Towards a unified framework for identity documents analysis and recognition

K.B. Bulatov [1,3], P.V. Bezmaternykh [1,3], D.P. Nikolaev [2,3], V.V. Arlazarov [1,3]
[1] Federal Research Center "Computer Science and Control" of RAS, Moscow, Russia;
[2] Institute for Information Transmission Problems of RAS (Kharkevich Institute), Moscow, Russia;
[3] Smart Engines Service LLC, Moscow, Russia

Abstract

Identity documents recognition is far beyond classical optical character recognition problems. Automated ID document recognition systems are tasked not only with the extraction of editable and transferable data but with performing identity validation and preventing fraud, with an increasingly high cost of error. A significant amount of research is directed to the creation of ID analysis systems with a specific focus for a subset of document types, or a particular mode of image acquisition, however, one of the challenges of the modern world is an increasing demand for identity document recognition from a wide variety of image sources, such as scans, photos, or video frames, as well as in a variety of virtually uncontrolled capturing conditions. In this paper, we describe the scope and context of identity document analysis and recognition problem and its challenges; analyze the existing works on implementing ID document recognition systems; and set a task to construct a unified framework for identity document recognition, which would be applicable for different types of image sources and capturing conditions, as well as scalable enough to support large number of identity document types. The aim of the presented framework is to serve as a basis for developing new methods and algorithms for ID document recognition, as well as for far more heavy challenges of identity document forensics, fully automated personal authentication and fraud prevention.

Keywords: optical character recognition, document recognition, document analysis, identity documents, recognition system, mobile recognition, video stream recognition.

Citation: 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.

Acknowledgements: This work was partially supported by the Russian Foundation for Basic Research (Project No. 18-29-03085 and 19-29-09055).

Introduction

OCR and document image analysis

Currently, Optical Character Recognition, or OCR, is a widespread term, but it is rather ambiguous. It originally denoted the problem of isolated character recognition: determining which character from a predefined alphabet is depicted in the provided image region. According to Eikvil [1], the first implementations of such recognizers were able to deal only with a single alphabet typed with a single specific font. The recognizers capable of handling multiple fonts and alphabets were highly desirable and appeared later, in the mid-sixties. Evidently, the focus of interest was not typically in recognition of isolated characters, but in the words or text fragments. Here, the importance of language models stepped in and the complexity of recognizers greatly increased. The path from the words recognition to whole page processing and layout analysis was short too. At that point, many page-by-page applications were in use, such as digitization of books or papers, check processing or postal recognition. The next step was to process structured documents: bank forms, invoices, questionnaires, tables, and many others. Thus, a consideration of the full document context became essential, and a field of Document Image Analysis has been established. It relates to many disciplines, such as image processing, pattern recognition, language theory and attracts a lot of attention. A consolidated source of this domain expertise is a volume edited by D. Doermann and K. Tombre [2].To sum it up, at the present time the term OCR mainly refers to a document image analysis technology that enables to transform various types of whole documents into transferable, editable and searchable data, thus defining a process that is far beyond original isolated characters recognition.

Among the rich variety of document types, identity documents play a crucial role. Many OCR systems are either tweaked, fine-tuned, or specifically designed to process these kinds of documents because a lot of specific information must be taken into account during their recognition and analysis. Such systems rely both on classical computer vision and document image analysis methods, as well as on recent advances in machine learning and deep learning-based methods.

The task of identity documents processing is complicated by a lot of issues (see fig. 1), not only related to the OCR, but to additional requirements often set for identity analysis systems, such as the requirements for validating and authenticating a document; the issues related to the different input sources and their characteristics; as well as various target devices on which such system should be executed. In this work, we will consider the issues facing

the identity documents recognition systems and explore the possibility of constructing a unified framework which would be capable of addressing them in a cohesive way.

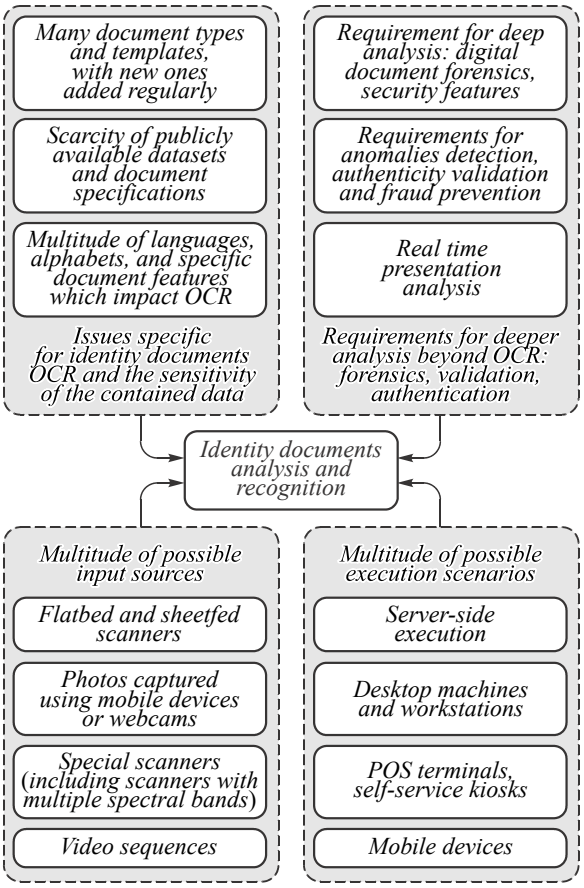


Fig. 1. Identity documents analysis and recognition problem, and the issues facing identity document processing systems

Recognition of identity documents

An identity document (ID document) is any document that can be used to confirm its owner's person and to prove their identity. Among the document fields, it has to contain various details about the person. The most common are the person's full name, birth date, address, identification number, gender, photo and an image of the personal signature. Some of these details are directly printed on the document, some may be represented in the machine-readable zone (MRZ) [3], [4] or encoded in a barcode form. Examples of identity documents are presented in fig. 2.

In general, ID documents can have a rather simple layout, especially if it does not contain many fields – an example could be custom identity badges that are issued by employers to their employees. However, in some areas, the ID documents are required to be more complex due to security concerns. Government-issued ID is often equipped with visual security elements and special non-visible tags such as RFID (Radio-frequency identification) or microchips with biometric data [5]. The emission of such documents is strictly regulated by governmental and executive organizations. A unique national identification number is usually assigned to a person and is embedded into the document, often alongside some biometric information, such as iris or a fingerprint [6], or with codes to access national biometric databases, as in the case of India's Aadhaar document [7].

The most common format of an identity document is an ID card, which corresponds in physical dimensions to a regular bank card. There is a series of international standards describing the characteristics of such documents, for instance, ISO / IEC 7810:2003 [8], which provides requirements for the physical features of an identity card. These standards became necessary due to widespread ID cards usage in multiple countries and are aimed to unify their characteristics and facilitate their processing. Another important identity document format is a passport or a travel document. Passports are usually issued in the form of a booklet, often equipped with embedded microchips for machine reading.

The amount of different identity documents issued around the world, naturally, is very vast. Special databases with examples of ID documents are collected and some are publicly available, such as the Public Register of Authentic identity and travel Documents Online (PRADO) [9]. However, only a portion of the total ID document variety can be found there. Moreover, many countries are subdivided into regions or states which are allowed to issue their own sets of document variations. For instance, each state of the United States of America issues a custom driver's license document and an American Association of Motor Vehicle Administrators (AAMVA) has to take care of their unification and control [10]. Besides, identity documents are changing from time to time due to renewed design and enhanced security requirements.

The scope of usage for automated systems for ID document data entry, from the industry perspective, could be divided into three main case groups. The first group is related to “offline” ID document processing automation. Here, the physical presence of a person together with their ID document is obligatory and the identity verification is carried out by trained staff or an operator. Examples of such cases are the offline opening of a bank account, receiving medical service, registration at a hotel desk, etc. A significant subset of such cases comprises variations of physical standardized access control systems in places where access is restricted or strictly regulated (government buildings, warehouses, etc.). ID document recognition systems allow to speed up identity document data entry and verification and facilitate more efficient service provision and queue management.

The second group is remote identity verification, which is a rapidly growing field in many areas of customer service. This group includes online banking, financial services, insurance services, remote government service, Know Your Customer (KYC) procedures, and much more. The physical presence of a person is no longer required, and the human operator either is eliminated or

performs verification through digital channels. Remote ID document verification makes the process more comfortable for many clients, however, at the same time, it raises additional challenges for the ID recognition systems, as well as security and privacy concerns. To pass an authentication step it is not enough to simply recognize the data from an ID document, but possible attempts at ID fraud should be detected and prevented. With the growing spread of remote ID document processing in a broad range of services, the cost of false identification becomes very high.

Fig. 2. Example image of identity documents: (a) Albanian ID card (2004); (b) French ID card (2021); (c) Serbian passport (2008). Images taken from WikiMedia Commons, each is in public domain according to the copyright laws of the corresponding issuing countries as being parts of the official regulation

The third group of cases deals with traveler documents, and it combines the features of the first two. Global international travel ought to be monitored by government services, with officials from various countries checking and validating ID documents issued in other countries. Passing the border control, boarding an airplane, train, or other types of transport, is almost universally accompanied by the necessity to prove the passenger identity, and remote identification processes are introduced in such use cases as well. A large flow of travellers with ID documents from all over the world must be serviced quickly, thus the format of ID documents eligible for usage during international travel is strictly regulated. The International Civil Aviation Organization Traveller Identification Programme (ICAO TRIP) was introduced specifically in order to enhance and regulate every aspect of traveller identification strategy [11]. It comprises five elements: credible evidence of identity; design and manufacture of standardized machine-readable travel documents; document issuance and control; inspection systems and tools; and interoperable applications that provide for quick, secure and reliable operations.

Every citizen holds several identity documents, each serving different purposes. The automatic identity document recognition systems ought to be prepared to process the enormous amount of document types to be scalable

and usable within multiple processes. The total number of ID document types worldwide is hard to estimate exactly, remote identification service providers report their support of 3500 – 6500 types of documents [12 – 14]. Thus, the design of an automatic ID document recognition systems is constrained by the vast number of target document layouts, languages, national specifics, changes in document appearance, as well as unavailability of training data – personal information stored in such documents can not generally be published for these purposes due to privacy and security concerns. In addition, progress in image capturing devices, new types of optical scanners and digital cameras, push the recognition systems to support more and more different types of image capture methods, as well as more challenging uncontrolled capturing conditions. Another aspect of identity document processing systems is the device they are executed on: while server-side systems and systems oriented on regular desktop machines and workstations do not lose their wide applicability, identity document processing has also became relevant when executed on low-end point of service terminals and mobile devices, where the computational power is severely limited, while transfer of images could be undesirable or impossible. Thus, it becomes apparent that it is necessary to formulate and formalize the methods and approaches to identity documents analysis and recognition in order to facilitate both theoretical research and practical implementation of modern, robust, and scalable recognition systems.

