

# Model-driven approach to creating ID document templates for localization and classification based on a single image

D.P. Matalov<sup>1,2</sup>, V.V. Arlazarov<sup>1,2</sup>

<sup>1</sup> Federal Research Center "Computer Science and Control" of Russian Academy of Sciences, Prospekt 60-Letiya Oktyabrya 9, Moscow, 119333, Russia;

<sup>2</sup> Smart Engines Service LLC, Prospekt 60-Letiya Oktyabrya 9, Moscow, 117312, Russia

## Abstract

ID document recognition systems are already deeply integrated into human activity, and the pace of integration is only increasing. The first and most fundamental problems of such systems are document image localization and classification. In this field, template matching-based approaches have become widely used. These methods offer industrial precision, require minimal training data, and provide real-time performance on mobile devices. However, these methods have a significant limitation in scalability: every document type represents a set of local features to store and process, which affects the required computing resources. Moreover, considering the number of different document types supported by modern industrial recognition systems, they become unusable. To mitigate the drawback, we propose a method to select a subset of the most "stable" keypoints. To estimate keypoints' stability we synthesize a dataset of images containing various distortions relevant to the process of taking photos of hand-held documents with a smartphone camera in uncontrolled lighting conditions. To perform experiments we use well-known MIDV datasets, which have been designed to benchmark modern ID document recognition. The experiments show that the proposed method allows for increased ID document detection performance with limited computing resources.

**Keywords:** one-shot learning, documents recognition, document processing, image augmentation, template matching, local features.

**Citation:** Matalov DP, Arlazarov VV. Model-driven approach to creating ID document templates for localization and classification based on a single image. *Computer Optics* 2025; 49(6): 1148-1155. DOI: 10.18287/2412-6179-CO-1762.

## Introduction

Identity document recognition systems have become an integral part of automated identification and authentication processes. Such processes include the KYC procedure in finance and the sharing economy, identification when crossing borders, authentication in access control systems, hospitality business, medicine, insurance, and other areas that require identification and verification of the authenticity of identity documents. The very fundamental processing stage in hierarchical recognition systems is determining the type and boundaries of the document [1, 2]. Identity documents are usually hard plastic cards or passport pages with a set of geometrically fixed static elements. Such static elements include a variety of logos, flags, background elements, and field annotations. Paper [3] states that the most useful methods for recognizing such documents are those based on a comparison of visual representations of documents. The statement is supported by numerous experimental results obtained on large public datasets, and the approach is actively being developed and refined [4, 5, 6, 7, 8]. Among the methods based on visual representation matching, the method of matching local features has become very popular among scientists and recognition systems engineers. Fig. 1 shows a diagram containing the main blocks of the method.

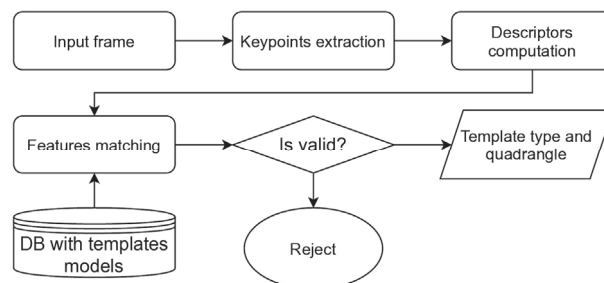


Fig. 1. Document retrieval scheme using local features template matching. Source: [9]

Despite the impressive list of such advantages as high accuracy sufficient for the operation of government systems or for business applications; no need for large arrays of training data; and real-time performance on energy-efficient devices, the method has some significant drawbacks [4]. One of them is limited scalability: each document type is represented as a set of local features by which the algorithm searches for a match with the features extracted from an input image. This significantly

affects memory consumption (both for storing templates in persistent storage and RAM), loading time into RAM, the speed of document localization, and type detection and recognition in general. At the same time, the method is widely used on mobile devices with limited computing power [1, 10, 4]. In this case, the method may not be applicable in the context of modern industrial recognition systems that support thousands of different types of documents.

