R. Sánchez-Rivero, P. V. Bezmaternykh, A. V. Gayer, A. Morales-González, F. J. Silva-Mata, K. B. Bulatov, A joint study of deep learning-based methods for identity document image binarization and its influence on attribute recognition, *Компьютерная оптика*, 2023, том 47, выпуск 4, 627–636

A joint study of deep learning-based methods for identity document image binarization and its influence on attribute recognition

R. Sánchez-Rivero [1], P.V. Bezmaternykh [2,3], A.V. Gayer [2,3],
A. Morales-González [1], F. José Silva-Mata [1], K.B. Bulatov [2,3]
[1] Advanced Technologies Application Center (CENATAV), Playa P.C.12200, Havana, Cuba, 7A, #21406 Siboney,;
[2] FRC "Computer Science and Control" RAS, 119333, Russia, Moscow, Vavilova st., 44 b.2;
[3] Smart Engines Service LLC, 117312, Russia, Moscow, 60th Anniversary of October avenue, 9

Abstract

Text recognition has benefited considerably from deep learning research, as well as the preprocessing methods included in its workflow. Identity documents are critical in the field of document analysis and should be thoroughly researched in relation to this workflow. We propose to examine the link between deep learning-based binarization and recognition algorithms for this sort of documents on the MIDV-500 and MIDV-2020 datasets. We provide a series of experiments to illustrate the relation between the quality of the collected images with respect to the binarization results, as well as the influence of its output on final recognition performance. We show that deep learning-based binarization solutions are affected by the capture quality, which implies that they still need significant improvements. We also show that proper binarization results can improve the performance for many recognition methods. Our retrained U-Net-bin outperformed all other binarization methods, and the best result in recognition was obtained by Paddle Paddle OCR v2.

Keywords: document image binarization, identity document recognition, optical character recognition, deep learning, U-Net architecture.

Citation: Sánchez-Rivero R, Bezmaternykh P, Gayer A, Morales-González A, José Silva-Mata F, Bulatov K. A joint study of deep learning-based methods for identity document image binarization and its influence on attribute recognition. Computer Optics 2023; 47(4): 627-636. DOI: 10.18287/2412-6179-CO-1207.

Introduction

Document image analysis and recognition is a rapidly growing domain that simultaneously relies on image processing techniques, pattern recognition approaches, and computer optic principles. The handbook [1] provides a gentle introduction to the subject. One of the latest surveys of document image recognition problems and existing solutions is presented in the paper [2]

Among the set of document types being analyzed, identity, or ID documents play a special role. They are utilized to confirm their owner's personality in a plenty of scenarios: usage of government services, banking, access granting, or travelling. The scope and context of their processing along with the corresponding recognition system design are addressed in works [3 – 4].

A typical ID document type can be defined by its "template" – a set of features shared by every document sample of this type. The list of common features includes static (known in advance) textual or graphic elements, the information about their relative location on the document, a set of keypoints and their descriptors, physical sizes and many other [3]. A set of ID document *attributes*, which vary from one sample to another and thus determine the identity, is known as the document's "content". Document number, surname, date of birth, owner's photo, are all examples of these attributes. In ID document recognition we are mainly interested in the "content" and the scope of this paper is limited to the recognition of printed textual attributes, placed in the positions predefined by the corresponding "template". Within this context, we consider two important stages of the typical document processing pipeline (DPP): DIB – document image binarization and OCR – optical character recognition. An OCR module is a common consumer of binarization outcome. Here, properly binarized document image can greatly simplify its recognition process. Some modern OCR modules are able to deal not only with binary images but also with colorful or grayscaled ones. This variation in DPP is displayed in Fig. 1. Having such a variation, it is important to assess the influence of the binarization stage upon the accuracy of OCR modules in DPP for ID documents.

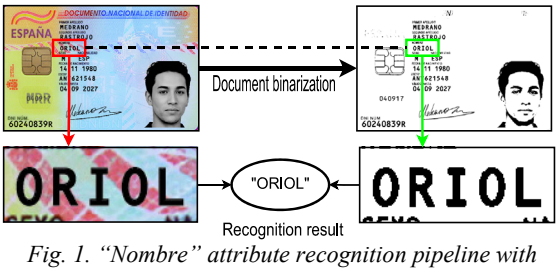

Fig. 1. "Nombre" attribute recognition pipeline with preliminary binarization stage (right branch) and without it (left branch)

Real ID documents contain a lot of personal data, which makes the task of creating a comprehensive publicly available dataset very difficult, thus, limiting the evaluation and benchmarking. Such a dataset, named

MIDV-500 [5], was published in 2018. Later, its successors, the MIDV-2019 and the MIDV-2020 [6], became available. They consist of video clips, scanned images, photos and templates of unique mock ID documents captured in various conditions. The ground truth is provided for some problems including the OCR one, but it is not available for the DIB.

Both DIB and OCR problems greatly benefited from the latest achievements in supervised learning and most part of recent algorithms is based on deep learning (DL) techniques. However, their applicability to the ID document analysis is not established. Thus, the goal of this work is to assess the accuracy of modern OCR modules with preliminary binarization stage and without it, within the field of ID document analysis on MIDV datasets. At the same time, the visual quality of captured ID images can vary a lot and it is also important to assess its impact on the final recognition accuracy.

This work is an extension of the study [7] with additional experimental details and insights. The contributions can be summarized as follows: (a) an experimental analysis of OCR modules accuracy over binarization outcome on the MIDV-500 and MIDV-2020 subsets; (b) an experimental analysis of input image quality influence over the performance of the reviewed modules on the MIDV-500; (c) an analysis of a U-Net based solution accuracy retrained with domain specific data from the MIDV-500; (d) a manual pixel-wise annotation of ID document templates from the MIDV-500; (e) a manual image quality annotation for the MIDV-500.

The remainder of this paper is organized as follows. Section 1 provides information about related work. Section 2 describes the set of algorithms surveyed in this work. Section 3 presents the proposed pipeline analysis and the experimental design and results. Section 4 gives the conclusion.