Related works

The task of automatic data extraction from images of identity documents became a topic of research in the early 2000s, in order to facilitate more efficient data entry and verification of personal information in such cases as checking-in in a hotel, boarding an airplane, and more [15]. While at first the major mode of input were flatbed and specialized scanners, camera-based document recognition also became the topic of study in the last 10 years [16] due to a wide spread of portable cameras and mobile devices such as smartphones. In recent years, a number of works have been published which describe systems and frameworks for identity documents recognition.

In [15] the task of identity cards recognition from images obtained using a flatbed scanner was proposed. The document was detected and deskewed in an image using Hough transform, and the document type identification was performed using colour histogram classification. Text detection was performed using connected components analysis, and OCR was performed given binarized text images, followed by post-processing with geometric and linguistic context.

Works [17, 18], and [19] describe systems for recognition of Indonesian identity cards. The workflow described in [17] is targeted on camera-captured documents, its processing steps include scaling, greyscaling and binarization of the document images, extracting of the text areas using connected component analysis, histogram-based per-character segmentation of text lines and template-based OCR. In [18] the characters of Indonesian identity cards were recognized using CNNs (Convolutional Neural Networks) and SVMs (Support Vector Machines) with pre-processing. The system described in [19] includes smoothing as one of the image pre-processing steps, morphological operations for text fields detection and uses Tesseract [20] for text line recognition. A similar workflow is described in [21] with regards to the camera-based recognition of various identity documents of Italy, however as a pre-processing step for document detection and identification vertices detection and analysis is used with CNN-based document type classification.

Works [22] and [23] describe systems for Vietnamese identity documents text fields detection and recognition. The pre-processing steps in [23] include greyscaling, tilt correction, smoothing and binarization, and the text fields are separately detected with ID card number detection on the one side of the document, and table structure analysis for the other side. Image pre-processing step in [22] include preliminary projective alignment of the camera-captured document using corner detection, corner classification, and geometric heuristics, SSD Mobilenet V2 [24] was used for text detection and Attention OCR architecture [25] was used for text lines recognition.

Papers [26, 27] and [28] describe systems of identity documents recognition with a focus on Chinese identity cards in a camera-based setting. The systems feature tilt correction using Hough transform, document image pre-processing steps such as brightness adjustment and greyscaling, projections and morphology-based text detection and text recognition using CNNs [26, 27] or template-based and SVM-based OCR [28]. Document type identification described in [26] also uses national emblem detection using AdaBoost-trained detection over Haar features.

The paper [29] describe a system for identity document analysis evaluated on ID cards of Colombia. The target goal of the described workflow is authentication of the ID card, and it includes deep learning-based background removal, corners and contours detection for projective alignment, checking of brightness and color coherence and aggregation greyscale histogram, face location, connected colour components and structural similarity markers for document authentication.

A crucial issue related to the research of new methods and algorithms for identity document processing is the availability of public datasets. Since identity documents contain personal information, the existence of a public dataset of real documents is impossible. Thus, to facilitate reproducible scientific research, a number of synthetic datasets of identity documents were developed in recent years and continue to be published. These include the Mobile Identity Document Video dataset family (MIDV) [30 – 32], which contain video clips of 50 identity document samples obtained from public sources and captured in various conditions, and

Brazilian Identity Document Dataset (BID Dataset) [33], which consists of images of Brazilian ID documents with blurred personal data and populated with synthetically generated fields. Some datasets were prepared to target specifically the task of identity document detection, such as LRDE Identity Document Image Database (LRDE IDID) [34], and more broadly targeted datasets of the document analysis community, such as the ones from SmartDoc family [35] feature examples of identity documents.

While the methods used to perform the individual processing steps differ from system to system, the general composition and ideas are shared. A typical composition of such identity document recognition system consists of the following steps (see fig. 3):

1) pre-processing of an input image. This step includes general image pre-processing steps, such as downscaling or colour scheme conversions, background removal, detection of contours, edges, and corners, or semantic segmentation;
2) preparation of the document image. This includes geometric rectification (tilt correction, projective restoration), brightness adjustments, etc.;
3) extraction of text fields and other important elements of identity documents, such as the photo of a face;
4) text fields pre-processing (such as binarization and skew correction), per-character segmentation, recognition and post-processing using language models.

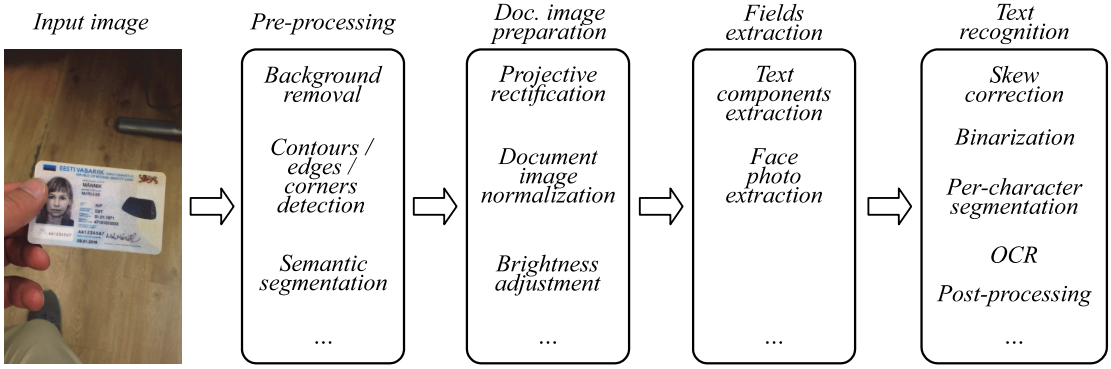


Fig. 3. General scheme of the identity document recognition pipeline

Beyond OCR

The task of ID document processing within automated systems is not limited only to text fields recognition. In addition to data extraction, an important aspect of identity processing is the validation of the document authenticity – determining whether a document is genuine or counterfeit or at least detecting anomalies in the images of the documents, which would indicate possibly malicious intent. These issues become more urgent in relation to the domain of remote identification and given the availability of a multitude of tools that allow to alter images or physical documents. This problem belongs to the field of digital image and document forensics, which is a subject of a separate branch of research [36 – 39]. In general, the tasks for digital document image forensics could be divided into three groups.

The first type of tasks is to confirm that the provided document contains the security features determined by its specification. The taxonomy of such features alongside some examples are described in PRADO's glossary [40]. The list includes such features as coloured security fibres, rainbow colouring, guilloche, holograms, bleeding inks, etc. The lack of these features helps to reveal illicit usage of fake documents or photocopied versions of fraudulently obtained real documents. Some security features, such as holographic security elements, may require video processing for reliable analysis [41]. Other features, such as coloured security fibres, may require high quality of source images or specific illumination for their successful detection [42].

The second group of tasks is related to the identification and validation of the image source. A remarkable example of impersonation attacks during remote identification is a presentation of recaptured images of an ID document. There is extensive research on this topic and several approaches to recaptured images detection [43, 44]. Advanced image editing software made the task of document image editing or tampering almost trivial. Through these tools, a problem of image region copy-pasting detection became important for ID documents analysis. Some methods designed to detect such attacks are presented in [45, 46]. Another approach that is used to validate document image source is estimation and control of lighting [47].

The third group of tasks deal with document content analysis in order to reveal data manipulation. ID document designers often introduce data duplication which can help to cross-check important textual fields and validate their correctness. Some document fields such as MRZ contain check digits [3] which should be verified. The usage of specific fonts, such as OCR-B [48] is required in many document types and therefore it is possible to validate font characteristics during document image analysis.

Many ID documents contain a photograph of the document holder's face. Forensic face matching is a

technology aiming to compare the person's face with the one depicted on their ID. The technology has been known for a long period [49], however the recent advances in face recognition domain made it much more practical. A review of forensic face matching is presented in the paper [50].

Given the multitude of security features used to manufacture ID documents and an increasing number of potential ways of attacking the presentation of ID document within identification processes, a high-end ID recognition system today is virtually inconceivable without application of digital image forensics methods.

Towards a unified framework

The published works on the composition of identity documents recognition, for the most part, consider a subset of identity document types (most commonly the types of one specific country); focus on a single particular type of input data (only scanned images, or only camera-based photographs) or imply a specific mode of system's execution; or do not target additional tasks of ID document image analysis, such as document forensics, anomaly detection and authenticity validation. In the paper [51], an identity documents recognition system is described which additionally considers the recognition in a video stream, with a per-frame combination of accumulated information. This article extends and revises the ideas of the system presented in [51], and sets a goal of developing a unified framework for identity document recognition and analysis, which would be applicable for multiple modes of image capture, such as scans, photos, or sequences of video frames; rise to the challenges of scalability; and serve as a basis for developing new methods and algorithms for solving not only the recognition problems but also more sophisticated challenges of identity document image forensics, automated personal authentication and fraud prevention.

1. Input characteristics

Let us consider the task of identity documents recognition with respect to the way the input image (or images) are produced.

Fig. 4. Examples of identity document images captured using a flatbed scanner [32]

Scanners

Traditionally documents are digitized using various types of scanners, such as flatbed, sheet-fed, or specialized ones [15, 52]. The flatbed scanners are used to capture identity document images in a lot of corporate and governmental use cases, where the time required to capture the image is less important than the cost of the required hardware. Flatbed scanners are usually designed to be able to scan documents with standard page sizes (such as B5, A4 or US Letter), thus the resulting image is typically larger than required for identity document processing. A comparatively small identity document (such as an ID card or passport) could be arbitrarily shifted or rotated in an image obtained using a flatbed scanner, as the strict compliance with document positioning on the part of the user is hardly enforceable. A class of specialized small-scale flatbed scanners designed for identity documents exists [53], however even using the small-scale scanners the document could still be shifted or slightly rotated when placed on the scanning surface. Thus, even the use of small-scale flatbed scanners does not change the geometric model, only imposes some additional constraints.

Sheet-fed scanners are typically used for batch scanning of multiple separated pages. They offer increased scanning speed, however are rarely applicable for the task of identity document processing. A separate class of sheet-fed scanners exists [54] which are designed to process identity cards and driving licences, however, the time required to produce a high-resolution image for such scanners is comparable with their flatbed analogues. Although, an important advantage of sheet-fed scanners is their ability to scan both sides of the card-like document. Book-like identity documents, such as passports, can not, in general, be captured using sheet-fed scanners, due to a risk of damaging the document.

