Algorithm for choosing the best frame in a video stream in the task of identity document recognition

M.A. Aliev [1,4], I.A. Kunina [1,2,3], A.V. Kazbekov [1], V.L. Arlazarov [4]
[1] Smart Engines Service LLC, Moscow, Russia;
[2] Institute for Information Transmission Problems (Kharkevich Institute) RAS, Moscow, Russia;
[3] Moscow Institute of Physics and Technology (State University), Moscow, Russia;
[4] Federal Research Center Computer Science and Control RAS, Moscow, Russia

Abstract

During the process of document recognition in a video stream using a mobile device camera, the image quality of the document varies greatly from frame to frame. Sometimes recognition system is required not only to recognize all the specified attributes of the document, but also to select final document image of the best quality. This is necessary, for example, for archiving or providing various services; in some countries it can be required by law. In this case, recognition system needs to assess the quality of frames in the video stream and choose the "best" frame. In this paper we considered the solution to such a problem where the "best" frame means the presence of all specified attributes in a readable form in the document image. The method was set up on a private dataset, and then tested on documents from the open MIDV-2019 dataset. A practically applicable result was obtained for use in recognition systems.

Keywords: human perception, quality assessment, document images, blur, sharpness, flares.

Citation: Aliev MA, Kunina IA, Kazbekov AV, Arlazarov VL. Algorithm for choosing the best frame in a video stream in the task of identity document recognition. Computer Optics 2021; 45(1): 101-109. DOI: 10.18287/2412-6179-CO-811.

Acknowledgements: This work was partially supported by the Russian Foundation for Basic Research (projects ## 17-29-03161, 18-07-01387).

Introduction

The steady growth of quality and reliability of automated text recognition algorithms over the past decade has led to an increase in demand for input and verification systems for various text documents [1, 2]. A classic source of document images in such systems are specialized flatbed scanners. However, with the development of modern mobile devices, input systems using small format cameras are gradually replacing traditional systems.

Images obtained from specialized scanners are characterized by a fairly uniform illumination of the document, high image resolution and the absence of projective distortions (see fig. 1a). At the same time, images of documents received from cameras of mobile devices can have a number of defects that are absent when working with a scanner: flares (see fig. 1d), "blurring" of the document area (see. fig. 1c), the document not completely present in the frame (see fig. 1e), etc. [3]. In this case, the use of several frames from the video stream with the subsequent combination of the recognition results obtained on those frames can significantly increase the recognition quality [4, 5].

In some cases, the recognition system is required to select one "good" frame from the video stream (see fig. 1b), which will either be shown to the operator to check the correctness of recognition, saved in a special database, or used to provide services (for example, issuing SIM cards). Hereinafter, such a frame will mean the document image that has in a readable form all the text attributes of the document with the owner's data.

In this paper we will consider the solution to the problem of evaluating the "goodness" of frames in a video stream and choosing the "best" frame. The possibility of cutting off "bad" frames based on the recognition results is investigated, provided that each given attribute corresponds to a certain recognition result. Since flare or camera focusing errors can lead to unpredictable recognition results [6], to assess the "goodness" of the frame the document image is additionally checked for flares and evaluated for blurs.

1. Related work

First of all, we note the works that directly analyze the quality of text areas in the document image. These methods can be conditionally divided into two groups: those that directly analyze the parameters of the font (typeface anatomy), and those that determine the readability of the font by indirect signs, for example, by the quality of recognition of OCR systems.

An example of the first group is the work [8], which uses the analysis of luminance gradients within zones containing individual characters or groups of characters. Another example is the work [9]. It describes three groups of features calculated for each symbol: morphological, anti-aliasing artifacts, and a group of spatial features that describe geometric distortions of the image. In [10] the assessment of the image quality of the document is the weighted sum of the image clarity assessment and the font parameters assessment. The latter, in turn, is the sum of three estimates: the number of dark specks around the text, imitating the speckle structure, the estimate of the inter-letter space and the estimate of the size of the inter-letter space to the total size of the letter, which were proposed in [11] and adapted by the authors of the article for their own document format.

Fig. 1. Example of difference between ideal document image from scanner and frames from video stream after localization stage: (a) document image from scanner; (b) relatively good shot; (c) blur; (d) flare; (e) document not fully presented. Frames are taken from video of Spanish id-documents from MIDV-2019 dataset [7] (folder 20_esp_id_new)

The second group of methods includes the work [12], which proposes a method for calculating the image quality assessment based on calculating the maps of the mean square deviations of the brightness gradient calculated on the text areas of the image. The calculated estimates were further correlated with the accuracy of the OCR systems on the same images. Another example is the work [13], where the image quality of the document was assessed using deep learning methods. For this, the input image using binarization methods was divided into sections containing text information of equal size, and the neural network was trained in such a way that the predicted quality of each section correlated with the recognition accuracy.

