# Weighted combination of per-frame recognition results for text recognition in a video stream

*O. Petrova [1,2], K. Bulatov [1,2,3], V.V. Arlazarov [1,2], V.L. Arlazarov [1,2,3]*
*[1] FRC CSC RAS, Moscow, Russia,*
*[2] Smart Engines Service LLC, Moscow, Russia,*
*[3] Moscow Institute of Physics and Technology (State University), Moscow, Russia*

***Abstract***

The scope of uses of automated document recognition has extended and as a result, recognition techniques that do not require specialized equipment have become more relevant. Among such techniques, document recognition using mobile devices is of interest. However, it is not always possible to ensure controlled capturing conditions and, consequentially, high quality of input images. Unlike specialized scanners, mobile cameras allow using a video stream as an input, thus obtaining several images of the recognized object, captured with various characteristics. In this case, a problem of combining the information from multiple input frames arises. In this paper, we propose a weighing model for the process of combining the per-frame recognition results, two approaches to the weighted combination of the text recognition results, and two weighing criteria. The effectiveness of the proposed approaches is tested using datasets of identity documents captured with a mobile device camera in different conditions, including perspective distortion of the document image and low lighting conditions. The experimental results show that the weighting combination can improve the text recognition result quality in the video stream, and the per-character weighting method with input image focus estimation as a base criterion allows one to achieve the best results on the datasets analyzed.

*Keywords*: mobile OCR, video stream, anytime algorithms, weighted combination, ensemble methods.

*Citation*: Petrova O, Bulatov K, Arlazarov VV, Arlazarov VL. Weighted combination of per-frame recognition results for text recognition in a video stream. Computer Optics 2021, 45(1): 77-89. DOI: 10.18287/2412-6179-CO-795.

*Acknowledgements*: This work is partially supported by the Russian Foundation for Basic Research (projects 17-29-03236 and 18-07-01387).

## *Introduction*

### *Document recognition in uncontrolled conditions*

Nowadays text object recognition is widely used not only in government and business processes but also in everyday life [1, 2]. One of the first problems in which optical character recognition (OCR) technologies found their application was automatic data entry. A few decades ago such problems required special equipment, knowledge of the used fonts, scanned image characteristics, etc. But today the scope of application of such technologies has expanded, and document recognition is increasingly carried out in uncontrolled capturing conditions. For instance, automatic personal data entry can be done without the use of specialized equipment, for example, when opening a bank account using a mobile application or when buying and registering SIM cards in a self-service mode [3]. Apart from the automatic input of personal data, text object recognition is essential in electronic document management systems, allows saving time, reducing expenses, and saving natural resources [4]. The development of hardware, such as personal mobile devices, has made it possible to expand the applicability of OCR technologies for recognizing text in natural scenes and use these technologies in such cases as driver assistance systems [5], assistance for people with visual impairments [6], online translators [7], government photo and video recording systems [8, 9], and many more. Along with the applicability of the text recognition technologies, the requirements for the quality and reliability of recognition results are increasing [10]. Besides, more and more cases require the possibility to use “improvised means” for the recognition, with input images captured using a smartphone camera or a web-camera [11, 12].

An important area of OCR technologies application is document recognition [13]. The translation of paper documents into electronic form allows quickly and conveniently to process and index them. A separate important subsection of document recognition is the identity documents recognition [14]. These technologies have found application when filling out various registration forms [15], identifying a person in security systems [10], filling out personal and sensitive information [16, 17], etc. In many of these applications recognition errors are extremely costly. Improving the recognition quality of identity documents captured using mobile devices is an important topic, and this paper will primarily consider identity documents as the target object for recognition.

Unlike images obtained with special scanners, for which it is possible to set up the lighting conditions be-

forehand, ensure the immobility of the recognized object and the recording matrix, etc., the frames received from a mobile camera can have low quality, contain highlights on the reflective surfaces of the object, be out of focus or blurry, target object can have strong projective distortions [13, 18, 19]. Especially often these difficulties arise when capture is performed in uncontrolled conditions [20]. Lighting problems can decrease text image quality and make it difficult to recognize. Uneven illumination can lead to sharp differences in brightness and to the appearance of false borders, which complicate text per-character segmentation [21]. To avoid highlights or shadows on the surface of the recognized document or the occlusion of recognized text fields by holograms and other security elements, the user can rotate the document during capture, thus some projective distortions of text objects may occur [22, 23]. This, as well as a complicated, cluttered, non-homogeneous background, can complicate the localization of the document in the image [24, 25]. Fig. 1 shows examples of document images with various types of distortions.


*a)* *b)* *c)*

*Fig. 1. Examples of document images: (a) with projective distortion; (b) with highlight; (c) defocused*

Despite the additional difficulties associated with the usage of mobile devices for recognition, the advantage of mobile cameras in comparison with scanners is that they allow to get not a single image of the recognized object, but a video stream, which makes it possible to get frames captured with different illumination, at different angles, with different focus characteristics, thus allowing to reduce sporadic errors of an OCR-system [16].

## *Scope*

After obtaining an image of a recognized document, the recognition process usually involves such stages as preprocessing input images, text fields localization, segmenting string image into characters, which then are submitted for recognition, post-processing of recognition results. Some of these steps may be absent. For example, in recognition systems where the text is analyzed in an end-to-end way [26], per-character segmentation is not required.

The purpose of input images preprocessing is to improve the accuracy of text detection and recognition. This stage includes, for example, contrasting, colored background removal [27], binarization [28], as well as removing image defects (noise, glare, overlaying holograms) using various types of filtering and the use of morphological operations [29].

The stage of document localization involves the precise detection of the document boundaries in an image. If the document has a fixed layout of fields, the document localization allows us to simplify and increase the accuracy of the fields localization and, as a consequence, text recognition quality. A common approach to document localization problem is to find the vanishing points using the straight lines present in an image (for example, document edges, baselines of text fields). In conditions of weak projective distortions, an approach based on the generalized Viola-Jones method [30] is also applicable. An approach based on the key points search is more robust to various kinds of image distortions [31]. Methods based on the fast Hough transform [14, 32], the RANSAC algorithm, or the least-squares method are used to search for straight lines in an image or to refine the search for feature points.

Algorithms for segmentation of the found fields into individual characters can be based both on the analysis of the horizontal projection [32], and use character candidates recognition methods with dynamic programming methods to determine the optimal set of cuts [33].

Approaches to the classification of individual symbols include pattern matching algorithms [34], support vector machine (SVM) based algorithms [32], artificial neural networks (ANN), and much more.

## *Text recognition in a video stream*

When using a video stream as input data for recognition, the problem arises of choosing methods for combining information obtained from different frames of a video sequence. The methods of combining per-frame information can be divided into two groups: methods, relying on image combination to obtain a higher quality object representation, and methods of combining the extracted text recognition results. The first group includes methods for selecting the most informative frame [35, 36], "super-resolution" methods that create a higher quality image based on several low-resolution frames [37 – 39], methods for tracking and combining images of a recognized object on a sequence of frames [40, 41], methods of blur compensation by replacing blurred areas in one frame with their clearer counterparts taken from other frames or using deep learning methods [42]. Also, for a better reconstruction of a recognized document image, it is possible to use the data obtained from various sensors of the recording device, such as, for example, an accelerometer or a gyroscope. However, for modern mobile devices, the error in their measurements can be quite significant and prevent using this data for image reconstruction [43]. The second group of methods involves combining the results of individual image recognition. The methods of the first group, which involve combination on the level of input images, could be time-consuming, sensitive to geometrical distortions between frames, and poorly scalable with regards to video sequences of arbitrary lengths. Thus, in this paper, we will consider methods of the second group,

i.e. the combination of the recognition results obtained for the individual frames.