For scanning identity documents in such cases as access control systems, border control applications, ticket sales kiosks or self-service kiosks for purchasing age-restricted products, a class of specialized identity document scanners is designed [52], [55 – 57]. The main motivation for this class is to reduce the time required to acquire the document image, while retaining the high resolution and provide additional functionality, such as reading an RFID chip containing biometric information, capturing infrared document images, or images under ultraviolet light (see fig. 5). Such specialized scanners are typically using a camera which allows to quickly obtain a high-resolution photo of a document with controlled lighting conditions. The camera could either point directly at the scanning surface, point at it through an angled mirror, or have an angular skew, depending on the particular method of minimizing the effect of highlights, glares, and data obstruction due to the holographic layer.

Though the size-scanning surface of the specialized scanners usually corresponds to the size of the target

identity document, some minor shifts and rotations are still possible, in the same way as with small-scale flatbed scanners mentioned above. Thus, the geometric model stays the same as with flatbed and small-scale sheetfed scanners, and the general characteristics of the document detection location problem do not change, only the capturing method does.

Fig. 5. Example of a (fake) identity document captured using a specialized scanner in three bands: (a) visible light, (b) infrared, and (c) ultraviolet

It is worth noting that one of the important problems which could present itself for a document recognition system, which deals with scanned images, is that the image resolution with respect to the captured document is not always known. For integrated systems, where image processing software and capturing hardware are both controlled components, the image resolution is known and regulated, whereas if the input images are obtained remotely, or preprocessed by an uncontrolled party (such as in the case of images uploaded by a remote operator or an end-user or a service), the scale might be unknown in advance.

Photographs

Global leap in communications and mobile technologies and the increased demand for fast and convenient provision of services, such as government, banking, insurance, and others, have led to the requirement to process images of documents remotely, with their images captured by end-users and uploaded for processing. Not all end-users have instant access to scanners, whereas mobile cameras are readily and almost universally available, and the quality of such cameras allow to obtain images of documents of sufficient quality, that is, enough at least for human analysis.

This trend led to a requirement for automatic document analysis and recognition systems to support photo inputs [17, 21, 26]. The complications of such input, in comparison with a scanned image, are quite numerous. Firstly, with a scanned image the background is typically homogeneous, whereas the photo could be made on an arbitrary one (see fig. 6). Different and uncontrolled background may prove to be an obstacle for precise detection and location of the document, especially if it is cluttered, has many high contrast lines or local regions, or has text which could be confused with parts of the document by the detection algorithm. The second important complication is uncontrolled lighting conditions. Images obtained from flatbed or sheet-fed scanners are always uniformly illuminated, specialized scanners typically have an integrated lighting system designed to control the illumination of the captured image, whereas the user-captured photo could have weak and inconsistent illumination, be over or underexposed [58]. Inconsistent lighting could present problems for the detection and location of the document on the photo, as well as for document layout analysis, text recognition, and other components of the automated document analysis system [31, 59]. Another serious problem with document images captured using cameras are the possibility of them being out of focus [60, 61], or containing motion blur [62].

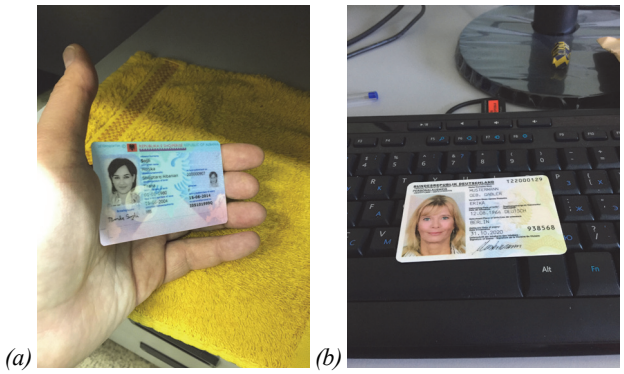
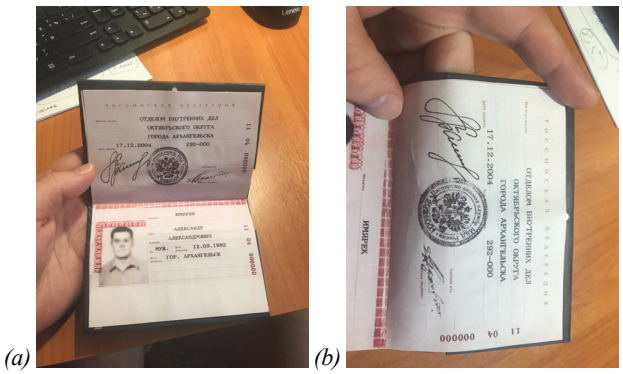


Fig. 6. Examples of identity document photos capturing using a smartphone [32]

Perhaps the most important distinction between images obtained using scanners and cameras is the geometric position of the document in an image. As was mentioned above, a small-scale document, such as an ID card or a driving li-

cence, could be rotated or shifted in an image obtained using a scanner, which limits the family of geometric transformations of a document to a subset of affine ones. If the document is captured using a web camera or a mobile device camera, the document could be rotated along the three Euler angles with respect to the optical system. This could be done both unintentionally and intentionally, for example, in an effort to prevent highlights on a reflective document surface. If a camera is regarded within a pinhole model, the family of possible geometric transformations of the document is now a subset of projective ones, which significantly complicates the task of precise document location [63, 64]. There could be even several projective transformations for different parts of the document in a single image, such as in the case of capturing both pages of a book-like document, e.g. a passport (see Fig. 7*a*).



(a) *(b)*

Fig. 7. Example photographs of Russian internal passport: (a) book-like spread, (b) warped page

Since the parameters of the camera lens could be unknown, the document images could also be affected with radial distortion [65]. Finally, if the document itself is not rigid, it may be subject to deformations of the document's medium itself, such as bending of paper pages of a passport (see fig. 7*b*).

Videostream

The usage of web cameras and mobile devices for capturing images of documents led to another mode of input data acquisition – using a sequence of video frames, or a video stream, instead of a single photo [35]. One of the considerations for using a video stream as an input is that it makes the input more resistant to tampering, as the video is harder to falsify in comparison with a single uploaded image, especially if the document analysis procedure is performed in real-time. From a document analysis and recognition perspective, using multiple input images of the same object presents several advantages: filtering and refinement techniques could now be employed for improving object detection and location accuracy [32, 66], the so-called "super-resolution" techniques [67] could be employed for obtaining images of higher quality, and text recognition results could be improved by means of accumulating per-frame recognition results in a single most reliable one [68]. Fig. 8 presents examples of video frames of an ID document.

From the scene geometry point of view, the video stream input has the same characteristics as single photos captured using a camera – the document could be arbitrarily positioned in the frame, and rotated along the three Euler angles with respect to the optical system. The document could be placed against an arbitrary and uncontrolled background, be inconsistently illuminated or blurry [31]. The geometric characteristics of the scene, along with such image properties as blur, lighting, presence of highlights, and others, could change from frame to frame. The addition of a temporal axis to the document recognition system input introduces redundancy, which could be exploited in order to increase the automatic analysis quality. Besides, if a video is regarded as a visual representation of an identification document instead of a single image, it can now be used to identify document elements that could hardly be detectable in a single image: holographic security elements and other optically variable devices (OVDs) [41]. Processing of multiple video frames, with the analysis of changes between the consecutive frames, is the only way, almost by definition, of accurately detecting the OVDs and distinguishing them from simple printed color regions of the document.

Fig. 8. Examples of video frames of an identity document capture, with changing scene from frame to frame

Recognition in 2D, 3D, and 4D

Having described the problem of identity document analysis with respect to the types of graphical input it receives, we can formulate an intuitive classification of the document recognition problem variations. Based on the input type the task can be roughly described as either "2D", "3D", or "4D" document recognition (see fig. 9).

The task of "2D" recognition deals with input images typically obtained using scanners and besides the common subtasks of visual document elements identification, layout analysis, text fields recognition, etc., it needs to deal with sometimes unknown image resolution, along with arbitrary shift and rotation of the document in an image. If the document image is obtained using a web camera or a camera of a mobile device, instead of analyzing a scanning surface the system has to analyze a three-dimensional scene, where the document has to be found taking into account projective transformations, possible non-linear distortions, and arbitrary background ("3D" recognition). Further document analysis methods need to deal with possible defocus, blur, inconsistent or inade-

quate illumination, and highlights on the reflective surface of the document. Finally, if the input consists not of a single image, but rather a sequence of video frames, the document needs to be tracked in time ("4D" recognition), with regards to changing capturing conditions. However, the visual information redundancy may be exploited in order to increase the reliability of the document analysis and recognition results, and the changing conditions may be utilized to detect and analyze optical variable devices, which are extensively used for identity document security.

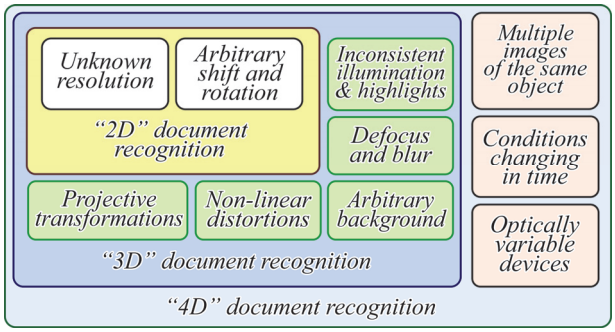


Fig. 9. Problems and features of "2D", "3D", and "4D" identity document recognition systems

It is worth mentioning that there are specific cases that could be tricky to classify with such descriptions of "2D", "3D", and "4D" document recognition systems. One example of such case is specialized video scanners [56] which use multiple frames in a "2D" setting and exploit the differences between the input images (originating from small re-positioning of the document on the scanning surface, or simply from a digital noise) to combine the per-frame results and improve the recognition accuracy. However, since the purpose of this paper is to construct a unified framework for an identity document analysis and recognition system which could encompass all of the described problem variations, the provided classification is sensible.

With different modes of acquiring input images or video frames, different models of a geometric position of the document and different sets of complications, such as the presence of blur, highlights, unconstrained lighting, etc., the actual target of the recognition – the identity document – remains the same, in terms of its structure and content. Hence, a universal framework for automatic data extraction from identity documents is needed, which would take into account the specifics of the target documents and would be applicable for different capturing modes. The individual components of such framework and their interrelation should allow for supporting a multitude of identity document types, and by richer specification of individual components it should be possible to add sophisticated document analysis methods, such as image forensics, into the processing pipeline.