In addition to directly determining the quality of the text areas of the document, in literature, one can single out the direction when the entire document image is analyzed. In the work [14], a neural network model is proposed that receives an image as input and returns a quality value. The model was trained on the following data: a pair (input image - target image) and the value of the quality score for this pair.

A number of methods have also been proposed that calculate a document quality score based on some sharpness score. In the work [15], two values are used to assess the sharpness: the maximum gradient and the standard deviation of the gradient calculated for the entire image. The first value characterizes the sharpest part of the frame, the second shows how uniform the image is as a whole. The estimate proposed in [16] is based on measuring the width of the gradient transition that forms the boundaries of objects in the image: the sharper the image, the narrower the gradient transitions, and the lower this indicator.

Most of the works use their own internal datasets, which makes it difficult to compare different approaches. Therefore, it is necessary to mention the existence of open datasets [17 – 19] containing document images. The

main purpose of their creation and application is both to ensure the possibility of correctly comparing the quality of different OCR systems on the same dataset, and to compare different approaches to assessing image quality among themselves.

As can be seen from the review, to determine the readability of textual information on the document image, either explicitly specified signs of text degradation (thinning of letter strokes or gaps in symbols) are used, or machine learning methods, which themselves formulate features based on a training sample. Taking into account that the concept of character readability is formalized rather poorly (as well as the concept of a high-quality image of a document in general), the latter will require a significantly larger amount of training data. It should also be noted that all the methods mentioned above do not take into account that the document image may not be entirely in the frame: then a situation is possible when the document image will have a high-quality rating, but it does not contain all the necessary details.

In this work, the result of document recognition is analyzed to check for the presence of all specified attributes on the document image in the frame. It is assumed that if a frame contains the document image with all the specified attributes in a readable text form, then the recognition network's confidence in its response will tend to 1.0 for each specified attribute. Therefore, in this work, the possibility of determining the readability of a symbol by the confidence of the recognition system in its answer will be investigated. This will allow us not to explicitly set a list of possible reasons for the poor readability of a single character, and will also reduce the total number of recognition networks in the document analysis system, which is especially important in conditions of recognition on low-power computing processors (for example, smartphones or tablets).

2. Proposed solution

2.1. Task formulation

Let, as a result of recognition of a sequence of frames, a sequence of projectively corrected and recognized images be obtained: $I=(I_1,\ldots,I_N)\in I^N$ of length $N(I(j)=I_j)$, and each $I(j)$ contains M_j fields (document attributes): $\{F_1,\ldots,F_{M_i}\}\in I_j$.

Let us define the indices array as $I(p)=\{1,\ldots,n\}$. As

$$\delta p\in P=\bigcup_{i=2}^{n}P^i$$

we will define permutation of a random subsequence p, $\forall i\in I(\delta p)\,\exists!\, j\in I(p):\delta p(i)\equiv p(j)$.

The function of choosing the best image in the sequence $Q:I^N\to\{I_q,q_{score}\}$, where the best frame estimate q_{score} takes the given values: $q_{score}\in\{$"*good*", "*bad*"$\}$.

Note that the choice of the best image does not depend on the order of images in the sequence: $\forall\delta p\; p(Q(p))\equiv\delta p(Q(\delta p))$.

Thus, the goal of this article is to construct a function Q for choosing the best image in the sequence.

2.2. Algorithm for choosing the best frame

The general scheme of the proposed decision-making algorithm is shown in the next paragraph in the form of pseudocode. Here, the algorithm receives as input a sequence of projectively corrected images of the document (the projective image is achieved using specialized algorithms such as [20, 21]) and the results of recognition of all specified attributes on each image. Each such result contains the coordinates of the text field bounding rectangle (for more information on this topic, see [22]), the field recognition result, and the neural network's confidence in its answer. The result of the algorithm is the best image from the input sequence and its evaluation in the form "good"/"bad". The assessment is carried out through a sequential analysis of three frame quality indicators, calculated by analyzing the confidences of the recognizer in its answer, searching for flare in the document image and evaluating the "blur" of the document image.

Algorithm: Best frame choosing.

