mathematics

MDPI

Article

A Method of Image Quality Assessment for Text Recognition on Camera-Captured and Projectively Distorted Documents

Julia Shemiakina [1,*], Elena Limonova [1,2,3], Natalya Skoryukina [1,2], Vladimir V. Arlazarov [1,2] and Dmitry P. Nikolaev [1,4]

[1] Smart Engines Service LLC, 117312 Moscow, Russia; limonova@smartengines.com (E.L.); skleppy.inc@smartengines.com (N.S.); vva@smartengines.com (V.V.A.); dimonstr@iitp.ru (D.P.N.)
[2] Federal Research Center Computer Science and Control RAS, 119333 Moscow, Russia
[3] Department of Innovation and High Technology, Moscow Institute of Physics and Technology, 117303 Moscow, Russia
[4] Institute for Information Transmission Problems (Kharkevich Institute) RAS, 127051 Moscow, Russia
[*] Correspondence: jshemiakina@smartengines.com

Abstract: In this paper, we consider the problem of identity document recognition in images captured with a mobile device camera. A high level of projective distortion leads to poor quality of the restored text images and, hence, to unreliable recognition results. We propose a novel, theoretically based method for estimating the projective distortion level at a restored image point. On this basis, we suggest a new method of binary quality estimation of projectively restored field images. The method analyzes the projective homography only and does not depend on the image size. The text font and height of an evaluated field are assumed to be predefined in the document template. This information is used to estimate the maximum level of distortion acceptable for recognition. The method was tested on a dataset of synthetically distorted field images. Synthetic images were created based on document template images from the publicly available dataset MIDV-2019. In the experiments, the method shows stable predictive values for different strings of one font and height. When used as a pre-recognition rejection method, it demonstrates a positive predictive value of 86.7% and a negative predictive value of 64.1% on the synthetic dataset. A comparison with other geometric quality assessment methods shows the superiority of our approach.

Keywords: projective distortion; image quality assessment; document analysis; text recognition

check for updates

Citation: Shemiakina, J.; Limonova, E.; Skoryukina, N.; Arlazarov, V.V.; Nikolaev, D.P. A Method of Image Quality Assessment for Text Recognition on Camera-Captured and Projectively Distorted Documents. *Mathematics* **2021**, *9*, 2155. https://doi.org/10.3390/math9172155

Academic Editor: Javier Martínez

Received: 1 June 2021
Accepted: 30 August 2021
Published: 3 September 2021

1. Introduction

The object recognition problem, which has been extensively studied in past decades, has a wide range of real applications. The accuracy of recognition systems is always the first priority. However, the error cost largely depends on the problem's specifics or on the particular application of the developed system. In many areas, such as identity verification [1–3], self-driving vehicles [4,5], and industrial diagnostics [6,7], incorrect recognition can cause financial loss or even harmful health outcomes. For this reason, the prediction of recognition reliability is vital for such systems. Obtaining an uncertain result should lead to the rejection of the image processing output or the transfer of control back to the user to prevent unfortunate situations. Therefore, modern recognition systems include different types of reliability assessment modules. There are three main approaches to reliability evaluation: recognition confidence analysis, pixel-based image quality assessment, and geometric image quality assessment. These approaches work at different recognition stages and, of course, can be applied together.

The first approach involves the estimation of the recognition confidence provided by the recognition module. Such systems aggregate the confidences of all recognized objects, such as text lines, and decide to accept or reject the recognition result depending on the error cost [8]. However, these methods have several problems. First, recognition neural networks can be unstable to changes in input images. An alteration in several pixels may lead to a change in the recognition result [9,10] and, thus, different confidences. Additionally, neural networks tend to be overconfident, i.e., return high confidences for incorrectly recognized images [11]. Moreover, this approach requires the entire recognition workflow to be performed for all objects/text fields or other data structures found, even if they are incorrectly segmented. Thus, it leads to unnecessary time spent on recognition with a high possibility of an incorrect result and, therefore, efficiency degradation, which can be a problem for mobile devices and embedded systems.

The other two approaches use image quality assessment because poor image quality is considered one of the most important sources of unstable recognition quality. Many works show a correlation between different image distortions and recognition accuracy [12–14]. These distortions can occur for many reasons, such as compression, transmission artifacts, or uncontrolled capturing conditions, with the possible presence of highlights, motion blur, defocus, or geometric distortions [15]. Quality assessment methods can be divided into two groups: pixel-based, which analyzes the image's pixel values, and geometric, which does not.

The main focus of recent research has been on pixel-based methods. These methods consider distortions such as blur, digital noise, compression, and transmission artifacts [16,17], given that these distortions are common in the majority of recognition system application fields. In addition, methods exist for evaluating a specular highlight saliency map by utilizing deep learning [18], an unnormalized form of Wiener entropy [19], and other approaches. The authors of [20] presented a detection method for holographic elements, which may significantly decrease text recognition accuracy.

Such methods can easily be incorporated into recognition systems. For example, in [21,22], the authors present a model of an optical recognition system with embedded image quality assessment and feedback modules. They rejected images of poor quality before recognition. This approach demonstrates an increase in recognition accuracy and reliability. Moreover, the authors showed that, in the case of recognition in a video stream, these modules provide new possibilities, such as selecting the best quality frames for further recognition or rejection of the worst frames. Considering the problem of document recognition, the assessment of text field images allows for a reevaluation of the confidence of the recognized text. The confidences obtained for one field in different frames can be further used as weights in the combination method for text field recognition in a video stream [23,24]. Pixel-based quality assessment methods need to analyze the whole input image, which can be time-consuming (especially if deep learning is used) and may be a problem for real-time recognition systems.

Geometric quality assessment methods analyze the geometric distortion of an object in an image. In the case of document recognition in images taken with a mobile device camera, the most common distortion is a projective transform of a plane. A user trying to avoid highlights may take a photo with a high projective distortion of the document. In this case, document text regions become poorly recognizable (Figure 1). Geometric quality assessment methods allow such regions to be rejected without analyzing their pixel intensities. This approach is fast and perfectly suited for the document recognition problem. However, there is a lack of research considering this subject.

In [25], the authors consider the recognition of rectangular documents. They obtained document quadrangles and checked three conditions: (1) at least one pair of opposed quadrangle edges is parallel, (2) the average difference in angles between each pair of opposed angles is relatively small, and (3) the average perpendicularity of the four vertices is less than 25°. In [26], the criterion includes the following conditions: (1) the ratio of a document quadrangle area to the area of the whole image must exceed a threshold, (2) the aspect ratio for the document quadrangle must fit some predefined interval, and (3) the angles of the document quadrangle must be close to 90°. Unfortunately, the authors do not report the thresholds and intervals used, so it is impossible to evaluate them experimentally. These empirical methods are reasonable. However, there is no theoretical proof or experimental evaluation of their connection to the level of projective distortion and recognition accuracy. For example, considering the relative area of a recognized object, images restored from one area may have significantly different quality (Figure 2).

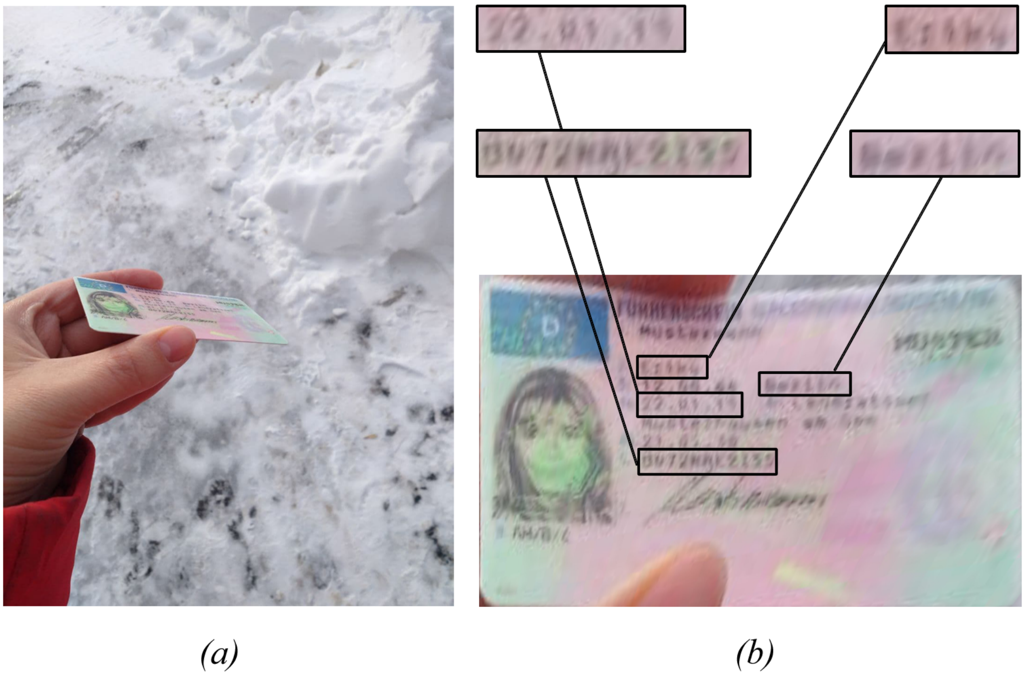


Figure 1. An example of document projective distortion (**a**) and the resulting quality of the extracted text regions (**b**).

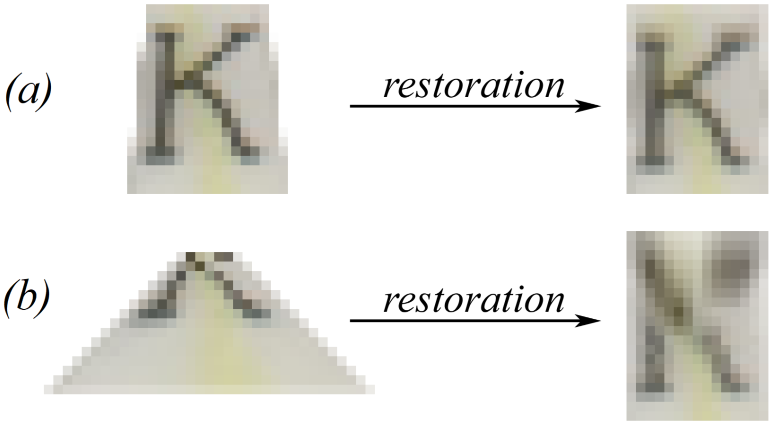


Figure 2. Restoration results of (**a**) a slightly projectively distorted image and (**b**) a highly distorted image of equal area. Restoration was conducted with bilinear interpolation.