2. Proposed framework

In this section, we will describe a novel framework for the automatic recognition of identity documents with a fixed layout, captured from a variety of input sources. The section is organized as follows: firstly, we define what is meant by "document" within the proposed framework; secondly, general composition of the proposed framework and its components is provided, with the description of their utility in the scope of "2D", "3D", and "4D" document recognition modes.

Documents with fixed layout

An identity document is a physical object, designed and issued by a legitimate authority according to a predefined set of rules and regulations, which purpose is to carry identification information of a specific person. From the perspective of a recognition system, a document is regarded as a logical entity comprising a set of named fields and elements, each with a clear semantic meaning. The basic high-level component of visual document representation is denoted as a "template" – a planar rectangular document page distinguishable by its static elements, such as background, immutable text, fiducial elements, national emblems, etc., as well as their positions on the page [69]. A class of documents with fixed layouts (in some literature referred to as semi-structured documents [70]) is characterized by the following three properties of their templates:

1) The positions and appearance of static template elements do not change between instances of the same template;
2) The positions and appearance of static elements of a template differ from those belonging to the other templates of the same document type, as well as from those belonging to other document types, and thus can be used to uniquely identify the template and the document type it represents;
3) The template defines the set of data objects (such as text fields) that can be extracted from the image of the template, along with the information of their locations and structure.

Fig. 10 shows as an example the main spread of the Russian national passport, which is used as the main identification document in the Russian Federation, composed of two rectangular templates.

Fig. 10. Example of Russian national passport with handwritten fields. The main spread is composed of two pages (2 and 3), each treated as a separate document template

The identity document recognition system presented in this paper assumes that in the input image there is only one document of the same type: if the system finds multiple templates and each of them could be a page of the same document type, then they are treated as different pages of the same physical document. Moreover, since the input of the system could be a sequence of images (e.g. video frames), the system assumes that all templates visible on all frames within a single recognition session correspond to the same document.

System composition

The general composition of the ID document recognition framework is presented in Figure 11. Its components could be subdivided into three categories: components, which process input images or video frames with a goal to find all visible document templates and determine their coordinates (shown as green blocks in fig. 11); components, which process each individual document template (shown as yellow blocks in fig. 11); and, finally, components, which collect template recognition results into logical representations of documents, perform post-processing, and output the recognition result (shown as blue blocks in fig. 11).

a) Persistent configuration

The system is designed to recognize identity documents from a predetermined set of types. The persistent configuration of the system could be divided into three blocks: the database of known document templates, with an index which is used during the templates detection and location (Block F5 in fig. 11); the database of recognition configuration for each template, which consists of the information about the layout, constituent fields and their properties, and other data required for extraction and recognition of template components (Block T10 in fig. 11); finally, the database of documents, which contains the information on how the recognition results of individual templates are combined and post-processed to produce a final document recognition result (Block D6 in fig. 11).

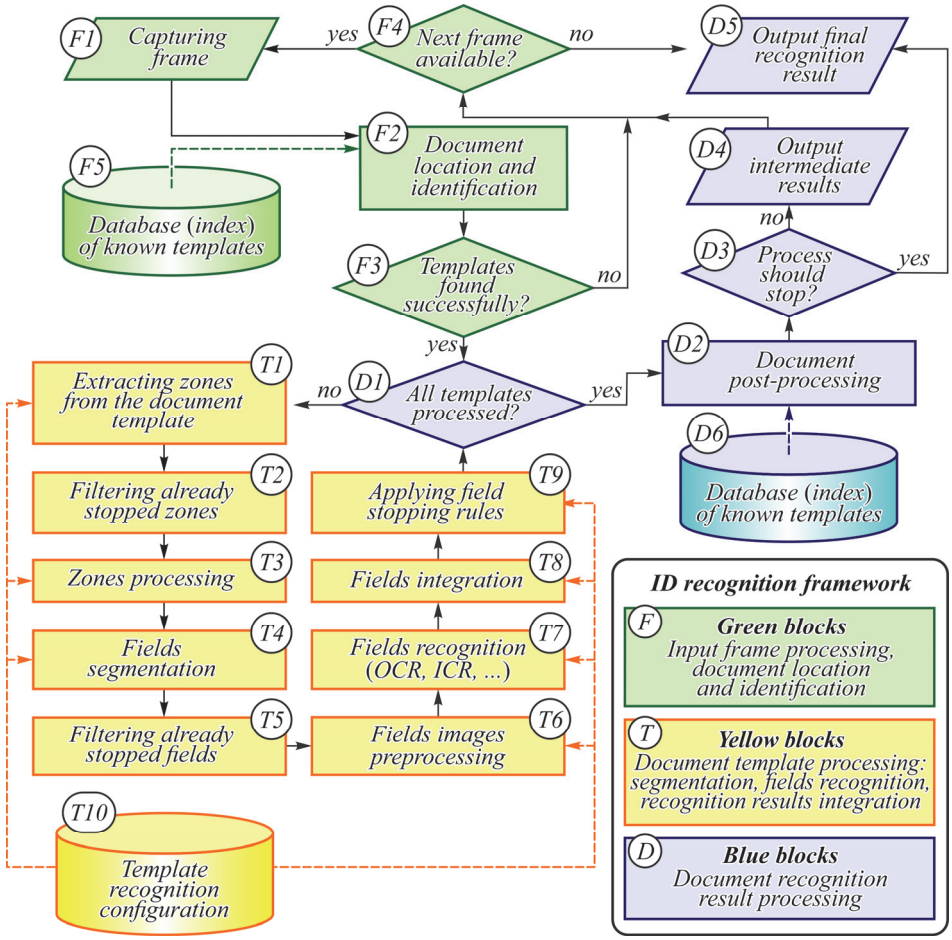


Fig. 11. Full scheme of the proposed framework

b) Templates detection and location

An input to the system is either a single image (photograph or a scan) or a series of images (video frames). Acquisition of each input image is performed in Block F1 (“Capturing frame”) in the fig. 11. The first step of input image processing within the proposed framework is the detection and location of the document templates (Block F2) given the index of known templates (Block F5). Despite the fact that a large number of document location methods exist which rely on preliminary text recognition [17, 71] since the focus of the system is

the recognition of documents with fixed layouts, it was designed having in mind the methods which perform templates location by their overall visual representation. The examples of such methods are the ones based on Viola and Jones approach generalized as a decision tree of strong classifiers [72], which could be applied for document page detection and location robust to moderate perspective distortions, as well as the methods based on document boundary detection [73], or deep learning-based method of segmenting the document from the background [74]. A more universal method of this class uses template identification and matching using feature points detection with descriptors indexing and RANSAC refinement [70, 75].

A formal problem statement of this stage can be divided into two separate tasks: identification of the document template (or a set of templates), and the location of the document in an image (or in a video frame) I from a set of all possible images $\mathbb{I}$. To formalize the former, we can assume that the database of the known templates (Block F5) defines a set $\mathbb{T}$ of template classes; each class $t \in \mathbb{T}$ represents either a distinct document template or a set of templates that are allowed to be visible simultaneously in a single image. Given a dataset $D_{\mathbb{T}} = \{(I_1, t_1), (I_2, t_2), \ldots, (I_n, t_n)\} \subset \mathbb{I} \times \mathbb{T}$ the problem is to find a mapping $f_{\mathbb{T}} : \mathbb{I} \times \mathbb{T}$ such as to maximize the classification accuracy:

$$\frac{100\%}{Card(D_{\mathbb{T}})} \cdot Card\left\{(I,t) \in D_{\mathbb{T}} \mid f_{\mathbb{T}}(I) = t\right\} \to \max_{f_{\mathbb{T}}}. \quad (1)$$

The problem of document templates location is more complicated to formalize, as it implies a specific model of a geometric template representation. If the real-world location target were a planar rectangular document template, under the general model of possible geometric distortions in a scan or in a photo captured with a pinhole camera, the representation of the template in an image would be a quadrilateral or a set of quadrilaterals if multiple templates were visible in an image. Let us denote a set of all possible template location results as $\mathbb{G}$ (the set of all possible sets of quadrilaterals), with a metric function $\rho_{\mathbb{G}} : \mathbb{G} \times \mathbb{G} \to \mathbb{R}_0^+$. Given a dataset $D_{\mathbb{G}} = \{(I_1, g_1), \ldots, (I_n, g_n)\} \subset \mathbb{I} \times \mathbb{G}$ the problem is to find a mapping $f_{\mathbb{G}} : \mathbb{I} \to \mathbb{G}$ such as to minimize the mean distance between the found templates location and the ground truth:

$$\frac{1}{Card(D_{\mathbb{G}})} \cdot \sum_{(I,g) \in D_{\mathbb{G}}} \rho_{\mathbb{G}}(f_{\mathbb{G}}(I), g) \to \min_{f_{\mathbb{G}}}. \quad (2)$$

Metric function $\rho_{\mathbb{G}}$ can be defined in multiple ways, depending on the specifics of the problem, the desired result, or on the specifics of the algorithms used in the later template processing. One of the most widely used metrics for document location is the Jaccard distance, defined as follows:

$$d_J(A,B) = 1 - \frac{Card(A \cap B)}{Card(A \cup B)}, \quad (3)$$

where A and B are regions of an image defined by sets of quadrilaterals $g_A, g_B \in \mathbb{G}$, and the cardinal numbers of their intersection or union represent the areas of the corresponding shapes.

Jaccard distance is used for identity document location in images, as well as other types of non-structured documents and arbitrary objects [74, 76]. While it is probably one of the most widely used metrics for document location, it has some flaws. Firstly, a small shift of the found quadrilateral can be almost identical to incorrect detection of a single document corner, while these types of errors very significantly differ with regards to rectangular documents with fixed layout − after perspective restoration, the latter error would result in a much higher skew of the document content, which could lead to problematic fields analysis and recognition. Secondly, in Jaccard distance terms there is no difference between a shift of the document boundaries outwards and inwards, while the latter is worse for further document analysis, as parts of the document content may be lost in such a way. Thus, some algorithms of fixed-layout documents location use different definitions of the metric function, for example, the maximal discrepancy of the corner position [77]:

$$d_C(g_A, g_B) = \frac{\max_{(p_A, p_B)} \| p_A - p_B \|_2}{\min_{(p_{B1}, p_{B2})} \| p_{B1} - p_{B2} \|_2}, \quad (4)$$

where (p_A, p_B) are pairs of the corresponding corners of the quadrilaterals from the representations g_A and g_B and (p_{B1}, p_{B2}) are pairs of the corners belonging to the same quadrilateral within the representation of g_B (which corresponds to the ground truth).