Input: N recognized images $I_1,\ldots,I_N$, thresholds T_{CS}, T_{FS}.
Output: Best image I_x, its grade as "good"/"bad"

```
1  rejected=[], accepted=[]
2  For each i⊂[1..N]:
3      CS_i = COMPUTE_CONFIDENCE_SCORE(I_i)
4      FS_i = COMPUTE_FLARE_SCORE(I_i)
5      DS_i = COMPUTE_DEFOCUS_SCORE(I_i)
6      If CS_i >T_CS and FS_i > T_FS then:
7          Add (Ij, DS_i) in accepted
8          Sort accepted by DS_i
9      else:
10         Add (Ij, DS_i) in rejected
11         Sort rejected by DS_i
12 If size(accepted)>0 then:
13     return accepted[0], "good"
14 else:
15     return rejected[0], "bad"
```

Document images for which the recognition system did not find all the attributes in the frame are not allowed to enter the module.

Next, the algorithms for calculating the three mentioned frame quality indicators will be considered in sections 2.3, 2.4, 2.5, and then, in section 3, all parameters and threshold values of the algorithm are adjusted.

2.3. Recognizer confidence analysis

Let there be a symbol image $x\in X$ in the string X and a finite set of classes $C=\{C_l\}_{l=1}^{M}$ of size M, also called the recognition alphabet.

The neural network implements the classifying function $A(x)$, which assigns the vector of alternatives $\vec{a}$ to the image x in such a way that:

$$A(x) = \vec{a} = (a_1, ..a_M),\ a_k = $$
$$= (C_{v_k}, p_k),\ p'_1 > p'_2 > \ldots > p'_M,\ \sum_{k=1}^{M} p'_k = 1,$$

where p'_k is an estimate of the recognition object x belonging to the class C_{v_k}, v_k is an index of class with k-th largest estimate. The recognition result of the symbol x is the class C_{v_1} with the maximal confidence p'_1 [23].

Let's define the recognition result of the string X as $\{(v_i, p_i)\}_{i=1}^{l_X}$, where (v_i, p_i) – the result of recognizing the i-th character in a string of l_X character long.

We will define the current frame as "good" if the following condition is met:

$$G = \underset{X \in X'}{\Xi} [\Psi(X)] > \eta,$$

where $\Psi(X)$ is some statistical function over the result of recognizing the string X, Ξ – over X', X' is the set of recognized lines on the document, η – experimentally selected threshold.

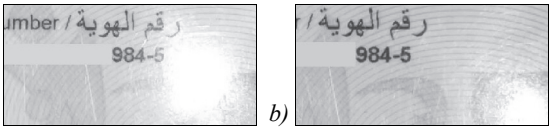
2.4. Image assessment for flares

We will consider a flare as a local spot with sharp edges and maximum possible brightness, that appears on laminated documents [24]. In this case, the value of image quality will be defined as the minimum among all estimates calculated as the ratio of the area of the attribute zone to the area of the flare that falls into the attribute zone. Thus, it is enough to overlap the area of one attribute with a flare to affect the quality assessment of the entire document. On the other hand, a flare is allowed if it does not interfere with the reading of the document attributes and does not affect the quality of recognition. However, it should be noted that if the flare is located in close proximity to the field, it is no longer possible to determine whether it covers part of the field or not without expert judgment (see fig. 2). Therefore, the requirement for the location of flares must be tighter: if the distance between the flare and the field in the direction of the text is less than one printed character, it is considered that the flare overlaps the field. In other words, the algorithm is applied to the widened bounding rectangles.

For segmentation of the document image into flare/non-flare, binarization is used with a threshold of T_{bin} (the choice of thresholds will be described in the section 3 of the algorithm settings). Further, for each zone of the text field on the image, bounded by the corresponding rectangle, the corresponding zone of the flare mask is considered:

1. for each pixel-width strip i of the field across the direction of the text, the ratio S_i of the area of the flare in this strip to the area of the strip, measured in the number of pixels, is calculated: $S_1, ..., S_{width}$;
2. the maximum among the calculated ratios is calculated $S_{max} = max\{S_i : i \in [1 \ldots width]\}$;
3. next, using the threshold T_{flare} flare score for the field is calculated FS_{field}:

$$\begin{cases} S_{max} < T_{flare} & \rightarrow FS_{field} = 1, \\ T_{flare} \le S_{max} & \rightarrow FS_{field} = 0. \end{cases}$$



a) *b)*

Fig. 2. The influence of the flare position on the ease of visual assessment: (a) the flare is close to the last character of the field and, probably, covers the field; (b) the flare is far from the field and does not interfere with the check

2.5. Calculating the sharpness score

To estimate the sharpness score, the modified algorithm described in [25] with the next steps was used:

1. get a one-channel image I_{1ch};
2. calculate gradient maps in two orthogonal directions – vertically G_H and horizontally G_V;
3. calculate given quantile T_q for each direction $Q_{T_q}(G_H)$ and $Q_{T_q}(G_V)$;
4. select the smallest of the obtained values $qsharp = \min(Q_{T_q}(G_H), Q_{T_q}(G_V))$.