A distinctive feature of text object recognition is that a text string is a composite object, i.e. consisting of multiple components (characters). Text recognition algorithms that analyze the text in an end-to-end way are more applicable for recognizing strings that are difficult to segment into characters (such as text written in Arabic script) or for recognizing large texts, where the majority of the words occur frequently, and there are fewer limitations to the processing speed [2, 44]. In a more general case, in particular, with regards to identity document recognition systems, the text recognition result is considered as a concatenation of character classification results. Such representation implies a preliminary text per-character segmentation procedure, i.e. the process of splitting the image of a text string into the images of special characters. With such text representation, the model of per-frame recognition results combination has to deal with strings obtained for different frames, which in the case of segmentation errors have different lengths, and the combination algorithm needs to be able to account for that. One of the combination approaches which allows variable-length input strings is the ROVER method (Recognizer Output Voting Error Reduction) [45]. This method was originally created to improve the quality of speech recognition by combining the recognition results received from different systems. This method includes two stages. At the first stage, all the combined recognition results are aligned by inserting an empty character in an optimal way and combined into a single transition network. In the second stage, using the voting procedure, the best recognition result for each element of the composite object is selected. The voting procedure can be considered as the task of combining classifiers and such classifier ensemble models as the rules of sum, product, maximum, median, etc. [46 – 48] can be used as an extension of the voting procedure in ROVER. Thus, using the ROVER method to combine the recognition results obtained from several frames allows producing correct recognition results even if in some frames the text field was incorrectly segmented into characters.

The combination algorithms for per-frame text recognition could be further improved by introducing weights of the input results. If a predictor could be constructed such that it would be possible to estimate the validity of the recognition result, such predictor can be used for weighting the per-frame results in the combination. This could include zero-valued weights for "rejecting" some of the per-frame results which could spoil the overall combined result, or select and combine only a few "best" results. The question, however, arises – which predictor to use to maximize the quality of the final result. The goal of this paper is to consider the weighting problem, investigate the functions of the input images or input recognition results which could be used as the quality predictors, and to propose the model and methods for weighted per-frame text recognition results combination.

The paper is structured as follows. Section 1 provides a detailed description of the difficulties that can arise at different stages of document recognition. Section 2 sets out the problem statement for the per-frame recognition results combination. In sections 3 and 4 a general weighing model and weighing criteria are proposed, respectively. In section 5 an approach to a weighted combination that takes into account the peculiarities of individual characters recognition is described. Section 6 describes the performed experimental evaluation. Section 7 provides an analysis of the obtained results. Finally, conclusions are drawn and the possible topics for future work are proposed in Section 8.

### *1. Error analysis*

Text field recognition errors can be caused both by the physical difficulty to read the entire field (when the data cannot be fully recognized even by a human) and by errors that have arisen at various stages of document recognition. Figure 2 shows examples of documents, with some parts of data which cannot be read. The first type of reasons is the occlusion of a text by a highlight or a holographic security element. If a highlight or a hologram appears as a bright spot that completely occludes a part of the text, then most likely this part will be completely discarded during text localization. If the highlight occludes the character partially (for example, if it is the edge of the highlighted region), then this could lead to a single character classification error – the character becomes similar to another (for example, partially occluded "B" or "8" may become similar to "3"). If this problem is not present on all frames of the video stream or the occluded areas are different in different frames, then combining the recognition results of individual frames can allow you to get the correct final result, even if there are no correctly recognized frames. Thus, the problem of combining text strings of different lengths arises.


*a)* *b)*

*Fig. 2. Example of images when part of the data is difficult to read: (a) with defocus, (b) with an occlusion by a highlight*

Image defocus or blur can significantly complicate the text segmentation into the separate characters and the characters classification, to the point that the text becomes unreadable. An example of identity document text strings with per-character segmentation errors is shown in Fig. 3. In contrast to the occlusion by highlights glare and holograms, blur and defocus often affect not the individual characters, but the entire text strings. Fig. 4 shows an example of extracted document text field images, most of which are unreadable due to defocus, and even when one

frame (frame number 3) was correctly recognized, the combination result became spoiled by irrelevant recognition results of low-quality frames (see tab. 1).

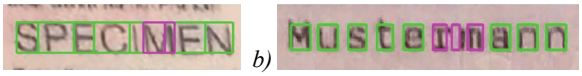

*Fig. 3. Example of per-character segmentation errors. Recognition results: (a) "SPECINAEN"; (b) "MUSTEI NANN"*

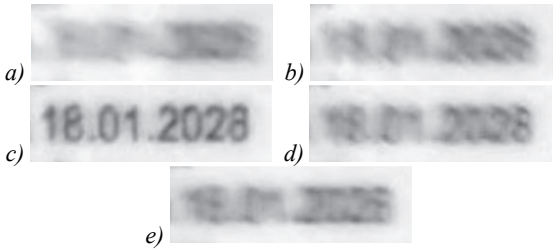

*Fig. 4. Field images from a video clip in which low-quality frame recognition results spoil the combined recognition result*

*Table 1. Example of a video clip in which low-quality frame recognition results spoil the combined recognition result*

| Frame number | Image | Frame recognition result | Combined recognition result |
|---|---|---|---|
| 1 | Fig. 4*a* | 33 | 33 |
| 2 | Fig. 4*b* | 8283-38 | 2838 |
| 3 | Fig. 4*c* | **18.01.2028** | **80.3028** |
| 4 | Fig. 4*d* | 88228 | 88328 |
| 5 | Fig. 4*e* | 8 – 2 4 | 88228 |

Recognition errors can be caused not only by an incorrect classification of individual characters but also by errors of document localization or determining its orientation in an image, for example, due to a complicated background. If the localization of the text fields of the document is based on the assumption of a fixed geometric layout [49], even a slight deviation of the found document boundaries from their actual position can lead to a noticeable distortion of local parts of the document and incorrect localization or cropping of the text field (Fig. 5). Even if the document fields are adequately found, then the incorrectly found document boundaries quadrangle leads to text distortions and, as a result, errors at the further stages of recognition. Serious errors in document search lead to incorrect localization of text fields and the appearance of recognition results that are far from the true values.

Incorrect classification of correctly segmented characters may be caused, for example, due to the similarity of some characters and the poor quality of the input images, as well as complicated document background. Examples of misclassification of individual characters are shown in Fig. 6. In this case, if recognition errors are sporadic (i.e. are not present on all frames), combining the recognition results of individual frames can also improve the recognition quality due to the fact that correct recognition results for individual parts of the text field can be obtained from different frames.

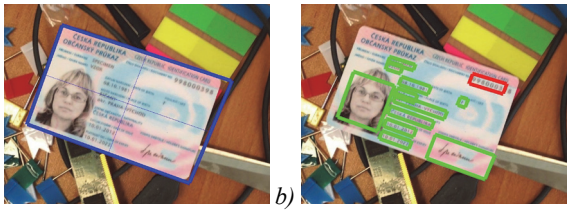

*Fig. 5. Example of a small document localization error (a) leading to incorrect text field localization (b)*

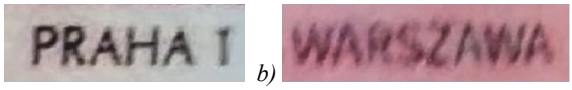

*Fig. 6. Example of character classification error. Recognition results: (a) "PRAHA I"; (b) "WARSZ/WA"*

### *2. Problem statement*

In this section, we will present a problem statement for the weighted text recognition results combination.

When talking about text recognition results, we will mean the recognition results of a text string composed of characters from a fixed finite alphabet. The recognition result $x$ of an individual character may be viewed as a sequence of membership estimations for each character class and represented as a vector:

$$x = (x_1, x_2, ..., x_K) \in [0.0, 1.0]^K, \quad \sum_{k=1}^{K} x_k = 1, \tag{1}$$

where $K$ is the number of character classes (i.e. the size of the alphabet), $x_k$ – membership estimation for the class $k$, which can take a real value in the range from 0.0 to 1.0. The recognition result $X$ of a text string can be represented as a matrix:

$$X = (x_{jk}) \in [0.0, 1.0]^{M \times K}, \quad \forall j : \sum_{k=1}^{K} x_{jk} = 1, \tag{2}$$

where $M$ is the length of the recognized string (in terms of the number of characters), $x_{jk}$ – membership estimation for the $j$-th symbol with regards to the $k$-th class. Such recognition result representation is commonly used at the text recognition post-processing stage to construct algorithms for correcting recognition errors based on a-prior information about the syntactic and semantic structure of the recognized data [50].