The output of the process F2 is a collection of found document templates $t \in \mathbb{T}$ each with a specified label (which allows to determine the configuration for its further processing) and geometric parameters $g \in \mathbb{G}$, such as the coordinates of its boundaries. If no known document templates are found (Block F3), the image is rejected by the system, which means that the process either ends with a null result (if only a single image is available for processing), or the next frame is acquired for processing (Block F4). From the point of view of the formal statement, the null result can be represented as a separate class $t_0 \in \mathbb{T}$, which corresponds to images that do not contain any known document templates.

If the input of the system is a sequence of video frames and each frame contains the templates of the same physical document, the document location and identification process F2 may benefit from the knowledge of the processing results on the previous frames. Thus, process F2 may have access to a non-persistent data storage, which accumulates processing results from multiple frames of a single document recognition session.

c) Template processing

After all templates discernible in the input image are located and identified, each of them is processed according to a predetermined workflow, the parameters of which are stored in the template recognition configuration (Block T10).

Since the template has been identified and its geometrical position in the image is determined, the position of most objects of interest can be calculated, and the images of most individual objects can be extracted with correction of angular rotation, projective distortions. Moreover, if the physical size of the document page corresponding to the processed template is known, the image of each individual object could be generated with a specified spatial resolution (for example, with a fixed number of pixels per inch), however, of course, the actual resolution in terms of the amount of information stored in each pixel will depend on the resolution of the initial image. Nevertheless, the exact knowledge (or, at least, a hypothesis) of the template coordinates does not always mean that the exact coordinates of each object of interest are known – while the static elements of the template are fixed in position (as per our definition of a template, see section 2a), the individual objects such as text fields could have variable length, and even variable position. For example, let us consider the third (main) page of the Russian national passport (see fig. 12). The first three fields from the top are Last name, Given name, and Patronymic. In the template, there are corresponding static labels and underlines which indicate where these fields are supposed to be located. However, the horizontal position of each field may vary from document to document. Moreover, due to printing defects, these fields could be shifted along the vertical axes (even in such a way as to intersect with the lines in the background), as well as have a slight angular skew.

Fig. 12. Field search zones on the photo page of Uruguay passport, sample image from WikiMedia Commons

This leads to the necessity of introducing an intermediate entity that would represent the localized area of the template which should be used to locate individual objects. We will refer to these areas as “zones” (represented as blue rectangles in fig. 12). A zone is a region of a template, with predefined coordinates, which can be processed by a single application of some predefined algorithm that would segment it and extract individual objects, such as text fields or other objects of interest. A zone on a template could comprise a single object of interest (e.g. a single text field), multiple fields or objects, and even correspond to the full document template − depending on its complexity, specifics of the extracted objects, and specifics of the algorithms employed to extract them.

Thus, the first step of template processing is the extraction of images of individual zones (Block T1), according to the information about zones and their positions encoded in the template processing configuration (Block T10). Each zone specified in the configuration defines a set of individual objects (such as text fields), which could be extracted from it. With a video stream as an input, the recognition or extraction results for individual objects are updated after each processed frame, and it is important to automatically determine when this process should be stopped for each object [78]. If the process of object extraction is terminated, there is no need to extract it on the next frame, as it would save processing time. Similarly, if all individual objects corresponding to a particular zone already satisfied their stopping rules, the zone itself can be skipped, thus saving the time otherwise spent on its analysis. Thus, after the set of zones of a currently considered document template is determined and the zones are extracted in process T1, the zones whose fields have already stopped are filtered out (Block T2).

The next step is the zone image processing (Block T3), which can include such operations as detection and rectification of angular skew, detection of specific security elements, such as holograms, suppression of background texture, and more [19]. While such image processing could be done on a level of individual object analysis, frequently the reliability of such image processing operations are improved if the whole zone context is available. A good example of that is the a whose text fields have the same angular skew (due to a printing defect) − while the angle could be determined at the level of each field, it could be beneficial to analyze the zone as a whole to obtain a more robust and consistent result. After the zone processing step, the zone is segmented into individual objects (Block T4). The method of segmentation could vary from zone to zone, depending on its structure: some zones could have fixed local coordinates of each individual object or field, specified in its configuration, and thus no additional search is required. Other zones may require a search of precise field coordinates with respect to a predefined pattern of fields [79], or even an application of a free-form text detection [80]. For the latter cases, similar quality metrics could be applied as for the task of document template location, such as Jaccard distance and other Intersection-Over-Union-based metrics [81] or, for independent analysis of text fields detection given their OCR results (see below), character-level metrics such as TedEval [82] or CLEval [83].

As an output of the process T4, there is a set of named individual objects (such as text fields, stamps, localized optical variable devices, graphical document zones such as signature or photo, etc.) with known coordinates in the zone (as well as in a document template and in the source image). Similarly to the zones filtering in the process T2, some of the extracted fields may already have stopped in the current video stream recognition session. Thus, the objects which have already satisfied their corresponding stopping rules, are filtered out (Block T5). Both processes of zone image processing T3 and object segmentation T4 may have access to non-persistent storage of the recognition session, to use the information gained at the previous processing stages in order to increase the robustness of the zone analysis.

The next steps of the template processing process are related to the analysis of individual objects. Firstly, the images of individual objects are preprocessed (Block T6) with a similar motivation as the zone image preprocessing (Block T3). This step could include rectification of specific features of objects, such as a shear of the text line [84]. Each individual object then undergoes the process of recognition (Block T7), if required by the nature of this object. It is worth noting that when the field undergoes the recognition process its nature is known beforehand, and stored information about its language, specifics of the used font, or other visual representation characteristics can be used to maximize the reliability and efficiency of the recognition. While the task for the recognition of such document objects as text fields might seem straightforward, to formalize such a problem a clear definition of what is considered the recognition result and a quality metric for it should be defined. Given a rectified image of a text field $I \in \mathbb{I}$ let us denote a recognition algorithm as a mapping $f_{\mathbb{X}} : \mathbb{I} \to \mathbb{X}$, where $\mathbb{X}$ represents the set of all possible recognition results. The most common option is to use the set of strings of characters to represent $\mathbb{X}$, however for some applications, and to enable further recognition results processing the text recognition results are also represented as sequences of character-level recognition results, each being a mapping from a predefined alphabet to a set of character membership estimations [85, 86]. Given a dataset of text field images with ground truth $D_{\mathbb{X}} = \{(I_1, x_1), (I_2, x_2), \ldots, (I_n, x_n)\} \subset \mathbb{I} \times \mathbb{X}$ the task of text fields recognition is to find a mapping $f_{\mathbb{X}}$ such as to minimize the mean distance between the text recognition results and the ground truth according to a predefined metric function $\rho_{\mathbb{X}} : \mathbb{X} \times \mathbb{X} \to \mathbb{R}_0^+$, similarly to the problem of document templates location (2):

$$\frac{1}{Card(D_{\mathbb{X}})} \cdot \sum_{(I,x) \in D_{\mathbb{X}}} \rho_{\mathbb{X}}(f_{\mathbb{X}}(I), x) \to \min_{f_{\mathbb{X}}}. \qquad (5)$$

A simplest metric $\rho_{\mathbb{X}}$ for text fields recognition quality evaluation is an end-to-end string results comparison:

$$\rho_E(x_1, x_2) = [x_1 \neq x_2] \qquad (6)$$

with which the mean distance on the dataset $D_{\mathbb{X}}$ (5) corresponds to the rate of incorrectly recognized input fields. A more character-level oriented metrics include Levenshtein distance (a minimal number of insertions, deletions, and substitutions required to convert the first string to the second), per-character recognition rate [85], normalized Levenshtein distance [87], and others. Notably, the Levenshtein and normalized Levenshtein distance can be generalized to be used for the representation of recognition results $\mathbb{X}$ with character membership estimations [86, 87].

If the video stream is used as an input for the system, the recognition results are then combined together in an integrated recognition result (Block T8) and a stopping rule is applied to determine whether additional observations of the same object are necessary (Block T9). The integration process T8 and the stopping rules T9 necessarily use the non-persistent storage of the recognition session, as they are required to read and update the current state of the video stream analysis.

The problem of field recognition results combination, given a sequence of field images and their corresponding recognition results obtained from different frames, can be formalized as the problem of finding a mapping (or, rather, a family of mappings) $c^{(k)} : \mathbb{X}^k \to \mathbb{X}$, where k stands for the number of per-frame field results in the sequence. Given a dataset of sequences of field images:

$$D_{\mathbb{X}^k} = \left\{(I_{11}, .., I_{1k}, x_1), ..., (I_{n1}, .., I_{nk}, x_n)\right\} \subset \mathbb{I}^k \times \mathbb{X} \qquad (7)$$

the task is to find a combination method $c^{(k)}$ such as to minimize the mean distance between the combined result and the ground truth:

$$\frac{\sum\limits_{(I_1, .., I_k, x) \in D_{\mathbb{X}^k}} \rho_{\mathbb{X}}(c^{(k)}(f_{\mathbb{X}}(I_1), .., f_{\mathbb{X}}(I_k)), x)}{Card(D_{\mathbb{X}^k})} \to \min_{c^{(k)}}, \qquad (8)$$

where $f_{\mathbb{X}}$ is the field recognition method and $\rho_{\mathbb{X}}$ is the metric function on the set of field recognition results, both corresponding to the text field recognition problem (5) stated above. Even simple selection strategies, such as selecting a single result with the maximum value of input image quality or of a recognition result confidence level [68] can be considered a combination method, along with alignment procedures such as ROVER (Recognizer Output Voting Error Reduction) [88] and its extension for text recognition results with per-character alternatives [86].