The choice of the T_q threshold is described in the section on algorithm setup 3.4.

3. Algorithm setup

3.1. Algorithm parameters setting on a training dataset

Adjustment of parameters of each of the modules responsible for the classification of the frame for good/bad was carried out by constructing the ROC curves corresponding to each specific module, followed by comparing the areas under them (area under the curve, AUC) and choosing the thresholds corresponding to the optimal ratio FPR/TPR.

To configure the final algorithm, an internal closed dataset of Arabic ID-documents (identifier ARE-BO-01001 in the PRADO [26] database) was selected, containing 1535 images from 26 video clips. Each image was marked good/bad – there were 723 "bad" and 812 "good" in total. Images were marked as follows: if at least one field was unreadable on the document image after localizing the document area and correcting its projective distortions, the document was marked as "bad".

Also, to compare the proposed algorithm with other approaches, the [27] approach was chosen and compared with.

3.2. Recognizer confidence analysis

The following statistical functions Ψ were considered in the paper:

1. $\Psi(X) = mean(X) = \dfrac{1}{l_X} \sum_{i=1}^{l_X} p_i$;
2. $\Psi(X) = median(X) = p_{(1/2)l_X}$, $\forall X : p_1 < p_2 < \ldots < p_{l_X}$;

3. $\Psi(X) = min(X) = p_1$, $\forall X : p_1 < p_2 < \ldots < p_{X_l}$.

For Ξ similar functions were taken, but they were considered over $\Psi(X) \forall X \in X'$.

By enumerating all possible combinations of Ψ and Ξ, ROC curves were constructed to select the most appropriate classifier.

Constructed ROC curves for the internal dataset of Arabic IDs are shown in the fig. 3a. As can be seen from the graphs, the most qualitative classifier turned out to be $\Xi = mean$ with $\Psi = mean$. We should also pay special attention to the behavior of the ROC curve with $\Psi = median$ (see fig. 3b). It can be seen from the graphs that a change in η by 0.1 can lead to a sharp change in the FPR / TPR ratio, which is not very convenient when setting up the algorithm and choosing the optimal threshold.

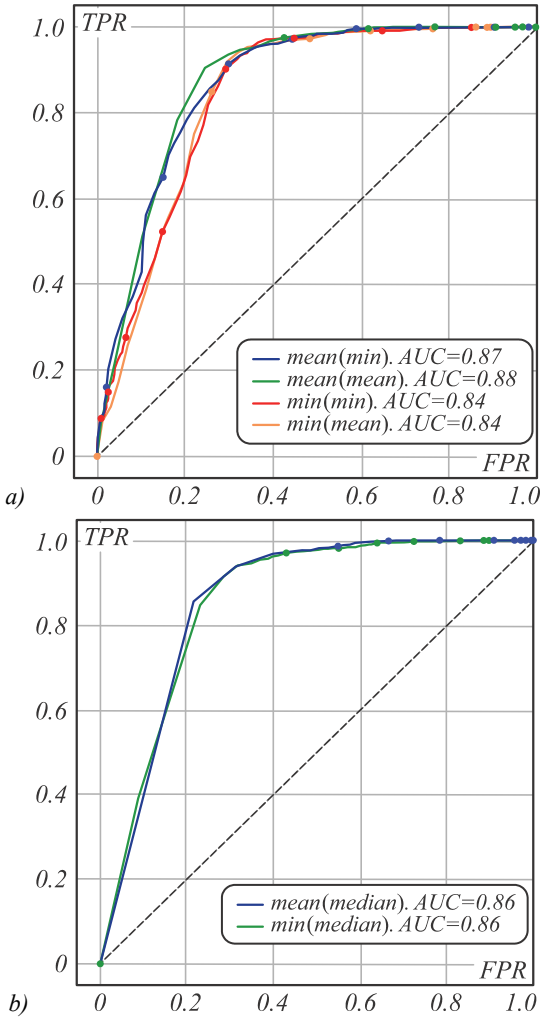


Fig. 3. ROC curves of six different methods of choosing "best" frame: 1) AUC[mean(min)] = 0.87, 2) AUC[mean(mean)] = 0.88, 3) AUC[min(min)] = 0.84, 4) AUC[min(mean)] = 0.84, 5) AUC[mean(median)] = 0.86, 6) AUC[min(median)] = 0.86. Points on graph corresponds to η, taken with step 0.1

Based on the results of the experiments, $\Xi = mean$, $\Psi = mean$, $\eta = 0.9$, were chosen to assess the frame quality (see. fig. 3*a*).