In this paper, we propose a novel, no-reference method for the quality assessment of images restored from projectively distorted sources. The image quality is considered in terms of the probability of correct text recognition. The proposed method was tested experimentally on synthetic data created from the publicly available dataset MIDV-2019 [27].

2. Document Image Quality Assessment Problem Statement

We consider the problem of document recognition in images obtained with a mobile camera. We use the pinhole camera model (Figure 3), so the camera is assumed to have no optical aberrations. Given that the document is a flat rectangular object, the document image is affected by projective distortion [28], and the document boundary is a quadrangle.

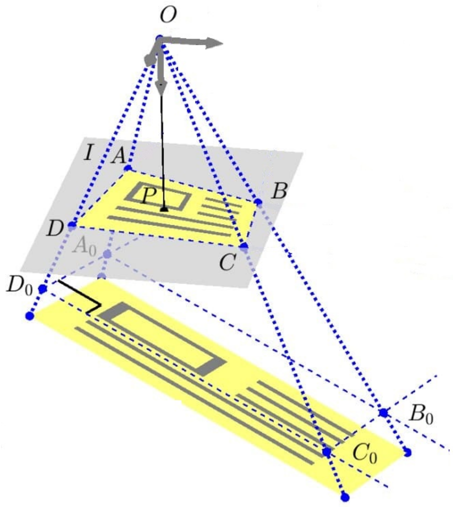


Figure 3. The captured image of the rectangular document in a pinhole camera model.

Document recognition systems commonly consist of several submodules: document localization in a source image, segmentation of required zones such as text and photo fields, and field image restoration and recognition (Figure 4). Considering the field segmentation step, the majority of the systems utilize document models. There are three general classes of models: templates, flexible forms, and end-to-end models. Templates define the strictest constraints on the location of each zone and are most commonly used for identity documents. In [25,29,30], document templates are used for the localization and classification of document images, but they are also helpful in field segmentation [31]. Flexible form models are based on text segmentation and recognition result analysis and describe documents with soft restrictions on their structure. This model may contain text feature points [32] or attributed relational graphs [33] as a structural representation of a document. End-to-end models involve the simultaneous segmentation and recognition of text field regions [34] and may not require any document structure.

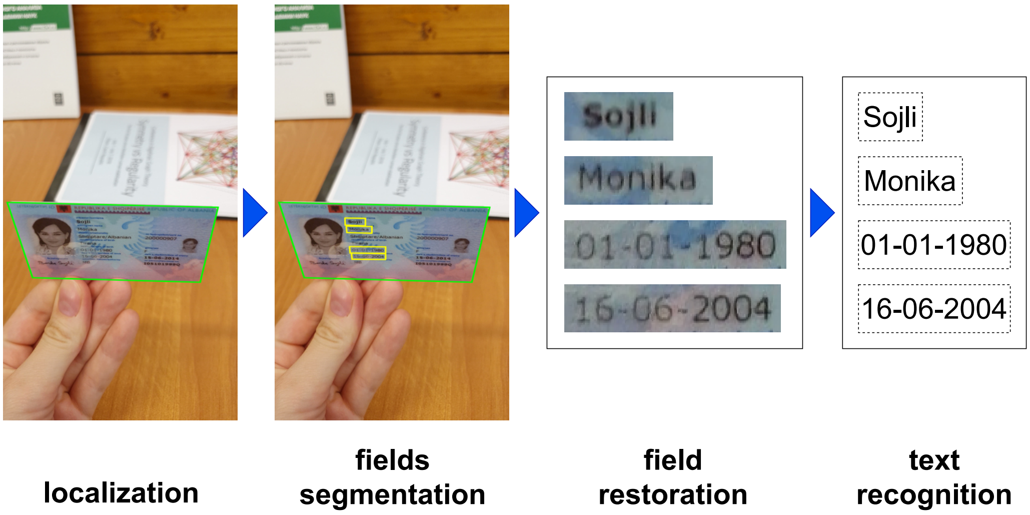


Figure 4. The general model of the document recognition system.

We consider identity document recognition systems such as [2] based on the template description of documents. For many identity documents, the regions of text and photo fields are fixed, and text fonts and font properties (size, boldness, etc.) for each of them are known. This information may be inserted into the document template description and used to further assess field image quality.

We assume that the results of the field segmentation are provided as a quadrangle in the source image. According to its coordinates, the field image should be restored and recognized. The main goal of this paper is to assess the quality of the restored image in terms of the reliability of recognition before the restoration itself (see Figure 5). If the quality is insufficient, then the system ceases further processing to prevent false recognition results. Moreover, the possibility of early rejection decreases the runtime of the system, as the restoration and recognition are not performed on images of low quality. Based only on the source field quadrangle and a priori information of its size and fon, the restored field image can be assessed by relying on known properties of further submodules. We briefly discuss the submodules below.

The restoration submodule resamples the source image according to the projective transform, which maps the quadrangle of the field in the source image to the rectangle of a predefined size in the document model. The resampling process is usually characterized by interpolation and antialiasing methods. Given that, under a high level of projectivity, the field image may have an arbitrary small area and consequently may not be recognizable, in this work, we focus only on the magnification problem when mapping magnifies a source region. In this particular case, anti-aliasing methods can be excluded, as after magnification, the restored image cannot contain high frequencies. The most well-known interpolation methods [35] are the nearest neighbor, bilinear, bicubic, and cubic B-spline methods. For all of the mentioned interpolation methods, except the nearest-neighbor algorithm, a small source area causes blur in the restored area, as shown in Figure 2. The images obtained by nearest-neighbor interpolation have comparatively low quality (see Figure 6), so we exclude this method from consideration.

The results presented in [21,36] show that the presence of blur in text field images decreases the quality of recognition. It should be noted that the recognition submodule may contain a pre-processing step that refines the image using a deblurring method, for example, [37]. However, its scope is limited, and a level of blurring exists under which the text field cannot be reliably recognized.

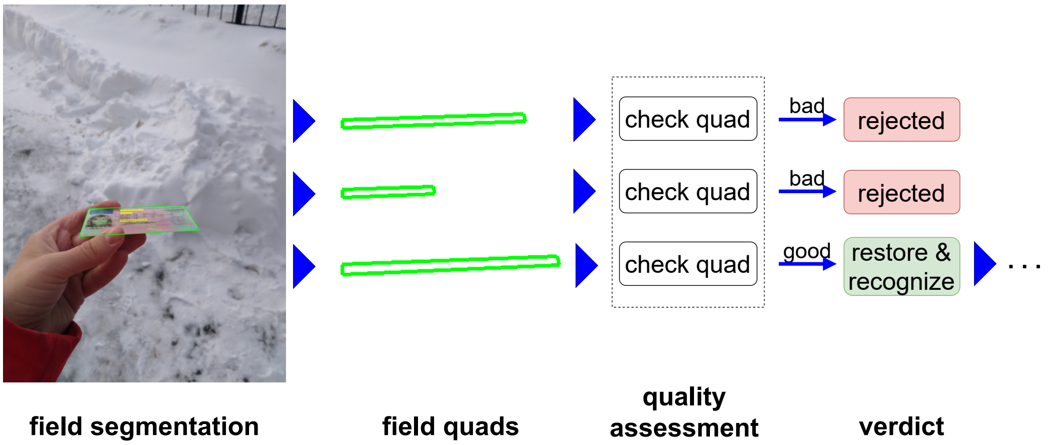


Figure 5. The model of the quality assessment submodule.

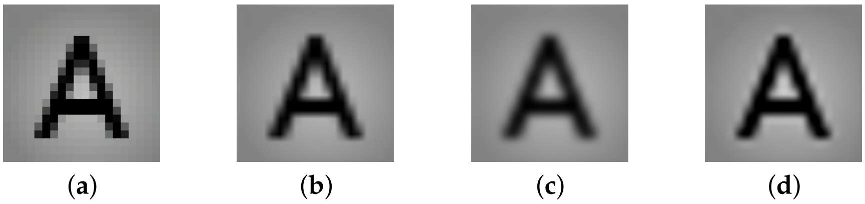


Figure 6. Interpolation examples [38]. (**a**) Nearest pixel, (**b**) bilinear, (**c**) B-spline, and (**d**) bicubic.

Assuming that the image restoration and the text recognition submodules can be predefined, we can use the given font and size of the field to estimate the maximum local distortion level that provides stable recognition of any text. We assume that the level can be evaluated as a rational value and denote the threshold distortion level as $\theta \in \mathbb{R}$. In this work, however, it is more convenient for us to use the inverse value $l \in \mathbb{R}, l = \frac{1}{\theta}$, which we call the minimum scaling coefficient threshold. It should be noted that this threshold value is presumed to be evaluated once while developing the recognition system.

Let us denote the source image as I_{src}, the segmented field quadrangle, i.e., four points of its corners in the source image, as F, and the rectangle of the restored image borders, defined in the document model, as R. We need to estimate whether the quality of the restored field image I_{rst} is sufficient in terms of reliability for further text recognition. For this purpose, let us denote the quality assessment function as Q. Q analyzes the source field quadrangle F, the restored field rectangle R, and the a priori minimum scaling coefficient threshold l and returns 1 if the image quality allows for reliable recognition and 0 otherwise:

$$Q(F, R, l) : \mathcal{F} \times \mathcal{R} \times \mathbb{R} \to \{0, 1\}, \tag{1}$$

where $\mathcal{F}$ is the set of all quadrangles lying inside the source image and $\mathcal{R}$ is the set of all possible rectangles. The function Q does not take the restored image itself as an argument. The evaluation process here is assumed to involve the analysis of a geometric transform rather than pixel intensities. Therefore, the quality assessment can be conducted before the use of the restoration submodule (Figure 3).

3. The Models of Distorted Field Image Acquisition and Restoration

First, let us briefly describe the model of projectively distorted text field image acquisition [35]. For simplification, we consider the one-dimensional case. Let us define the undistorted field image signal as a continuous bounded function $I(x)$:

$$0 < I(x) < B,\ x \in \mathbb{R},\ B \in \mathbb{R}, \tag{2}$$

where B is the upper bound of $I(x)$.