### 1. Related Work

Enormous number of papers are devoted to the selection of the optimal set of features for solving pattern recognition problems. Considering specifics of identity documents retrieval with local features matching, various researchers have tried to mitigate the lack of scalability of the method in different ways. Paper [4] shows that the use of document boundaries can significantly improve the quality of document localization, which can potentially reduce the number of keypoints in a document template. Paper [11] shows that the use of 128-bit binary descriptors BinBoost [12], instead of SURF [13], can significantly reduce memory consumption. Document-oriented descriptors [14] can simultaneously improve the quality of document localization and further reduce memory consumption. For individual scenarios, for example, images obtained using flatbed scanners, 64-bit descriptors are sufficient to ensure localization quality at the level of 95 %. In any case, a document template consists of a set of pairs of keypoints and descriptors, and the size of the template description linearly depends on the size of the keypoints set. Therefore, choosing an optimal subset of relevant and high-quality keypoints considering the specifics of the task allows for significantly reducing the computing resources and expanding the scope of the method. In paper [3], a document template is computed according to local features specific to a certain document type detected outside the pre-marked document zones. In Fig. 2, the document exemplar's unique zones (photo, personal data, document number, issuing authority identifier, etc.) are selected with red rectangles, and zones containing the same data in all the exemplars of the particular type are in green

The paper states that a document template can be computed either from a set of document copies of a certain type (set of matched features during the "inference" stage in the dataset) or from a single image.

To mitigate the scalability drawback, we propose to select a limited subset of the most "stable" keypoints which is used to compute a document template. We elaborate the idea proposed in [3] and propose selecting the "best-quality" keypoints using a dataset of images.

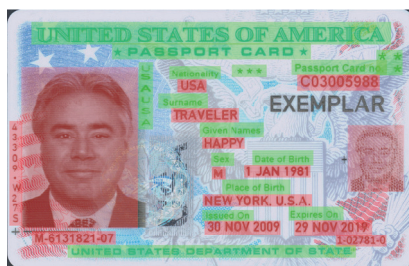


Fig 2. The personal data zones of a document exemplar are highlighted with red rectangles, while the data annotations are highlighted with green ones. Source: [9]

Despite significant progress in the area of access to public datasets [10, 15, 16], identity documents are highly sensitive data, and personal data (which include document images) is regulated by many laws around the world. We propose to synthesize a set of document images relevant to capturing documents in uncontrolled conditions, including "handheld" capturing with a smartphone camera, using only one ideal document image of identity document exemplar.

## 2. Proposed Method

### 2.1. Document Image Synthesis

To research a set of distortions typical for capturing hand-held document with a smartphone camera in uncontrolled conditions, we investigated public datasets MIDV [10, 17, 18]. The observed distortions can be grouped by the source of their formation, as follows:

- Geometric - distortions caused by the geometric model of image formation (Fig. 3a);
- Brightness - distortions caused by uneven and insufficient lighting (Fig. 3b);
- Blurring - distortions caused by defocus and motion blur (Fig. 3c);
- Noise - signal sampling and discretization errors caused by camera sensor (Fig. 3d).

Let us consider the models and methods we use to synthesize distorted document images in more detail.

#### 2.1.1. Camera shift and rotation simulation

To synthesize images with geometric distortions, we use the model presented in the paper [19]. The paper presents a method for rectifying a document plane in an image (in the rectified image, the sides of the document are parallel to the sides of the image) by two vanishing points, and is based on the pinhole camera model. The method tests a number of hypotheses about potential camera rotations and shifts in three-dimensional space at which the observed image of the document quadrangle is consistent with the expected rectangular image. We solve the inverse problem - having the camera

rotations and shifts parameters and the "original" camera image, where the ideal document image is placed in the center of the camera, we calculate the corresponding homography. The process of computing a geometrically distorted document template is shown in Fig. 4.



Fig. 3. Distortions from MIDV. (a) Geometric, (b) Lighting, (c) Motion-blur, (d) Noise



Fig. 4. The process of synthesizing document images with geometric distortions. (a) Ideal document, (b) Camera image, (c) Simulated distortion

### 2.1.2. Uncontrolled lighting

To synthesize images of uneven and insufficient lighting, we use various pixel brightness transformations. Fig. 5a shows an example of global darkening using a linear transformation of the brightness of all pixels. Fig. 5b shows an example of a 'shadow' simulation, where the document image is divided by a line, and all pixels of the plane on one side of the line are linearly darkened. Fig. 5c shows an example of a synthesized image emulating uneven illumination caused by light diffusion.



Fig. 5. Simulation of different lighting conditions. (a) Global and local darkening, (b) Shadow, (c) Local lighting

### 2.1.3. Defocus and Motion blur

To simulate defocus and motion blur we use classical methods of image smoothing with a Gaussian filter with various  $\sigma$  parameters. To simulate motion blur, a motion map was generated along a straight line and then convolved with the motion and Gaussian kernels.

