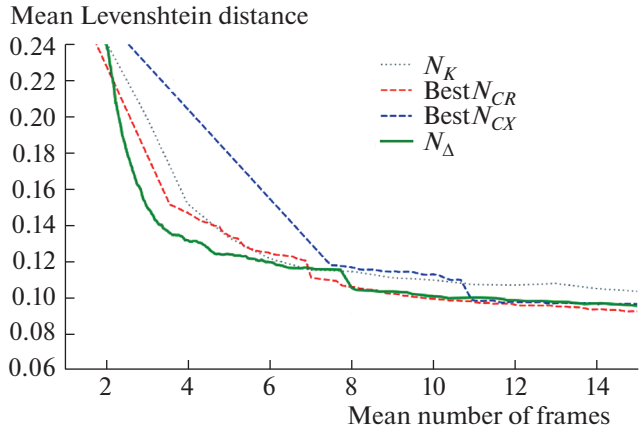

Fig. 5. Expected performance profiles for the baseline stopping method (simple integration, N_K), best versions of the clustering stopping methods, and the stopping method N_Δ, described in section 2. MIDV-2019.

Table 1 and 2 show the achieved mean integrated result accuracy (in terms of distance to the correct value) at stopping time, using the evaluated stopping methods and with restrictions to the mean number of processed observations. It can be seen that the method based on modelling the next integrated result and thresholding the estimation of the expected distance from the current result to the next one (N_Δ) outperforms the other methods. In particular, it allows to achieve higher result quality with the same average number of processed observations even when compared with the best achievable version of the previously proposed method (Arlazarov, 2018).

The stopping method N_Δ has a disadvantage to the clustering methods with regards to the computational efficiency when making the stopping decision. The computations required for the stopping method N_Δ to make the stopping decision at stage n grows linearly relatively to the number of processed frames, whereas for the clustering methods this dependence is virtually constant. Figure 6 illustrates the mean time per decision relative to the number of frames for the three eval-

Table 1. Achieved values of average distance from the integrated result to the correct field value at stopping time with restricted average number of processed observations at MIDV-500 dataset

Stopping method	Limitation to the average number of observations							
	≤3	≤4	≤5	≤6	≤7	≤8	≤9	≤10
best N_{CX}	0.161	0.084	0.080	0.078	0.074	0.072	**0.069**	**0.069**
best N_{CR}	0.096	0.084	0.080	0.077	0.074	0.072	0.071	**0.069**
N_K	0.115	0.104	0.097	0.089	0.084	0.082	0.078	0.074
N_Δ	**0.092**	**0.082**	**0.076**	**0.073**	**0.071**	**0.070**	**0.069**	**0.069**

Table 2. Achieved values of average distance from the integrated result to the correct field value at stopping time with restricted average number of processed observations at MIDV-2019 dataset

Stopping method	Limitation to the average number of observations							
	≤3	≤4	≤5	≤6	≤7	≤8	≤9	≤10
best N_{CX}	0.278	0.278	0.278	0.278	0.278	0.116	0.114	0.113
best N_{CR}	0.278	0.147	0.136	0.125	**0.111**	0.106	0.102	**0.099**
N_K	0.200	0.152	0.133	0.122	0.115	0.114	0.111	0.110
N_Δ	**0.150**	**0.123**	**0.119**	**0.116**	0.115	**0.103**	**0.101**	0.100

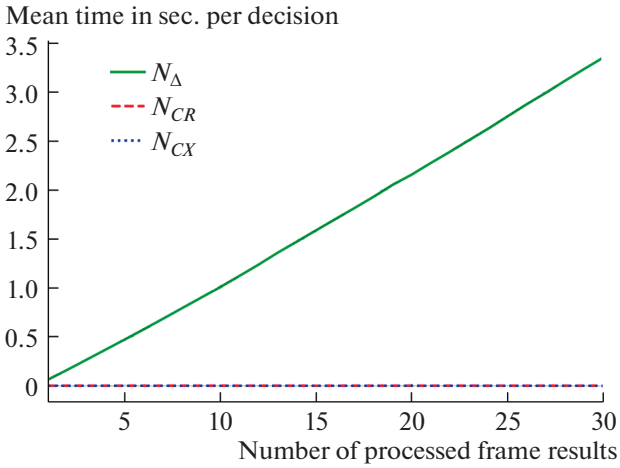

Fig. 6. Mean time per decision for the clustering stopping methods, and the stopping method N_Δ, described in section 2. MIDV-2019.

uated methods. Table 3 shows the tabular representation of the measured time characteristics.