While being captured with a camera, the signal is distorted with a projective transform $u = H(x)$:

$$I^c_{src}(u) = I(H^{-1}(u)),\ u \in \mathbb{R}, \tag{3}$$

where $I^c_{src}(u)$ is the continuous projectively distorted signal. Then, the signal $I^c_{src}(u)$ is sampled by a function $s(u)$ with a known sampling pitch Δu_s to obtain a discrete image $I_{src}(k),\ k \in \mathbb{Z}$:

$$\begin{aligned} I^d_{src}(u) &= I^c_{src}(u)s(u), \\ I_{src}(k) &= I^d_{src}(k\Delta u_s),\ k \in \mathbb{Z}, \end{aligned} \tag{4}$$

where $I^d_{src}(u)$ is the sampled distorted signal defined on $\mathbb{R}$. We consider ideal sampling with the following:

$$s(u) = \sum_{n \in \mathbb{Z}} \delta(u - n\Delta u_s), \tag{5}$$

where $\delta(u)$ is the Dirac delta function.

The image $I_{src}(k)$ is the input of the recognition system. Before the final text recognition, the image should be restored to compensate for the projective distortion. In the image restoration process, the image is resampled with the inverse of the original projective transform $x = H^{-1}(u)$. This transformation can be evaluated based on the source field quadrangle F obtained in the field segmentation step and the rectangle of the restored field R defined by the template description: $R = H^{-1}(F)$. The resampling model is as follows. The discrete image $I_{src}(k)$ is reconstructed to obtain a continuous signal $I^c_{src}(u)$ through convolution with a reconstruction filter $r(u)$:

$$I^c_{src}(u) = \sum_{k \in \mathbb{Z}} I_{src}(k) r(u - k\Delta u_s),\ u \in \mathbb{R}. \tag{6}$$

After that, the domain of the continuous signal $I^c_{src}(u)$ is warped with the projective transform $x = H^{-1}(u)$:

$$I^c_{rst}(x) = I^c_{src}(H(x)),\ x \in \mathbb{R}, \tag{7}$$

where $I^c_{rst}(x)$ is the restored continuous signal.

Depending on the mapping function $H^{-1}(x)$, $I^c_{rst}(x)$ may have arbitrary high frequencies. To conform to the Nyquist rate, the signal should be bandlimited by a prefilter function $h(x, y)$ that prevents aliasing:

$$\hat{I}^c_{rst}(x) = I^c_{rst}(x) \circledast h(x) = \int\limits_{\mathbb{R}} I^c_{rst}(t)h(x-t)dt, \tag{8}$$

where $\hat{I}^c_{rst}(x)$ is the bandlimited restored signal and $\circledast$ denotes convolution. Then, the obtained signal is sampled with the same sampling pitch $\Delta x_s = \Delta u_s$:

$$\begin{aligned} I^d_{rst}(x) &= \hat{I}^c_{rst}(x)s(x), \\ I_{rst}(j) &= I^d_{rst}(j\Delta x_s),\ j \in \mathbb{Z}, \end{aligned} \tag{9}$$

where $I^d_{rst}(x)$ is the sampled restored signal on $\mathbb{R}$ and $I_{rst}(j)$ is the discrete restored signal.

In this paper, we consider only the magnification case, when the source region is stretched:

$$|H(x) - H(x + \Delta x)| < \Delta x \tag{10}$$

In this scenario, the signal mapping cannot provide high frequencies. Therefore, the prefilter has little impact on the restored image signal and can be ignored. Then, the restored image $I_{rst}(j)$ is as follows:

$$I_{rst}(j) = \sum_{k \in \mathbb{Z}} I_{src}(k)r(H(j\Delta x_s) - k\Delta u_s),\ j \in \mathbb{Z}. \tag{11}$$

Let us define the sample pitches as equal to 1: $\Delta x_s = \Delta u_s = 1$. For simplicity, we refer to discrete images $I_{src}(k)$ and $I_{rst}(j)$ as $I_{src}(u)$ and $I_{rst}(x)$ and specify $u, x \in \mathbb{Z}$. Then, Formula (11) can be rewritten as follows:

$$I_{rst}(x) = \sum_{k \in \mathbb{Z}} I_{src}(k)r(H(x) - k),\ x \in \mathbb{Z}. \tag{12}$$

The ideal reconstruction filter $r(u)$, $u \in \mathbb{R}$, is an ideal low-pass filter sinc $= {\sin \pi x}/{\pi x}$ according to the cardinal theorem of interpolation [39]. However, in practice, one uses its approximations with a finite window radius R:

$$r(u) = 0, |u| \geq R. \tag{13}$$

The bilinea, bicubic B-spline, and bicubic reconstruction functions have finite windows of radii $R = 1$ and $R = 2$.

We also assume that the reconstruction function is Lipschitz continuous with constant M:

$$|r(u) - r(u + \Delta u)| < M\Delta u,\ M > 0. \tag{14}$$

Hypothesis 1. *The bilinear, bicubic B-spline, and bicubic reconstruction functions (see Figure 7) are Lipschitz continuous.*

Let us consider the bilinear reconstruction function.

Lemma 1. *The bilinear interpolation function $r_l(u)$ is Lipschitz continuous (14), where $r_l(u)$ is defined as follows:*

$$r_l(u) = \begin{cases} 1 - |u|, & |u| < 1, \\ 0, & |u| \geq 1. \end{cases} \tag{15}$$

Proof. Let us consider a pair of arbitrary points x, y. Due to the piecewise nature of $r_l(u)$, we have three cases.

Case 1: $\forall x, y \in (-\infty, -1] \cup [1, \infty)$

$$|r_l(x) - r_l(y)| = 0 \leq M|x - y|, \ \forall M > 0. \tag{16}$$

Case 2: $\forall x, y \in (-1, 1)$
By the reverse triangle inequality, we can obtain:

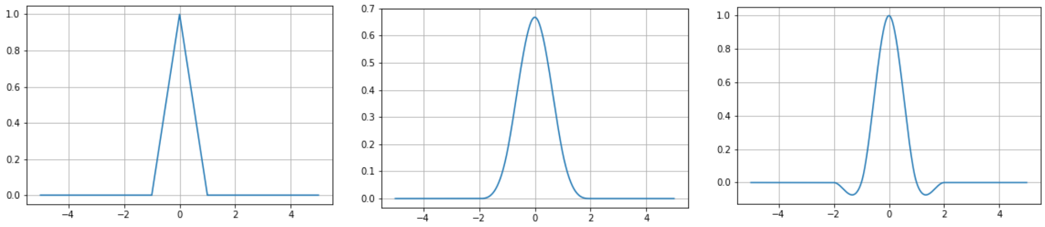
$$|r_l(x) - r_l(y)| = ||1 - |x|| - |1 - |y||| \leq ||x| - |y|| \leq |x - y|. \tag{17}$$

Case 3: $\forall x \in (-1, 1), \forall y \in (-\infty, -1] \cup [1, \infty)$

$$|r_l(x) - r_l(y)| = |1 - |x|| = 1 - |x| \leq |y| - |x| \leq |x - y|. \tag{18}$$

Hence, the bilinear reconstruction function $r_l(u)$ is Lipschitz continuous with the constant $M = 1$. □

The bicubic B-spline and bicubic reconstruction functions are shown in Figure 7b,c. We can see that they are continuous and have a bounded value increment and are thus Lipschitz continuous. The direct proof of Hypothesis 1 falls outside of the scope of this work.



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

Figure 7. Reconstruction functions: (**a**) bilinear, (**b**) B-spline, and (**c**) bicubic.

4. The Minimum Scaling Coefficient Assessment at a Restored Image Point

In [21,36], the authors incorporated an estimation of image blur into the algorithms of a combination of text recognition results in a video stream. Since an unblurred text image has high contrast in regions corresponding to strokes, they assume that the level of image blur is inversely related to the sharpness (called focus in the cited papers), which represents the directional minimum of the highest local contrasts of the image. In these papers, the blur is caused by defocusing or motion blur and is constant for the whole image. The sharpness is calculated based on the intensities in the source image. For this purpose, gradient images are calculated in different directions; for each of them, a 0.95 quantile of the gradient image is obtained, and their minimum represents the sharpness estimation.

In our case, the blurring distortion of the restored image is caused by projective mapping and, hence, is uneven over different points of the image. Let us consider the original undistorted image $I(x)$ and denote its local contrast in a region between neighboring sampling points $[\bar{x}, \bar{x} + \Delta x_s]$ as $L(\bar{x}, \Delta x_s) : 0 < L(\bar{x}, \Delta x_s) \leq B/\Delta x_s$:

$$L(\bar{x}, \Delta x_s) = \frac{|I(\bar{x}) - I(\bar{x} + \Delta x_s)|}{\Delta x_s}. \tag{19}$$

One can verify whether the restored image is able to provide the expected contrast in this region. Let us denote the local contrast of the restored image as $L_{rst}(\bar{x}, \Delta x_s)$. It can be calculated as follows:

$$\begin{aligned} L_{rst}(\bar{x}, \Delta x_s) = \frac{1}{\Delta x_s}|I_{rst}(\bar{x}) - I_{rst}(\bar{x} + \Delta x_s)| \overset{(12)}{=} \frac{1}{\Delta x_s}\Bigg| \sum_{k_1 \in K_1} I_{src}(k_1) r(H(\bar{x}) - k_1) - \\ - \sum_{k_2 \in K_2} I_{src}(k_2) r(H(\bar{x} + \Delta x_s) - k_2)\Bigg|, \end{aligned} \tag{20}$$

where $I_{rst}(x)$ is the discrete restored signal; $I_{src}(k)$ is the discrete distorted signal; and $K_1 = |H(\bar{x}) - t_1| < R,\ t_1 \in \mathbb{Z}$ and $K_2 = |H(\bar{x} + \Delta x_s) - t_2| < R,\ t_2 \in \mathbb{Z}$ are sets of samples in the source image that are used for the reconstruction of samples $\bar{x}$ and $\bar{x} + \Delta x_s$, respectively.

According to (10), the distance between points $H(\bar{x})$ and $H(\bar{x} + \Delta x_s)$ in the source image is less then its sampling pitch:

$$|H(\bar{x}) - H(\bar{x} + \Delta x_s)| < \Delta x_s = \Delta u_s \tag{21}$$

Then, in the worst case, the points $H(\bar{x})$ and $H(\bar{x} + \Delta x_s)$ have the same set of samples used for reconstruction, i.e., $K_1 = K_2$. In that case, the contrast (20) provided by the restored image can be estimated as follows:

$$\begin{aligned} L_{rst}(\bar{x}, \Delta x_s) &= \frac{1}{\Delta x_s}|I_{rst}(\bar{x}) - I_{rst}(\bar{x} + \Delta x_s)| = \\ &= \frac{1}{\Delta x_s}\Bigg| \sum_{k_1 \in K_1} I_{src}(k_1)(r(H(\bar{x}) - k_1) - r(H(\bar{x} + \Delta x_s) - k_1))\Bigg| \leq \\ &\leq \frac{1}{\Delta x_s} \sum_{k_1 \in K_1} I_{src}(k_1)|r(H(\bar{x}) - k_1) - r(H(\bar{x} + \Delta x_s) - k_1)| \overset{(14)}{\leq} \\ &\leq \sum_{k_1 \in K_1} I_{src}(k_1) \frac{M|H(\bar{x}) - H(\bar{x} + \Delta x_s)|}{\Delta x_s} \overset{(2)}{\leq} |K_1| BM \frac{|H(\bar{x}) - H(\bar{x} + \Delta x_s)|}{\Delta x_s}, \end{aligned} \tag{22}$$

where $|K_1|$ is the size of set K_1.