3.3. Image assessment for flares

To determine the binarization threshold T_{bin} for all binarization thresholds T_{bin} with a step of 5 (for the range of values of the original image [0, 255]), ROC curves were constructed for the cutoff thresholds T_{flare}. The fig. 4*a* shows four curves with the maximum area, the rest are omitted for clarity. As you can see from the graph, the maximum values are reached at T_{bin} thresholds equal to 235, 240, 245. The average value of 240 was chosen for the algorithm. The fig. 4*b* shows the ROC curve for this threshold separately.

For the flare estimation $T_{bin} = 240$ and $T_{flare} = 0.33$ were chosen.

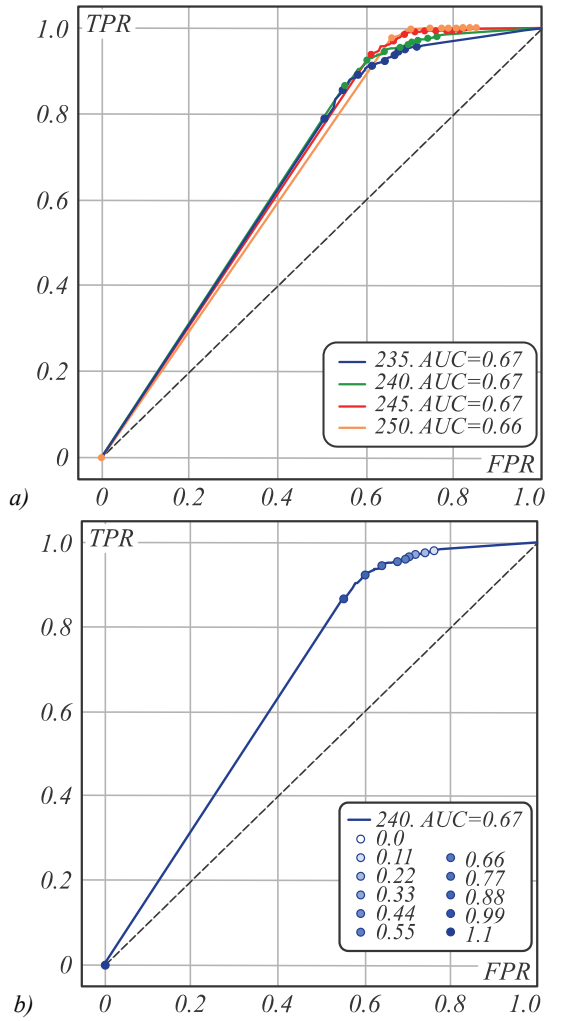


Fig. 4. (a) ROC curves of different binarization thresholds; (b) ROC curve for cutoff threshold for binarization threshold $T_{bin} = 240$

3.4. Sharpness analysis

To check the performance of the algorithm, the following was done. A sequence of frames [28] was taken, on which certain conditions of blur were reproduced: camera shift in different directions in combination with a slow shutter speed, focusing error, document capturing at an angle at low apertures (uneven sharpness across the

field frame). On each frame, the document was localized and projective distortions were corrected. Within the series, the images were sorted by the degree of sharpness: out of 25 frames, the first 15 images were the sharpest, then the sharpness gradually decreases with increasing frame number. The graphs of the dependence of the sharpness estimate on the frame number for different quantiles were built (fig. 5*b*).

Based on results of the experiments, the following conclusions were drawn:

1. The contrast of the image significantly affects the absolute value of the sharpness score. The image can be visually sharper, but due to the low contrast, have a lower sharpness score (for example, as in fig. 6). In this algorithm, the value of the sharpness score is not normalized in any way and is not tied to the image contrast. This is done because the proposed algorithm does not need the score of "absolute" sharpness: the sharpness scores of the images of the same document, taken in the same sequence under similar conditions, are compared with each other, that is, within the task under consideration, such big changes within the video stream are not assumed.
2. Flare of a relatively small size (less than 5 % of the frame area) does not affect the value of the score. Below, in fig. 5*a* graphs of the sharpness score for two series – with and without flares are shown. All images within the series are visually sharp, the series differ only in the presence/absence of flare. As you can see from the graphs, the range of values for the series is the same.
3. Visual assessment of sharpness is in better agreement with the calculated value of the sharpness score for the 95 % quantile than for other values of the quantile (fig. 5*b*).