### 2.2. Stability of keypoints

Feature point detectors provide some estimate of the "goodness" of each detected point based on intermediate calculations performed by the algorithm. Thus, in the Harris corner detector [20], the Harris estimate is used, which is closely related to the model of a corner that the algorithm uses. A local contrast estimation is used both in SURF [13] and SIFT [21]. Modern machine-learning based detectors such as TILDE [22], SuperPoint [23] and LF-Net [24] use estimates computed according to the architectures of the models and seen training data. In any case, every keypoint extraction algorithm, especially handcrafted, addresses a certain class of problems and objects. There exist several published

methods for evaluating the "goodness" of a keypoint, without being tied to a specific detector that extract it. In papers [25, 26], the authors propose to estimate the "goodness" of a keypoint based on the analysis of neighbourhood pixels, which may conflict with the criteria used for its detection. This approach could lead to a situation where a keypoint extracted infrequently is assigned the highest quality score. We propose to estimate the "goodness" of a keypoint considering stability [27] of its detection on distorted images. Let there be a sequence of images  $\{I_m, m = 0, 1, \dots, M\}$  containing the same object under different geometric or brightness distortions. Let image  $I_0$  be a reference image, and  $H_m$  be a projective transformation from  $I_0$  to  $I_m$ . Then the keypoint stability  $R(p \in I_0)$  of the keypoint  $p$  is defined as

$$R(p \in I_0) = \sum_m [\min_{q \in I_m} \|H_m(p) - q\|_2 < \varepsilon], \quad (1)$$

where  $[\cdot]$  is the indicator function,  $\|\cdot\|_2$  is the Euclidean distance,  $q \in I_m$ . Points  $q$  and  $p$  are considered matched if the distance between them is less than the specified threshold  $\varepsilon$ . This approach allows to treat a keypoint detector as a "black box" and is independent of the specific detection algorithm used. To describe a document template, we select keypoints with the largest  $R(p)$  on the synthesized dataset of documents, and the algorithm can be represented as follows:

*Algorithm. 1. Algorithm for selecting keypoints for a document template*

---

```

Input: Ideal template image  $I_0$ , set of rectangles  $S$ , number of
images to synthesize  $M$ , maximum number of keypoints to
select  $N$ , maximum matching distance  $\varepsilon$ 
Output: Set of selected keypoints  $T$ 
1 Detect a set of keypoints  $P = \{(x_1, y_1, r_1), (x_2, y_2, r_2), \dots, (x_n, y_n, r_n)\}$ 
in  $I_0$  using any keypoint extraction algorithm;
2  $P = P \setminus \{p \in P, p \notin S\}$ ; // Remove keypoints that lie in
personal data rectangles
3  $m \leftarrow 1$ 
4 while  $m \leq M$  do
5   Generate a synthetic ID image  $I_m$  using  $I_0$  and get homography
matrix  $H_m : I_m \rightarrow I_0$ 
6   Detect a set of keypoints  $P_m = \{(x, y)\}$  in  $I_m$ 
7   foreach  $p_m \in P_m$  do
8      $p_{m0} \leftarrow H_m(p_m)$ 
9      $match_{id} \leftarrow -1$ 
10     $err \leftarrow \varepsilon$ 
11    foreach  $i \in \{1, \dots, |P|\}$  do
12       $err_{mi} \leftarrow \|p_{m0} - p_i\|_2$ 
13      if  $err_{mi} < err_{id}$  then
14         $err \leftarrow err_{mi}$ 
15         $match_{id} \leftarrow i$ 
16      end
17    end
18    if  $match_{id} \neq -1$  then
19       $r(P[match_{id}]) \leftarrow r(P[match_{id}]) + 1$ ; // Increase
repeatability score of the  $match_{id}$ -th point
20    end
21  end
22 end
23 Select a subset  $T$  of size  $|N|$  from  $P$  with the highest  $r$ 
24 return  $T$ 