If the restored local contrast is much lower than the contrast in the original undistorted image $L(\bar{x}, \Delta x_s)$, then the restored image edges are highly blurred or even undetectable. As we can see, the upper bound of the restored image local contrast $L_{rst}(\bar{x}, \Delta x_s)$ depends on the distance between the corresponding points in the source image $|H(\bar{x}) - H(\bar{x} + \Delta x_s)|$. Thus, the smaller the distance, the higher the level of blur distortion in the considered region. Then, the ratio of the distance between source points to the sampling pitch can be used to estimate the maximum achievable sharpness of the restored region. Let us denote this function as the scaling coefficient $s(\bar{x}, \Delta x)$:

$$s(\bar{x}, \Delta x_s) = \frac{|H(\bar{x}) - H(\bar{x} + \Delta x_s)|}{\Delta x_s} \tag{23}$$

Above, we considered the one-dimensional case; however, the image is a two-dimensional function. The projective transform of the plane $(u, v) = H(x, y)$ is determined as follows:

$$\begin{cases} u = h_x(x, y) = \dfrac{h_{0,0}x + h_{0,1}y + h_{0,2}}{h_{2,0}x + h_{2,1}y + h_{2,2}} \\ v = h_y(x, y) = \dfrac{h_{1,0}x + h_{1,1}y + h_{1,2}}{h_{2,0}x + h_{2,1}y + h_{2,2}} \end{cases}, \ (x, y) \in \mathbb{R}^2,\ (u, v) \in \mathbb{R}^2, \tag{24}$$

where $h_{q,w},\ q, w \in \{0, 1, 2\}$ are the coefficients of the projective transform H.

The projective transform map points unevenly. For a fixed point (x, y) and several shifts $(\Delta x_m, \Delta y_m)$ of one length: $|(\Delta x_m, \Delta y_m)| = const\ \forall m$, the distance $|H(x, y) - H(x + \Delta x_m, y + \Delta y_m)|$ can significantly vary. Since the directions of text strokes causing high local contrast in the image are arbitrary, the sharpness should be estimated for all possible shifts. Image sampling is conducted with a grid, so the sampling pitch in different directions also varies. However, the function $s(\bar{x}, \Delta x_s)$ is the length ratio, so a useful simplification is to consider Δx_s equal for all of them. It should be noted that, here, we implicitly change the domain of the function $s(\bar{x}, \Delta x_s)$ from $\mathbb{Z}^2$ to $\mathbb{R}^2$. This can be performed because the image function is no longer used, and the projective transform H is defined on a set of real numbers.

Then, the scaling coefficient function $s(\bar{p}, \Delta x_s)$ defined in (23) should be rewritten for the two-dimensional case as follows:

$$s(\bar{p}, \Delta x_s) = \min_{\Delta p : |\Delta p| = \Delta x_s} \frac{|H(\bar{p}) - H(\bar{p} + \Delta p)|}{\Delta x_s}, \bar{p} \in \mathbb{R}^2 \tag{25}$$

Let us denote this function as the *minimum scaling coefficient*. The region under consideration is a circle with the center at the point $\bar{p} = (\bar{x}, \bar{y})$ and the radius equal to Δx_s:

$$|\bar{p} - p| = \Delta x_s,\ p \in \mathbb{R}^2. \tag{26}$$

The projective transform $(u, v) = H(x, y)$ maps the points of the infinity line l_∞ : $h_{2,0}x + h_{2,1}y + h_{2,2} = 0$ to infinity. If the line crosses or touches the circle, then some points of its inner region become infinite, which is not possible in image restoration. Consequently, we can assume that the circle is not crossed by the l_∞ line and is mapped onto an ellipse. Then, the length a_{min} of the ellipse semi-minor axis is the minimum distance between pairs of projected points:

$$\min_{\Delta p : |\Delta p| = \Delta x_s} |H(\bar{p}) - H(\bar{p} + \Delta p)| = a_{min} \tag{27}$$

Since Δp is assumed to be small, one can locally approximate the projective transform H with an affine transform. In this approach, it can be shown that, for a unit circle, the lengths of the ellipse semi-axes are equal to the roots of eigenvalues λ_{min} and λ_{max} of the matrix $\bar{J}^T \bar{J}$, where $\bar{J}$ is the Jacobian matrix of the transform H at the point $\bar{p}$ [40]. Then, for the circle with the radius Δx_s, the lengths of the semi-minor and semi-major axes for the restored point $\bar{p}$, a_{min} and a_{max}, respectively, are calculated as follows:

$$\begin{aligned} &\bar{J} = Jacobian(H, \bar{p}) = \left. \begin{pmatrix} \frac{\partial h_x(x,y)}{\partial x} & \frac{\partial h_x(x,y)}{\partial y} \\ \frac{\partial h_y(x,y)}{\partial x} & \frac{\partial h_y(x,y)}{\partial y} \end{pmatrix} \right|_{(x,y)=(\bar{x},\bar{y})} \\ &\lambda_{min,max}(\bar{p}) :\ \bar{J}^T \bar{J} p = \lambda p, \\ &a_{min}(\bar{p}) = \Delta x_s \sqrt{\lambda_{min}(\bar{p})}, \\ &a_{max}(\bar{p}) = \Delta x_s \sqrt{\lambda_{max}(\bar{p})}. \end{aligned} \tag{28}$$

It should be noted that the points on the infinity line l_∞ become infinite under the transformation, so eigenvalues are not defined on this line. Then, the length function domain is $\mathbb{R}^2 \setminus l_\infty$.

It is a well-known fact that the eigenvalues are the roots of the characteristic equation. Then, the lengths of the semi-minor and semi-major axes can be calculated as follows:

$$a_{min,max}(\bar{p}) = \Delta x_s \sqrt{\lambda_{min,max}(\bar{p})} = \Delta x_s \sqrt{\frac{trace(\bar{J}^T \bar{J}) \pm \sqrt{trace(\bar{J}^T \bar{J})^2 - 4det(\bar{J})^2}}{2}}, \bar{p} \in \mathbb{R}^2 \setminus l_\infty \tag{29}$$

One can derive the values of the trace and the determinant of the matrix $\bar{J}^T\bar{J}$ expressed in terms of coefficients of the homography H:

$$\begin{aligned}
det(\bar{J}) &= \frac{det(H)}{(h_{2,0}\bar{x} + h_{2,1}\bar{y} + h_{2,2})^3} \\
trace(\bar{J}^T\bar{J}) &= \frac{T_1^2 + T_2^2 + T_3^2 + T_4^2}{(h_{2,0}\bar{x} + h_{2,1}\bar{y} + h_{2,2})^4} \\
T_1 &= \alpha\bar{y} + c_1,\ T_2 = -\alpha\bar{x} + c_2,\ T_3 = \beta\bar{y} + c_3,\ T_4 = -\beta\bar{x} + c_4 \\
\alpha &= h_{0,0}h_{2,1} - h_{0,1}h_{2,0},\ \beta = h_{1,0}h_{2,1} - h_{1,1}h_{2,0} \\
c_1 &= h_{0,0}h_{2,2} - h_{0,2}h_{2,0},\ c_2 = h_{0,1}h_{2,2} - h_{0,2}h_{2,1}, \\
c_3 &= h_{1,0}h_{2,2} - h_{1,2}h_{2,0},\ c_4 = h_{1,1}h_{2,2} - h_{1,2}h_{2,1}
\end{aligned} \tag{30}$$

In this work, we use only the values of the semi-minor axis length. However, the other lengths may be helpful in the problem of image decimation estimation. To illustrate the behavior of the semi-minor and semi-major length functions, we constructed heatmaps for a synthetic example. An arbitrary source quadrangle F (Figure 8a) and a restored rectangle R (Figure 8b,c) were used to estimate semi-minor (Figure 8b) and semi-major (Figure 8c) axis lengths at grid points on the restored plane. As we can see, the values increase as we approach the infinity line l_∞, shown as a blue line in the figure. The region inside rectangle R with semi-minor axis lengths less than the threshold appears to be connected.

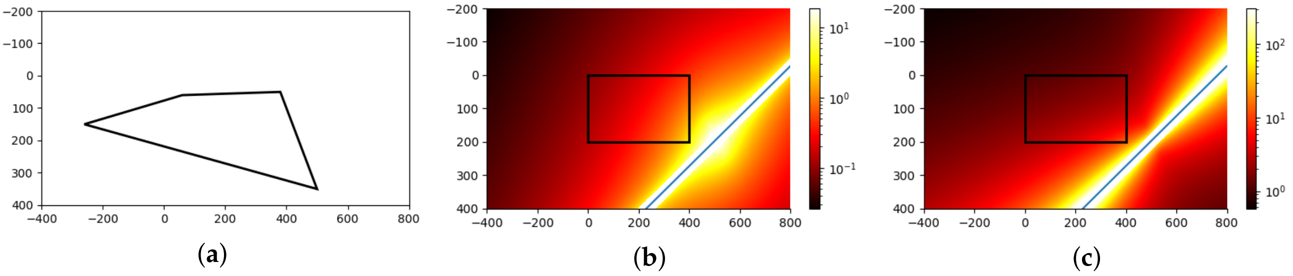


Figure 8. The heatmaps of the semi-minor and semi-major axis lengths: (**a**) the source quadrangle F, (**b**) semi-minor axis lengths, and (**c**) semi-majorFL axis lengths.