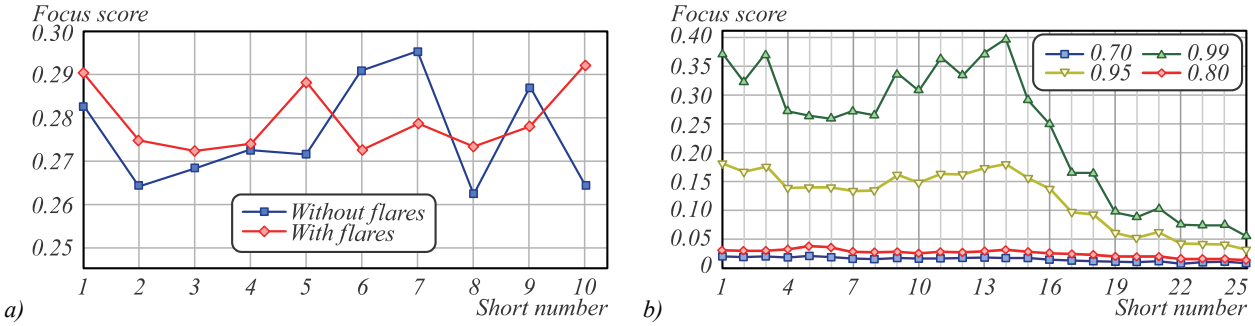


Fig. 5. (a) Influence of flare on the sharpness score (two series of 10 frames, shot in identical conditions and differing only in the presence/absence of flare); (b) Influence of the threshold (quantile) of gradients on the value of the sharpness score for the same series of 25 images

Sharpness score is 0.055 *Sharpness score is 0.052*

Fig. 6. A blurry but high contrast image (left) may have a higher sharpness score than a visually sharper but low-contrast image (right)

3.5. Comparison with other approaches

To compare the proposed algorithm with other approaches, a ROC curve was constructed for frame evaluation only by the sharpness assessment proposed in [27]. As you can see from the graph in fig. 7, the area under the ROC curve for evaluating the frame quality based on the sharpness assessment is less than the area for the recognizer confidence.

3.6. The result of setting the algorithm on the training dataset

Let us introduce terminology: True Positive (TP) is number of correct images on which zone was found correctly, True Negative (TN) is number of incorrect images which were correctly rejected, False Positive (FP) is number of correct images on which the target zone was found incorrectly or incorrect images that were accepted,

False Negative (FN) is number of correct images which were mistakenly rejected.

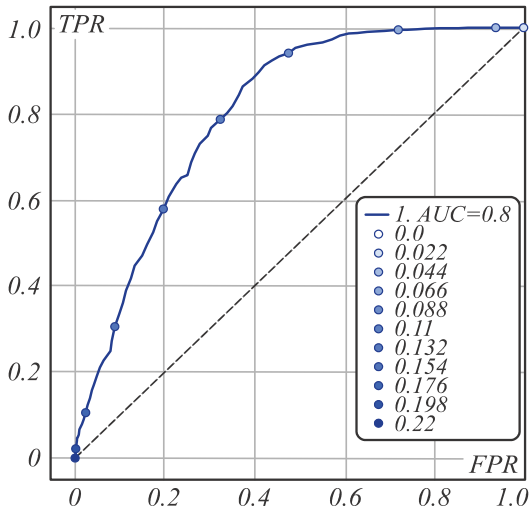


Fig. 7. ROC-curves for different sharpness thresholds, AUC[sharpness] = 0.8

The results on the described dataset after adjustment are presented in Tab. 1.

4. Experimental results

The fine-tuned algorithm was tested on the following documents from the MIDV-2019 reference dataset: German, Spanish, Slovak, Turkish and Czech id documents (folders "14_deu_id_new", "20_esp_id_new", "42_svk_id", "43_tur_id" and "10_cze_id"), as well Algerian passports (folder "18_dza_passport") and Italian driving licenses (folder "30_ita_drvlic"). Total of 256 images were used, of which 157 were "good" and 99 "bad". The images were marked up in the same way and could be downloaded from [28].

The results for this dataset are presented in Table 2. The result of algorithm with disabled separate parts of the algorithm is also presented.

An experiment was also carried out when a system configured to recognize a new type of German IDs was given "good" images of old German IDs as input. Even in the case of an erroneous linking of documents, the result of their recognition was ultimately assessed as "bad". Thus, even if "good" images of documents are submitted for recognition, but not of the type for which the recognition system is configured, the proposed algorithm will reject them.