1. Related work

1.1. Document image binarization

DIB algorithms have already been studied for a long time and many surveys and comparative reviews have been published [8-9]. Nevertheless, most of them are focused on classic methods such as Otsu, Niblack, Wolf, Nick, Sauvola and many others. In most cases, their parameters should be carefully tuned to get appropriate result on target data. However, recent advances in machine learning, especially DL, have revolutionized the domain and gave rise to a multitude of methods based on end-to-end pixel classification using trained artificial neural networks (ANNs). The performance and limitations of such methods are studied at this moment. For instance, the recent review [10] takes in to account some of them. Therefore, the goal of this paper is to address not classical, but DL-based methods of DIB.

The training process requires the presence of consistent pixel-wise ground truth annotations (PWGT). A number of them have been collected and published. Since 2009, Document Image Binarization Contest (DIBCO) regularly provided such annotated data for benchmark and track of DIB progress. Another examples of relevant datasets are the Palm Leaf Manuscripts and the Persian Heritage Image Binarization Dataset. All these datasets are mainly focused on historical document analysis, so they contain a lot of handwritten data [11]. ID documents, on the contrary, contain mostly printed texts with well-defined characteristics. So, the set of problems is different from those seen on historical document images. Several issues can negatively interfere with the binarization and further OCR output, for instance, special security objects and marks, diversity of colors and backgrounds, printing methods, diversity of sources and other special characteristics that depend on the country and its emission. Thus, it is important to evaluate whether the trained solutions are applicable for these documents or not. The performance of ANNs is heavily influenced by the training data. As shown in [12], their application for images with minimal similarity to the data from the training dataset might provide extremely poor outcomes.

1.2. Evaluation of binarization method performance

The performance of DIB method can be evaluated using different strategies, depending on the final objective. These strategies are mainly divided into two groups, “direct” and “indirect”, based on the presence of PWGT. For the first group, the presence of such well-established ground truth is essential, but its creation is very resource-consuming procedure which is mostly performed in a semi-manual way by domain experts. Moreover, classification results may vary from one expert to another. Many aspects of PWGT creation for binarization needs are covered in paper [13]. Main performance metrics using PWGT are examined in work [14].

The “indirect” group relies on the evaluation of binary document visual appearance or its recognition performance [15, 16]. This approach was popular before datasets with PWGT like DIBCO appeared [17, 18] and it is still employed in some scenarios [15]. In such a case, the final result depends on the used OCR method and in fact the performance of the pair “binarization method × OCR method” is evaluated. To better understand each binarization method behavior, several OCR methods should be considered.

When the binarization outcome is fixed, the common way to evaluate the performance of a single OCR method is to calculate some Levenshtein-based metrics between the obtained results and the textual ground truth [16]. Some insights about experimental evaluation of OCR methods performance are presented in paper [19]. For the task of ID document binarization the “indirect” strategy seems to be a better choice for two reasons: (a) the recognition quality is the real final objective for the majority of applications; (b) there is no relevant dataset with well-established PWGT for this document class.

1.3. Document image quality analysis

Image quality assessment is another important task in the area of image processing with a lot of applications [20]. The quality of ID document images can suffer from multiple distortions, especially when capturing conditions are not controlled. The photometric quality and geometry of the document image are affected by the presence of specular light, shape and motion distortions, defocusing and many others [21]. These factors normally lead to poor document analysis results. Clearly, they strongly affect the recognition stage [22]. Thus, a common step in ID document recognition system is to evaluate the quality score of every input image. This score helps to filter out evidently bad images, choose the best image from video stream and increase the reliability of the final recognition result [23]. Taking in to account only the images with high scores allows to improve the system's performance in terms of speed and recognition quality. The maximum level of distortion acceptable for a recognition algorithm can be determined by the method from paper [24].

2. Algorithm selection

Several document binarization contests have been held, providing quantitative evaluations of various binarization methods, including DL-based ones. The higher the method's ranking, the more interest it attracts. However, most participants do not publish their methods or make it difficult to reproduce results under different conditions. Despite this, some methods, such as U-Net-bin [25] and Gallego's autoencoder [26], are top-ranked and provide their solutions and training procedures publicly. For this reason, we selected these methods, along with ROBIN [27] and the popular Otsu method [28] as a non-DL baseline as used in [5, 12, 29], for our study.

As for recognition methods, their choice is also based on participation in contests and benchmarks, as well as their relevance and recentness. The availability of source codes and pre-trained models, which allows for performance evaluations was also a significant factor in the selection process.

Semantic Reasoning Networks (SRN) [30]. It is a four-stage DL-based system that won the first place in the ICDAR 2013 competition. It creates a 2D feature map by combining a ResNet50 backbone network with a Feature Pyramid Network and two transformer units. The authors developed a novel attention mechanism called Parallel Visual Attention, which surpasses previous attention mechanisms in terms of efficiency.

Paddle Paddle OCR v2 (PPOCR2) [31]. This framework uses the same architecture as SRN, and it is a new version of it focusing on improving the training process using novel strategies like Collaborative Mutual Learning (CML) and new data augmentation techniques. It uses the new LCNet network as backbone which is a modification of MobileNet v1.

Self-Attention Text Recognition Network (SATRN) [32]. It is an autoencoder influenced by Transformers, which exploits the 2D spatial dependencies of characters in a text image.

Baek et al. [33]. This approach presents four stages in which authors combine text normalization, feature extraction (ResNet), sequence modeling (BiLSTM) and character sequence prediction.

ResNet CTC. It is a mix of DL approaches that includes a ResNet backbone as a feature extractor and a Connectionist Temporal Classification (CTC) module that uses the features to forecast the text's characters.

ResNet FC. A straightforward DL technique including a ResNet backbone for extraction and a fully connected layer for character prediction.

CSTR [34]. This is a classification-based process that incorporates a two-network/stage framework: a core network for classification based on classification perspective network and a second stage prediction based on separated convolutions with global average pooling prediction network.