Then, according to (27)–(29), the minimum scaling coefficient (25) does not depend on the sampling pitch Δx_s and can be redefined as $s(\bar{p})$:

$$\begin{aligned}
s(\bar{p}) \equiv s(\bar{p}, \Delta x_s) &\overset{(27,28)}{=} \frac{\Delta x_s \sqrt{\lambda_{min}(\bar{p})}}{\Delta x_s} = \\
&\overset{(29)}{=} \sqrt{\frac{trace(\bar{J}^T\bar{J}) - \sqrt{trace(\bar{J}^T\bar{J})^2 - 4det(\bar{J})^2}}{2}},\ \bar{p} \in \mathbb{R}^2 \setminus l_\infty.
\end{aligned} \tag{31}$$

This function can be used to estimate the local sharpness at each point of the restored image and is directly related to the local image quality. It should be noted that, if the transformation H is affine, i.e., $h_{2,0}^2 + h_{2,1}^2 = 0$, then the Jacobian matrix and the minimum scaling coefficient are constant for the whole plane. Thus, only one value at an arbitrary point can be calculated.

5. The Proposed Method of Projectively Distorted Image Quality Assessment

Next, we define the quality assessment method Q, which provides a binary estimation of the whole restored image in terms of recognition reliability. Considering that incorrect recognition of any character leads to incorrectness of the whole recognized field text, the image quality can be estimated according to the region with the lowest quality.

For this purpose, we can estimate the maximum level of local distortion θ that enables stable recognition of the restored image. The threshold depends on the recognition subsystem and on the chosen interpolation algorithm. Since the function $s(\bar{p})$ is inversely proportional to the local distortion level, for simplification, we use the minimum scaling coefficient threshold l, which is the inverse of the level of distortion θ:

$$l = 1/\theta \tag{32}$$

Then, we can construct the level curve of the minimum scaling coefficient function as follows:

$$s(p) = l,\ p \in \mathbb{R}^2 \setminus l_\infty. \tag{33}$$

If the level curve intersects the restored rectangle R, then one of two corresponding parts of the restored field image is not recognized reliably. Otherwise, if there is no intersection, we can calculate the value for one arbitrary point inside the rectangle to check whether the whole restored image has sufficient quality.

According to (31) and (30), the level curve (33) can be written as follows:

$$\begin{aligned} l^4(h_{2,0}x + h_{2,1}y + h_{2,2})^6 - l^2(h_{2,0}x + h_{2,1}y + h_{2,2})^2(T_1^2 + T_2^2 + T_3^2 + T_4^2) + \\ + det(H)^2 = 0,\ (x, y) \in \mathbb{R}^2 \setminus l_\infty. \end{aligned} \tag{34}$$

This equation holds for both the minimum scaling coefficient function $s(p)$ and the maximum scaling coefficient function $s_{max}(p)$, which is defined as the ratio of the semi-major axis length to the sampling pitch Δx_s:

$$s_{max}(p) \equiv \frac{a_{max}(p)}{\Delta x_s},\ p \in \mathbb{R}^2 \setminus l_\infty. \tag{35}$$

If the $s_{max}(p)$ branch intersects the rectangle, then both of its parts have low quality.

For simplification, Equation (34) is translated to the new coordinate system by transformation T:

$$(X, Y) = T(x, y) = (h_{2,1}x - h_{2,0}y, h_{2,0}x + h_{2,1}y + h_{2,2}). \tag{36}$$

Under this transform, the infinity line l_∞ is mapped to the line $Y = 0$. After the substitution of (36) into the level curve in Equation (34), we obtain the following:

$$\begin{aligned} & l^4Y^6 - l^2\frac{\alpha^2 + \beta^2}{h_{2,0}^2 + h_{2,1}^2}(Y^4 + \frac{(\alpha\delta - \beta\gamma)^2}{(\alpha^2 + \beta^2)^2}Y^2 + (X - \frac{\alpha\gamma + \beta\delta}{\alpha^2 + \beta^2})^2Y^2) + det(H)^2 = 0, \\ & Y \neq 0. \end{aligned} \tag{37}$$

where $\gamma = h_{2,0}c_1 + h_{2,1}c_2$, $\delta = h_{2,0}c_3 + h_{2,1}c_4$ and $\alpha, \beta, c_1, c_2, c_3, c_4$ are defined in (30).

As we can see, Equation (37) is quadratic in terms of X and, hence, symmetric. Then, we can approximate it by a piecewise linear curve. For this purpose, the minimum and maximum Y values of the rectangle R are calculated. After that, we choose several values $Y_i, i = \{0, n-1\}, Y_i \neq 0$ between them and, for each Y_i, calculate two corresponding X coordinates of the curve according to the following equality:

$$\begin{aligned} X_{i,j} = \frac{\alpha\gamma + \beta\delta}{\alpha^2 + \beta^2} \pm \sqrt{D_i},\ i = \{0, n-1\},\ j = \{1, 2\}, \\ D_i = l^2\frac{h_{2,0}^2 + h_{2,1}^2}{(\alpha^2 + \beta^2)}Y_i^4 - Y_i^2 - \frac{(\alpha\delta - \beta\gamma)^2}{(\alpha^2 + \beta^2)^2} + \frac{det(H)^2(h_{2,0}^2 + h_{2,1}^2)}{l^2Y_i^2(\alpha^2 + \beta^2)},\ Y_i \neq 0. \end{aligned} \tag{38}$$

We should also take into account that, for semi-minor and semi-major axis lengths, both branches of the curve may intersect the rectangle simultaneously. In order to construct the curve approximation correctly, we need to separate the points related to different branches. Then, moving along the Y-axis for each Y_i value, we compare the corresponding discriminant D_i with zero. If it is positive, then the obtained points lie on one branch. If the discriminant for a Y_i value is equal to zero, then there is an inflection point in the current branch and the following values Y_{i+k}, $k = \{1, n-i-1\}$ relate to another branch of the curve. Similarly, the negative discriminant D_i implies a gap between branches, and points calculated for further values Y_{i+k} lie on another branch.

As soon as the curve is obtained, we should decide whether the considered field quality is sufficient. There are several possible approaches. For example, we can calculate the ratio of sufficient and insufficient quality areas inside the restored rectangle. However, in this work, we mark the quality of the whole image as insufficient if there is a low-quality region of any area. The whole procedure for evaluating the restored image quality has $O(1)$ complexity because it is not dependent on the input image size but only on the number of points in the curve approximation, which we assume to be predefined. The procedure is summarized in Algorithm 1.

Algorithm 1 For quality assessment of a projectively distorted field quadrangle.

```
Input:
$F$—field quadrangle in source image;
$R$—rectangle of restored field;
$l$—minimum scaling coefficient threshold;
$n$—vertex number of curve approximation.
Output:
True $\equiv$ 1, if the restored field is predicted as recognizable;
False $\equiv$ 0, otherwise.
procedure Q($F, R, l, n$)
    calculate coefficients of a projective transform $H : H(R) = F$
    $center \leftarrow$ center point of $R$
    if $h_{2,0}^2 + h_{2,1}^2 = 0$ then                         ▷ affine transformation
        $s_c \leftarrow s(center)$ according to (31)
        return $s_c \geq l$
    $R' \leftarrow T(R)$ according to (36)            ▷ calculate new coordinates of rectangle
    $Y_{min} \leftarrow \min\{R'_{iY}\}_{i=\{1..4\}}$
    $Y_{max} \leftarrow \max\{R'_{iY}\}_{i=\{1..4\}}$
    calculate $\alpha, \beta, \gamma, \delta$ according to (30)
    $X_{sym} \leftarrow \dfrac{\alpha\gamma + \beta\delta}{\alpha^2 + \beta^2}$
    $no\_roots\_prev \leftarrow True$
    $one\_root\_prev \leftarrow False$
    $curve \leftarrow \{\}$
    for $i = \{0, ..n-1\}$ do
        $Y_i \leftarrow Y_{min} + i\dfrac{Y_{max} - Y_{min}}{n-1}$
        calculate $D_i$ according to (38)
        if $D_i > 0$ then
            $X_{i1,2} \leftarrow X_{sym} \pm \sqrt{D_i}$
            if $NOT(no\_roots\_prev)$ then
                $Insert(curve, Segment\{(X_{i1}, Y_i),\ (X_{i-1,1}, Y_{i-1})\})$
                $Insert(curve, Segment\{(X_{i2}, Y_i),\ (X_{i-1,2}, Y_{i-1})\})$
            if $no\_roots\_prev\ AND\ i \neq 0$ then
                $Insert(curve, Segment\{(X_{i1}, Y_i),\ (X_{i2}, Y_i)\})$
            $one\_root\_prev \leftarrow False$
            $no\_roots\_prev \leftarrow False$
        else if $D_i = 0$ then
            $X_{i1,2} \leftarrow X_{sym}$
            if $NOT(one\_root\_prev)\ AND\ NOT(no\_roots\_prev)$ then
                $Insert(curve, Segment\{(X_i, Y_i),\ (X_{i-1,1}, Y_{i-1})\})$
                $Insert(curve, Segment\{(X_i, Y_i),\ (X_{i-1,2}, Y_{i-1})\})$
```

Algorithm 1 *Cont.*

```
            $one\_root\_prev \leftarrow True$
            $no\_roots\_prev \leftarrow False$
        else
            if $NOT(one\_root\_prev)$ $AND$ $NOT(no\_roots\_prev)$ then
                $Insert(curve, Segment\{(X_{i-1,1}, Y_{i-1}),\ (X_{i-1,2}, Y_{i-1})\}$
            $one\_root\_prev \leftarrow False$
            $no\_roots\_prev \leftarrow True$
    for each $segment$ from $curve$ do
        if $segment$ intersects $R'$ then
            return $False$
    $s_c \leftarrow s(center)$ according to (31)
    return $s_c \geq l$
```

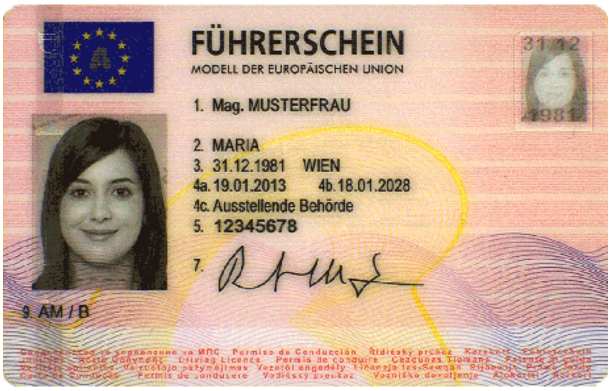
6. Experimental Results

In this section, the experimental results obtained using the proposed algorithmfor the quality assessment of projectively distorted field images are presented and compared with the performance of the algorithm described in [25]. In the recognition system workflow, in order to obtain a quadrangle of the field to be restored and recognized, we need to perform document localization and field segmentation. Evaluating quality assessment methods, we had to eliminate the errors that occurred in these stages. For this purpose, datasets that provide ground truth for field quadrangles are commonly used. To the best of our knowledge, the only publicly available dataset with at least mild projective distortions is MIDV-2019 [27]. However, a preliminary experiment showed that it does not contain images with enough projectivity to produce insufficient restored image quality. For this reason, we created a dataset with synthetically distorted images of text fields.