Table 1. Results for different settings of the algorithm for the training dataset

Run type	Total #	Correct #	TP		TN		FP		FN		Precision %	Recall %	Accuracy %
			#	%	#	%	#	%	#	%			
NN confidence	1535	1195	796	51.9	399	26.0	324	21.1	16	10.0	71.1	98.0	77.9
Flares	1535	1077	708	46.1	369	24.0	354	23.1	104	6.8	66.7	87.2	70.2
All	1535	1334	753	49.1	581	37.9	142	9.3	59	3.8	84.1	92.7	86.9

Table 2. Results for different settings of the algorithm for the test dataset

Run type	Total #	Correct #	TP		TN		FP		FN		Precision %	Recall %	Accuracy %
			#	%	#	%	#	%	#	%			
NN confidence	256	220	143	55.9	77	30.1	22	8.6	14	5.5	86.7	91.1	85.9
Flares	256	203	138	53.9	65	25.4	34	13.3	19	7.4	80.2	87.9	79.3
Both	256	227	141	55.1	86	33.6	13	5.1	16	6.3	91.6	89.8	88.7

It should be noted that the average number of frames in a clip when recognizing from a video stream is 4–8 frames. With the obtained precision value of 91.6 for frame-by-frame evaluation, we can assume that the method allows to select the best frame in the video stream with high accuracy.

5. Further research

In further work the authors plan to improve the algorithm for flare detection: first, use adaptive flare threshold (this is especially needed for black–white document copies) [29], and second, use clustering approach for understanding if a flare was found or just a white part of the document.

It is also necessary to expand the amount of data - to increase the number of document types both for setting up the algorithm and for testing.

6. Conclusion

This work considered the problem of choosing the best frame and its assessment. The main factors for assessing the quality of the frame were the confidence of the neural network's response to the recognized text, as well as the presence of flares in the document image and the defocus degree of the frame. The choice of the best image is proposed to be considered as the problem of ranking images by quality.

The reference markup for a part of the MIDV-2019 open dataset has been prepared and made publicly available.

A practical method is proposed for choosing the best frame when recognizing a document in a video stream and the results of its application on the selected dataset are obtained: accuracy is 88.7%. Also, the proposed method can be considered suitable for verifying the correctness of the input data and settings of the recognition system: if the system receives a document unfamiliar to it, the recognition result of such a document will correspond to the system's low confidence in the response and all frames in the stream will be marked as "bad". This situation can serve as a signal to the user of the system about the occurrence of an emergency situation.

References

[1] Bulatov KB, Arlazarov VV, Chernov TS, Slavin OA, Nikolaev DP. Smart idreader: Document recognition in video stream. ICDAR 2017: 39-44. DOI: 10.1109/ICDAR.2017.347.

[2] Puybareau É, Géraud T. Real-time document detection in smartphone videos. 25th IEEE ICIP 2018: 1498-1502. DOI: 10.1109/ICIP.2018.8451533.

[3] Polevoy DV, Bulatov KB, Skoryukina NS, Chernov TS, Arlazarov VV, Sheshkus AV. Key aspects of document recognition using small digital cameras. Russian Foundation for Basic Research Journal 2016; 4: 97-108. DOI: 10.22204/2410-4639-2016-092-04-97-108.

[4] Bulatov K. Selecting optimal strategy for combining per-frame character recognition results in video stream. ITiVS 2017; 3: 45-55.

[5] Bulatov KB. A method to reduce errors of string recognition based on combination of several recognition results with per-character alternatives. Vestnik YuUrGU MMP 2019; 12(3): 74-88. DOI: 10.14529/mmp190307.

[6] Dodge S, Karam L. Understanding how image quality affects deep neural networks. Eighth International Conference on Quality of Multimedia Experience (QoMEX) 2016: 1-6.

[7] Bulatov K, Matalov D, Arlazarov VV. MIDV-2019: Challenges of the modern mobile-based document OCR. Proc SPIE 2020; 11433:114332N. DOI: 10.1117/12.2558438.

[8] Li H, Zhu F, Qiu J. CG-DIQA: No-reference document image quality assessment based on character gradient. 24th International Conference on Pattern Recognition (ICPR) 2018: 3622-3626. DOI: 10.1109/ICPR.2018.8545433.

[9] Obafemi-Ajayi T, Agam G. Character-based automated human perception quality assessment in document images. IEEE Transactions on Systems, Man, and Cybernetics (TSMC) 2012; 42: 584-595. DOI: 10.1109/TSMCA.2011.2170417.

[10] Nayef N, Ogier J-M. Metric-based no-reference quality assessment of heterogeneous document images. Proc SPIE 2015; 9402: 94020L.

[11] Cannon M, Hochberg J, Kelly P. Quality assessment and restoration of typewritten document images. Int J Doc Anal Recognit 1999; 2(2-3): 80-89.

[12] Alaei A, Conte D, Raveaux R. Document image quality assessment based on improved gradient magnitude similarity deviation. 13th ICDAR 2015: 176-180.