To formalize a statement for the stopping problem (Block T9), a notion of observation cost needs to be introduced. Let us denote as $\gamma_k(I_1, I_2, \ldots, I_k)$ the cost of acquiring k observations of a field (that is, acquiring images $I_1, I_2, \ldots, I_k$, performing their pre-processing and recognition), in relation to the potential recognition error of the combined result $c^{(k)}(f_{\mathbb{X}}(I_1), \ldots, f_{\mathbb{X}}(I_k))$.The observation cost in the simplest terms can be expressed simply as the number of processed frames k, or, in a more general case,

the time or computational resources required to acquire the frames and perform their recognition. The total loss after acquiring and processing of k images of the field can then be expressed as follows:

$$L_k(I_1,..,I_k,x) = \\ = \rho_{\mathbb{X}}(c^{(k)}(f_{\mathbb{X}}(I_1),..,f_{\mathbb{X}}(I_k)),x) + \gamma_k(I_1,..,I_k), \quad (9)$$

where x is the ground truth value of the recognized field. The stopping rule defines a stopping time (or stopping frame) K which can be considered as a random variable whose distribution depends on the observations $I_1, I_2, \ldots$. The stopping problem implies minimization of the expected loss at stopping time:

$$E(L_K(I_1,I_2,...,I_K,x) \mid D_{\mathbb{X}^k}) \to \min_K, \quad (10)$$

where $E(\cdot)$represents an expected value, $D_{\mathbb{X}^k}$ represents a dataset of sequences of field images (7), and the sample space of K is the set $\{1,2,\ldots,k\}$ of all possible stopping points. The simplest stopping rule stops the process after a fixed number of observations are processed; however, there exist more effective approaches such as thresholding of the maximal cluster of identical results [89], or modelling of the combined recognition result at the next stage of the process [78].

Non-text objects such as personal signatures of photos can also undergo the process of integration T8. For example, this process may perform a selection of a single best image by analyzing the focus score [68, 90], consistency of illumination or the presence of highlights [91]. Alternatively, the images extracted from individual frames may be combined into a single image of a higher quality using methods of video super-resolution [92, 93]. Optically variable devices such as holographic security elements could be analyzed in this process to verify that their variation between frames is consistent and corresponds to how a true object should behave. It is also worth noting that it could be feasible to use the techniques such as video super-resolution to combine the images of text fields prior to their recognition, in which case the process T7 could be skipped for such fields and the actual recognition could be performed at the end of the integration process T8.

d) Collecting the document recognition result

After all individual templates are processed (at the end of the cycle with precondition D1) the next step is to collect the found documents, which are composed of these templates. The information on the known document types, their constituent templates, and the sets of objects expected at the output is stored in the document configuration (Block D6).

The final stage of document analysis is the post-processing stage (Block D2). The text fields of identity documents usually have a specific syntactic and semantic structure thatis known in advance. Such structure typically comprises the following components:

1) Syntax: rules regulating the structure of text fields representation. For example, a “birth date” field of a machine-readable zone of international passports consists of six characters, each of which can take one of only 11 possible values (decimal digits and a filler character);
2) Field semantics: rules representing a semantic interpretation of the text field or its constituents. For example, a “birth date” field of a machine-readable zone of international passports is written in a fixed format “YYMMDD”, where “YY” are the last two digits of the year, “MM” is the month 01 to 12, “DD” is the day (in accordance with the month number), and unknown components of the date are replaced with pairs of filler characters “<<”;
3) Semantic relationships: rules representing the structural or semantic relationships between different fields of the same document. For example, the value of the “issue date” field cannot represent a moment in time preceding the value of the “birth date” field in valid documents.

Having the information about the syntax and semantics of a text field a language model can be built in form of a mapping $\lambda : \mathbb{X} \to \mathbb{R}_0^+$ from the set of all possible recognition results to the set of real-valued language membership estimations. From the point of view of data structures, the language model λ can be represented as a dictionary, finite-state automaton, validation grammar (based on a text field validity predicate), an N-gram model, and other methods [94, 95]. The problem statement for language-dependent recognition results correction based on a combination of hypotheses (encoded in the text field recognition result $f_{\mathbb{X}}(I)$) language model λ and an error model which defines possible modifications, is presented in [96] using WFSTs (Weighted Finite-State Transducers). An alternative approach that is based on representing λ as a validation grammar with a custom predicate is described in [97, 98].

If the system’s input is a video stream, after the document post-processing stage D2 a final decision should be made whether the full result could be considered terminal (Block D3). This decision is influenced by not only the stopping rules for individual objects (which are checked in Block T9) but also by the presence of required document templates and required objects in the formed document recognition result. If the result could be considered terminal, it is given as a system’s output (Block D5) and the recognition process finishes. If the result is not considered terminal, the next frame is captured (Block F1) and the process continues. The intermediate non-terminal result can also be returned to the caller for visualization and external control purposes (Block D4).

Universality for 2D, 3D, and 4D recognition

The document recognition process described in section 2b and represented in fig. 11 was designed having in mind the variations of input data sources, corresponding

to the notions of "2D", "3D", and "4D" recognition (see section 1d, fig. 9). The possibly unknown input resolution and arbitrary shifts and rotation of the document pages, which characterize the "2D" recognition case are handled in the first stages of the process, where the document templates are preliminary located and identified in every input image (Block F2 in fig. 11), and the local zones and objects are analyzed after the document templates are located, thus such analysis is performed on rectified and correctly scaled images.

Projective transformations, non-linear distortions, or arbitrary background, which complicate the problem of "3D" document recognition, could also be handled on the template identification and location stage before any further processing takes place. The issues of defocus, blur, highlights or inconsistent illumination, which characterize the "3D" document recognition case, are some of the reasons why specific processes of zone image preprocessing (Block T3) and field image preprocessing (Block T6) are stated as separate components − as the image processing of these modules could either help to analyze the complex cases or help to perform an early rejection of the bad quality input. Even with blurry and inconsistently illuminated images of the document fields, their recognition (Block T7) and forensic analysis benefit from the prior knowledge of the target field nature, as the recognition with this document processing scheme is always performed after the document template is classified, and the target field or zone identified.

In a "4D" document recognition case the capturing conditions changing in time can be accounted for in individual processing stages by having access to non-persistent storage of the recognition session and being able to use the extraction results from the previous video frames. The addition of the integration modules (Block T8) allows to increase the expected recognition accuracy of the document fields by using the information from multiple frames, as well as perform analysis of optically variable devices, and the application of stopping rules (Block T9) allows reducing the number of processed observations, as well as saving time by skipping the objects which have already terminated (Blocks T2 and T5) during the analysis of the next frames.

Conclusion

The research of modern identity document recognition systems touches multiple subdisciplines of computer science, including image processing and image analysis, pattern recognition, computer vision, information security, computational photography and optics. Societal and industrial challenges imposed to such systems require not only the development of methods for solving particular sub-tasks of ID documents analysis, such as text recognition and document detection but also the creation of new approaches and methodologies for system composition and operation. To answer these challenges, firstly a common language needs to be established which would facilitate research of the application of image analysis, recognition, and forensics methods, to the tasks of identity documents processing.

In this paper, we presented an analysis of the characteristics of images that serve as input for ID document recognition systems, and presented a framework for ID document analysis, which is designed to be applicable for different types of image capture, and be scalable both in terms of the range of supported documents types and the range of information which can be extracted from them. The separation of specific components of the framework, such as document location and identification steps, zones extraction and processing, fields processing, etc., was intended in order to have access to as much information as possible at the level of analysis of any single component of an ID document. Such an approach allows to use more specialized recognition methods (without restricting the usage of more generalized ones) and to construct systems with the capacity to perform deep forensic analysis of the input image, document substrate, or its visual components. Preliminary document location and identification, in our view, are especially crucial for performing meaningful document validity and authenticity checks, since given real-world capturing conditions the sensitive security details of ID documents can be robustly and reliably analyzed only after their precise location with respect to the analyzed document.

The framework was designed with mostly fixed-layout identity document templates in mind, however, there exist ID document types that can hardly be described as documents with a fixed layout, due to variations either of their subtypes or of the fields positions. Nevertheless, the framework could be used to describe and process such documents, with a possible specification of the definition of a document template and with richer template features, for example, by using static text guides as part of the template description.

An important aspect of identity document recognition systems, which is included in the presented framework, is the analysis of video capture and processing of sequences of video frames. When the target is an ID document, processing of video stream not only serves the purpose of enhancing the reliability and improving the accuracy of the recognition result, but also allows to perform specific analysis of the scene in order to detect fraudulent identification attempts, as well as to detect and analyze highlights, reflection patterns, and optically variable devices, such as holograms. For the application of ID document analysis systems to the remote identification process, the analysis of video stream becomes crucial even if the recognition module itself does not accumulate the per-frame information.

The aim of the presented framework is to unite various approaches to solving specific problems of identity document analysis; serve as a basis for developing methodologies, approaches, and new algorithms for ID document processing and automated personal identification;

and address the technical and industrial challenges imposed by the usage of identity document recognition frameworks in real-world scenarios.

References

[1] Eikvil L. OCR – Optical Character Recognition. 1993. Source: ⟨https://www.nr.no/~eikvil/OCR.pdf⟩.

[2] Doermann D, Tombre K, eds. Handbook of document image processing and recognition. London: Springer; 2014. ISBN: 978-0-85729-858-4.

[3] International Civil Aviation Organization. ICAO Doc 9303 – Machine readable travel documents. Source: ⟨https://www.icao.int/publications/pages/publication.aspx?docnum=9303⟩.

[4] Hartl A, Arth C, Schmalstieg D. Real-time detection and recognition of machine-readable zones with mobile devices. In Book: Braz J, Battiato S, Imai F, eds. Proceedings of the 10th International Conference on Computer Vision Theory and Applications. Volume 1: VISAPP. Berlin, Germany: 2015: 79-87. DOI: 10.5220/0005294700790087.

[5] Avoine G, Kalach K, Quisquater J-J. ePassport: Securing International contacts with contactless chips. In Book: Tsudik G, ed. Financial cryptography and data security. Berlin, Heidelberg: Springer; 2008: 141-155. DOI: 10.1007/978-3-540-85230-8_11.

[6] Buchmann N, Rathgeb C, Wagner J, Busch C, Baier H. A preliminary study on the feasibility of storing fingerprint and iris image data in 2d-barcodes. 2016 International Conference of the Biometrics Special Interest Group (BIOSIG) 2016: 1-5. DOI: 10.1109/BIOSIG.2016.7736904.

[7] Agrawal H. Aadhaar enabled applications. 2015. Source: ⟨https://darpg.gov.in/sites/default/files/Aadhaar.pptx⟩.

[8] ISO/IEC 7810:2003: Identification cards – Physical characteristics. 2003. Source: ⟨https://www.iso.org/standard/31432.html⟩.

[9] Council of the European Union. PRADO – Public Register of Authentic identity and travel Documents Online. Source: ⟨https://www.consilium.europa.eu/prado/en/prado-start-page.html⟩.

[10] American Association of Motor Vehicle Administrators. AAMVA DL/ID card design standard (CDS). Source: ⟨https://www.aamva.org/DL-ID-Card-Design-Standard⟩.

[11] International Civil Aviation Organization. Traveller identification programme – ID management solutions for more secure travel documents. Source: ⟨https://www.icao.int/security/FAL/TRIP/Pages/default.aspx⟩.