```

---

### 3. Experimental Evaluation

#### 3.1. Datasets and metrics

To evaluate the performance of the proposed keypoints scoring and selection method in the identity document recognition problem, we use the pipeline of the method presented in [3] and measure the quality of document typing and localization according to [4]. Unlike [3], we perform brute force matching instead of FLANN [28] during descriptors search to get reproducible results and exclude potential collisions. As for test data, we use the well-known MIDV [10, 17, 18] datasets. MIDV-500 [10] contains video clips of 50 different document types taken by a smartphone camera, presented on various and complex backgrounds under projective distortions. MIDV-2019 [17] includes documents captured under extreme projective distortion and low light in 4K resolution. MIDV-2020 [18] contains only 10 document types, but this dataset has even greater variability in capturing conditions and consists of 1000 fictitious ID cards with unique personal data. From MIDV-2020, we consider only the images obtained from the scanner, since we consider video frames from MIDV-500 and MIDV-2019 to be sufficiently relevant for evaluating our method.

#### 3.2. Image synthesis and keypoints stability estimation

To generate projective distortions, we generate 3 parameters of camera rotation angles relative to the optical axis according to the uniform distribution  $X \sim U(-45^\circ, 45^\circ)$ . To simulate the illumination changes, we used the following transformations:

- Monotone transformation of brightness by random function [29]

- Global linear brightness transform
- Smooth brightness change along a random line on the image
- Normally distributed noise
- Darkening of one of the half-planes of the image divided by a random line
- Morphological closing
- Gaussian blur

To compute keypoints stability for further selection we synthesize 1000 images for each target document type and calculate the keypoint stability using  $\epsilon = 5$ . Fig. 6 shows the keypoints stability computation process. Fig. 6a shows all the keypoints detected in an ideal template image and filtered according to the personal data zones. Fig. 6b shows the keypoints detected in a distorted synthesized image. Fig. 6c shows matching verdict for every template point (indicator function result in Eq. 1) – matched points are highlighted with green color, while unmatched is in red.

### 3.3. Experimental results

The size of the document template directly affects the localization and classification quality. To assess this dependency, we performed a series of experiments on 50 types of documents using MIDV datasets [10] and estimated document detection quality following the pipeline noted in Sec. 3.1. The algorithm of assessing localization performance is based on the calculation of the maximal deviation of the computed document corner coordinates divided by the length of the shortest document boundary side [4]. We use SURF as a keypoint detector and descriptor and assessed the document localization and classification quality using different limits on the maximum number of keypoints to describe a document template. We compare two methods for keypoints scoring: a method based on the responses provided by the SURF algorithm on an ideal template image, and the proposed method, which is based on stability estimates obtained from a synthesized dataset. Fig. 7 shows the graph of this dependency obtained on MIDV-500 dataset.

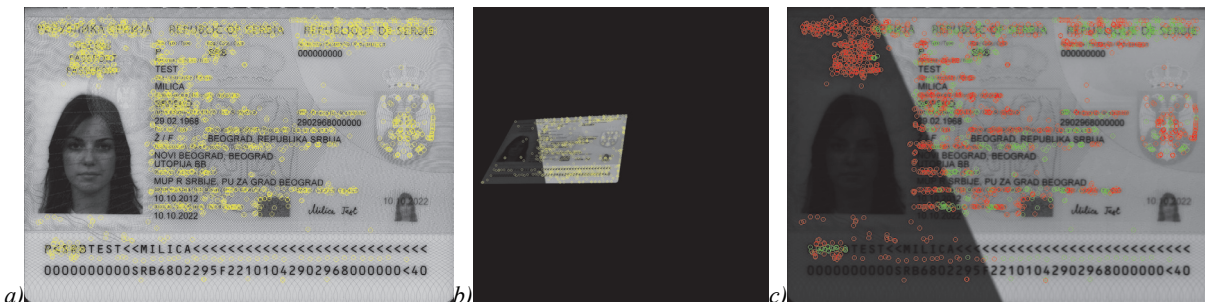


Fig. 6. The visualization of the point stability calculation. (a) Template image, (b) Synthesized image, (c) Matched points

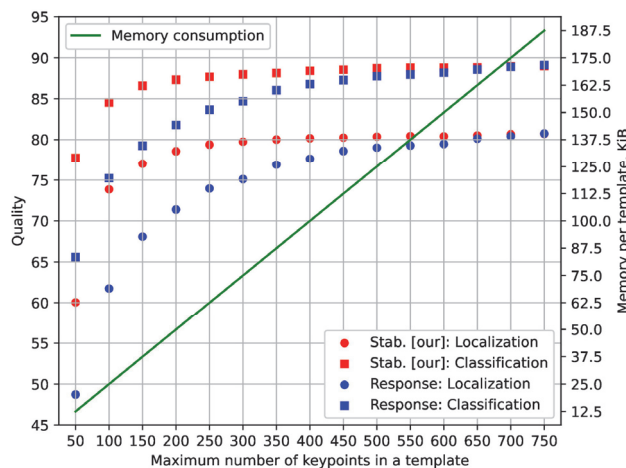


Fig. 7. The dependence of the quality of document detection on the maximum number of points describing the template