[13] Kang L, Ye P, Li Y, Doermann D. A deep learning approach to document image quality assessment. IEEE ICIP 2014: 2570-2574.

[14] Singh P, Vats E, Hast A. Learning surrogate models of document image quality metrics for automated document image processing. 13th IAPR International Workshop on Document Analysis Systems (DAS) 2018: 67-72.

[15] Zhan Y, Zhang R. No-reference image sharpness assessment based on maximum gradient and variability of gradients. IEEE Transactions on Multimedia2018; 20(7): 1796-1808.

[16] Marziliano P, Dufaux F, Winker S, Ebrahimi T. Perceptual blur and ringing metrics: Applications to jpeg2000. Signal Process Image Commun 2004; 19: 163-172.

[17] SmartDoc-QA: A dataset for quality assessment of smartphone captured document images – single and multiple distortions. 2015. Source: ⟨https://hal.archives-ouvertes.fr/hal-01319900⟩.

[18] Kumar J, Ye P, Doermann D. A dataset for quality assessment of camera captured document images. In Book: Iwamura M, Shafait F, eds. Camera-based document analysis and recognition. Cham: Springer International Publishing; 2014: 113-125.

[19] Chabchoub F, Kessentini Y, Kanoun S, Eglin V, Lebourgeois F. SmartATID: A mobile captured arabic text images dataset for multi-purpose recognition tasks. 15th ICFHR 2016: 120-125.

[20] Skoryukina N, Shemiakina J, Arlazarov VL, Faradjev I. Document localization algorithms based on feature points and straight lines. Proc SPIE 2018; 10696: 106961H. DOI: 10.1117/12.2311478.

[21] Shemyakina J, Zhukovskiy A, Nikolaev D. The method for homogrpaphy estimation between two planes based on lines and points. Proc SPIE 2018; 10696: 106961G. DOI: 10.1117/12.2310111.

[22] Povolotskiy MA, Tropin DV, Chernov TS, Savelev BI. Dynamic programming approach to textual structured objects segmentation in images. ITiVS 2019; 69(3): 66-78. DOI: 10.14357/20718632190306.

[23] Arlazarov VV, Bulatov KB, Karpenko SM. Recognition confidence determining method for embossed symbol recognition problem [In Russian]. Trudy ISA RAN 2013; 63(3): 117-122.

[24] Bulatov KB, Ilin DA, Polevoy DV, Chernyshova YS. Recognition problems of machine-readable zones using small-format digital cameras of mobile devices [In Russian]. Trudy ISA RAN 2015; 65(3): 85-93.

[25] Bulatov K, Polevoy D. Reducing overconfidence in neural networks by dynamic variation of recognizer relevance. ECMS 2015: 488-491. DOI: 10.7148/2015-0488.

[26] Council of the european union. 2020. Source: ⟨https://www.consilium.europa.eu/prado/en/search-by-document-country.html⟩.

[27] Chernov TS, Razumnuy NP, Kozharinov AS, Nikolaev DP, Arlazarov VV. Image quality assessment for video stream recognition systems. Proc SPIE 2018; 10696: 106961U. DOI: 10.1117/12.2309628.

[28] Article experimental data. 2020. Source: ⟨ftp://vis.iitp.ru/best_frame_article_data/⟩.

[29] Lange H. Automatic glare removal in reflectance imagery of the uterine cervix. Proc SPIE 2005: 5747: 2183-2192. DOI: 10.1117/12.596012.

Authors' information

Mikhail Aleksandrovich Aliev, graduated from Moscow Institute of Steel and Alloys in 2008, majoring in Applied Mathematics. He works as a programmer at Smart Engines, and a researcher in FRC "Computer Science and Control" of RAS. Research interests are computer vision, Hough transform, pattern recognition, image processing.

Irina Andreevna Kunina, a research assistant at the IITP RAS. She received the S.S. and M.S. degrees correspondingly from NUST MISiS in 2014 and from Moscow Institute of Physics and Technology (State University), Moscow, Russia in 2016. Research interests are image processing, computer vision, calibration of visual systems.
E-mail: *kunina@iitp.ru* .

The information about author **Alan Valentinovich Kazbekov** you can find on page 90 of this issue.

Vladimir Lvovich Arlazarov (b. 1939), Dr. Sc., corresponding member of the Russian Academy of Sciences, graduated from Lomonosov Moscow State University in 1961. Currently he works as head of sector 9 at FRC CSC RAS. Research interests are machine learning, computer vision and artificial intelligence.
E-mail: *vladimir.arlazarov@smartengines.com* .

Received September 17, 2020. The final version – December 31, 2020.