If the recognition result is represented as a matrix of membership estimations, the ROVER method can be generalized [51] as follows. Firstly, the set of possible classes is expanded by the "empty" class λ (with a class number $k = 0$), such that its membership estimations for all characters of a frame recognition results will have zero value. In terms of the matrix, this corresponds to adding a zero-valued column at the beginning of the matrix. The distance between the recognition results of two characters $x^1$ and $x^2$ can be determined as:

$$\rho(x^1, x^2) = \frac{1}{2} \sum_{k=0}^{K} \left| x_k^1 - x_k^2 \right|. \tag{3}$$

Using this metric, the two text string recognition results can be aligned with each other so as to minimize the total pairwise distance between characters. At the voting stage of the ROVER method, membership estimations for the combined recognition result $r$ of matching characters of the strings with aligned characters can be calculated as the weighted average of membership estimations for $x^1$ and $x^2$:

$$r = (r_k) \in [0.0, 1.0]^{K+1}, \quad \forall k : r_k = \frac{x_k^1 \cdot w(x^1) + x_k^2 \cdot w(x^2)}{w(x^1) + w(x^2)}, \quad (4)$$

where $w(x^1)$ and $w(x^2)$ – weights with which the recognition results are included in the combination.

Consider the problem of a weighted combination of text field recognition results in a video stream as follows: an object $\overline{X}$ is recognized in a sequence of $N$ frames $I_1(\overline{X}), I_2(\overline{X}), ..., I_N(\overline{X})$, $I_i \in I$, where $I$ is a set of frames, $X_i \in \chi$ is a recognition result of $\overline{X}$ on a frame $I_i(\overline{X})$, $\chi$ is a set of all possible text string recognition results. Let $X^* \in \chi$ be the correct value of the text string. Let us define the quality of a frame $I_i(\overline{X})$ recognition as a distance $\rho(X_i, X^*)$ according to some pre-defined metric $\rho : \chi \times \chi \to \mathbb{R}_0^+$. The base weighing function $w : \mathrm{I} \times \chi \to \mathbb{R}_0^+$ assigns to the pair $(I_i(\overline{X}), X_i)$ a posterior quality assessment of recognition result $X_i$ for the frame $I_i(\overline{X})$. The combination function

$$R^{(N)}(X_1, X_2, ..., X_N, w_1, w_2, ..., w_N) : \chi^N \times (\mathbb{R}_0^+)^N \to \chi$$

takes as an input the recognition results of the sequence of frames and their weights and outputs the combined recognition result (which will be treated as the recognition result of the whole video sequence). With a fixed sequence of frames $I_1(\overline{X}), I_2(\overline{X}), ..., I_N(\overline{X})$ and a fixed combination function $R^{(N)}$, the task is to assign weights $w_1, w_2, ..., w_N$ to the frame recognition results $X_1, X_2, ..., X_N$ such that to minimize the expected distance $\rho(R^{(N)}(X_1, X_2, ..., X_N, w_1, w_2, ..., w_N), X^*)$.

### *3. Weighting model*

Low-quality recognition result of the individual frame can decrease the quality of the combined result. Therefore, one of the questions is what strategy is better – a weighted combination of several recognition results or selection of the single best result. This issue was considered in [52] with regard to individual character recognition. In the context of text field recognition on identity documents and bank cards in a video sequence, it was shown that in the absence of localization and segmentation errors, i.e. when a document was found incorrectly or text fields were incorrectly split into characters, the strategy of combining several of the most "competent" classifiers according to the product rule (the product of membership estimations for each class) or a voting procedure shows the best result. However, it is not clear whether such strategy is applicable in the case of a full-text string recognition problem. Unlike the individual characters combination problem, in the case of text strings recognition, a correctly recognized single frame could absent, but, at the same time, combining the recognition results can give the correct result (for example, in the case of a "sliding" highlight). Therefore, even in the absence of localization and segmentation errors, taking into account the recognition results from all frames may turn out to be essential.

To generalize and unify the combination approach and the selection of the best frame, the weighting model can be specified as follows: Let us set the order $\pi \in S_N$ ($S_N$ being the set of permutations with the length $N$) of the recognition results according to a non-decreasing value of the basic weighting function:

$$\pi(i) < \pi(j) \Leftrightarrow w(I_i(\overline{X}), X_i) \geq w(I_j(\overline{X}), X_j)$$

and the cut-off threshold $t \in \{1, ..., N\}$. Then the weights can be defined with the following function:

$$w_i^{(t)} = \begin{cases} w(I_i(\overline{X}), X_i), & \text{if } \pi(i) \leq t, \\ 0, & \text{if } \pi(i) > t. \end{cases} \quad (5)$$

This weighting model can be used to generalized both the selection of the single best result according to the quality predictor $w$ (if the threshold value is $t=1$), and the full weighted combination of all input samples (with $t=N$), as well as the weighted combination of a few best frame results. Given such weighting model, the task is now to determine the best combination strategy, the best weighting criterion $w$ and the threshold $t$.

### *4. Weighting criteria*

In this paper, we considered two weighing criteria. The first is a focus estimation $F(I_i(\overline{X}))$, calculated using an algorithm proposed in [53]. This criterion was also used to control the input frame quality in video stream document recognition systems [54]. First, the values of the image gradients are calculated in four directions (vertical, horizontal, and two diagonals):

$$\begin{aligned} G_{\mathrm{r,c}}^{V}(I_i(\overline{X})) &= \left|\mathrm{I}_{r+1,\mathrm{c}} - \mathrm{I}_{\mathrm{r,c}}\right|, \\ G_{\mathrm{r,c}}^{H}(I_i(\overline{X})) &= \left|\mathrm{I}_{\mathrm{r,c+1}} - \mathrm{I}_{\mathrm{r,c}}\right|, \\ G_{\mathrm{r,c}}^{D_1}(I_i(\overline{X})) &= (1/\sqrt{2})\left|\mathrm{I}_{r+1,\mathrm{c+1}} - \mathrm{I}_{\mathrm{r,c}}\right|, \\ G_{\mathrm{r,c}}^{D_2}(I_i(\overline{X})) &= (1/\sqrt{2})\left|\mathrm{I}_{\mathrm{r,c+1}} - \mathrm{I}_{r+1,\mathrm{c}}\right|, \end{aligned} \quad (6)$$

where $\mathrm{I}_{\mathrm{r,c}}$ is the intensity value of the pixel with coordinates $(r, c)$ of the image $I_i(\overline{X})$.

The image focus estimation is then calculated as the minimum 0.95-quantile of the obtained derivatives:

$$\begin{aligned} F(I_i(\overline{X})) = \min\{ & q(G^V(I_i(\overline{X}))), q(G^H(I_i(\overline{X}))), \\ & q(G^{D_1}(I_i(\overline{X}))), q(G^{D_2}(I_i(\overline{X})))\}, \end{aligned} \quad (7)$$

where $q(G)$ is a 0.95-quantile of the gradient image $G$.

It was assumed that the weighting method based on the text field image focus estimation will allow to reduce the significance of frames in which the image of the recognized field is of poor quality due to defocus, smears,

and blur, that can lead to errors in text localization and per-character segmentation, as well as low quality of individual characters recognition.

The second weighting criterion used for recognition quality estimation was a-posteriori recognition confidence $Q(X)$, where $X$ is the text recognition result (2). The text string recognition result confidence value is calculated as a minimal value of the highest membership estimation across all string character classification results:

$$Q(X) = \min_{j=1}^{M} \left\{ \max_{k=1}^{K} x_{jk} \right\}. \quad (8)$$

This weighting criterion was based on the assumption that with correct recognition of the text field, the "best" membership estimations will have a higher value than with an inappropriate recognition when the classifier cannot determine the recognized character with high confidence.

### 5. Per-character weighting

Proposed weighting model and weighting criteria were evaluated in [55] for the problem of per-frame combining of identity documents text fields recognition results. It has been shown that the weighted combination actually improves the recognition quality. However since the result of the text string recognition depends on the results of individual characters classification, and, in some cases, the quality of the character images in the same frame can vary greatly (for example, in the case of highlight, partial defocus, mechanical occlusions of a part of the text string, etc.) or weighting criterion may not always correctly represent the quality of recognition of individual characters (for example, the confidence value criterion in the case of incorrect per-character segmentation), an additional question arises – how correct is it to assign the combination weights based on the characteristics calculated over the entire text string. Therefore it is sensible to introduce a weighting model considering each individual character with its own weight.