As it can be seen from the performed experiments, the modelling of the next result according to (7) is very time-consuming, especially if an extended string recognition result representation is considered. This issue could be mitigated by approximate calculation of the next integrated results without complete integration, or alternative estimations of the expected distance to the next integrated result, such as using time series predictions. The essence of the stopping method does not change, as the main idea would still be to threshold the estimated expected distance to the next integrated observation, however the computational efficiency could be significantly improved.

CONCLUSION

The paper describes the problem of stopping the process of text line recognition in a video stream. Previously presented stopping methods were described and their properties analyzed. A method based on modelling of the next integrated result is described and applied to the model of text recognition result as an alternatives matrix with extended per-character classification results. The applicability of the stopping method in these conditions is shown, and the comparative evaluation is performed against previously published methods. It was shown that the next integrated result modelling method outperforms the previously published clustering methods, even in their best achievable configurations, but requires more time to make a decision. As future work it is planned to evaluate different ways to estimate the expected distance between the current integrated result and the next one, either with more efficient approximate modelling, or using

Table 3. Mean time in sec. per decision at MIDV-2019 dataset

Stopping method	Number of processed frame results									
	1	2	3	4	5	6	7	8	9	10
N_{CX}	5.7×10^{-4}	5.6×10^{-4}	5.5×10^{-4}	5.7×10^{-4}	5.4×10^{-4}	5.3×10^{-4}	5.6×10^{-4}	5.6×10^{-4}	5.8×10^{-4}	5.4×10^{-4}
N_{CR}	6.2×10^{-4}	6.1×10^{-4}	5.8×10^{-4}	6.2×10^{-4}	6.1×10^{-4}	6.4×10^{-4}	5.9×10^{-4}	6.3×10^{-4}	6.3×10^{-4}	6.5×10^{-4}
N_Δ	0.06	0.16	0.27	0.37	0.47	0.58	0.68	0.79	0.9	1.01
Stopping method	Number of processed frame results									
	11	12	13	14	15	16	17	18	19	20
N_{CX}	5.4×10^{-4}	5.7×10^{-4}	5.8×10^{-4}	5.3×10^{-4}	5.4×10^{-4}	5.8×10^{-4}	5.6×10^{-4}	5.5×10^{-4}	5.4×10^{-4}	5.5×10^{-4}
N_{CR}	6.0×10^{-4}	6.2×10^{-4}	6.3×10^{-4}	6.3×10^{-4}	6.2×10^{-4}	6.0×10^{-4}	6.4×10^{-4}	6.4×10^{-4}	6.1×10^{-4}	6.1×10^{-4}
N_Δ	1.12	1.24	1.36	1.47	1.59	1.7	1.82	1.93	2.05	2.16
Stopping method	Number of processed frame results									
	21	22	23	24	25	26	27	28	29	30
N_{CX}	5.6×10^{-4}	5.7×10^{-4}	5.6×10^{-4}	5.5×10^{-4}	5.4×10^{-4}	5.8×10^{-4}	5.5×10^{-4}	5.5×10^{-4}	5.5×10^{-4}	5.7×10^{-4}
N_{CR}	6.4×10^{-4}	6.2×10^{-4}	6.4×10^{-4}	6.1×10^{-4}	6.1×10^{-4}	6.3×10^{-4}	6.4×10^{-4}	6.1×10^{-4}	6.2×10^{-4}	6.2×10^{-4}
N_Δ	2.28	2.39	2.51	2.63	2.75	2.88	2.99	3.12	3.24	3.35

time series analysis, in order to improve on the computational efficiently of the stopping decision.

SOURCE OF FINANCING

This work is partially financially supported by Russian Foundation for Basic Research (projects 18-07-01387 and 19-29-09055).