6.1. Data Generation

In order to generate the data, we used the MIDV-2019 dataset. This dataset contains 50 different types of annotated identity documents (ID cards, passports, driving licenses, etc.). It consists of 50 template images (original high-quality document images used for creating physical document copies, one per document type) and video clips of these documents acquired in different conditions. An example of a template document is shown in Figure 9.

All of the images were annotated manually. The video frames have a ground truth for their type and document quadrangle. The template images have a ground truth description consisting of field rectangles and their text content.



FÜHRERSCHEIN
MODELL DER EUROPÄISCHEN UNION

1. Mag. MUSTERFRAU
2. MARIA
3. 31.12.1981 WIEN
4a. 19.01.2013 4b. 18.01.2028
4c. Ausstellende Behörde
5. 12345678
7.
9. AM / B

Figure 9. The template of the new Austrian driving license document.

We considered template images only and scaled them to 300 dpi to obtain comparable pixel sizes for all documents. We used ground truth field rectangles to extract undistorted images of text fields with an additional 10% margin of their size. We only considered numeric fields and fields written with the Latin alphabet: dates, document numbers, machine-readable zone (MRZ) lines, document holder name, and surname. We recognized text in the obtained field images with Tesseract Open Source OCR Engine 4.1.1, which employs the LSTM neural network [41]. Incorrectly recognized fields were eliminated from further processing. In our experiments, we used 184 fields collected from all document templates. Since the text in the fields may have different fonts, font sizes, and other properties, we considered them separately in our experiments. Here, we describe synthetic data generation for one field.

We denote an original image of a field f as D_f and a rectangle bounding the field as R_f.

To test our algorithm, we generated a set of N projectively distorted field images $\{I^i_{src,f}\}_{i=1..N}$ with bounding quadrangles $\{F^i_f\}_{i=1..N}$ and corresponding projective transforms $\{H^i_f\}_{i=1..N}$: $F^i_f = H^i_f(R_f)$. In order to generate a distorted quadrangle F^i_f, we added random shifts to the corners of R_f. Then, the quadrangle F^i_f was downscaled to approximately the same size as the original field image for a more representative dataset. We also ensured that the obtained distorted quadrangle F^i_f and corresponding quadrangle of the whole distorted document were convex. Then, the homography H^i_f was calculated, and the original field image was transformed to obtain the distorted field image: $I^i_{src,f} = H^i_f(D_f)$. Algorithm 2 shows the procedure of distorted image generation.

Then, the restoration process was conducted. The distorted images $\{I^i_{src,f}\}_{i=1..N}$ were rectified with projective transforms that map their bounding quadrangles F^i_f to the rectangles R_f: $R_f = {H^i_f}^{-1}(F^i_f)$. Thus, we obtained a set of restored images $\{I^i_{rst,f}\}_{i=1..N}$: $I^i_{rst,f} = {H^i_f}^{-1}(I^i_{src,f})$. The projective mapping of images was conducted using the bilinear interpolation method.

Finally, we generated the ground truth for our problem of the binary quality assessment. We consider it to be a binary classification problem, with a positive case when «field image is recognizable» and a negative case otherwise. Thus, we used Tesseract to recognize the restored field images $I^i_{rst,f}$ and compared the results with the annotation from MIDV-500. If the recognition was correct, then the restored image was marked as recognizable.

6.2. Performance Metrics

To evaluate the performance of quality assessment algorithms, we calculated the positive and negative predictive values, PPV and NPV, respectively, as follows:

$$\begin{aligned} PPV &= \frac{TP}{TP+FP}, \\ NPV &= \frac{TN}{TN+FN}, \end{aligned} \tag{39}$$

where TP is the number of true-positive samples (restored field images were correctly recognized by Tesseract and marked as recognizable by the quality assessment algorithm under evaluation), TN is the number of true-negative samples (fields were not recognized by Tesseract and marked as non-recognizable by the algorithm), FP is the number of false-positive samples (fields were not correctly recognized by Tesseract but marked as recognizable by the algorithm), and FN is the number of false-negative samples (fields were correctly recognized by Tesseract but marked as non-recognizable by the algorithm).

We also had to ensure the balance of data used to evaluate the algorithm. The decision made by the proposed quality assessment algorithm Q depends on the minimum scaling coefficient threshold l. Hence, the probability of randomly generating a sample predicted to be positive or negative varies when l changes. To overcome this issue, for each l, we took 1000 restored field images marked as positive and 1000 restored field images marked as negative by the algorithm.

Algorithm 2 Generation of projectively distorted images of a field.

Input:

D—an undistorted field image;

$R(A, B, C, D)$—a bounding rectangle of an undistorted field, where A, B, C, and D are points of its corners from top left to bottom left clockwise;

$T(A_t, B_t, C_t, D_t)$—a bounding rectangle of a whole undistorted document, where A_t, B_t, C_t, and D_t are points of its corners from top left to bottom left clockwise;

N—the number of samples to generate.

Output:

$\{I^i_{src}\}_{i=1..N}$—a set of distorted field images;

$\{F^i(A'_i, B'_i, C'_i, D'_i)\}_{i=1..N}$—a set of bounding quadrangles of distorted field;

$\{H^i\}_{i=1..N}$—a set of corresponding projective transforms.

```
procedure G($D$, $R$, $T$, $N$)
    calculate width of $R$: $w \leftarrow B.x - D.x$
    calculate height of $R$: $h \leftarrow A.y - D.y$
    set a uniform real random number generator: $rand = uniform(0, 5\min(w, h))$;
    set generated number of samples $n \leftarrow 0$
    while $n < N$ do
        $A' \leftarrow A + (-rand(), rand())$
        $B' \leftarrow B + (rand(), rand())$
        $C' \leftarrow C + (rand(), -rand())$
        $D' \leftarrow D + (-rand(), -rand())$
        generated quadrangle $F' = (A', B', C', D')$
        $w' = \max(A'.x, B'.x, C'.x, D'.x) - \min(A'.x, B'.x, C'.x, D'.x)$
        $h' = \max(A'.y, B'.y, C'.y, D'.y) - \min(A'.y, B'.y, C'.y, D'.y)$
        calculate scale factor $s = \min\left(1.5\frac{w}{w'}, 1.5\frac{h}{h'}\right)$
        if $s < 1$ then
            $F' \leftarrow sF'$
        if $F'$ is not convex then
            continue
        calculate projective transform $H'$: $F' = H'(R)$
        if quadrangle $H'(T)$ is not convex then
            continue
        $F^i \leftarrow F'$
        $H^i \leftarrow H'$
        $I^i_{src} \leftarrow H^i(D)$
        $n \leftarrow n + 1$
    return $\{I^i_{src,f}\}_{i=1..N}$, $\{F^i_f\}_{i=1..N}$, $\{H^i_f\}_{i=1..N}$
```

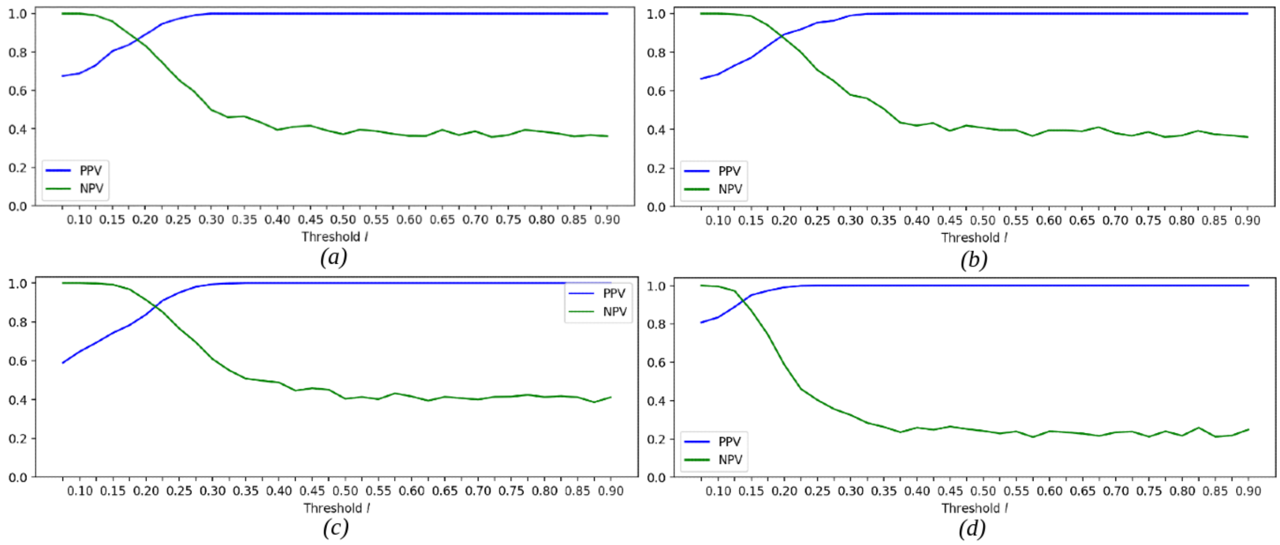
6.3. Behavior of the Proposed Method for Fields of Same and Different Fonts

In the framework of the first experiment, we estimated the variations in the PPV and NPV for the proposed algorithm Q, depending on the minimum scaling coefficient threshold l. We calculated the PPV and NPV functions separately for each field f. We varied the threshold l values from 0.075 to 0.9 with a step of 0.025. For each threshold l, we generated 1000 positively and 1000 negatively marked images and calculated the predictive values. The parameter n of the algorithm Q that defines the vertex number of the level curve approximation was set to 100. Figure 10 shows an example of the estimated PPV and NPV curves that were calculated for several text fields of the new Austrian driving license document, which is shown in Figure 9.

While developing the assessment method, we assumed that the threshold is equal for all characters of one font. Thus, the predictive value functions should be close for different fields of one font and may vary if the font or font properties (size, boldness, etc.) are changed. As we can see, the curves for the date fields with the same font (Figure 10a–c) show almost equal predictive values, as was expected. This means that we can estimate the valid threshold for all possible text fields of one font in advance. At the same time, PPV and NPV differ for a document number field that has a bold font (Figure 10d). Comparing them, we can infer that bold text can be more projectively distorted while still being reliably recognized. Thus, the minimum scaling coefficient threshold should be chosen separately for each font and font property.

