[12] On the quantitative performance evaluation of image analysis algorithms

P.P. Koltsov [1], A.S. Osipov [1], A.S. Kutsaev [1], A.A. Kravchenko [1], N.V. Kotovich[1], A.V. Zakharov [1]
[1] Scientific-Research Institute for System Analysis of the Russian Academy of Sciences

Abstract

The paper contains a brief review of main approaches to the comparative performance evaluation of image analysis algorithms. Some empirical methods used for the comparative evaluation of edge detectors and image segmentation algorithms are considered and quantitative criteria employed in these methods are studied. Problems associated with the use of these criteria are described. Finally, using the edge detector evaluation as an example, we propose an empirical method, called EDEM, which is implemented using our proprietary software system PICASSO.

Keywords: *comparative study, image analysis, edge detectors, image segmentation, performance measures, ground truth image, fuzzy sets.*

Citation: *Koltsov PP, Osipov AS, Kutsaev AS, Kravchenko AA, Kotovich NV, Zakharov AV. On the quantitative performance evaluation of image analysis algorithms. Computer Optics, 2015; 39(4): 542-56. DOI: 10.18287/0134-2452-2015-39-4-542-556.*

Introduction

As known, the scanning and computerized processing of images had started in 1956 at the U.S. National Bureau of Standards. At this time, the design of the image enhancement algorithms had begun [1]. Sixty years later, thousands of various image processing algorithms have been developed. Some of them have been specific to certain applications (such as the enhancement of latent fingerprints), whereas the others have been more generic in nature, often without information on their best application area. The scope of these algorithms is rather large: from automatically extracting and depicting regions of interest such as in the case of segmentation, to improving the perceived quality of an image by image enhancement methods. Since the early years of computer vision (as in the other subfields of software design), a part of the design process has been dedicated to algorithm testing. Such testing serves a double purpose. Firstly, it gives either a qualitative or quantitative method of evaluating an algorithm. Secondly, it provides a comparative measure of the algorithm against similar algorithms in terms of the same criteria. The design and choice of proper criteria of evaluation is a difficult task. Do we use a criterion which measures accuracy, robustness, or sensitivity? Performance evaluation, in a broad sense, is a measure of some required behavior of an algorithm, whether it is achievable accuracy, robustness and adaptability. It allows one to emphasize the most essential properties of an algorithm, to evaluate its advantages and limitations. The analysis of algorithm's failures is closely related to such evaluation. This analysis primarily requires a definition of characteristics of success. Such failure analysis, during the design stage of an algorithm is of high importance. In this case, it can be considered as initial testing.

This thorough testing has not yet become a common practice. Part of this is attributable to the lack of formal process used in performance evaluation, from the establishment of testing regimes to the design of performance metrics. Also, in the last half-century, different approaches to performance evaluation of image processing algorithms have been poorly covered in the literature. At the same time, it should be noted that the choice of an appropriate evaluation methodology depends on the objective of the task.

As noted in [2], the purpose of evaluating an algorithm is to understand its behavior on different categories of images and help in choosing the best parameters for different applications. In its final stage this involves some comparison with similar algorithms in order to provide practical guidelines for choosing algorithms on the basis of application do-

main. Accessing the performance of any algorithm depends on several factors [3]:
- the algorithm itself,
- the nature of images used to measure the performance of the algorithm,
- the algorithm parameters used in the evaluation,
- the method used for evaluating the algorithm.

The difficulty in evaluating of an algorithm is directly proportional to the number of parameters it requires. For the purpose of performance optimization, a certain selection of parameters is required, which is not an easy task by itself. Besides that, the optimal parameters may vary on different test images. As to the influence of test images on performance evaluation accuracy, evaluation with a set of "easy" images may often produce a higher accuracy than the use of images containing difficult objects or situations.

Nowadays there are no rigid guidelines characterizing the process of performance evaluation, however there are a number of factors to be considered: testing protocol, testing regime, performance indicators, performance metrics and image [4]. The testing protocol relates to the successive approach used to perform testing. Next is the testing regime which relates to the strategy used for testing the images. There are four main testing categories. The first of these is exhaustive testing, which is a crude approach to testing based upon the use of every image in a database. Such an approach can be excessive and should be limited to the verification stage of the design process. Next is boundary value testing which evaluates an algorithm on a pre-defined representative subset of the database images. The third regime is random testing in which images are indiscriminately selected. Compared with the previous case, under this testing regime more diverse situations may occur, because boundary-value testing deals with subjective selection of images which might not take into account the diversity of practical situations. For example, it is realistic to test a mass-detection algorithm on a database of mammograms containing mainly images with malignant masses, whereas clean mammograms or mammograms with benign masses are predominant in practice. The fourth test regime concerns worst-case testing. It mainly focused on the situations when the test image contains rare or unusual features.