The left vertical axis  $OY$  in Fig. 7 represents the detection quality, while the horizontal axis  $OX$  displays the explored restrictions on the maximum number of keypoints  $N$  to describe a document template. The right vertical axis  $OY$  displays amount of memory required to store SURF descriptors of a single document's template. Table 1 shows the exact values of the obtained document localization quality with various restrictions on the maximum template size obtained for video clips from MIDV-500 and MIDV-2019. To improve readability, some values have been omitted due to their low informational content. The last row in the table shows the performance when all the detected points were used for a document template construction.

Tab. 2 displays the average amount of memory required to store SURF descriptors for a document template from MIDV datasets.

Tab. 1. Document classification and localization performance in video clips from MIDV-500 and MIDV-2019

N	MIDV-500				MIDV-2019			
	Localization, %		Typization, %		Localization, %		Typization, %	
	Resp.	Stab.	Resp.	Stab.	Resp.	Stab.	Resp.	Stab.
50	48.73	<b>60.03</b>	65.56	<b>77.76</b>	32.21	<b>34.42</b>	52.58	<b>57.25</b>
100	61.74	<b>73.86</b>	75.23	<b>84.55</b>	46.29	<b>48.62</b>	63.75	<b>67.63</b>
150	68.07	<b>77.02</b>	79.25	<b>86.59</b>	51.52	<b>54.73</b>	68.62	<b>71.85</b>
200	71.39	<b>78.57</b>	81.8	<b>87.33</b>	54.94	<b>57.57</b>	69.83	<b>72.83</b>
250	73.97	<b>79.39</b>	83.67	<b>87.69</b>	56.78	<b>58.35</b>	71.17	<b>73.53</b>
300	75.12	<b>79.75</b>	84.69	<b>87.99</b>	57.68	<b>59.33</b>	71.6	<b>73.78</b>
350	76.91	<b>80.01</b>	86.05	<b>88.15</b>	58.64	<b>59.5</b>	72.38	<b>74.05</b>
400	77.65	<b>80.17</b>	86.79	<b>88.43</b>	59.7	<b>60.2</b>	72.87	<b>74.15</b>
...								
∞	82.89	-	91.48	-	62.22	-	75.42	-

Tab. 2. The average memory consumption of a document template for indexing SURF descriptors depends on the number of points in the template

N	50	100	150	200	250	300	350	400	∞
KiB	12.5	25	37.5	50	62.5	75	87.5	100	1478.3

One may see from the Fig. 7 that starting from a certain value of the maximum number of keypoints for a document template, the performance gain becomes extremely insignificant regardless the keypoints selection method. At the same time, the scalability of the method decreases remarkably. For example, to store SURF descriptors for 11,000 document templates considering the keypoints limit at the value of 750, one may need ~ 2 GiB. However, if one chooses a value of 500, there is a need in ~ 1.3 GiB of memory without significant quality loss. It can also be noted that under more stringent memory consumption constraints, the proposed method of estimating keypoints stability allows to select the most efficient subset of points in terms of quality compared to selecting based on SURF response.

The number of keypoints in a template directly affects the required computing resources and document recognition performance. The choice of the maximum number of keypoints for a template (parameter *N* in Algorithm 1) is primarily determined by the class of the executable device and the number of target document types. Considering an energy-efficient single-board computer with a RAM capacity of 1-4 GB and several thousand target document types, the algorithm may exceed the memory limit. The paper [9] provides a measurement of the memory consumption for storing descriptors for 50 document types from MIDV using an unlimited number of SURF keypoints. In this case, the average size of one template takes ~1.64 MiB of memory, and in the case of 1000 document types, more than 1.5 GiB of memory is required just to store descriptors. Considering memory-efficient document-oriented RFDoc descriptors [9], to store 11,000 templates which are supported by an industrial recognition system [30] will require ~1900 MiB of memory.