The ROVER method in this case needs to be modified. Before adding the recognition result $X_i$ to the combination result, weights $w_1^i, w_2^i, ..., w_{M_i}^i$ have to be assigned to each character component $x_1^i, x_2^i, ..., x_{M_i}^i$ of the string recognition result $X_i$. For the weighting criterion based on focus estimation, the character weight is calculated as focus estimation of the image of character submitted to the recognition module. For the confidence value weighting criterion, the character weight can be assigned simply as the highest membership estimation of this character. For the "empty" character λ the weight coincides with the weight with which the text string recognition result as a whole is included in the combined result, i.e. $F(I_i(\bar{X}))$ or $Q(X_i)$.

In the first step, the text string recognition result $X_1$ of the first frame is stored as the combined result $R$, with the corresponding combined character weights $w(r_1), w(r_2), ..., w(r_M)$ and the full result weight $W_R$ taken from the weights of the first result. At the next steps, when adding the recognition result $X_i$ with weight to the combined result $R$, alignment is performed so as to minimize the total pairwise distance between characters, calculated according to the character recognition results distance function (3). After the combined result and the new frame result are aligned, their characters are combined according to the combination rule (4).

If the character $x_j^i$ of the added recognition result was not matched with any combined result character during alignment, then it is combined with an empty symbol λ with weight $W_R$. If the character $r_j$ of the combined result did not match with any character of the per-frame result, then it is combined with an empty symbol λ with weight $w_i$. When combining two characters with weights $W_1$ and $W_2$ the weight of the combination character result is determined as $W_1+W_2$. After combining all text string recognition results, the updated weight of the new combined result is calculated as $W_R+w_i$. The diagram of the modified ROVER algorithm is shown in Fig. 7.

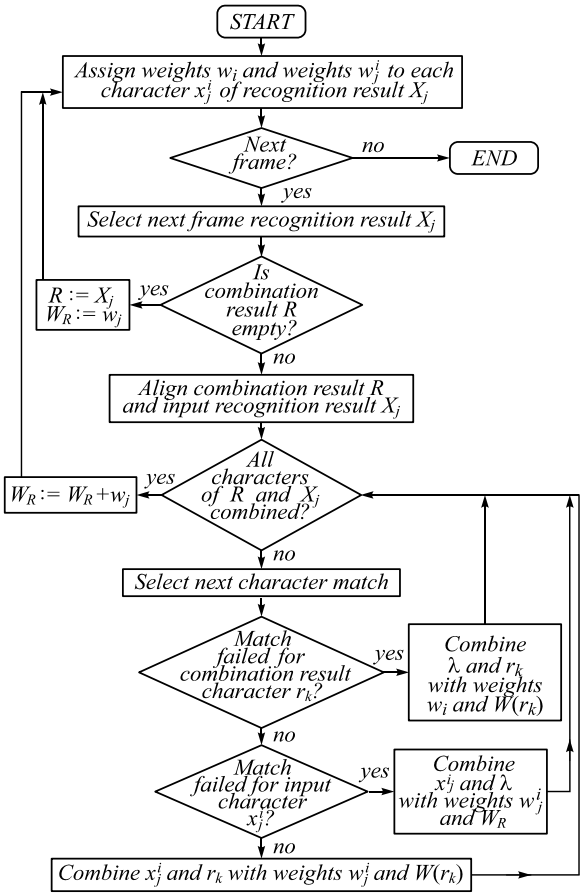


*Fig. 7. Diagram of combining text recognition results with per-character weighting*

### 6. Experimental evaluation

#### Full string weighting

After the definition of the weighting model, weighting criteria, and the algorithm for the combination of text string recognition results with per-character weighting, we can proceed to the experimental evaluation.

Previous work [55] described experiments performed on the MIDV-500 [16] dataset. This dataset contains 500 video clips of identity documents captured with mobile cameras without strong distortion. However, it seems important to evaluate the quality of the proposed method and weighing criteria in various, including challenging conditions. Therefore, experiments were also performed on the MIDV-2019 [22], which contains 200 video clips of identity documents. A feature of the MIDV-2019 dataset is that video clips were captured in low lighting conditions (subset MIDV-2019-L) and with strong projective distortions of document image (subset MIDV-2019-D). Each video clip contains 30 frames, but only the frames on which the document is fully visible were considered; if the resulting clip length had fewer than 30 frames, the frames were repeated in a loop, following the experimental procedure set up in other papers using this dataset [51]. Four field types were analyzed: document numbers, numeric dates, Latin name components, and machine-readable zone lines. The fields were recognized using the method described in [56]. The comparison with the correct text field values was case-insensitive, and the letter "O" was considered identical to the digit "0". Normalized Generalized Levenshtein Distance [57] was used as a metric function for the set of text string recognition results.

On the first stage for each basic weighting function we considered five weighted combination strategies: combination without weighting (i.e. using a constant value as a basic weighting function and threshold parameter $t=N$), choosing the single best result (threshold parameter $t=1$), weighted combinations of the 3 best (threshold parameter $t=3$), of the best 50 % (threshold parameter $t=N/2$), and of all frames (threshold parameter $t=N$).

Fig. 8 shows the rate of combined text recognition result error decrease after the addition of new per-frame results for the various approaches to weighted combination using a focus estimation (7) and a recognition result confidence value (8), as measured on all analyzed field groups of the MIDV-500 dataset. Such plotted rates can be viewed as performance profiles [58] for the process of text recognition in a video stream as an anytime-algorithm (i.e. the algorithm with results increasing their quality over time).

It can be seen that weighted integration improves the quality of recognition, and in the case of using the focus estimation as a weighting criterion, noticeable improvements are achieved regardless of the number of combined frames. When using the confidence value of the recognition result for weighting, the selection of the few best frames to combine does not improve the recognition, in particular at the later stages of the process, i.e. with a higher number of combined per-frame results.

Fig. 9 and 10 demonstrate similar performance profiles for recognition results on MIDV-2019 dataset, for subsets with low lighting conditions and with strong projective distortions respectively.

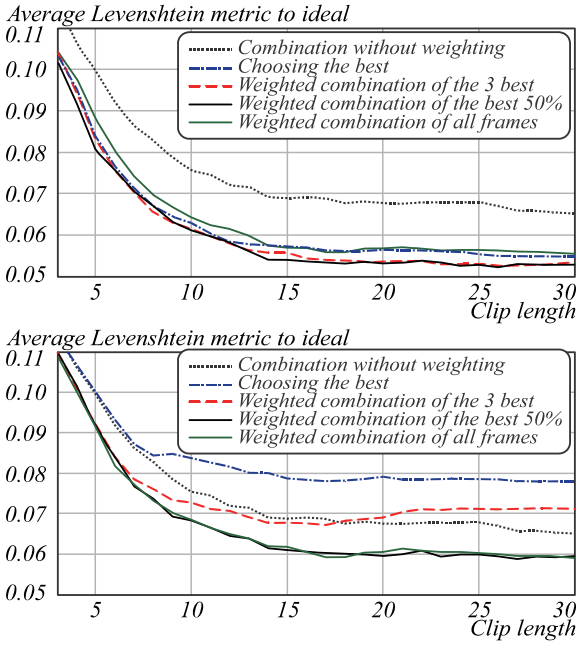


*Fig. 8. Performance profiles for weighted combination based on focus estimation (top) and confidence value (bottom) for MIDV-500 dataset*

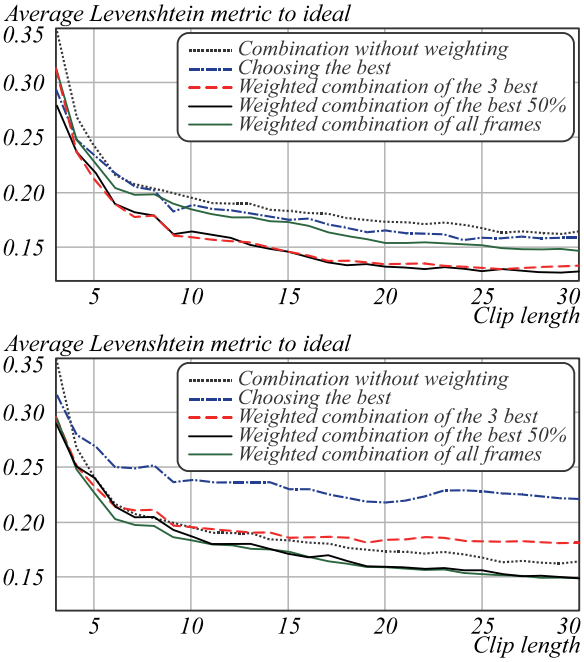


*Fig. 9. Performance profiles for weighted combination based on focus estimation (top) and confidence value (bottom) for MIDV-2019-L dataset*

According to the experimental results, it can be seen that on both datasets, weighting according to focus estimation criterion allows achieving a higher recognition quality than when using confidence value as a weighting criterion. It can also be noticed that, on average, the best result is achieved by a weighted combination of the best 50 % of frames. Figures 11, 12, and 13 show comparative profiles for combining the best 50 % frames for different combining strategies.