For all fields, the specific behavior of the curves is similar. The greater the l, the sharper the restored image should be to be marked as «recognizable». Indeed, in Section 2, we define the minimum scaling coefficient threshold to be inverse to the level of distortion θ. As the threshold l increases, rejection occurs at a lower level of distortion. The threshold value can be chosen according to the cost of false-positive and false-negative errors. In the case of equal cost, the PPV and NPV are higher than 80% for all four considered fields.

It should be noted that the obtained predictive value curves are non-monotonic. This effect occurs because the OCR is not strictly monotonic with the projective distortion level. However, the tendency toward reduced recognition accuracy is evident.



(a) (b) (c) (d)

Figure 10. Positive and negative predictive values for varied minimum scaling coefficient thresholds estimated on the set of distorted images for fields: (**a**) "31.12.1981", (**b**) "19.01.2013", (**c**) "18.01.2028", and (**d**) "12345678".

6.4. Recognition System Simulation

In the second experiment, we estimated the recognition system's performance with the incorporated reject submodule. We compared the results obtained for the proposed algorithm with the rejection criterion presented in [25], which assesses the whole distorted document quadrangle. In addition, we estimated the same algorithm applied to each field quadrangle separately.

The geometric criterion presented in [25] is based on the analysis of the quadrangle angles. The document quadrangle is rejected if not satisfying the following conditions:

1. At least one pair of the opposed edges is parallel with a tolerance of 5°:

$$\left[\begin{array}{l}(\measuredangle[\overrightarrow{AB}] - \measuredangle[\overrightarrow{CD}]) < 5^\circ \\ (\measuredangle[\overrightarrow{AD}] - \measuredangle[\overrightarrow{BC}]) < 5^\circ\end{array}\right., \tag{40}$$

where A, B, C, and D are the corners of the document quadrangle and $\measuredangle[\overrightarrow{AB}], \measuredangle[\overrightarrow{CD}]$, $\measuredangle[\overrightarrow{AD}]$, and $\measuredangle[\overrightarrow{BC}]$ denote the edges' angles with the horizontal axis defined in the range $[-90^\circ, 90^\circ]$.

2. The average difference in angles between each pair of opposed angles is less than 10°:

$$\begin{cases} \dfrac{|\hat{A}-\hat{B}|+|\hat{C}-\hat{D}|}{2} < 10^\circ \\ \dfrac{|\hat{A}-\hat{D}|+|\hat{B}-\hat{C}|}{2} < 10^\circ \end{cases}, \tag{41}$$

   where $\hat{A}, \hat{B}, \hat{C}$, and $\hat{D}$ are the angles of the quadrangle defined in the range $[0^\circ, 180^\circ]$.
3. The average perpendicularity of the four corners is less than 25°:

$$\left|\frac{\hat{A}+\hat{B}+\hat{C}+\hat{D}}{4} - 90^\circ\right| < 25^\circ. \tag{42}$$

In order to estimate the system performance and to avoid errors that may occur in the document localization and segmentation stages, we synthesized distorted field images, as described in Section 6.1. Before the experiment evaluating the performance of the proposed algorithm, we automatically estimated the field thresholds for the proposed algorithm as follows. Each of the 184 original field images D_f was gradually uniformly downscaled from 0.9 to 0.1 of its size with a step of 0.025. The smallest scale that provided a correct recognition result was chosen as the threshold l_f. Then, for each field f and threshold l_f, we generated 1000 positively and 1000 negatively marked restored field images. The proposed algorithm parameter n defining the vertex number of the level curve approximation was set to 100. All positive images for all fields were contained in the overall positive set with a size of 184,000. The overall negative set was obtained similarly. The restored images of both sets were recognized using Tesseract, and the cumulative PPV and NPV values were calculated.

In the experiments conducted to evaluate the algorithm [25], we used two versions of the criterion. The first, original criterion assesses the document quadrangle and, thus, ceases further processing of all document fields simultaneously. Additionally, we evaluated the strategy of applying the criterion to each distorted field quadrangle. For both versions, we used the same processes of data generation and performance evaluation, except that the set of 1000 images predicted to be recognized was constructed based on the algorithm under evaluation. The same applies to the set predicted to be unrecognized.

The results of the conducted experiments are shown in Table 1. It can be seen that the thresholds of the algorithm in [25] were defined under the assumption of a much higher cost of false-positive error. However, the proposed algorithm outperforms both versions of the algorithm from [25] not only in NPV but also in PPV.

Table 1. The cumulative predictive values, positive (PPV) and negative (NPV), of the quality assessment algorithms.

	PPV	NPV	TP	FP	TN	FN
The proposed algorithm	86.7%	64.1%	159,622	24,378	117,998	66,002
The algorithm [25] for document quadrangle	79.2%	24.1%	145,672	38,328	44,332	139,668
The algorithm [25] for field quadrangle	77.2%	24.6%	142,098	41,902	45,331	138,669

Examples of false-positive and false-negative field images for the proposed method are shown in Figures 11 and 12, respectively. As we can see, for some, the recognition error is due to the OCR submodule, while the images themselves can be easily read. In the examples of false-negative images, the level of corruption differs. For example, field (b) is barely recognizable, while field (e) has adequate sharpness. The main reason is that we estimated the minimum possible sharpness in all directions. However, if the image is scaled orthogonally to the stroke, the blurring effect is small, which is seen in Figure 12e.

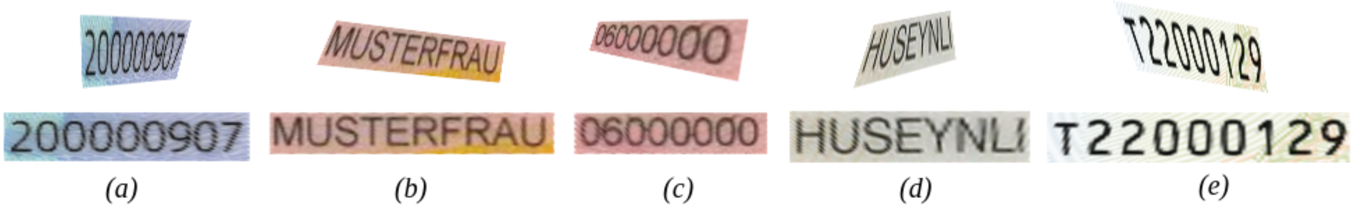


Figure 11. Examples (**a**–**e**) of false-positive distorted and restored field images for the second experiment.

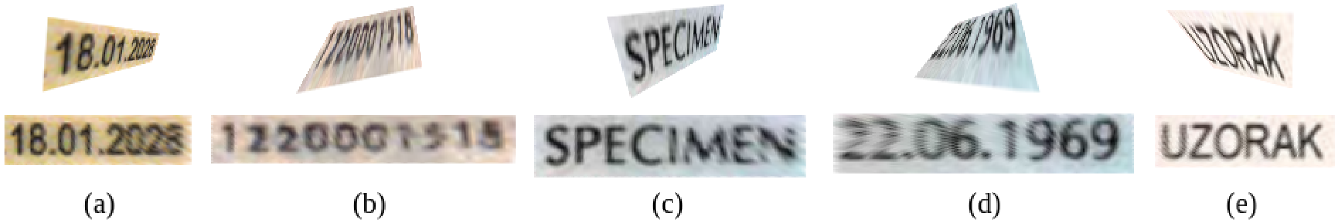


Figure 12. Examples (**a**–**e**) of false-negative distorted and restored field images for the second experiment.

Another possible reason for the proposed algorithm errors is the chosen approach for the estimation of the threshold. Due to the errors of the recognition submodule, some of the fields may overestimate the minimum scaling coefficient threshold. Moreover, in real applications, the text of a considered document field differs from the text in the template image. The current threshold estimation method is limited to only one possible text version. Thus, a more stable approach to threshold estimation needs to be developed to increase the performance of the algorithm. However, the presented results show that the proposed algorithm for text field quality assessment can already be successfully exploited for recognition reliability prediction.

7. Conclusions

In this paper, we consider the problem of quality assessment of a field image restored from a projectively distorted source document image. The quality is interpreted in terms of text recognition reliability. The results show that, by using a priori information about the field font, the restored field image quality can be estimated based only on the projective transform analysis. We present a theoretically based method for evaluating the distortion level at a point in the restored image. Moreover, we propose a novel algorithm of binary quality assessment that does not depend on the image size, i.e., it has $O(1)$ complexity. We also discuss the model of the reject submodule embedded in the document recognition system.

The algorithm was tested on synthetically distorted field images. The dataset was created based on document template images from the publicly available dataset MIDV-2019. According to the obtained results, the algorithm provides equal predictive value curves, both positive and negative, for different text strings of one font and one font size. For dissimilar fonts, these curves differ. Thus, the assumption is confirmed that the maximum level of distortion that enables reliable recognition depends on the font of the recognized text. Therefore, the threshold of the algorithm can be estimated in advance for each font, regardless of the text that may occur in the input distorted field images.

In the experiment evaluating the performance of the reject submodule, we compared the proposed algorithm with the rejection criterion presented in [25]. This algorithm is designed to assess the whole document quadrangle and, therefore, to reject or accept all document fields simultaneously. Additionally, we applied the same criterion separately for each distorted field image. The thresholds for the proposed algorithm were estimated in advance for each field by iterative downscaling of the undistorted field image and by recognizing the obtained image. The results show the superiority of our algorithm. The cumulative positive predictive value (PPV) for the proposed algorithm equals 86.7%, which is 7.5% higher than the best PPV value of other compared algorithms. The cumulative negative predictive value (NPV) estimated for our algorithm is 64.1%, and the difference from the best value of the other algorithm is 39.5%.

For future work, a more stable method for estimating the threshold should be developed. It should utilize all alphabet characters of an estimated font and projective distortions in addition to the scaling transform. Additionally, the current approach may be improved by relying on the ratio of sufficient and insufficient region areas defined by the constructed level curve.

It should be noted that the proposed method may be exploited not only for the reject submodule. The other possible application field is combination methods for text field recognition in video streams. The binary quality estimation can be used to reevaluate the confidence of the recognition result for one frame. Moreover, as the method also provides the level curve that bounds the low-quality region, we can utilize it to reevaluate the confidence of each recognized character according to its location. This may increase the video stream recognition accuracy.

For future work, a stable method of threshold estimation should be developed. It has to analyze the recognition correctness after restoration from different levels of projective distortions instead of only scaling transformations. The whole alphabet of the font considered is to be included to provide a stable threshold for all possible text strings in the field. Additionally, an experimental comparison should be conducted for the approaches to image quality estimation according to the constructed level curve. The current approach may be improved by relying on the ratio between sufficient and insufficient region areas.