[12] Global coverage for identity verification. Source: ⟨https://www.jumio.com/global-coverage⟩.

[13] Onfido. Supported documents. Source: ⟨https://onfido.com/supported-documents⟩.

[14] Keesing Technologies. Unrivaled coverage of international ID documents. Source: ⟨https://www.keesingtechnologies.com/documentchecker/id-documents⟩.

[15] Llados J, Lumbreras F, Chapaprieta V, Queralt J. ICAR: Identity card automatic reader. Proc Sixth Int Conf on Document Analysis and Recognition 2001: 470-474. DOI: 10.1109/ICDAR.2001.953834.

[16] Mollah AF, Majumder N, Basu S, Nasipuri M. Design of an optical character recognition system for camera-based handheld devices. Int J Comput Sci Appl 2011; 8(4): 283-289.

[17] Ryan M, Hanafiah N. An examination of character recognition on ID card using template matching approach. Procedia Computer Science 2015; 59: 520-529. DOI: 10.1016/j.procs.2015.07.534.

[18] Pratama MO, Satyawan W, Fajar B, Fikri R, Hamzah H. Indonesian ID card recognition using convolutional neural networks. 5th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI) 2018: 178-181. DOI: 10.1109/EECSI.2018.8752769.

[19] Satyawan W, Pratama MO, Jannati R, Muhammad G, Fajar B, Hamzah H, Fikri R, Kristian K. Citizen ID card detection using image processing and optical character recognition. J Phys Conf Ser 2019; 1235: 012049. DOI: 10.1088/1742-6596/1235/1/012049.

[20] Smith R. An overview of the Tesseract OCR engine. Ninth Int Conf on Document Analysis and Recognition (ICDAR 2007) 2007; 2: 629-633. DOI: 10.1109/ICDAR.2007.4376991.

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

[22] Viet HT, Hieu Dang Q, Vu TA. A robust end-to-end information extraction system for vietnamese identity cards. 6th NAFOSTED Conf on Information and Computer Science (NICS) 2019: 483-488. DOI: 10.1109/NICS48868.2019.9023853.

[23] Thanh TNT, Trong KN. A method for segmentation of vietnamese identification card text fields. Int J Adv Comput Sci Appl 2019; 10(10): 415-421. DOI: 10.14569/IJACSA.2019.0101057.

[24] Sandler M, Howard A, Zhu M, Zhmoginov A, Chen L. Mobilenetv2: Inverted residuals and linear bottlenecks. 2018 IEEE/CVF Conf on Computer Vision and Pattern Recognition 2018: 4510-4520. DOI: 10.1109/CVPR.2018.00474.

[25] Guo Q, Deng Y. Attention OCR. 2017. Source: ⟨https://github.com/da03/Attention-OCR⟩.

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

[27] Wu X, Xu J, Wang J, Li Y, Li W, Guo Y. Identity authentication on mobile devices using face verification and id image recognition. Procedia Computer Science 2019; 162: 932-939. DOI: 10.1016/j.procs.2019.12.070.

[28] Fang X, Fu X, Xu X. Id card identification system based on image recognition. 2017 12th IEEE Conf on Industrial Electronics and Applications (ICIEA) 2017: 1488-1492. DOI: 10.1109/ICIEA.2017.8283074.

[29] Castelblanco A, Solano J, Lopez C, Rivera E, Tengana L, Ochoa M. Machine learning techniques for identity document verification in uncontrolled environments: A case study. In Book: Mora KMF, Marín JA, Cerda J, Carrasco-Ochoa JA, José Martínez-Trinidad JF, Olvera-López JA, eds. MCPR 2020: Pattern Recognition. Cham, Switzerland: Springer Nature; 2020: 271-281. DOI: 10.1007/978-3-030-49076-8_26.

[30] 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.

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

[32] Skoryukina N, Arlazarov V, Nikolaev D. Fast method of id documents location and type identification for mobile and

server application. 2019 Int Conf on Document Analysis and Recognition (ICDAR) 2019: 850-857. DOI: 10.1109/ICDAR.2019.00141.
[33] de Sá Soares Á, das Neves Junior R, Bezerra B. BID Dataset: a challenge dataset for document processing tasks. Anais Estendidos do XXXIII Conference on Graphics, Patterns and Images 2020: 143-146. DOI: 10.5753/sibgrapi.est.2020.12997.
[34] Ngoc MOV, Fabrizio J, Géraud T. Saliency-based detection of identy documents captured by smartphones. 13th IAPR International Workshop on Document Analysis Systems (DAS) 2018: 387-392. DOI: 10.1109/DAS.2018.17.
[35] Chazalon J, Gomez-Krämer P, Burie J, Coustaty M, Eskenazi S, Luqman M, Nayef N, Rusiñol M, Sidère N, Ogier J. SmartDoc 2017 video capture: Mobile document acquisition in video mode. 14th IAPR International Conference on Document Analysis and Recognition (ICDAR) 2017; 4: 11-16. DOI: 10.1109/ICDAR.2017.306.
[36] Sencar HT, Memon N. Overview of state-of-the-art in digital image forensics. In Book: Bhattacharya BB, Sur-Kolay S, Nandy SC, Bagchi A, eds. Statistical science and interdisciplinary research: Volume 3. Algorithms, architectures and information systems security. Singapore: World Scientific Publishing Co Pte Ltd; 2009: 325-347. DOI: 10.1142/9789812836243_0015.
[37] Piva A. An overview on image forensics. ISRN Signal Process 2013; 2013: 68-73. DOI: 10.1155/2013/496701.
[38] Centeno AB, Terrades OR, Canet JL, Morales CC. Identity document and banknote security forensics: A survey. arXiv preprint, 2019. Source: ⟨https://arxiv.org/abs/1910.08993⟩.
[39] Ferreira WD, Ferreira CB, da Cruz Júnior G, Soares F. A review of digital image forensics. Comput Electr Eng 2020; 85: 106685. DOI: 10.1016/j.compeleceng.2020.106685.
[40] Council of the European Union. PRADO Glossary – Technical terms related to security features and to security documents in general (in alphabetical order) 2021. Source: ⟨https://www.consilium.europa.eu/prado/en/prado-glossary/prado-glossary.pdf⟩.
[41] Arlazarov VV, Chernov TS, Nikolaev DP, Skoryukina NS, Slavin OA. Method for holographic elements detection in video stream. 2017, US Patent US10354142B2 of July 16, 2019. Source: ⟨https://patents.google.com/patent/US10354142B2/en⟩.
[42] Kunina IA, Aliev MA, Arlazarov NV, Polevoy DV. A method of fluorescent fibers detection on identity documents under ultraviolet light. Proc SPIE 2020; 11433: 114330D. DOI: 10.1117/12.2558080.
[43] Li H, Wang S, Kot AC. Image recapture detection with convolutional and recurrent neural networks. Electronic Imaging 2017; 2017(7): 87-91. DOI: 10.2352/ISSN.2470-1173.2017.7.MWSF-329.
[44] Sun Y, Shen X, Liu C, Zhao Y. Recaptured image forensics algorithm based on image texture feature. Intern J Pattern Recognit Artif Intell 2020; 34(03): 2054011. DOI: 10.1142/S0218001420540117.
[45] Warbhe AD, Dharaskar R, Thakare V. A scaling robust copy-paste tampering detection for digital image forensics. Procedia Computer Science 2016; 79: 458-465. DOI: 10.1016/j.procs.2016.03.059.
[46] Yusoff N, Alamro L. Implementation of feature extraction algorithms for image tampering detection. Int J Adv Comput Res 2019; 9(43): 197-211. DOI: 10.19101/IJACR.PID37.
[47] Kumar M, Rani A, Srivastava S. Image forensics based on lighting estimation. Int J Image Graph 2019; 19(03): 1950014. DOI: 10.1142/S0219467819500141.
[48] ISO 1073-2:1976: Alphanumeric character sets for optical recognition – Part 2: Character set OCR-B – Shapes and dimensions of the printed image. International Organization for Standardization; 1976. Source: ⟨https://www.iso.org/standard/5568.html⟩.
[49] Starovoitov V, Samal D, Sankur B. Matching of faces in camera images and document photographs. IEEE Int Conf on Acoustics, Speech, and Signal Processing 2000; 4: 2349-2352. DOI: 10.1109/ICASSP.2000.859312.
[50] Fysh MC, Bindemann M. Forensic face matching: A review. In Book: Bindemann M, Megreya AM, eds. Face processing: Systems, disorders and cultural differences. New York: Nova Science Publishing Inc; 2017: 1-20.
[51] Bulatov K, Arlazarov VV, Chernov T, Slavin O, Nikolaev D. Smart IDReader: Document recognition in video stream. 14th Int Conf on Document Analysis and Recognition (ICDAR) 2017; 6: 39-44. DOI: 10.1109/ICDAR.2017.347.
[52] Valentín K, Wild P, Štolc S, Daubner F, Clabian M. Optical benchmarking of security document readers for automated border control. Proc SPIE 2016; 9995: 999503. DOI: 10.1117/12.2241169.
[53] Fujitsu fi-65F: Flatbed scanner for passports, ID cards. Spigraph catalogue, 2021. Source: ⟨http://www.spigraph.com/Scanners/Catalogue-scanner/Documents/Specifics/Fujitsu/fi-65F⟩.
[54] PS667 Simplex ID Card Scanner with AmbirScan. Ambir Technology. Source: ⟨https://www.ambir.com/product/simplex-id-card-scanner-ambirscan-ps667-as⟩.
[55] Talwerdi M. Apparatus and method for reading a document and printing a mark on the document. 2018, Japan patent JP6314332B2 of July 4, 2017. Source: ⟨https://patents.google.com/patent/JP6314332B2/en⟩.
[56] Bocharov NA, Limonova EE, Nikolaev DP, Paramonov NB, Slavin OA, Usilin SA. Automatized workplace for passport documents control. Pat RF of Invent N RU 182557 U1 of August 22, 2018. Source: ⟨https://yandex.ru/patents/doc/RU182557U1_20180822/⟩.
[57] Volonkin VM, Evstafjev EN, Nikonorov MV, Podoljskii AD, Stolyarov EV. Universal reader of passport and visa documents. 2013, Pat RF of Invent N RU 127977 U1 of May 10, 2013. Source: ⟨https://patents.google.com/patent/RU127977U1/en⟩.
[58] Arlazarov VV, Zhukovskiy AE, Krivtsov VE, Nikolaev DP, Polevoy DV Analysis of the usage specifics of stationary and small-scale mobile video cameras for documents recognition [In Russian]. Information Technologies and Computing Systems (ITiVS) 2014; 3: 71-81.
[59] Li X, Zhang B, Liao J, Sander PV. Document rectification and illumination correction using a patch-based CNN. ACM Trans Graph 2019; 38(6): 168. DOI: 10.1145/3355089.3356563.
[60] Asad F, Ul-Hasan A, Shafait F, Dengel A. High performance OCR for camera-captured blurred documents with LSTM networks. 12th IAPR Workshop on Document Analysis Systems (DAS) 2016: 7-12. DOI: 10.1109/DAS.2016.69.
[61] Chernov TS, Razumnuy NP, Kozharinov AS, Nikolaev DP, Arlazarov VV. Image quality assessment for video stream recognition systems. Proc SPIE 2017; 10696: 106961U. DOI: 10.1117/12.2309628.
[62] Nunnagoppula G, Deepak KS, Harikrishna G, Rai N, Krishna PR, Vesdapunt N. Automatic blur detection in mobile captured document images: Towards quality check in mobile based document imaging applications. IEEE