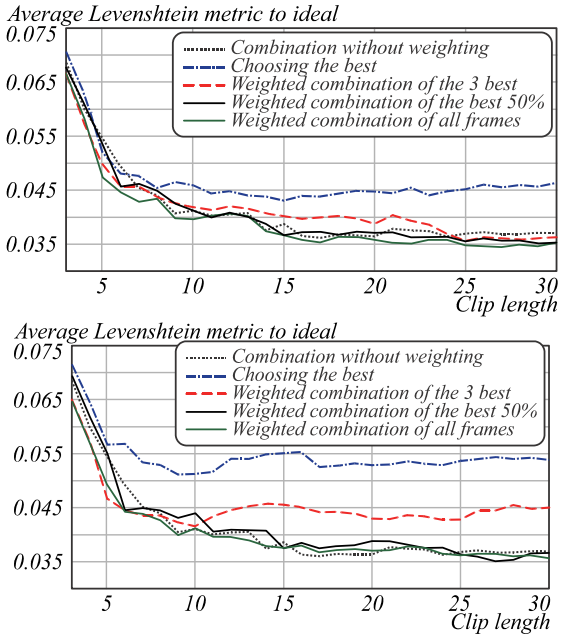


*Fig. 10. Performance profiles for weighted combination based on focus estimation (top) and confidence value (bottom) for MIDV-2019-D dataset*

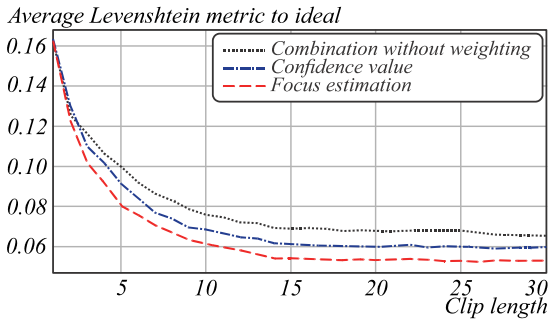


*Fig. 11. Criteria comparison on MIDV-500*

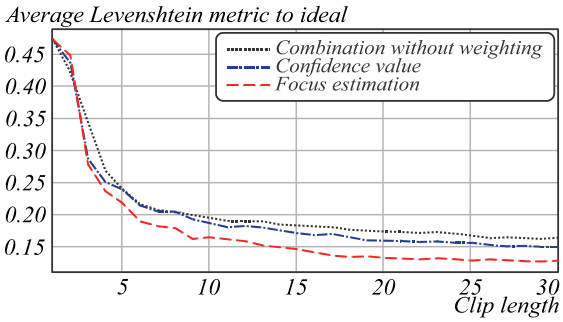


*Fig. 12. Criteria comparison on MIDV-2019-L*

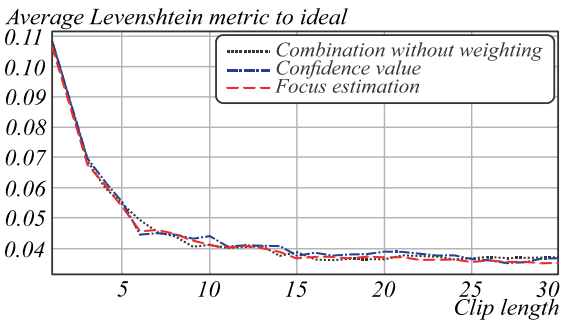


*Fig. 13. Criteria comparison on MIDV-2019-D*

Tab. 2, 3, and 4 demonstrate the mean distance metric value to the correct results for different number of combined frame results using the evaluated weighting strategies, for clips from datasets MIDV-500, MIDV-2019-L, and MIDV-2019-D, respectively.

*Table 2. Mean Normalized Levenshtein metric distance to the correct result on MIDV-500*

| Combination method | Mean Normalized Levenshtein metric | | | | | |
|---|---|---|---|---|---|---|
| | 5 frames | 10 frames | 15 frames | 20 frames | 25 frames | 30 frames |
| Without weighting | 0.0995 | 0.0756 | 0.0689 | 0.0677 | 0.0680 | 0.0652 |
| Confidence value: best 50 % | 0.0911 | 0.0684 | 0.0612 | 0.0598 | 0.0601 | 0.0597 |
| Focus estimation: best 50 % | **0.0804** | **0.0612** | **0.0541** | **0.0533** | **0.0529** | **0.0529** |

*Table 3. Mean Normalized Levenshtein metric distance to the correct result on MIDV-2019-L*

| Combination method | Mean Normalized Levenshtein metric | | | | | |
|---|---|---|---|---|---|---|
| | 5 frames | 10 frames | 15 frames | 20 frames | 25 frames | 30 frames |
| Without weighting | 0.2392 | 0.1950 | 0.1833 | 0.1732 | 0.1675 | 0.1643 |
| Confidence value: best 50 % | 0.2392 | 0.1868 | 0.1708 | 0.1594 | 0.1561 | 0.1497 |
| Focus estimation: best 50 % | **0.2175** | **0.1647** | **0.1461** | **0.1327** | **0.1283** | **0.1283** |

*Table 4. Mean Normalized Levenshtein metric distance to the correct result on MIDV-2019-D*

| Combination method | Mean Normalized Levenshtein metric | | | | | |
|---|---|---|---|---|---|---|
| | 5 frames | 10 frames | 15 frames | 20 frames | 25 frames | 30 frames |
| Without weighting | 0.0546 | **0.0412** | 0.0388 | **0.0365** | 0.0370 | 0.0371 |
| Confidence value: best 50% | 0.0551 | 0.0441 | 0.0376 | 0.0390 | 0.0365 | 0.0368 |
| Focus estimation: best 50% | **0.0535** | 0.0413 | **0.0367** | 0.0371 | **0.0356** | **0.0353** |

## *Per-character weighting*

At the second stage, experiments with a per-character weighing model were performed. The main attention was paid to using focus estimation as a base weighting function, which gave the best results in the previous experiments. Fig. 14 shows performance profiles for combination without weighting, weighted combination of all frames, and half of the best frames both for the combination with weighting of the entire text field and for the per-character weighting modification.

Fig. 15 shows performing profiles for clips with low light conditions, Fig. 16 represents similar plots for clips with strong projective distortions.