Tesseract [35] It is a popular open-source engine which is commonly used as a baseline for recognition accuracy evaluation in competitions [29] and surveys [36]. We used version 4.1.1 utilizing an LSTM (Long short-term memory) ANN for recognition, making it ideal for this study as it focuses on DL approaches. Additionally, the engine's sensitivity to image preprocessing techniques can help in the analysis of the binarization step's influence.

All the provided models were trained mostly on synthetic images. The global architecture of their networks use the same structure. In the most simple framework they employ a feature extractor with a prediction layer, and then they add a sequence modeling stage, some attention mechanisms, or some angle correction steps. Only Tesseract, PPOCR2, SRN and SATRN models are trained for recognizing punctuation characters (back and forward slashes, dots, commas, hyphens and other symbols) which are regularly presented in the ID document images, thus affecting the recognition accuracy.

3. Experimental analysis

To objectively measure the influence of DIB on the task of textual field recognition in ID document analysis domain, we designed a set of four experiments. With these experiments, we expect to address the following questions:

1) Is the binarization stage relevant within the text recognition pipeline for ID documents?
2) Is it possible to improve text recognition accuracy for ID documents using a binarization algorithm retrained with domain specific data?
3) How do text recognition errors behave for individual fields of ID documents? Which are the most problematic fields in this context?

4) How and to what extent do image quality variations influence the recognition accuracy?

In this section, these experiments are carried out and their results and discussions are presented. We used the MIDV-500 [5] and the MIDV-2020 [6] datasets, which are among the few available datasets that focus on ID documents. The MIDV-500 dataset comprises 50 document types, each with a corresponding "template" – the best quality document image sample that is not affected by any capture problems. The set of templates is denoted as T_{500}. The dataset also contains 10 videos (for each type of document) of 30 frames each, captured with two different devices in 5 scenarios with varying conditions.

A ground truth annotation is provided for every textual attribute. It consists of bounding rectangle and the text string. The MIDV-2020 dataset is organized similarly to the MIDV-500 but contains more data. It comprises data for 100 templates T_{2020} from each of ten of the documents in the MIDV-500, for a total of 1000 unique templates of ten kinds.

The basic scheme of every proposed experiment $\mathcal{E}$ is as follows: the set of binarization algorithms $\mathbb{B}_{\mathcal{E}}$ is applied to a subset of images from MIDV-500 or MIDV-2020 denoted as $\mathbb{I}_{\mathcal{E}}$. From the binary outcome of every $B \in \mathbb{B}_E$, textual field images are extracted according to the provided ground truth. The sets of retrieved textual fields $\mathbb{F}_B$ are input for every recognition algorithm $R \in \mathbb{R}_{\mathcal{E}}$. The recognition error E is evaluated for every algorithm R and set of processed fields $\mathbb{F}_B$. The set of all recognition algorithms is denoted as $\mathbb{R}$.

In this work, we also use two special binarization methods, B_{gt} and B_{id}. The first one is employed to evaluate the optimal result that can be achieved with binarization, using the PWGT of the image set $\mathbb{I}_{\mathcal{E}}$ (note that such PWGT is not available for either MIDV-500 or MIDV-2020). B_{id} preserves the original image and is used to evaluate the recognition error in the absence of binarization, helping determine the necessity of this step. The set of all binarization algorithms is denoted as $\mathbb{B}$.

In this work, the OCR module error R is evaluated over the given dataset $\mathbb{D} = \{(f,g) | f \in \mathbb{F}, g \in \mathbb{G}, |\mathbb{F}| = |\mathbb{G}| = N\}$. Here, $\mathbb{F}$ and $\mathbb{G}$ represent sets of textual image fields and corresponding ground truth. The recognition result r of the textual field $f \in \mathbb{F}$ is a string, so does the corresponding value $g \in \mathbb{G}$. To compare r with g, string matching approach, based on Levenshtein distance $L_{dist}(r,g)$ calculation, is used. It is known as "normalized Levenshtein distance" (Eq. 1) and is described in details in work [37].

$$V(r,g) = \frac{2 \cdot L_{dist}(r,g)}{|r| + |g| + L_{dist}(r,g)}. \quad (1)$$

The overall error of recognition algorithm R over the dataset $\mathbb{D}$ is denoted as $E(R, \mathbb{D})$. It is calculated as mean value of all the distances $V(r,g)$ (Eq. 2).

$$E(R, \mathbb{D}) = \frac{1}{N} \cdot \sum_{i=1}^{N} V(R(f_i), g_i). \quad (2)$$

3.1. Experiment 1: Impact of binarization algorithms in the text recognition pipeline

The goal of the first experiment is to establish the real effect of the DIB stage within the context of the recognition process on the MIDV-500 dataset.

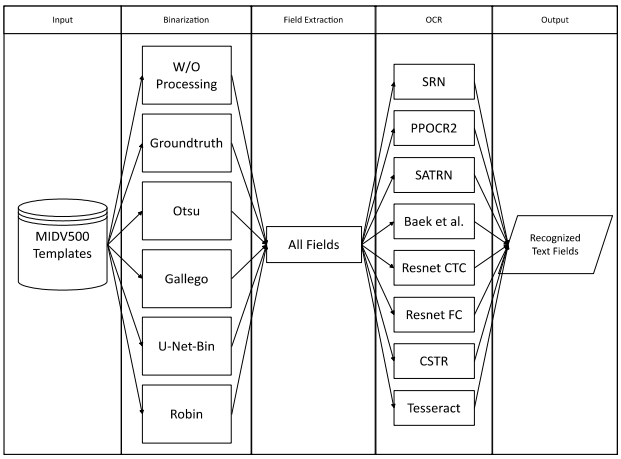

Fig. 2. Description of the first experiment

In this experiment, the image data is a subset of T_{500}, with ten documents written in non-Latin alphabets excluded. The best character size for each DIB algorithm, as determined in [7], is used, and all the template images are accordingly resized before being processed by the binarization algorithms. The set of filtered and preprocessed templates is $T^*_{500} = \mathbb{I}_{\mathcal{E}}$, with $\mathbb{B}_{\mathcal{E}} = \mathbb{B} \cup \{B_{id}, B_{gt}\}$ and $\mathbb{R}_{\mathcal{E}} = \mathbb{R}$. The full pipeline is illustrated in Fig. 2.