REFERENCES

Polevoy D.V. *Ispol'zovanie mobil'nyh ustrojstv dlja vyjavlenija priznakov fabrikacii dokumentov, udostoverjajushhih lichnost'* [Identity documents forgery detection with mobile devices]. *Sensornye sistemy* [Sensory systems]. 2019. T. 33 (2). C. 142–156 (In Russian).

Slugin D., Arlazarov V.V. *Poisk tekstovyh polej dokumenta s pomoshh'ju metodov obrabotki izobrazhenij* [Text fields extraction based on image processing]. *Trudy ISA RAN* [Proc. Institute for Systems Analysis RAS]. 2017. V. 67 (4). P. 65–73 (In Russian).

Arlazarov V.V., Bulatov K., Chernov T., Arlazarov V.L. MIDV-500: A Dataset for Identity Documents Analysis and Recognition on Mobile Devices in Video Stream. *Computer optics*. 2019. V. 43 (5). P. 818–824.

Arlazarov V.V., Bulatov K., Manzhikov T., Slavin O., Janiszewski I. Method of determining the necessary number of observations for video stream documents recognition. *In Proc. SPIE (ICMV 2017)*. 2018. V. 10696. https://doi.org/10.1117/12.2310132

Berezovskij B.A., Gnedin A.V. Theory of choice and the problem of optimal stopping at the best entity. *Automation and Remote Control*. 1981. V. 42. P. 1221–1225.

Bulatov K. A method to reduce errors of string recognition based on combination of several recognition results with per-character alternatives. *Bulletin of the South Ural State University. Ser. Mathematical Modelling, Programming & Computer Software*. 2019a. V. 12 (3). P. 74–88. https://doi.org/10.14529/mmp190307

Bulatov K., Arlazarov V.V., Chernov T., Slavin O., Nikolaev D. Smart IDReader: Document recognition in video stream. *In 14th International Conference on Document Analysis and Recognition (ICDAR)*. 2017. V. 6. P. 39–44. https://doi.org/10.1109/ICDAR.2017.347

Bulatov K., Matalov D., Arlazarov V.V. MIDV-2019; challenges of the modern mobile-based document OCR. Twelfth International Conference on Machine Vision (ICMV 2019). 2020a. V. 11433. P. 717–722. https://doi.org/10.1117/12.2558438

Bulatov K., Razumnyi N., Arlazarov V.V. On optimal stopping strategies for text recognition in a video stream as an application of a monotone sequential decision model. *International Journal on Document Analysis and Recognition (IJDAR)*. 2019b. V. 22. P. 303–314. https://doi.org/10.1007/s10032-019-00333-0

Bulatov K., Savelyev B., Arlazarov V.V. Next integrated result modelling for stopping the text field recognition process in a video using a result model with per-character alternatives. *Proc. SPIE 11433, Twelfth International Conference on Machine Vision (ICMV 2019)*. 2020b. V. 114332M. https://doi.org/10.1117/12.2559447

Chernyshova Y., Aliev M., Gushchanskaia E., Sheshkus A. Optical font recognition in smartphone-captured images and its applicability for id forgery detection. *In Proc. SPIE (ICMV 2018)*. 2019. V. 11041. https://doi.org/10.1117/12.2522955

Chow Y.S., Robbins H. A martingale system theorem and applications. *Proceedings of the 4th Berkeley Symposium on Mathematics, Statistics and Probability*. 1961. V. 1. P. 93–104. University of California Press, Berkeley, CA.

Christensen S., Irle A. The monotone case approach for the solution of certain multidimensional optimal stopping problems. 2019. arXiv.1705.01763

Dangiwa B.A., Kumar S.S. A business card reader application for iOS devices based on Tesseract. *2018 International Conference on Signal Processing and Information Security (ICSPIS)*. 2018. P. 1–4. https://doi.org/10.1109/CSPIS.2018.8642727

Esser D., Muthmann K., Schuster D. Information extraction efficiency of business documents captured with smartphones and tablets. *In Proceedings of the 2013 ACM Symposium on Document Engineering*. 2013. P. 111–114. ACM, New York, NY, USA. https://doi.org/10.1145/2494266.2494302

Ferguson T.S. *Optimal stopping and applications*. 2006. URL: https://www.math.ucla.edu/~tom/Stopping/Contents.html (accessed 03.05.2020).