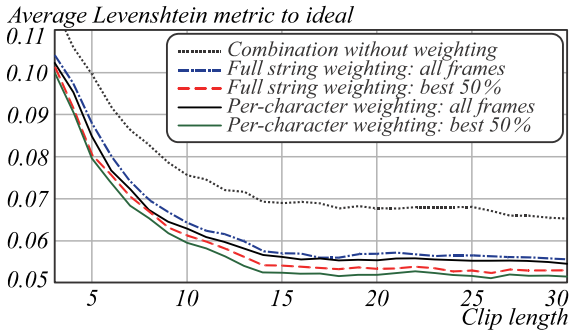


*Fig. 14. Performance profiles for focus estimation weighting on MIDV-500*

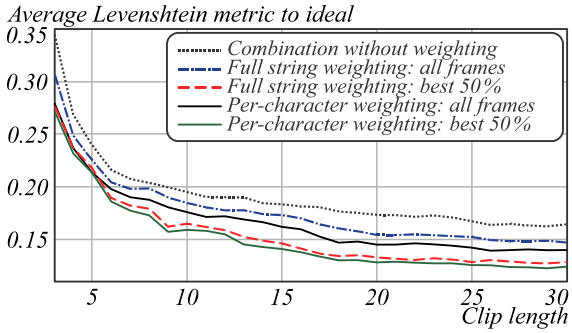


*Fig. 15. Performance profiles for focus estimation weighting on MIDV-2019-L*

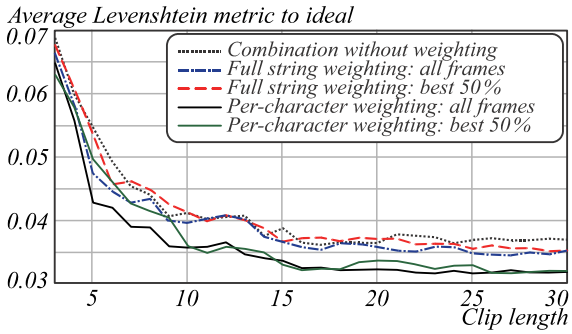


*Fig. 16. Performance profiles for focus estimation weighting on MIDV-2019-D*

Tab. 5, 6, and 7 represent the mean distance metric value to the correct results for different combination strategies, for clips from datasets MIDV-500, MIDV-2019-L, and MIDV-2019-D, respectively.

From the results of the experiments, it follows that per-character weighing allows improving the quality of the text string recognition in a video stream, regardless of the features of capture clips.

### *7. Discussion*

From the results of the first series of experiments, it can be seen that for clips without significant projective distortions (Figs. 8 and 9), the weighted combination allows to improve the recognition precision, and with using focus estimation as a weighting criterion noticeable improvements are achieved regardless of the number of combined frames. For frames with strong projective distortions of the document, the image quality can be noticeably different for different parts of the text strings. Therefore, a predictor constructed over the entire text field may not fully adequately reflect the quality of the recognition. This is especially important for long text fields, such as the machine-readable zone. Fig. 17 shows an example of a document and its machine-readable zone lines, with visibly uneven image quality.

*Table 5. Mean Normalized Levenshtein metric distance to the correct result on MIDV-500 using focus estimation*

| Combination method | Mean Normalized Levenshtein metric | | | | | |
|---|---|---|---|---|---|---|
| | 5 frames | 10 frames | 15 frames | 20 frames | 25 frames | 30 frames |
| Without weighting | 0.0995 | 0.0756 | 0.0689 | 0.0677 | 0.0680 | 0.0652 |
| Full string weighting: all frames | 0.0879 | 0.0643 | 0.0570 | 0.0569 | 0.0565 | 0.0555 |
| Full string weighting: best 50 % | 0.0804 | 0.0612 | 0.0541 | 0.0533 | 0.0529 | 0.0529 |
| Per-character weighting: all frames | 0.0847 | 0.0628 | 0.0561 | 0.0553 | 0.0552 | 0.0545 |
| Per-character weighting: best 50 % | **0.0795** | **0.0595** | **0.0524** | **0.0518** | **0.0516** | **0.0515** |

*Table 6. Mean Normalized Levenshtein metric to the correct result on MIDV-2019-L using focus estimation*

| Combination method | Mean Normalized Levenshtein metric | | | | | |
|---|---|---|---|---|---|---|
| | 5 frames | 10 frames | 15 frames | 20 frames | 25 frames | 30 frames |
| Without weighting | 0.2392 | 0.1950 | 0.1833 | 0.1732 | 0.1675 | 0.1643 |
| Full string weighting: all frames | 0.2250 | 0.1845 | 0.1730 | 0.1540 | 0.1520 | 0.1469 |
| Full string weighting: best 50 % | 0.2175 | 0.1647 | 0.1461 | 0.1327 | 0.1283 | 0.1283 |
| Per-character weighting: all frames | 0.2133 | 0.1757 | 0.1617 | 0.1448 | 0.1420 | 0.1398 |
| Per-character weighting: best 50 % | **0.2131** | **0.1589** | **0.1406** | **0.1281** | **0.1255** | **0.1239** |

*Table 7. Mean Normalized Levenshtein metric distance to the correct result on MIDV-2019-D using focus estimation*

| Combination method | Mean Normalized Levenshtein metric | | | | | |
|---|---|---|---|---|---|---|
| | 5 frames | 10 frames | 15 frames | 20 frames | 25 frames | 30 frames |
| Without weighting | 0.0546 | 0.0412 | 0.0388 | 0.0365 | 0.0370 | 0.0371 |
| Full string weighting: all frames | 0.0475 | 0.0396 | 0.0367 | 0.0358 | 0.0348 | 0.0353 |
| Full string weighting: best 50 % | 0.0535 | 0.0413 | 0.0367 | 0.0371 | 0.0356 | 0.0353 |
| Per-character weighting: all frames | **0.0428** | **0.0357** | 0.0336 | **0.0323** | **0.0317** | **0.0319** |
| Per-character weighting: best 50 % | 0.0497 | 0.0360 | **0.0330** | 0.0337 | 0.0330 | 0.0320 |


*Fig. 17. Example of uneven image quality with strong projective distortion*

The results of the conducted experiments also show that for longer video streams (i.e. with a higher number of frames) the selection of a fixed small number of the best frames for weighting combination using a confidence value as a weighting criterion does not improve recognition (the blue and red performance profiles for the confidence value criterion in Figures 8–10 are above the gray performance profile corresponding to the baseline). The cases when the weighting is performed using the focus estimation show quality improvement, but at the same time, the weighting according to the confidence value does not seem to be robust for clips in which some of the characters were lost due to per-character segmentation errors, clips with low quality of the original image or with highlights. Highlights lower the frame focus estimation criterion as computed according to the method (7), whereas for the confidence value, on the contrary, the loss of several characters of the text line may even increase the weight of the text string. On the other hand, the predictor built on the character membership estimations could better reflect the text field localization errors, as one can expect that the incorrectly localized field the maximal membership estimations of the characters will be lower than for the recognition of a correctly localized text field. It should be noted that in the performed experiments this case was not considered, since the document and text field coordinates were taken from the ground truth, and some additional investigations should be performed to analyze the impact of localization errors.

Thus, weighting combination using confidence values allows to potentially discard precarious recognition results (for example, from frames with strong defocus or incorrect text field localization), but does not work very well in the case where some of the characters are lost. The influence of this can be seen in the fact that, according to the results of the experiments on all datasets, weighing according to the focus estimation allows achieving a higher recognition quality than using the confidence value as a weighing criterion.

Also, it can be noticed that, on average, the best results are obtained using a weighted combination of the best three frames and a weighted combination of the best 50 % of frames. This can be explained by the fact that some frames, obtained using a mobile camera in uncontrolled conditions, are recognized with poor quality. The weighting model which allows to cut off the least significant per-frame result, which could worsen the overall combined recognition result, while still retaining the benefit of a weighted combination over the selection of a single best per-frame result.

To take into account the possible unevenness of the text field recognition quality, the second series of experiments was carried out. In these experiments, in addition to the total weight for the text field, we calculated weights for each individual character. The results of the experiments showed that per-character weighting allows to increase the recognition quality when compared with a full-string weighting model. It should be noted that if for clips without strong projective distortions (Fig. 14 and 15), for which the full-string weighted combination made it possible to improve the recognition quality, per-character weighting only slightly improves the result. However, for clips with strong projective distortions (Fig. 16), per-character weighting significantly increases the combined result precision. Moreover, such improvement occurs regardless of the number of combined per-frame results. From this, we can conclude that for clips with strong projective distortions, if the weighted combination is performed using the focus estimation criterion, it is important to account for the local features of the text field. This can be explained in particular by the fact that, as mentioned earlier, for such clips, the image quality (and hence the character classification precision) of the text line may be uneven.