To use B_{gt}, we created PWGT for all 50 templates from the MIDV-500 dataset (Fig. 9b). For every template this annotation represents a binary image that delimits the background from the texts and any attributes containing printed or handwritten characters, including signatures. During this process, texts over any support (ink, printed, sublimated, optically variable, etc), in any alphabet, and with different sizes, colors and typefaces were taken into account. This PWGT is not restricted only to the training and evaluating some binarization algorithm, it extends the dataset for future experiments and research like signature detection, segmentation and recognition.

The annotation was performed by multiple specialists in order to obtain more variability in the resulting data, given that some pixels can receive different classification by different persons. Some of the templates presented low resolution, *complex* backgrounds, overlapping texts and zones with text occlusions (because of security and information printings like photos, watermarks and seals). The annotation is freely available online on *ftp://smartengines.com/midv-500-extra-annotations*.

The conducted experiment results are presented in Tab. 1.

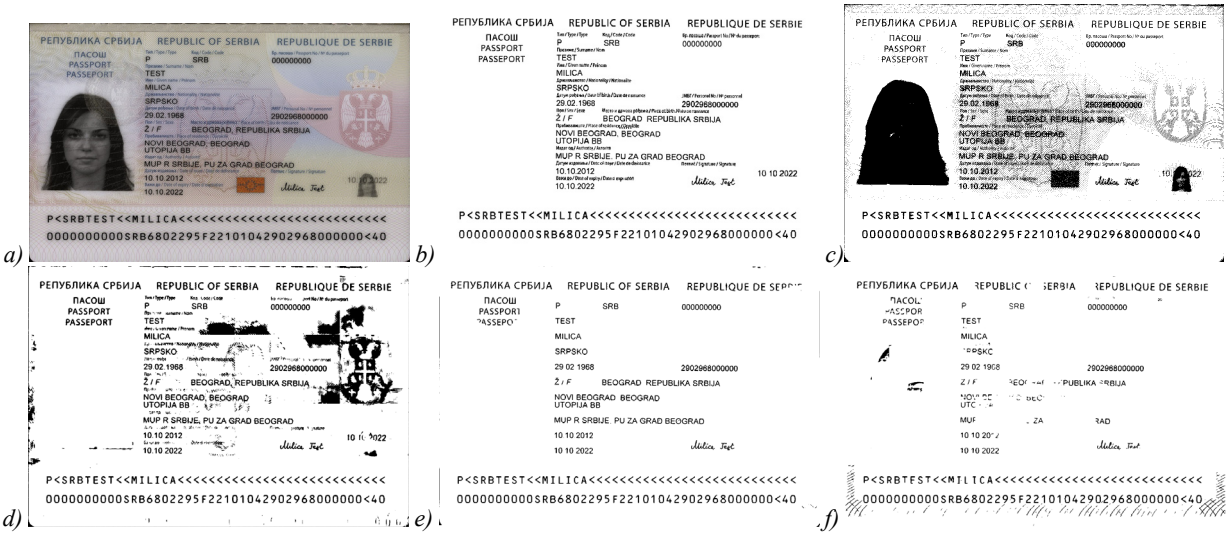

Fig. 3. Original template and different binarization outputs, a) Template from MIDV, b) Prepared ground truth, c) Otsu output, d) Gallego output, e) U-Net-Bin Output, f) ROBIN output

Tab. 1. Recognition error E over binarization outcomes for T^*_{500}

Recognition	B_{gt}	B_{id}	Otsu	Gallego	U-Net-bin	Robin
PPOCR2	**0.049**	0.052	0.102	0.105	0.135	0.149
SRN	**0.056**	0.059	0.098	0.105	0.146	0.159
CSTR	0.239	**0.233**	0.272	0.283	0.295	0.308
SATRN	**0.160**	0.173	0.204	0.206	0.223	0.246
ResNet CTC	0.217	**0.202**	0.301	0.267	0.288	0.300
Baek et al.	0.198	**0.188**	0.259	0.244	0.258	0.273
Tesseract	**0.081**	0.092	0.162	0.139	0.160	0.182
ResNet FC	0.294	**0.284**	0.337	0.325	0.339	0.354

These results show that all recognition algorithms perform better on non-preprocessed images or on ground truth sources. The poor performance of binarization on the text recognition pipeline suggests that current DL-based binarization results are not goodenough or current DL-based OCR algorithms are already good enough without the need for preprocessing their input. However, when the same recognition algorithms were run on PWGT (Tab. 1 column B_{gt}), lower error rates were obtained comparing to original images, suggesting that binarization can improve recognition accuracy and that the used algorithms have room for improvement.

It can be observed that the Otsu algorithm consistently outperforms other methods, particularly for the uniform background and high-quality templates found in ID documents. The DL-based methods, such as the Gallego autoencoder, also show promising results, but are not able to reach the level of performance of Otsu or non-binarized images in some cases. Additionally, the PPOCR2 and SRN recognition algorithms stand out as the top performers among the recognition methods tested.

3.2. Experiment 2: Retraining binarization network on specific domain data

The goal of this experiment is to establish the effect of the DIB within the context of the whole recognition process on the subset MIDV-2020 dataset after retraining one of the binarization solutions using domain data taken from the MIDV-500 dataset.

Here, the image data is the subset of T_{2020}. Since there are two document types filled entirely in non-Latin alphabet, the 200 corresponding templates are excluded from the evaluation for a total of 800 images. The filtered set of templates is designated as $T^*_{2020} = \mathbb{I}_{\mathcal{E}}$. As for binarization methods, $\mathbb{B}_{\mathcal{E}} = \{B_{id}, B_{Otsu}, B_U, B^R_U\}$. Here, B_U is original U-Net-bin method and B^R_U – its retrained version which used some domain data from the newly annotated PWGT for MIDV-500. The choice of Otsu stems from its results in the first experiment and the fact that it is still a common baseline for the task of binarization. Unfortunately, the size of T_{2020} is too big, so PWGT preparation is too resource consuming. Thus, B_{gt} is not available for this experiment. The set of recognition algorithms contains only the recognizer with the lowest error according to the first experiment: $\mathbb{R}_{\mathcal{E}} = \{R_{PROCR2}\}$.