To verify the effectiveness of the image synthesis methods we use, which mainly address distortions related to documents captured in video stream in uncontrolled conditions, we also assessed the detection quality on images produced by the flatbed scanner. Tab. 3 shows the experimental results obtained on MIDV-2020 dataset. The experimental results show that the proposed method also allows for increased detection performance in high-resolution images.

Tab. 3. Document classification and localization performance in the scan parts of the MIDV-2020

N	Scan upright				Scan rotated			
	Localization, %		Typization, %		Localization, %		Typization, %	
	Resp.	Stab.	Resp.	Stab.	Resp.	Stab.	Resp.	Stab.
50	48.95	<b>49.94</b>	80.1	<b>88.3</b>	33.05	<b>33.14</b>	71.3	<b>74.1</b>
100	67.5	<b>70.74</b>	89.1	<b>95.4</b>	57.42	<b>60.01</b>	84.2	<b>92.6</b>
150	76.51	<b>77.5</b>	93.5	<b>97.8</b>	69.1	69.1	91.4	<b>97.4</b>
200	79.53	<b>86.86</b>	98.5	<b>99.8</b>	73.82	<b>78.0</b>	95.2	<b>98.9</b>
250	83.39	<b>87.65</b>	100	100	78.43	<b>83.18</b>	97.9	<b>99.3</b>
300	85.74	<b>89.16</b>	100	100	81.7	<b>83.98</b>	98.8	<b>99.6</b>
350	87.4	<b>90.68</b>	100	100	83.49	<b>84.49</b>	99.3	<b>99.9</b>
400	88.23	<b>90.95</b>	100	100	83.83	<b>88.94</b>	99.8	99.8
...								
∞	95.63	-	100.0	-	93.62	-	100.0	-

#### 4. Conclusion

In this paper, we proposed a method to improve the scalability of the method of localization and identification of identity documents with local features matching. The proposed method consists of selecting the most effective set of

keypoints for describing a document template using one ideal document image. The method consists of a preliminary assessment of the stability of extracting a template keypoint on a set of synthesized document images relevant to the problem of recognizing identity documents under uncontrolled conditions and is independent of the keypoint detection algorithm. Experimental results show, that the proposed method significantly increases the quality of localization and identification of documents compared with the keypoint selection based on estimates provided by the keypoint detection algorithm while working with a large number of document types and limited computational resources.

For future work extension, we plan to research more advanced methods of image synthesis, including JPEG compression, dithering, and color distortions, and to investigate the independence of the method for template keypoint stability estimation from the used keypoint extraction algorithm. Another possible research is to study the correlation of the distributions of the keypoints' responses provided by different detectors and their stability in the context of identity document recognition in uncontrolled conditions.