Thus, from the results of the conducted experiments, it can be concluded that the evaluation and use of the image quality of the recognized field when combining the image quality makes it possible to improve the recognition precision of the text in a video sequence. The best result was achieved using the focus estimation as a weighting criterion. Due to the fact that the input frames obtained using a mobile device camera in uncontrolled conditions may not be of very high quality, the best combination result is obtained using the strategy of combining 50% of the highest scoring frames. Such approach allows simultaneously to cut off the outliers (lowest quality input frames) and to accumulate useful information from multiple well-recognized frames. Since the image quality can be uneven, the use of per-character weighting, which accounts for the local features of the recognized text field, makes it possible to additionally improve the recognition quality, in particular, if the text is captured with significant projective distortions.

## *8. Conclusion*

The paper considered the problem of the combination of text recognition results in a video stream to improve recognition quality. A weighting model and two weighing criteria were proposed: an assessment of the focus esti-

mation of the text image and an a-posteriori text string recognition confidence value. The experiments were carried out on two open datasets containing video clips of identity documents captured with a mobile camera in various conditions.

The results of the first series of experiments have shown that a weighted combination of the recognition results of individual frames can improve the overall recognition quality in the absence of strong projective distortions of the text image. The combination of the best 50% of input frames weighted using an image focus estimation was shown to increase the precision of text recognition in a video stream, as such approach both filters the low-quality outliers and accumulates information from multiple input frames. However, in the case of uneven image quality, in particular on clips with high projective distortion of the recognized text, assigning weights based on characteristics calculated over the entire text string image could be inadequate. Therefore, a per-character weighting procedure was proposed. Experimental results show that the per-character weighting improves the recognition accuracy for all types of the analyzed video clips, including the clips with sliding highlights, long text strings, and the clips with high projective distortions.

Thus, it can be concluded that the combination of the best 50 % frames with per-character weighting according to the input image focus estimation can be applied for video stream recognition systems for increasing the text recognition result precision.

In future research, we plane to explore other possible weighting criteria for the individual frames recognition results, explore the possibility of using deep learning methods to solve the problem of combining recognition results in a video stream, as well as evaluate the proposed weighting model and methods for other domains of application, such as road scene objects recognition.

### *References*


[1] Singh A, Bacchuwar K, Bhasin A. A survey of OCR applications. Int J Mach Learn Comput 2012; 2(3): 314-318. DOI:10.7763/IJMLC.2012.V2.137.

[2] Soheili MR, Yousefi MR, Kabir E, Stricker D. Merging clustering and classification results for whole book recogni-tion. 10th Iranian Conference on Machine Vision and Image Processing (MVIP) 2017: 134-138.

[3] Digitising the real-world: Transforming scanned text into digital data. Source: ⟨https://www.itproportal.com/features/digitising-the-real-world-transforming-scanned-text-into-digital-data/⟩.

[4] Optical character recognition: how using ocr software can increase business efficiency. Source: ⟨https://suscosolutions.com/optical-character-recognition-using-ocr-software-can-increase-business-efficiency/⟩.

[5] Mir AW, Ahmed H, Shah AA. Automated speed limit iden-tification for efficient driving system. International Conference on Communication, Computing and Digital Systems (C-CODE) 2017: 299-303. DOI: 10.1109/C-CODE.2017.7918946.

[6] Jabnoun H, Benzarti F, Amiri H. A new method for text detection and recognition in indoor scene for assisting blind people. Proc SPIE 2017; 10341: 1034123. DOI: 10.1117/12.2268399.

[7] Saudagar A, Habeebvulla M. Augmented reality mobile application for arabic text extraction, recognition and translation. J Stat Manage Syst 2018; 21: 617-629. DOI: 10.1080/09720510.2018.1466968.

[8] License plate recognition systems. Source: ⟨https://epic.org/privacy/licenseplates/⟩.

[9] Optical character recognition system hits the highway. Source: ⟨https://www.vision-systems.com/non-factory/security-surveillance-transportation/article/16737819/optical-character-recognition-system-hits-the-highway⟩.

[10] Arora K, Bist A, Prakash R, Chaurasia S. Custom OCR for identity documents: OCRXNet. Aptisi Transactions On Technopreneurship 2020; 2: 112-119. DOI: 10.34306/att.v2i2.87.

[11] Panchal RB, Sonawane, RG, Shaikh H, Gawali PP. Design of text detection and translation system for camera based android smartphone. International Journal for Scientific Research & Development 2015; 3(1): 4 p.

[12] Ôn Vũ Ngoc M, Fabrizio J, Géraud T. Document detection in videos captured by smartphones using a saliency-based method. ICDARW 2019: 19-24. DOI: 10.1109/ICDARW.2019.30059.

[13] Esser D, Muthmann K, Schuster D. Information extraction efficiency of business documents captured with smartphones and tablets. Proc ACM Symposium on Document Engineering 2013: 111-114.

[14] Xu J, Wu X. A system to localize and recognize texts in Oriented ID card images. IEEE International Conference on Progress in Informatics and Computing (PIC) 2018: 149-153. DOI: 10.1109/PIC.2018.8706303.

[15] Attivissimo F, Giaquinto N, Scarpetta M, Spadavecchia M. An automatic reader of identity documents. 2019 IEEE International Conference on Systems, Man and Cybernetics (SMC) 2019: 3525-3530. DOI: 10.1109/SMC.2019.8914438.

[16] Arlazarov V, Bulatov K, Chernov T, Arlazarov VL. MIDV-500: a dataset for identity document analysis and recognition on mobile devices in video stream. Computer Optics 2019; 43(5): 818-824. DOI: 10.18287/2412-6179-2019-43-5-818-824.

[17] Myasnikov E, Savchenko A. Detection of sensitive textual information in user photo albums on mobile devices. International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON, Novosibirsk, Russia) 2019: 0384-0390. DOI: 10.1109/SIBIRCON48586.2019.8958325.

[18] Arlazarov VV, Zhukovsky A, Krivtsov V, Nikolaev D, Polevoy D. Analysis of using stationary and mobile small-scale digital cameras for documents recognition [In Russian]. Information Technologies and Computing Systems 2014; 3: 71-81.

[19] Alaql O, Ghazinour K, Lu CC. Classification of image distortions for image quality assessment. International Conference on Computational Science and Computational Intelligence (CSCI) 2016: 653-658. DOI: 10.1109/CSCI.2016.0129.

[20] Puybareau É, Géraud T. Real-time document detection in smartphone videos. 25th IEEE International Conference on Image Processing (ICIP 2018: 1498-1502. DOI: 10.1109/ICIP.2018.8451533.

[21] Chernov TS, Ilin DA, Bezmaternykh PV, Faradjev IA, Karpenko SM. Research of methods for segmentation of document text block images using algorithms of structure