Now let describe the retraining process of U-Net based binarization solution with domain specific data from the MIDV-500 dataset, in order to contrast it with the original solution trained on general image data. For document templates binarization we used a DL approach, provided by DIBCO-2017 competition winners and based on the U-Net model. Compared to DIBCO-2017 challenge, binarization of ID documents is easier than arbitrary historical handwritten documents. Also, there is a difference in the amount and variability of training data: in case of MIDV-500, the number of training images is two times less, and some of them are similar in many respects. Thus, to reduce the effect of overfitting and improve performance, we reduced the number of training parameters in the network by reducing the number of filters in all convolutional layers by 2 times.

The model was trained from scratch using MIDV-500 templates as a training set along with their newly

annotated PWGT, B_{gt} MIDV-500 contains 50 template images in different quality and resolution, which differ from the MIDV-2020 test set. The intersected set of document types was removed from training data to eliminate the biased estimation. To overcome this, we scaled template images to widths: 930, 1100, 1400, 1800 and 2160 (original aspect ratio was preserved). The obtained images were sliced into 128×128 grayscale patches with step size 64, and also with random shifts from 0 to 32 pixels. During training, we used the following augmentation on the fly: (a) cutting out region with a size of 0.6 to 0.9 from patch size followed by scaling to initial size, probability 0.1; (b) random rotations of 90 or 180 degrees with a probability of 0.1; (c) autocontrast, probability 0.2; (d) adding lines, probability 0.05; (e) Gaussian noise, probability 0.15. The U-Net model was trained for 80 epochs using SGD optimizer (learning rate 1e-6, momentum 0.99) and batch size 128.

The data annotated in the MIDV-2020 is more detailed than in the MIDV-500 even for the same kind of document. The MIDV-2020 contains some extra annotated fields, and some of them are difficult to binarize, for instance, holographic texts and vertically oriented texts. In this work, this set of fields is called "problematic" and its complement – "regular". Let denote the whole set of all textual fields as $\mathbb{F}_B^{All}$ and the set of "regular" textual fields as $\mathbb{F}_B^{Reg}$. In this experiment, error is measured over these two groups of fields.

The results of this experiment are shown in Tab. 2. The first row corresponds to the set of all fields, the second one only to "regular" fields.

Tab. 2. Recognition error $E\{R_{PROCR2}\}$ over $\mathbb{F}_B^{All}$ and $\mathbb{F}_B^{Reg}$ sets of fields

Dataset	B_{gt}	B_{id}	B_U^R	B_U	B_{Otsu}
$\mathbb{F}_B^{All}$	N/A	**0.074**	0.089	0.609	0.101
$\mathbb{F}_B^{Reg}$	N/A	**0.044**	0.047	0.602	0.064

It can be observed that even if the retrained network still does not outperform the non binarized images as in the first experiment, Otsu is no longer the one with best binarization results, which may indicate that for this domain, specific data training gets better results than general purpose algorithms.

3.3. Experiment 3: Binarization and recognition of individual ID document fields

To further investigate the behaviour of the PPOCR2 recognizer jointly with the retrained version of U-Net-bin binarizer, another experiment was designed. The recognition error is evaluated for every previously binarized field of each document. It sheds light on how recognition errors behave inside a single document type.

The Finnish ID templates from the MIDV-2020 dataset, denoted as T_{2020}^{Fin}, were chosen as input image data for this experiment. This document type is indicative since it simultaneously contains two kinds of problematic fields: holographic and vertically oriented (see Fig. 4*a*). Finally, $\mathbb{I}_\varepsilon = T_{2020}^{Fin}$, $\mathbb{B}_\varepsilon = \{B_U^R\}$, $\mathbb{R}_\varepsilon = \{R_{PROCR2}\}$. In this experiment, the resulting measurements are integrated over field types.

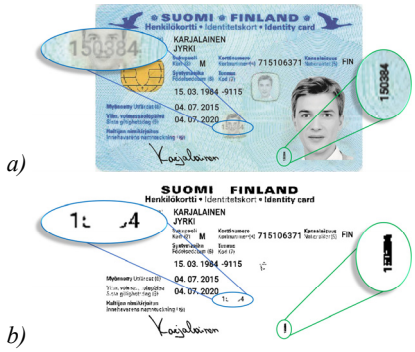

Fig. 4. Holographic (in blue) and vertically oriented (in green) fields: (a) original, (b) binarized

As observed in Tab. 3, the difference in recognition error between problematic and regular fields is significant. It means that there is room for improvement in the binarization and recognition algorithms in these special cases that are common in this domain. This behaviour also supports the results shown in the second row of Table 2, displaying lower recognition error if these fields are not taken in to account.

Tab. 3. Recognition error $E(R_{PPOCR2})$ over B_U^R for T_{2020}^{Fin} dataset

Field name	**Regular field?**	$E(R_{PPOCR2})$
Surname	✓	0.051
Name	✓	0.020
Gender	✓	0.005
Number	✓	0.000
Nationality	✓	0.120
Birth Date	✓	0.167
Issue Date	✓	0.167
Expiry Date	✓	0.167
Code	✓	0.006
Birth Date (holographic)	✗	**0.744**
Birth Date (vertical)	✗	**0.921**

3.4. Experiment 4: Influence of image capture quality on the document attributes recognition

Previous experiments were performed on ID document images with the best possible quality. However, in real-world applications, ID document analysis is often performed using a video stream as input, resulting in varying image quality. A method for a-prioriquality estimation is desirable in these circumstances. Currently, there is no universal solution to this problem, but there are domain-specific methods such as document image quality assessment (IQA) [20] and ID document IQA [38]. However, the goal of this experiment is not to evaluate the quality of the image itself, but to illustrate its influence on the entire DPP the

binarization and recognition steps using the data and methodology established earlier in this work.