Second Int Conf on Image Information Processing (ICIIP-2013) 2013: 299-304. DOI: 10.1109/ICIIP.2013.6707602.
[63] Miao L, Peng S. Perspective rectification of document images based on morphology. 2006 Int Conf on Computational Intelligence and Security 2006; 2: 1805-1808. DOI: 10.1109/ICCIAS.2006.295374.
[64] Takezawa Y, Hasegawa M, Tabbone S. Robust perspective rectification of camera-captured document images. 14th IAPR Int Conf on Document Analysis and Recognition (ICDAR) 2017; 06: 27-32. DOI: 10.1109/ICDAR.2017.345.
[65] Kunina I, Gladilin S, Nikolaev D. Blind radial distortion compensation in a single image using fast Hough transform. Computer Optics 2016; 40(3): 395-403. DOI: 10.18287/2412-6179-2016-40-3-395-403.
[66] Zhukovsky A, Nikolaev D, Arlazarov V, Postnikov V, Polevoy D, Skoryukina N, Chernov T, Shemiakina J, Mukovozov A, Konovalenko I, Povolotsky M. Segments graph-based approach for document capture in a smartphone video stream. 14th IAPR Int Conf on Document Analysis and Recognition (ICDAR) 2017; 01: 337-342. DOI: 10.1109/ICDAR.2017.63.
[67] Haris M, Shakhnarovich G, Ukita N. Recurrent back-projection network for video super-resolution. IEEE/CVF Conf on Computer Vision and Pattern Recognition (CVPR) 2019: 3892-3901. DOI: 10.1109/CVPR.2019.00402.
[68] 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.
[69] Awal AM, Ghanmi N, Sicre R, Furon T. Complex document classification and localization application on identity document images. 14th IAPR Int Conf on Document Analysis and Recognition (ICDAR) 2017; 01: 426-431, DOI: 10.1109/ICDAR.2017.77.
[70] Augereau O, Journet N, Domenger J-P. Semi-structured document image matching and recognition. Proc SPIE 2013; 8658: 865804. DOI: 10.1117/12.2003911.
[71] Slavin OA. Using special text points in the recognition of documents. In Book: Kravets AG, Bolshakov AA, Shcherbakov MV. Cyber-physical systems: Advances in design & modelling. Cham: Springer International Publishing; 2020: 43-53. DOI: 10.1007/978-3-030-32579-4_4.
[72] Minkina A, Nikolaev D, Usilin S, Kozyrev V. Generalization of the Viola-Jones method as a decision tree of strong classifiers for real-time object recognition in video stream. Proc SPIE 2015; 9445: 944517. DOI: 10.1117/12.2180941.
[73] Puybareau E, Geraud 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.
[74] das Neves Junior RB, Lima E, Bezerra BL, Zanchettin C, Toselli AH. HU-PageScan: a fully convolutional neural network for document page crop. IET Image Process 2020; 14: 3890-3898. DOI: 10.1049/iet-ipr.2020.0532.
[75] Loc CV, Cao De T, Burie JC, Ogier JM. Content region detection and feature adjustment for securing genuine documents. 12th Int Conf on Knowledge and Systems Engineering (KSE) 2020: 103-108. DOI: 10.1109/KSE50997.2020.9287382.
[76] Forman S, Samanthula BK. Secure similar document detection: Optimized computation using the Jaccard coefficient. IEEE 4th Int Conf on Big Data Security on Cloud, IEEE Int Conf on High Performance and Smart Computing, (HPSC) and IEEE Int Conf on Intelligent Data and Security (IDS) 2018: 1-4. DOI: 10.1109/BDS/HPSC/IDS18.2018.00015.
[77] Skoryukina N, Nikolaev DP, Sheshkus A, Polevoy D. Real time rectangular document detection on mobile devices. Proc SPIE 2015; 9445: 94452A. DOI: 10.1117/12.2181377.
[78] Bulatov K, Razumnyi N, Arlazarov VV. On optimal stopping strategies for text recognition in a video stream as an application of a monotone sequential decision model. Int J Doc Anal Recognit 2019; 22(3): 303-314. DOI: 10.1007/s10032-019-00333-0.
[79] Povolotskiy MA, Tropin DV. Dynamic programming approach to template-based OCR. Proc SPIE 2019; 11041: 110411T. DOI: 10.1117/12.2522974.
[80] Zhou X, Yao C, Wen H, Wang Y, Zhou S, He W, Liang J. EAST: An efficient and accurate scene text detector. IEEE Conf on Computer Vision and Pattern Recognition (CVPR) 2017: 2642-2651. DOI: 10.1109/CVPR.2017.283.
[81] Wolf C, Jolion J-M. Object count/area graphs for the evaluation of object detection and segmentation algorithms. Int J Doc Anal Recognit 2006; 8(4): 280-296.
[82] Lee CY, Baek Y, Lee H. TedEval: A fair evaluation metric for scene text detectors. arXiv preprint, 2019. Source: ⟨https://arxiv.org/abs/1907.01227⟩.
[83] Baek Y, Nam D, Park S, Lee J, Shin S, Baek J, Lee CY, Lee H. CLEval: Character-level evaluation for text detection and recognition tasks. arXiv preprint, 2020. Source: ⟨https://arxiv.org/abs/2006.06244⟩.
[84] Bezmaternykh PV, Nikolaev DP, Arlazarov VL. Textual blocks rectification method based on fast Hough transform analysis in identity documents recognition. Proc SPIE 2018; 10696: 1069606. DOI: 10.1117/12.2310162.
[85] 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.
[86] Bulatov KB. A method to reduce errors of string recognition based on combination of several recognition results with per-character alternatives. Bulletin of the South Ural State University, Series: Mathematical Modelling, Programming and Computer Software 2019; 12(3): 74-88. DOI: 10.14529/mmp190307.
[87] 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.
[88] 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 1997: 347-354. DOI: 10.1109/ASRU.1997.659110.
[89] Arlazarov VV, Bulatov K, Manzhikov T, Slavin O, Janiszewski I. Method of determining the necessary number of observations for video stream documents recognition. Proc SPIE 2018; 10696: 106961X. DOI: 10.1117/12.2310132.
[90] Tolstov I, Martynov S, Farsobina V, Bulatov K. A modification of a stopping method for text recognition in a video stream with best frame selection. Proc SPIE 2021; 11605: 116051M. DOI: 10.1117/12.2586928.
[91] Polevoy DV, Aliev MA, Nikolaev DP. Choosing the best image of the document owner's photograph in the video stream on the mobile device. Proc SPIE 2021; 11605: 116050F. DOI: 10.1117/12.2586939.
[92] Shi W, Caballero J, Huszár F, Totz J, Aitken AP, Bishop R, Rueckert D, Wang Z. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. IEEE Conf on Computer Vision and Pat-

tern Recognition (CVPR) 2016: 1874-1883. DOI: 10.1109/CVPR.2016.207.
[93] Ren H, El-Khamy M, Lee J. Video super resolution based on deep convolution neural network with two-stage motion compensation. IEEE Int Conf on Multimedia Expo Workshops (ICMEW) 2018: 1-6. DOI: 10.1109/ICMEW.2018.8551569.
[94] Mei J, Islam A, Wu Y, Moh'd A, Milios EE. Statistical learning for OCR text correction. arXiv preprint, 2016. Source: ⟨https://arxiv.org/abs/1611.06950⟩.
[95] Nguyen T, Jatowt A, Coustaty M, Nguyen N, Doucet A. Post-OCR error detection by generating plausible candidates. Int Conf on Document Analysis and Recognition (ICDAR) 2019: 876-881. DOI: 10.1109/ICDAR.2019.00145.
[96] Llobet R, Cerdan-Navarro J, Perez-Cortes J, Arlandis J. OCR post-processing using weighted finite-state transducers. 20th Int Conf on Pattern Recognition 2010: 2021-2024. DOI: 10.1109/ICPR.2010.498.
[97] Bulatov KB, Nikolaev DP, Postnikov VV. Universal algorithm for post-processing of recognition results based on validation grammars [In Russian]. Trudy ISA RAN 2015; 65(4): 68-73.
[98] Petrova O, Bulatov K. Methods of machine-readable zone recognition results post-processing. Proc SPIE 2019; 11041: 110411H. DOI: 10.1117/12.2522792.

Authors' information

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* .

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 and text layout analysis.
E-mail: *bezmaternyh@isa.ru* .

Dmitry PetrovichNikolaev, (b. 1978), obtained his master's degree in physics in 2000 from Moscow State University and the Ph.D. degree in computer science from Moscow State University in 2004. Since 2007 he has been a Head of the Vision Systems Laboratory at the Institute for Information Transmission Problems of RAS and since 2016 he is a CTO of Smart Engines Service LLC. He is an author of over 220 papers and 6 patents. Research interests: computer vision, algorithms for fast image processing, pattern recognition. E-mail: *dimonstr@iitp.ru* .

Vladimir Viktorovich Arlazarov, (b. 1976), received a specialist degree in Applied Mathematics from the Moscow Institute of Steel and Alloys in 1999 and the Ph.D. degree in Computer Science in 2005. Since 1999 he has been working at the Institute for Systems Analysis of RAS (now – FRC "Computer Science and Control" of RAS) as a Researcher, Senior Researcher and Head of Laboratory. Since 2016 he is a General Director of Smart Engines Service LLC. He has published over 90 papers and authored seven patents. Research interests: computer vision and document analysis systems. E-mail: *vva@smartengines.com* .

Received August 25, 2021. The final version – October 5, 2021.