analysis and machine learning. RFBR Journal 2016; 4(92): 55-71. DOI: 10.22204/2410-4639-2016-092-04-55-71.
[22] Bulatov K, Matalov D, Arlazarov V. MIDV-2019: challenges of the modern mobile-based document OCR. Proc SPIE 2020: 11433: 717-722. DOI: 10.1117/12.2558438.
[23] Tropin DV, Shemyakina YA, Konovalenko IA, Faradzhev IA. Localization of planar objects on the images with complex structure of projective distortion. Informatsionnye Protsessy 2019; 19(2): 208-229.
[24] Javed K, Shafait F. Real-time document localization in natural images by recursive application of a CNN. 14th IAPR International Conference on Document Analysis and Recognition (ICDAR) 2017: 105-110. DOI: 10.1109/ICDAR.2017.26.
[25] Zhu A, Zhang C, Li Z. et al. Coarse-to-fine document local-ization in natural scene image with regional attention and recursive corner refinement. IJDAR 2019; 22: 351-360. DOI: 10.1007/s10032-019-00341-0.
[26] Cheng Z, Lu J, Niu Y, Pu S, Wu F, Zhou S. You only recognize once: Towards fast video text spotting. Proceedings of the 27th ACM International Conference on Multimedia 2019: 855-863. DOI: 10.1145/3343031.3351093.
[27] Brisinello M, Grbić R, Stefanovič D, Pečkai-Kovač R. Optical character recognition on images with colorful background. IEEE 8th International Conference on Consumer Electronics (ICCE-Berlin) 2018: 1-6. DOI: 10.1109/ICCE-Berlin.2018.8576202.
[28] Mustafa WA, Aziz H, Khairunizam W, Ibrahim Z, Shahriman A, Razlan ZM. Review of different binarization approaches on degraded document images. International Conference on Computational Approach in Smart Systems Design and Applications (ICASSDA) 2018: 1-8. DOI: 10.1109/ICASSDA.2018.8477621.
[29] Islam N, Islam Z, Noor N. A survey on optical character recognition system. ITB Journal of Information and Communication Technology 2016; 10(2).
[30] Tereshin A, Usilin S, Arlazarov VV. Performance improvement of multi-class detection using greedy algorithm for Viola-Jones cascade selection. Proc SPIE 2018; 10696: 106960D. DOI: 10.1117/12.2310101.
[31] Skoryukina N, Arlazarov VV, Nikolaev D. Fast method of ID documents location and type identification for mobile and server application. 15th International Conference on Document Analysis and Recognition (ICDAR) 2019: 850-857. DOI: 10.1109/ICDAR.2019.00141.
[32] Fang X, Fu X, Xu X. ID card identification system based on image recognition. 12th IEEE Conference on Industrial Electronics and Applications (ICIEA) 2017: 1488-1492. DOI: 10.1109/ICIEA.2017.8283074.
[33] Volkova V, Deriuga I, Osadchyi V, Radyvonenko O. Improvement of character segmentation using recurrent neural networks and dynamic programming. IEEE Second International Conference on Data Stream Mining AND Processing (DSMP) 2018: 218-222. DOI: 10.1109/DSMP.2018.8478457.
[34] Ryan M, Hanafiah N. An examination of character recognition on ID card using Template Matching Approach. International Conference on Computer Science and Computational Intelligence (ICCSCI) 2015; 59: 520-529. DOI: 10.1016/j.procs.2015.07.534.
[35] Anantharajah K, Denman S, Sridharan S, Fookes C, Tjondronegoro D. Quality based frame selection for video face recognition. 6th International Conference on Signal Processing and Communication Systems 2012: 1-5. DOI: 10.1109/ICSPCS.2012.6507950.
[36] Zhanzhan C, Jing L, Yi N, Shiliang P, Fei W, Shuigeng Z. You only recognize once: Towards fast video text spotting. 27th ACM International Conference 2019: 855-863. DOI: 10.1145/3343031.3351093.
[37] Haris M, Shakhnarovich G, Ukita N. Recurrent back-projection network for video super-resolution. Proc IEEE Conference on Computer Vision and Pattern Recognition 2019: 3897-3906. DOI: 10.1109/CVPR.2019.00402.
[38] Deudon M, Kalaitzis A, Goytom I, Arefin MdR, Lin Z, Sankaran K, Michalski V, Kahou SE, Cornebise J, Bengio Y. HighRes-net: Multi-Frame Super-Resolution by Recursive Fusion. ICLR 2020 Conference. Source: ⟨https://openreview.net/forum?id=HJxJ2h4tPr⟩.
[39] Mehregan K, Ahmadyfard A, Khosravi H. Super-resolution of license-plates using frames of low-resolution video. 5th Iranian Conference on Signal Processing and Intelligent Systems (ICSPIS) 2019: 1-6. DOI: 10.1109/ICSPIS48872.2019.9066104.
[40] Merino-Gracia C, Mirmehdi M. Real-time text tracking in natural scenes. IET Comput Vis 2014; 8(6): 670-681. DOI: 10.1049/iet-cvi.2013.0217.
[41] Cheng Z, Lu J, Xie J, Niu Y, Pu S, Wu F. Efficient video scene text spotting: Unifying detection, tracking, and recognition. arXiv e-prints 2019. Source: ⟨https://arxiv.org/abs/1903.03299⟩.
[42] Zhang S, Li P, Meng Y, Li L, Zhou Q, Fu X. A video deblurring algorithm based on motion vector and an encorder-decoder network. IEEE Access 2019; 7: 86778-86788. DOI: 10.1109/ACCESS.2019.2923759.
[43] Myasnikov VV, Dmitriev EA. The accuracy dependency investigation of simultaneous localization and mapping on the errors from mobile device sensors. Computer Optics 2019; 43(3): 492-503. DOI: 10.18287/2412-6179-2019-43-3-492-503.
[44] Sankar K, Jawahar C, Manmatha R. Nearest neighbor based collection OCR. Proceedings of the 9th IAPR International Workshop on Document Analysis Systems 2010: 207-214. DOI: 10.1145/1815330.1815357.
[45] Fiscus JG. A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER). IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings 1997: 347-354. DOI: 10.1109/ASRU.1997.659110.
[46] Zhou ZH. Ensemble methods: Foundations and algorithms. New York: Chapman and Hall/CRC; 2012. ISBN: 978-1-4398-3003-1.
[47] Polikar R. Ensemble based systems in decision making. IEEE Circuits and Systems Magazine 2006; 6(3): 21-45. DOI: 10.1109/MCAS.2006.1688199.
[48] Kittler J. On combining classifiers. IEEE Trans Pattern Anal Mach Intell 1998; 20(3): 226-239.
[49] Bulatov K, Arlazarov V, Chernov T, Slavin O, Nikolaev D. Smart IDReader: Document recognition in video stream. 14th IAPR International Conference on Document Analysis and Recognition (ICDAR) 2017; 6: 39-44. DOI: 10.1109/ICDAR.2017.347.
[50] Llobet R, Navarro Cerdán J, Perez-Cortes J-C, Arlandis J. OCR Post-processing using weighted finite-state transducers. Proc ICPR 2010: 2021-2024. DOI: 10.1109/ICPR.2010.498.
[51] Bulatov K. A method to reduce errors of string recognition based on combination of several recognition results with per-character alternatives. Bulletin SUSU MMCS 2019; 12(3): 74-88. DOI: 10.14529/mmp190307.
[52] Bulatov K, Lynchenko A, Krivtsov V. Optimal frame-by-frame result combination strategy for OCR in video stream. Proc SPIE 2017; 10696: 106961Z. DOI: 10.1117/12.2310139.

[53] Bulatov KB, Polevoy DV. Reducing overconfidence in neural networks by dynamic variation of recognizer relevance. Proc ECMS 2015, 488-491. DOI: 10.7148/2015-0488.
[54] Chernov TS, Ilyuhin SA, Arlazarov VV. Application of dy-namic saliency maps to video stream recognition systems with image quality assessment. Proc SPIE 2018; 11041: 110410T. DOI: 10.1117/12.2522768.
[55] Petrova O, Bulatov K, Arlazarov VL. Methods of weighted combination for text field recognition in a video stream Proc SPIE 2020; 11433: 114332L. DOI: 10.1117/12.2559378.
[56] Chernyshova YS, Sheshkus AV, Arlazarov VV. Two-step CNN framework for text line recognition in camera-captured images. IEEE Access 2020; 8: 32587-32600. DOI: 10.1109/ACCESS.2020.2974051.
[57] Yujian L, Bo L. A normalized Levenshtein distance metric. IEEE Trans Pattern Anal Mach Intell 2007; 29(6): 1091-1095. DOI: 10.1109/TPAMI.2007.1078.
[58] Zilberstein S. Using anytime algorithms in intelligent systems. AI Magazine 1996; 17(3): 73-83.

### *Authors' information*

**Olga Olegovna Petrova** (b. 1994), graduated from Moscow Institute of Physics and Technology (State University) in 2019, majoring in Applied Mathematics and Informatics. Currently she is a postgraduate student at FRC CSC RAS. Research interests are computer science, document analysis and recognition. E-mail: *opetrova@smartengines.com* .

**Konstantin Bulatovich Bulatov** (b. 1991), PhD, graduated from National University of Science and Technology "MISiS" in 2013, majoring in Applied Mathematics. Currently he works as a senior researcher at FRC CSC RAS. Research interests are pattern recognition, combinatorial algorithms and computer vision.
E-mail: *kbulatov@smartengines.com* .

**Vladimir Viktorovich Arlazarov** (b. 1976), PhD, graduated from Moscow Institute of Steel and Allows in 1999, majoring in Applied Mathematics. Currently he works as head of division 93 at FRC CSC RAS. Research interests are pattern recognition and machine learning. E-mail: *vva@smartengines.com* .

**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 August 3, 2020. The final version – December 30, 2020.*