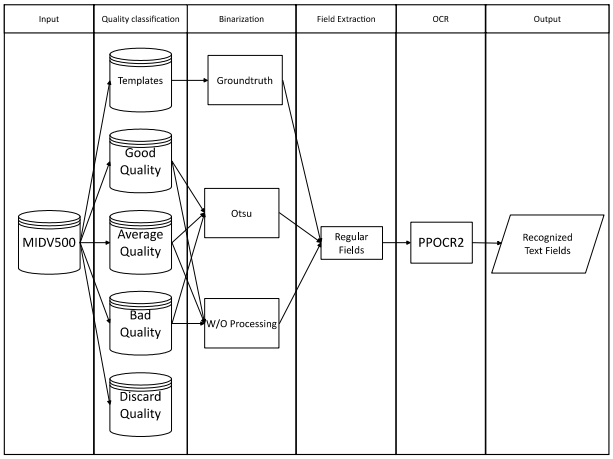

Fig. 5. Description of the fourth experiment

For this experiment, the image set of of all video frames from the MIDV-500 dataset was used. We conducted an evaluation based on qualitative values determined by expert personnel, aided by algorithms for focus analysis and presence of specular light. Given this evaluation, four document image quality groups are established, which are used to assess the recognition error. All the frames from F_{500} are divided into these four groups: (*a*) "good" – without any visible problem and close to template images quality (Fig. 6); (*b*) "average" – with almost no incidences on their fields (Fig. 6); (*c*) "bad" – with very low photometrical quality, but readable with effort (Fig. 6); (*d*) "discard" – with unreadable fields due to motion blur, occlusion or specular light (Fig. 4). From initial 15000 frames 5476 were discarded, 2294 denoted as "bad", 3160 as "average", and 4070 as "good". The obtained dataset is denoted as F_{500}^{G} and this annotation is available on ftp://smartengines.com/midv-500-extra-annotations. Recognition error is calculated for every group using the same methodology as in the previous experiments. Finally, $\mathbb{I}_{\mathcal{E}} = F_{500}$, $\mathbb{B}_{\mathcal{E}} = \{B_{id}, B_{gt}, B_{Otsu}\}$, $\mathbb{R}_{\mathcal{E}} = \{R_{PROCR2}, R_{SRN}, R_{Tesseract}\}$.

According to the proposed annotation, all non-discarded frames were fed to the top three ranked recognition algorithms from the first experiment. Additionally, all the frames were binarized using Otsu algorithm, since it was the best binarization algorithm in the MIDV-500 templates experiments. The pipeline is presented in the Fig. 13.

The results are presented in Tab. 4. The results measured for the templates from the first experiment are added as baseline reference. As expected, there is a direct relationship between the quality of the frames and the recognition error. The best case scenario are the templates (3 first rows), which depict the best possible capture quality. Other rows indicate how the recognition error behave when the quality is degraded in video streams. The first row corresponds to the binarization ground-truth, representing the ideal binarization output.

Fig. 6. Quality annotated frame samples from MIDV-500 dataset: (a) "Good", (b) "Average", (c) "Bad", (d) "Discard"

Tab. 4. Recognition error E over binarization outcomes for F_{500}^{G} and T_{500}^{*}

Group	Binarization	PPOCR2	SRN	Tesseract
Ground-truth	B_{gt}	0.049	0.056	0.081
Templates	B_{id}	**0.052**	0.060	0.092
Templates	B_{Otsu}	0.102	**0.098**	0.162
"Good" frames	B_{id}	**0.084**	0.099	0.271
"Good" frames	B_{Otsu}	0.286	**0.277**	0.401
"Average" frames	B_{id}	**0.143**	0.161	0.369
"Average" frames	B_{Otsu}	0.449	**0.444**	0.515
"Bad" frames	B_{id}	**0.446**	0.471	0.589
"Bad" frames	B_{Otsu}	**0.686**	0.690	0.689

Even for algorithms like PPOCR2, which obtained the best results in Experiment 1 without binarization, using an appropriate binarization process could improve its results on images with lower quality or captured in the wild. In this context, even the bestquality images have room for improvement in order to achieve the results obtained for the binarization ground truth.

Although the Otsu method obtained good results in the previous experiments, in this one the error rates raised significantly even for the "Good" quality batch, which visually is not so distant from the templates. This is consistent with Otsu's algorithm well-known drawback

when dealing with cluttered or non-uniform backgrounds, as well as its dependence on illumination, focus and photometrical quality. This algorithm is not recommended for processing ID documents captured from video streams.

The overall best recognition algorithm for video frames with lower quality was PPOCR2, which maintains good results even for the "Average" batch.

This analysis can provide basis for future workflows to include IQA as an intermediate step in the recognition pipeline. Also, the negative impact of bad quality samples on the recognition process can justify the need for more robust binarization methods to handle these real scenarios artifacts.

Conclusions

In this paper, we performed a comparative joint study of DIB and OCR stages within the ID document analysis domain. This subject has been poorly addressed in the literature, lacking studies of this type. We conducted our experiments on the ID document image datasets MIDV-500 and MIDV-2020.

Two new ground-truth annotations were obtained: one related to the quality of ID document images captured from video stream and another one is a pixelwise ground truth for 50 good document images. A trained model, specifically for ID documents binarization was obtained, which could serve as baseline for future studies.

We could observe that all recognition algorithms seem to behave better on non-binarized images, except when the input was the image binary ground truth. This means that if binarization algorithms improve their results, they can be helpful for the recognition task. A valuable observation in all the experiments was that for this domain, the PPOCR2 algorithm outperformed all the other evaluated methods in terms of recognition rate. It was shown that Otsu algorithm outperformed all DL methods in many cases while using the image templates, but the domain-specific retrained U-Net network obtained lower error rates than Otsu. As expected, it also improved the rates obtained by the same network pretrained on general data.

We conducted a study regarding the recognition by field within each document (instead of the global document image). We found that fields with holographic and vertical characteristics are the ones with greater influence in dropping the recognition rates. This may indicate that this kind of fields requires especial attention in this research domain.