### References

- [1] 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.
- [2] V. V. Arlazarov, Key stages of document template processing in modern identification document recognition systems, *Trudy ISA RAN (Proceedings of ISA RAS)* 72 (3) (2022) 19–25, doi: 10.14357/20790279220303.
- [3] A. M. Awal, N. Ghanmi, R. Siere, T. Furon, Complex document classification and localization application on identity document images, in: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), IEEE, 2017, p. 426–431. doi:10.1109/icdar.2017.77. URL <http://dx.doi.org/10.1109/ICDAR.2017.77>
- [4] N. Skoryukina, V. V. Arlazarov, D. P. Nikolaev, Fast method of id documents location and type identification for mobile and server application, in: ICDAR 2019, The Institute of Electrical and Electronics Engineers (IEEE), Manhattan, New York, U.S., 2019, pp. 850–857, doi: 10.1109/ICDAR.2019.00141.
- [5] G. Chiron, N. Ghanmi, A. M. Awal, Id documents matching and localization with multi-hypothesis constraints, in: 2020 25th International Conference on Pattern Recognition (ICPR), IEEE, 2021, p. 3644–3651. doi:10.1109/icpr48806.2021.9412437. URL <http://dx.doi.org/10.1109/ICPR48806.2021.9412437>
- [6] N. Z. Valishina, A. V. Gayer, N. S. Skoryukina, V. V. Arlazarov, Fast keypoint filtering for feature-based identity documents classification on complex background, in: W. Osten, D. Nikolaev, J. Debayle (Eds.), *ICMV 2023*, Vol. 13072, Society of Photo-Optical Instrumentation Engineers (SPIE), Bellingham, Washington 98227-0010 USA, 2024, pp. 1307205–1–1307205–9. doi:DOI: 10.1117/12.3023194.
- [7] D. V. Tropin, I. A. Konovalenko, N. S. Skoryukina, D. P. Nikolaev, V. V. Arlazarov, Improved algorithm of id card detection by a priori knowledge of the document aspect ratio, in: *ICMV 2020*, Vol. 11605, Society of Photo-Optical Instrumentation Engineers (SPIE), Bellingham, Washington 98227-0010 USA, 2021, pp. 116051F1–116051F9. doi:DOI: 10.1117/12.2587029.
- [8] N. S. Skoryukina, E. A. Shalnova, V. V. Arlazarov, Method for detecting false responses of localization and identification algorithms using global features, *ITiVS* (4) (2023) 28–36, doi: 10.14357/20718632230403.
- [9] D. P. Matalov, E. E. Limonova, N. S. Skoryukina, V. V. Arlazarov, Rfdoc: memory efficient local descriptors for id documents localization and classification, in: J. Lladós, D. Lopresti, S. Uchida (Eds.), *ICDAR 2021*, 2nd Edition, Vol. 12822 of Lecture Notes in Computer Science (LNCS), Springer Nature Group, London, UK (main office), 2021, pp. 209–224, doi: 10.1007/978-3-030-86331-9\_14.
- [10] V. V. Arlazarov, K. Bulatov, T. Chernov, V. L. Arlazarov, Midv-500: A dataset for identity document analysis and recognition on mobile devices in video stream, *Computer Optics* 43 (5) (2019) 818–824, doi: 10.18287/2412-6179-2019-43-5-818-824.
- [11] N. Skoryukina, V. V. Arlazarov, A. Milovzorov, Memory consumption reduction for identity document classification with local and global features combination, in: W. Osten, J. Zhou, D. P. Nikolaev (Eds.), *Thirteenth International Conference on Machine Vision*, SPIE, 2021, p. 36. doi:10.1117/12.2587033. URL <http://dx.doi.org/10.1117/12.2587033>
- [12] T. Trzcinski, M. Christoudias, V. Lepetit, Learning image descriptors with boosting, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37 (3) (2015) 597–610. doi:10.1109/tpami.2014.2343961. URL <http://dx.doi.org/10.1109/TPAMI.2014.2343961>
- [13] H. Bay, A. Ess, T. Tuytelaars, L. Van Gool, Speeded-up robust features (surf), *Computer Vision and Image Understanding* 110 (3) (2008) 346–359. doi:10.1016/j.cviu.2007.09.014. URL <http://dx.doi.org/10.1016/j.cviu.2007.09.014>
- [14] D. P. Matalov, E. E. Limonova, N. S. Skoryukina, V. V. Arlazarov, Memory efficient local features descriptor for identity document detection on mobile and embedded devices, *IEEE Access* 11 (2022) 1104–1114, doi: 10.1109/ACCESS.2022.3233463.
- [15] J. Lerouge, G. Betmont, T. Bres, E. Stepankevich, A. Bergès, Docxpan-25k: a large and diverse benchmark dataset for identity documents analysis, *arXiv preprint arXiv:2407.20662* (2024).
- [16] L. Xie, Y. Wang, H. Guan, S. Nag, R. Goel, N. Swamy, Y. Yang, C. Xiao, J. Prisby, R. Maciejewski, J. Zou, Idnet: A novel identity document dataset via few-shot and quality-driven synthetic data generation, in: 2024 IEEE International Conference on Big Data (BigData), 2024, pp. 2244–2253. doi:10.1109/BigData62323.2024.10825017.
- [17] K. Bulatov, D. Matalov, V. V. Arlazarov, Midv-2019: Challenges of the modern mobile-based document ocr, in: W. Osten, D. Nikolaev, J. Zhou (Eds.), *ICMV 2019*, Vol. 11433, Society of Photo-Optical Instrumentation Engineers (SPIE), Bellingham, Washington 98227-0010 USA, 2020, pp. 114332N1–114332N6, doi: 10.1117/12.2558438.
- [18] K. B. Bulatov, E. V. Emelyanova, D. V. Tropin, N. S. Skoryukina, Y. S. Chernyshova, A. V. Sheshkus, S. A. Usilin, Z. Ming, J.-C. Burie, M. M. Luqman, V. V. Arlazarov, Midv-2020: A comprehensive benchmark dataset for identity document analysis, *Computer Optics* 46 (2) (2022) 252–270, doi: 10.18287/2412-6179-CO-1006.
- [19] J. Shemiakina, I. Konovalenko, D. Tropin, I. Faradjev, Fast projective image rectification for planar objects with manhattan structure, in: W. Osten, D. P. Nikolaev (Eds.), *Twelfth International Conference on Machine Vision (ICMV 2019)*, SPIE, 2020, p. 123. doi:10.1117/12.2559630. URL <http://dx.doi.org/10.1117/12.2559630>