Performance indicators specify the qualities of an algorithm. They are often loose characterizations and in themselves are difficult to measure. Typical performance indicators are [4]:
- Accuracy: how well the algorithm has performed with respect to some reference.
- Robustness: an algorithm's capacity for tolerating various conditions.
- Sensitivity: how responsive an algorithm is to small changes in input data.
- Adaptivity: how the algorithm deals with variability of images.
- Reliability: the degree to which the algorithm, when repeated using the similar data, yields the similar results.
- Efficiency: the practical viability of an algorithm (convenience, cost, modifiability, etc.).

Finally there is the notion of image database: which images should be selected to test an algorithm. This relates to the diversity and complexity of the selected images and the significance of the images to the algorithm's purpose (e. g. segmentation or edge detection).

1. Current evaluation methods

As mentioned above, by now, thousands of image processing algorithms have been proposed. Many of them have multiple software implementations (some of them are available in the public domain, like e. g. the famous Canny edge detector). As a result, the developer of computer vision system faces a difficult task of choosing the most appropriate algorithms for his practical purposes.

Due to the above reasons, testing of image processing algorithms for practical purposes has no unique method. The main differences between the methods used for comparative evaluation of algorithms of the same class (e. g. edge detectors, segmentation algorithms, texture finders, etc.) are the following:
- different sets of test images which differ both in the type of images (real or synthesized) and in the size, amount and their sources (user-created or from an available database);
- different procedures of choosing the optimal parameters of the algorithms;
- different evaluation criteria (quantitative or qualitative, using reference images or not).

To date, several attempts to classify these methods have been made. In [5] the following classification of evaluation methods for image segmentation algorithms has been offered:
1. subjective evaluation
2. objective evaluation
2.1. system level evaluation
2.2. direct evaluation
2.2.1. analytical methods
2.2.2. empirical methods
2.2.2.1. supervised methods

2.2.2.2. unsupervised methods.
In principle, such classification is appropriate to classify evaluation methods for another class of algorithms (e. g. edge detectors).
The most widely used type of evaluation method is subjective (or visual) evaluation. The disadvantage of such methods (reflected in their name) is that visual or qualitative evaluation is inherently subjective. Subjective evaluation scores may vary significantly from one human evaluator to another.
Objective evaluation methods do not use visual assessment of images. They are divided into system level evaluation methods and direct evaluation methods.
System level evaluation methods assess an algorithm on the basis of overall performance of the computer vision system which contains this algorithm. We mention, as an example, the work of Shin, Goldgof and Bowyer [6] which presents a task-oriented evaluation methodology for edge detectors. Such assessment does not necessarily indicate the flaws in the algorithm itself; it may indicate the algorithm, which output is most suitable for further processing.
The direct objective evaluation can be divided into analytical methods and empirical methods, based on whether the algorithm itself, or the results generated by the algorithm are being examined.
Analytical methods assess algorithms independently of their output [7]. The evaluation is based on such properties of the algorithms as processing strategy (parallel, sequential, iterative or mixed), processing complexity, resource efficiency, etc. These properties are generally independent of the quality of the algorithm's results Analytical methods considered in literature usually deal with some special tasks (see e. g. [8]).
On the contrary, empirical methods assess the results of the algorithm on a set of test images. They are divided into supervised and unsupervised methods.
Supervised methods are also known in literature as empirical discrepancy methods (see e. g. [9]). The latter definition is probably more appropriate, since such methods perform a comparison between a processed image (algorithm's output) against a reference image which is often referred to as a *ground-truth*, by using discrepancy measures. The ground truth images are often manually created and contain the features which are ideal from the evaluator's viewpoint. For instance, if we evaluate the edge detectors, then for every test image there is a matching ground truth image containing ideal (user-defined) edges. A situation is possible when we study several features of each algorithm. In this case, several ground truths may correspond to a single test image (see also [23]). In most cases, such methods can provide a fair evaluation. However, for many test images the creation of matching ground truths is labor-intensive (e. g. in case of segmentation of real-world images) and is subjective.
Unsupervised (or goodness [9]) methods evaluate a processed image based on how well it matches some set of characteristics as desired by humans. Perhaps the most distinct advantage of unsupervised evaluation is that it requires no reference image. This feature enables to perform control and self-learning in real time systems.
One of the key elements of comparative evaluation methods, are their criteria of evaluation (also called as performance criteria, performance metrics, performance measures, performance indices [10]).
Our paper mainly deals with the empirical supervised evaluation methods. In particular, in next section we consider the quantitative performance measures, which are mainly used for performance evaluation of edge detection algorithms, discuss their application features and their disadvantages. Then we analyze the main criteria for evaluating the image segmentation algorithms. In section 4 we consider our method EDEM for comparative evaluation of image processing algorithms implemented within the program system PICASSO. We demonstrate the principles of our approach considering evaluation of edge detectors.
Before considering the available quantitative performance measures, used for evaluation of a certain class of image processing algorithms, it is desirable to formulate the requirements for their output images. For example, for the case of image segmentation algorithms, such requirements (in the form of qualitative criteria) have been formulated in [11] (see section 3). In most cases the evaluator tries to find the measures most suitable for the achievement of these objectives.