For ID documents captured from video streams, we measured how quality of frames affects the recognition rates. For good and average image quality groups there is still room for improvement if a good binarization could be obtained from them, since the best recognition methods degrade their results with respect to that obtained over the binarization ground-truth andimage templates (not affected by quality artifacts). In this case, Otsu's algorithm obtained worse error rates, thus we recommend using DL-based solutions. As future research we plan to develop new binarization methods more robust to the quality issues presented in video environments, as well as new methods that take in to account the problematic fields we studied.

References

[1] Doermann D, Tombre K. Handbook of document image processing and recognition. Springer Publishing Company Inc; 2014.

[2] Arlazarov VV, Andreeva EI, Bulatov KB, Nikolaev DP, Petrova OO, Savelev BI, Slavin OA. Document image analysis and recognition: a survey. Computer Optics 2022; 46(4): 567-589. DOI: 10.18287/2412-6179-CO-1020.

[3] Bulatov KB, Bezmaternykh PV, Nikolaev DP, Arlazarov VV. Towards a unified framework for identity documents analysis and recognition. Computer Optics 2022; 46(3): 436-454. DOI: 10.18287/2412-6179-CO-1024.

[4] Arlazarov VL, Arlazarov VV, Bulatov KB, Chernov TS, Nikolaev DP, Polevoy DV, Sheshkus AV, Skoryukina NS, Slavin OA, Usilin SA. Mobile ID document recognition-coarse-to-fine approach. Pattern Recognit Image Anal 2022; 32(1): 89-108. DOI: 10.1134/S1054661822010023.

[5] Arlazarov VV, 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.

[6] Bulatov K, Emelianova E, Tropin D, et al. MIDV-2020: A comprehensive benchmark dataset for identity document analysis. arXiv Preprint. 2021. Source: ⟨https://arxiv.org/abs/2107.00396⟩.

[7] Sánchez-Rivero R, Bezmaternykh P, Morales-González A, Silva-Mata FJ, Bulatov K. Assessing the relationship between binarization and ocr in the context of deep learning-based id document analysis. In Book: Heredia YH, Núñez VM, Shulcloper JR, eds. Progress in artificial intelligence and pattern recognition. Cham: Springer International Publishing; 2021: 134-144.

[8] Lins RD, Almeida MMD, Bernardino RB, Jesus D, Oliveira JM. Assessing binarization techniques for document images. DocEng 2017: Proc 2017 ACM Symposium on Document Engineering 2017: 183-192. DOI: 10.1145/3103010.3103021.

[9] Mustafa WA, Kader MMMA. Binarization of document images: A comprehensive review. J Phys: Conf Ser 2018; 1019: 012023. DOI: 10.1088/1742-6596/1019/1/012023.

[10] Tensmeyer C, Martinez T. Historical document image binarization: A review. SN Comput Sci 2020; 1(3): 173. DOI: 10.1007/s42979-020-00176-1.

[11] Pratikakis I, Zagoris K, Barlas G, Gatos B. Icfhr2016 handwritten document image binarization contest (h-dibco 2016). 2016 15th Int Conf on Frontiers in Handwriting Recognition (ICFHR) 2016: 619-623.

[12] Pratikakis I, Zagoris K, Karagiannis X, Tsochatzidis L, Mondal T, Marthot-Santaniello I. Document image binarization (dibco 2019). 2019 Int Conf on Document Analysis and Recognition (ICDAR) 2019: 1547-1556. DOI: 10.1109/ICDAR.2019.00249.

[13] Smith EHB. An analysis of binarization ground truthing. Proc 8th IAPR Int Workshop on Document Analysis Systems (DAS '10) 2010: 27-34. DOI: 10.1145/1815330.1815334.

[14] Ntirogiannis K, Gatos B, Pratikakis I. Performance evaluation methodology for historical document image