Author Contributions: Conceptualization, research and experiments, writing, and editing: J.S.; methodology, review, and editing: E.L.; software, experiments, and visualization: N.S.; conceptualization and supervision: D.P.N.; methodology and review: V.V.A. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partially supported by the Russian Foundation for Basic Research (19-29-09066, 18-29-26035).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Restrictions apply to the availability of these data. Data was obtained from Smart Engines Service LLC and are available from the authors with the permission of Smart Engines Service LLC.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Attivissimo, F.; Giaquinto, N.; Scarpetta, M.; Spadavecchia, M. An Automatic Reader of Identity Documents. In Proceedings of the IEEE International Conference on Systems, Man and Cybernetics (SMC), Bari, Italy, 6–9 October 2019; pp. 3525–3530.
2. Bulatov, K.; Arlazarov, V.V.; Chernov, T.; Slavin, O.; Nikolaev, D. Smart IDReader: Document recognition in video stream. In Proceedings of the 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), Kyoto, Japan, 9–15 November 2017; Volume 6, pp. 39–44.
3. Li, H.; Zhu, F.; Qiu, J. Towards Document Image Quality Assessment: A Text Line Based Framework and a Synthetic Text Line Image Dataset. In Proceedings of the International Conference on Document Analysis and Recognition (ICDAR), Sydney, Australia, 20–25 September 2019; pp. 551–558.
4. Liu, P.; Yang, R.; Xu, Z. How Safe Is Safe Enough for Self-Driving Vehicles? *Risk Anal.* **2019**, *39*, 315–325. [CrossRef] [PubMed]
5. Rao, Q.; Frtunikj, J. Deep learning for self-driving cars: Chances and challenges. In Proceedings of the 1st International Workshop on Software Engineering for AI in Autonomous Systems, Gothenburg, Sweden, 28 May 2018; pp. 35–38.
6. Wu, Z.; Lin, W.; Ji, Y. An Integrated Ensemble Learning Model for Imbalanced Fault Diagnostics and Prognostics. *IEEE Access* **2018**, *6*, 8394–8402. [CrossRef]
7. Rajagopal, H.; Khairuddin, A.; Mokhtar, N.; Ahmad, A.; Yusof, R. Application of image quality assessment module to motion-blurred wood images for wood species identification system. *Wood Sci. Technol.* **2019**, *53*, 967–981. [CrossRef]

8. Kimura, S.; Tanaka, E.; Sekino, M.; Sakurai, T.; Kubota, S.; So, I.; Koshi, Y. A Man-Machine Cooperating System Based on the Generalized Reject Model. In Proceedings of the 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), Kyoto, Japan, 9–12 November 2017; pp. 1324–1329.
9. Moosavi-Dezfooli, S.; Fawzi, A.; Frossard, P. DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2574–2582.
10. Su, J.; Vargas, D.; Sakurai, K. One Pixel Attack for Fooling Deep Neural Networks. *IEEE Trans. Evol. Comput.* **2019**, *23*, 828–841. [CrossRef]
11. Pereyra, G.; Tucker, G.; Chorowski, J.; Kaiser, L.; Hinton, G. Regularizing Neural Networks by Penalizing Confident Output Distributions. *arXiv* **2017**, arXiv:abs/1701.06548.
12. Dodge, S.; Karam, L. A Study and Comparison of Human and Deep Learning Recognition Performance Under Visual Distortions. In Proceedings of the 26th International Conference on Computer Communication and Networks (ICCCN), Vancouver, BC, Canada, 31 July–3 August 2017.
13. Alsmirat, M.A.; Al-Alem, F.; Al-Ayyoub, M.; Jararweh, Y.; Gupta, B. Impact of digital fingerprint image quality on the fingerprint recognition accuracy. *Multimed. Tools Appl.* **2019**, *78*, 3649–3688. [CrossRef]
14. Ferrara, M.; Franco, A.; Maltoni, D. on the Effects of Image Alterations on Face Recognition Accuracy. In *Face Recognition Across the Imaging Spectrum*; Bourlai, T., Ed.; Springer: Cham, Switzerland, 2016; pp. 195–222.
15. Arlazarov, V.; Zhukovsky, A.; Krivtsov, V.; Nikolaev, D.; Polevoy, D. Analysis of specifics of using stationary and mobile small-sized digital cameras for document recognition. *J. Inf. Technol. Comput. Syst.* **2014**, *3*, 71–81.
16. Athar, S.; Wang, Z. A Comprehensive Performance Evaluation of Image Quality Assessment Algorithms. *IEEE Access* **2019**, *7*, 140030–140070. [CrossRef]
17. Zhai, G.; Min, X. Perceptual image quality assessment: A survey. *Sci. China Inf. Sci.* **2020**, *63*, 11. [CrossRef]
18. Fu, G.; Zhang, Q.; Lin, Q.; Zhu, L.; Xiao, C. Learning to Detect Specular Highlights from Real-world Images. In Proceedings of the 28th ACM Multimedia Conference, Seattle, WA, USA, 12–16 October 2020; pp. 1873–1881.
19. Tian, Q.; Clark, J.J. Real-time Specularity Detection Using Unnormalized Wiener Entropy. In Proceedings of the International Conference on Computer and Robot Vision, Regina, SK, Canada, 28–31 May 2013.
20. Chernov, T.S.; Kolmakov, S.I.; Nikolaev, D.P. An algorithm for detection and phase estimation of protective elements periodic lattice on document image. *Pattern Recognit. Image Anal.* **2017**, *27*, 53–65. [CrossRef]
21. Chernov, T.S.; Razumnuy, N.P.; Kozharinov, A.S.; Nikolaev, D.P.; Arlazarov, V.V. Image quality assessment for video stream recognition systems. In Proceedings of the 10th International Conference on Machine Vision (ICMV), Vienna, Austria, 13–15 November 2017; Volume 10696.
22. Chernov, T.S.; Ilyuhin, S.A.; Arlazarov, V.V. Application of dynamic saliency maps to video stream recognition systems with image quality assessment. In Proceedings of the 11th International Conference on Machine Vision (ICMV), Munich, Germany, 1–3 November 2018; Volume 11041.
23. Bulatov, K.B. A Method to Reduce Errors of String Recognition Based on Combination of Several Recognition Results with Per-Character Alternatives. *Bull. South Ural. State Univ. Ser. Math. Model. Program. Comput. Softw.* **2019**, *12*, 74–88. [CrossRef]
24. Petrova, O.; Bulatov, K.; Arlazarov, V.L. Methods of weighted combination for text field recognition in a video stream. In Proceedings of the 12th International Conference on Machine Vision (ICMV), Amsterdam, The Netherlands, 16–18 November 2019; Volume 11433.
25. Awal, A.M.; Ghanmi, N.; Sicre, R.; Furon, T. Complex document classification and localization application on identity document images. In Proceedings of the 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), Kyoto, Japan, 9–15 November 2017; pp. 427–432.
26. Sánchez-Rivero, R.; Silva-Mata, F.; Morales-Quevedo, A. Capture of identity document images in the wild: Detection and quality assessment. In Proceedings of the 18th International Convention and Fair Informática 2020 (CICCI 2020), Nairobi, Kenya, 18–20 November 2020.
27. Bulatov, K.; Matalov, D.; Arlazarov, V.V. MIDV-2019: Challenges of the modern mobile-based document OCR. In Proceedings of the 12th International Conference on Machine Vision (ICMV), Amsterdam, The Netherlands, 16–18 November 2019; Volume 11433.
28. Hartley, R.; Zisserman, A. *Multiple View Geometry in Computer Vision*, 2nd ed.; Cambridge University Press: New York, NY, USA, 2004.
29. Skoryukina, N.; Arlazarov, V.V.; Nikolaev, D.P. Fast method of ID documents location and type identification for mobile and server application. In Proceedings of the 15th International Conference on Document Analysis and Recognition (ICDAR), Sydney, Australia, 20–25 September 2019; pp. 850–857.
30. Li, J.; Mei, Z.; Zhang, T. A method for document image enhancement to improve template-based classification. In Proceedings of the 2020 4th High Performance Computing and Cluster Technologies Conference & 2020 3rd International Conference on Big Data and Artificial Intelligence, Qingdao, China, 3–6 July 2020.
31. Povolotskiy, M.; Tropin, D. Dynamic Programming Approach to Template-based OCR. In Proceedings of the Eleventh International Conference on Machine Vision (ICMV 2018), Munich, Germany, 1–3 November 2018. [CrossRef]
32. Slavin, O. Using Special Text Points in the Recognition of Documents. *Cyber-Phys. Syst. Adv. Des. Model.* **2020**, *259*, 43–53.

33. Bagdanov, A.D.; Worring, M. Fine-grained document genre classification using first order random graphs. In Proceedings of the 6th International Conference on Document Analysis and Recognition (ICDAR), Seattle, WA, USA, 10–13 September 2001; pp. 79–83.
34. Ryan, M.; Hanafiah, N. An examination of character recognition on ID card using template matching approach. *Procedia Comput. Sci.* **2015**, *59*, 520–529. [CrossRef]
35. Wolberg, G. *Digital Image Warping*, 1st ed.; IEEE Computer Society Press: Los Alamitos, CA, USA, 1990.
36. Petrova, O.; Bulatov, K.; Arlazarov, V.V.; Arlazarov, V.L. Weighted combination of per-frame recognition results for text recognition in a video stream. *Comput. Opt.* **2021**, *45*, 77–89. [CrossRef]
37. Hanjing, A.; Suantai, S. A Fast Image Restoration Algorithm Based on a Fixed Point and Optimization Method. *Mathematics* **2020**, *8*, 378. [CrossRef]
38. Trusov, A.; Limonova, E. The analysis of projective transformation algorithms for image recognition on mobile devices. In Proceedings of the 12th International Conference on Machine Vision (ICMV), Amsterdam, The Netherlands, 16–18 November 2019; Volume 11433.
39. Marks, R. *Introduction to Shannon Sampling and Interpolation Theory*; Springer: New York, NY, USA, 1991.
40. Shemiakina, J.; Zhukovsky, A.; Konovalenko, I.; Nikolaev, D.P. Automatic cropping of images under projective transformation. In Proceedings of the 11th International Conference on Machine Vision (ICMV), Munich, Germany, 1–3 November 2018; Volume 11041.
41. Tesseract OCR. Available online: https://github.com/tesseract-ocr/ (accessed on 2 September 2021).