- [20] C. Harris, M. Stephens, A combined corner and edge detector , in: Proceedings of the Alvey Vision Conference 1988, AVC 1988, Alvey Vision Club, 1988, pp. 23.1–23.6. doi:10.5244/c.2.23 . URL <http://dx.doi.org/10.5244/C.2.23>
- [21] D. Lowe, Object recognition from local scale-invariant features , in: Proceedings of the Seventh IEEE International Conference on Computer Vision, IEEE, 1999, p. 1150–1157 vol.2. doi:10.1109/iccv.1999.790410 . URL <http://dx.doi.org/10.1109/ICCV.1999.790410>
- [22] Y. Verdier, K. Yi, P. Fua, V. Lepetit, Tilde: A temporally invariant learned detector, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.
- [23] D. DeTone, T. Malisiewicz, A. Rabinovich, Superpoint: Self-supervised interest point detection and description, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2018.
- [24] Y. Ono, E. Trulls, P. Fua, K. M. Yi, Lf-net: Learning local features from images, Advances in neural information processing systems 31 (2018).
- [25] C. Schmid, R. Mohr, C. Bauckhage, Evaluation of interest point detectors , International Journal of Computer Vision 37 (2) (2000) 151–172. doi:10.1023/a:1008199403446 . URL <http://dx.doi.org/10.1023/A:1008199403446>
- [26] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, L. V. Gool, A comparison of affine region detectors , International Journal of Computer Vision 65 (1–2) (2005) 43–72. doi:10.1007/s11263-005-3848-x . URL <http://dx.doi.org/10.1007/s11263-005-3848-x>
- [27] B. Li, R. Xiao, Z. Li, R. Cai, B.-L. Lu, L. Zhang, Rank-sift: Learning to rank repeatable local interest points , in: CVPR 2011, IEEE, 2011. doi:10.1109/cvpr.2011.5995461 . URL <http://dx.doi.org/10.1109/CVPR.2011.5995461>
- [28] M. Muja, D. G. Lowe, Fast approximate nearest neighbors with automatic algorithm configuration., VISAPP (1) 2 (331-340) (2009) 2.
- [29] S. O. Emelyanov, A. A. Ivanova, E. A. Shvets, D. P. Nikolaev, Methods of training data augmentation in the task of image classification, Sensory systems 32 (3) (2018) 236–245. doi:DOI: 10.1134/S0235009218030058 .
- [30] 11,000 document templates from all over the world, <https://web.archive.org/web/20250118171700/https://regulaforensics.com/news/regula-has-added-another-thousand-to-its-database-of-document-templates/> [Accessed: 01.03.2025] (2022).

---

#### *Authors' information*

**Matalov Daniil Pavlovich** (b. 1996) received his bachelor's degree in applied mathematics from the National University of Science and Technology "MISiS", Moscow, Russia, in 2017, and the master's degree in physics, mathematics and computer science from the Moscow Institute of Physics and Technology, in 2019. Since 2020, he has been employed with the Federal Research Center "Computer Science and Control", Russian Academy of Sciences. His research interests include object detection, machine learning, pattern recognition, efficient image recognition algorithms. E-mail: [d.matalov@smartengines.com](mailto:d.matalov@smartengines.com); ORCID: <https://orcid.org/0000-0003-3260-9104>

**Vladimir Viktorovich Arlazarov** (b. 1976) studied at Moscow Institute of Steel and Alloys where he completed his Specialist degree in 1999. Received his Ph.D. degree in Computer Science in 2005, and Doctor of Sciences degree in computer science in 2023. Currently he works as a lead researcher and head of a department at the FRC "Computer Science and Control" RAS. Since 2016, he has been a general director of Smart Engines Service LLC. Research interests: computer vision and document analysis systems. E-mail: [vva@smartengines.com](mailto:vva@smartengines.com); ORCID: <https://orcid.org/0000-0002-1079-2414>

---

*Received July 01, 2025. The final version – August, 18 2025.*

---