binarization. IEEE Trans Image Process 2013; 22(2): 595-609. DOI: 10.1109/TIP.2012.2219550.
[15] Rani U, Kaur A, Josan G. A new binarization method for degraded document images. Int J Inf Technol 2019; 15(1): 1035-1053. DOI: 10.1007/s41870-019-00361-3.
[16] Milyaev S, Barinova O, Novikova T, Kohli P, Lempitsky V. Image binarization for end-to-end text understanding in natural images. 2013 12th Int Conf on Document Analysis and Recognition 2013: 128-132. DOI: 10.1109/icdar.2013.33.
[17] Chou C-H, Lin W-H, Chang F. A binarization method with learning-built rules for document images produced by cameras. Pattern Recogn 2010; 43(4): 1518-1530. DOI: 10.1016/j.patcog.2009.10.016.
[18] Wen J, Li S, Sun J. A new binarization method for non-uniform illuminated document images. Pattern Recogn 2013; 46(6): 1670-1690. DOI: 10.1016/j.patcog.2012.11.027.
[19] Tafti AP, Baghaie A, Assefi M, Arabnia HR, Yu Z, Peissig P. OCR as a service: An experimental evaluation of google docs OCR, tesseract, ABBYY FineReader, and transym. In Book: Bebis G, Boyle R, Parvin B, Koracin D, Porikli F, Skaff S, Entezari A, Min J, Iwai D, Sadagic A, Scheidegger C, Isenberg T, eds. Advances in visual computing. Cham, Switzerland: Springer International Publishing AG; 2016: 735-746. DOI: 10.1007/978-3-319-50835-1_66.
[20] Li Z, Yang C, Shen Q, Wen S. A document image dataset for quality assessment. J Phys: Conf Ser 2021; 1828(1): 012033. DOI: 10.1088/1742-6596/1828/1/012033.
[21] Ye P, Doermann D. Document image quality assessment: A brief survey. 2013 12th Int Conf on Document Analysis and Recognition 2013; 723-727. DOI: 10.1109/ICDAR.2013.148.
[22] Polevoy DV, Bulatov KB, Skoryukina NS, Chernov TS, Arlazarov VV, Sheshkus AV. Key aspects of document recognition using small digital cameras. RFBR J 2016; 4: 97-108. DOI: 10.22204/2410-4639-2016-092-04-97-108.
[23] Chernov T, Ilyuhin S, Arlazarov VV. Application of dynamic saliency maps to the video stream recognition systems with image quality assessment. Proc SPIE 2019; 11041: 110410T. DOI: 10.1117/12.2522768.
[24] Shemiakina J, Limonova E, Skoryukina N, Arlazarov VV, Nikolaev DP. A method of image quality assessment for text recognition on camera-captured and projectively distorted documents. Mathematics 2021; 9(17): 2155. DOI: 10.3390/math9172155.
[25] Bezmaternykh PV, Ilin DA, Nikolaev DP. U-Net-bin: hacking the document image binarization contest. Computer Optics 2019; 43(5): 825-832. DOI: 10.18287/2412-6179-2019-43-5-825-832.
[26] Calvo-Zaragoza J, Gallego AJ. A selectional auto-encoder approach for document image binarization. Pattern Recogn 2019; 86: 37-47. DOI: 10.1016/j.patcog.2018.08.011.
[27] Masyagin M. Robust document image binarization tool. 2021. Source: ⟨https://github.com/masyagin1998/robin⟩.
[28] Otsu N. A threshold selection method from gray-level histograms. IEEE Trans Syst Man Cybern Syst 1979; 9(1): 62-66. DOI: 10.1109/TSMC.1979.4310076.
[29] Lins RD, Simske SJ, Bernardino RB. Doceng'2020 time-quality competition on binarizing photographed documents. Proc ACM Symposium on Document Engineering 2020; 2020: 2. DOI: 10.1145/3395027.3419578.
[30] Yu D, Li X, Zhang C, Liu T, Han J, Liu J, Ding E. Towards accurate scene text recognition with semantic reasoning networks. Computer Vision and Pattern Recognition (CVPR) 2020: 12113-12122.
[31] Du Y, Li C, Guo R, Cui C, Liu W, Zhou J, Lu B, Yang Y, Liu Q. Pp-ocrv2: Bag of tricks for ultra lightweight ocr system. arXiv Preprint. 2021. Source: ⟨https://arxiv.org/abs/2109.03144⟩.
[32] Lee J, Park S, Baek J, Oh SJ, Kim S, Lee H. On recognizing texts of arbitrary shapes with 2d self-attention. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition Workshops 2020: 546-547.
[33] Baek J, Kim G, Lee J, Park S, Han D, Yun S, Oh SJ, Lee H. What is wrong with scene text recognition model comparisons? dataset and model analysis. 2019 IEEE/CVF Int Conf on Computer Vision (ICCV) 2019: 4714-4722. DOI: 10.1109/ICCV.2019.00481.
[34] Cai H, Sun J, Xiong Y. Revisiting classification perspective on scene text recognition. arXiv Preprint. 2021. Source: áhttps://arxiv.org/abs/2102.10884ñ.
[35] Smith R. An overview of the tesseract ocr engine IEEE Int conf on Document Analysis and Recognition (ICDAR'07) 2007; 2: 629-633. DOI: 10.1109/ICDAR.2007.4376991.
[36] Michalak H, Okarma K. Robust combined binarization method of non-uniformly illuminated document images for alphanumerical character recognition. Sensors 2020; 20(10): 2914. DOI: 10.3390/s20102914.
[37] 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.
[38] Schulz D, Maureira J, Tapia J, Busch C. Identity documents image quality assessment. 2022 30th European Signal Processing Conf (EUSIPCO) 2022: 1017-1021.

Authors' information

Rubén Sánchez-Rivero, (b. 1993), graduated in Software Engineering from the Havana University of Technologies "José Antonio Echeverría", Cuba, in 2017. He is a researcher at the Advanced Technologies Application Center (CENATAV). He is author of more than 5 scientific publications. Research interest: computer vision, digital image processing, document analysis. E-mail: *rsanchez@cenatav.co.cu* .

Pavel Vladimirovich Bezmaternykh, (b. 1987), received a specialist degree in Applied Mathematics from the Moscow Institute of Steel and Alloys in 2009. Since 2016 he is employed at Smart Engines Service LLC, and since 2019 he is employed at the FRC "Computer Science and Control" of RAS. He is an author of more than 10 scientific publications. Research interests: image processing, document recognition. E-mail: *bezmaternyh@isa.ru* .

Alexander Vyacheslavovich Gayer, (b. 1995), received a master degree in Applied Informatics from the National University of Science and Technology "MISiS" in 2019. Since 2017 he is employed at Smart Engines Service LLC, and

since 2019 he is employed at the FRC "Computer Science and Control" of RAS. He is an author of more than 5 scientific publications. Research interests: computer vision, deep learning, object detection.
E-mail: agayer@smartengines.com .

Annette Morales-González, (b. 1982) graduated in Software Engineering (2005) and received her Ph.D. in Technical Sciences (2014) from the Havana University of Technologies "José Antonio Echeverría", Cuba. She is a researcher at the Advanced Technologies Application Center (CENATAV). She has authored more than 30 scientific publications. Research interests: computer vision, video surveillance, biometrics. E-mail: amorales@cenatav.co.cu .

Francisco José Silva-Mata, (b. 1959) graduated in Electronic Engineering (1982) and received his Ph.D. in Technical Sciences (2017) from the Havana University of Technologies "José Antonio Echeverría", Cuba. He is a researcher at the Advanced Technologies Application Center (CENATAV). He has authored more than 40 scientific publications. Research interests are image processing, computer vision. E-mail: fjsilva@cenatav.co.cu .

Konstantin Bulatovich Bulatov, (b. 1991) received a specialist degree in Applied Mathematics from the National University of Science and Technology "MISiS" in 2013. He obtained his Ph.D. degree in 2020 from the FRC "Computer Science and Control"of RAS. Since 2014 he is employed at the FRC "Computer Science and Control" of RAS and since 2016 he is employed at Smart Engines Service LLC. He is the author of more than 30 scientific publications. Research interests: computer vision, image processing, and document recognition systems.
E-mail: kbulatov@smartengines.com .

Received September 13, 2022. The final version – February 20, 2023.