Ferguson T., Klass M. House-hunting without second moments. *Sequential Analysis*. 2010. V. 29 (3). P. 236–244. https://doi.org/10.1080/07474946.2010.487423

Fiscus J.G. A post-processing system to yield reduced word error rates: Recognizer Output Voting Error Reduction (ROVER). *In IEEE Workshop Automatic Speech Recognition and Understanding*. 1997. P. 347–354. https://doi.org/10.1109/ASRU.1997.659110

Llobet R., Cerdan-Navarro J., Perez-Cortes J., Arlandis J. OCR post-processing using weighted finite-state transducers. *In 2010 20th International Conference on Pattern Recognition*. 2010. P. 2021–2024. https://doi.org/10.1109/ICPR.2010.498

Povolotskiy M., Tropin D. Dynamic programming approach to template-based OCR. *In Proc. SPIE (ICMV 2018)*. 2019. V. 11041. https://doi.org/10.1117/12.2522974

Ravneet K. Text recognition applications for mobile devices. *Journal of Global Research in Computer Science*. 2018. V. 9(4). P. 20–24.

Skoryukina N., Shemiakina J., Arlazarov V.L., Faradjev I. Document localization algorithms based on feature points and straight lines. *In Proc. SPIE (ICMV 2017)*. 2018. V. 10696. https://doi.org/10.1117/12.2311478

Smith R. An overview of the Tesseract OCR engine. In Proceedings of the *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*. 2007. V. 02. P. 629–633.

Van Phan T., Cong Nguyen K., Nakagawa M. A nom historical document recognition system for digital archiving. *International Journal on Document Analysis and Recognition (IJDAR)*. 2016. V. 19 (1), P. 49–64. https://doi.org/10.1007/s10032-015-0257-8

Yujian L., Bo L. A normalized levenshtein distance metric. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2007. V. 29 (6). P. 1091–1095. https://doi.org/10.1109/TPAMI.2007.1078

Zilberstein S. Using anytime algorithms in intelligent systems. *AI Magazine*. 1996. V. 17 (3). P. 73–83. https://doi.org/10.1609/aimag.v17i3.1232

Анализ метода останова распознавания текста в видеопотоке с использованием расширенной модели результата с посимвольными альтернативами

К. Б. Булатов[a,b], Б. И. Савельев[a,b,#], В. В. Арлазаров[a,b], Н. В. Федотова[b]

[a] *Федеральное государственное учреждение "Федеральный исследовательский центр "Информатика и управление" Российской академии наук", 117312 Москва, проспект 60-летия Октября, д. 9, Россия*

[b] *ООО "Смарт Энджинс Сервис", 121205, Москва, Инновационный центр Сколково, улица Нобеля, д. 7, 132, Россия*

[#]*E-mail: bsaveliev@smartengines.com*

В сфере анализа и распознавания документов на мобильных устройствах, а также распознавания объектов в видеопотоке, задача определения момента времени, когда необходимо остановиться, является очень важной. Эффективность останова влияет не только на время, затраченное на распознавание и ввод данных, но и на ожидаемую точность результата. Данная работа направлена на расширение метода останова, основанного на моделировании следующего результата интеграции, с целью использования результата распознавания в виде строки с посимвольными альтернативами. Описаны метод и примечания по его расширению, произведена экспериментальная оценка на открытых наборах данных MIDV-500 и MIDV-2019. Рассматриваемый метод был сравнен с методами, опубликованными ранее и основанными на кластеризации входных наблюдений. Полученные результаты указывают на то, что метод останова, основанный на моделировании следующего результата интеграции, позволяет достигать более высокой точности, даже по сравнению с наилучшей достижимой конфигурацией конкурирующих методов. Однако данный метод обладает высокой вычислительной трудоемкостью и существует необходимость в оптимизации его реализации.

Ключевые слова: распознавание в видеопотоке, мобильный OCR, правила останова, принятие решения, мобильное распознавание документа, "anytime" алгоритмы