2. Quantitative performance criteria for edge detectors

It is agreed that the main requirements for the edge detection algorithms were first formulated by J. Canny in his classical work [12]. The author first managed to define a comprehensive set of goals for the computation of edge points, to formulate them in the form of a certain optimization problem, and,

finally, to solve this problem. According to Canny, these requirements (or performance criteria) are as follows:
1. Good signal-to-noise ratio. There should be a high probability of failing to mark real edge points, and low probability of falsely marking non-edge points.
2. Good localization. The points marked as edge points by the detector should be as close as possible to the center of the true edge.
3. Only one response to a single edge. This is implicitly contained in the first criterion (if there are two responses to the same edge, one of them must be considered false). However, the mathematical form of the criterion does not capture the multiple response requirement and it has to be made explicit.
J. Canny considered the mathematical problem of deriving an optimal smoothing filter (performing a preprocessing stage in edge detection) given the above criteria. He showed that this filter is a sum of four exponential terms, and that it can be well approximated by first-order derivatives of Gaussians. Although the work of Canny was done in the early days of computer vision, his edge detector is still among the best and mostly used ones.
The Canny's requirements for a "good" edge detector may be considered as an example of requirements mentioned at the end of the previous section. Hence, to find out how these requirements are met for the evaluation of edge detectors, some matching quantitative criteria are required. In terms of the empirical supervised evaluation methods these requirements mean that the main quantitative features of a good edge detector are a high percentage of correctly detected edge pixels (good detection level) and a high level of localization accuracy (points identified as edge pixels should be as close as possible to the centre of the matching edge on the ground truth image). Note, that as mentioned by Canny, there is a sort of uncertainty principle between good detection and good localization. Perhaps this explains the fact that, until now, no measure which evaluates accurately both of the above features, has been developed. Accordingly, in a number of papers on the subject (see [13] and references thereafter), the considered quantitative criteria were divided into two classes: detection performance, or "statistical" measures and localization performance, or "distance" measures.
We mention here several widely used detection performance measures (see also [13]). Namely, let X be the image raster (containing N pixels), B – the estimated image (the output of evaluated edge detector) and A – the corresponding true image (the ground truth edge map).Then define the type I error rate as:

$$\alpha(A,B)=\frac{n(B\setminus A)}{n(X\setminus A)},$$

where $n(S)$ – number of pixels in S; i. e. as the ratio of the number of incorrectly detected edge pixels to the number of non-edge pixels.
The type II error rate is defined as:

$$\beta(A,B)=\frac{n(A\setminus B)}{n(A)},$$

i. e. as the ratio of the number of non-detected edge pixels to the number of edge pixels.
Also, quite common are such measures as Sensitivity:

$$Se=\frac{n(B\cap A)}{n(A)}=1-\beta,$$

(the ratio of the number of correctly detected edge pixels to the number of edge pixels) and Specificity:

$$Sp=\frac{n(X/B\cap A)}{n(X/A)}=1-\alpha,$$

Initially, these measures were used in medical statistics as measures of risk.
The mean squared Eucledian distance (mainly used for comparison of grayscale images) is another example of detection performance measures as well as the signal to noise ratios (both pick and mean square).
The above measures have a wide practical application, and at the same time their disadvantages are widely known. Perhaps the most significant disadvantage is that discrepancies between A and B are measured by the number of disagreements regardless to the pattern. For instance, errors which affect a small number of pixels but severely affect 'shape', such as the deletion of a linear feature, filling-in of small holes, etc. have high values of these measures. As an example, consider a test image shown in Figure 1*a* and a corresponding ground truth image in Figure 1*b*. We apply two edge detection algorithms to the test image; the results are presented in Figures 1*c* and 1*d*, respectively.

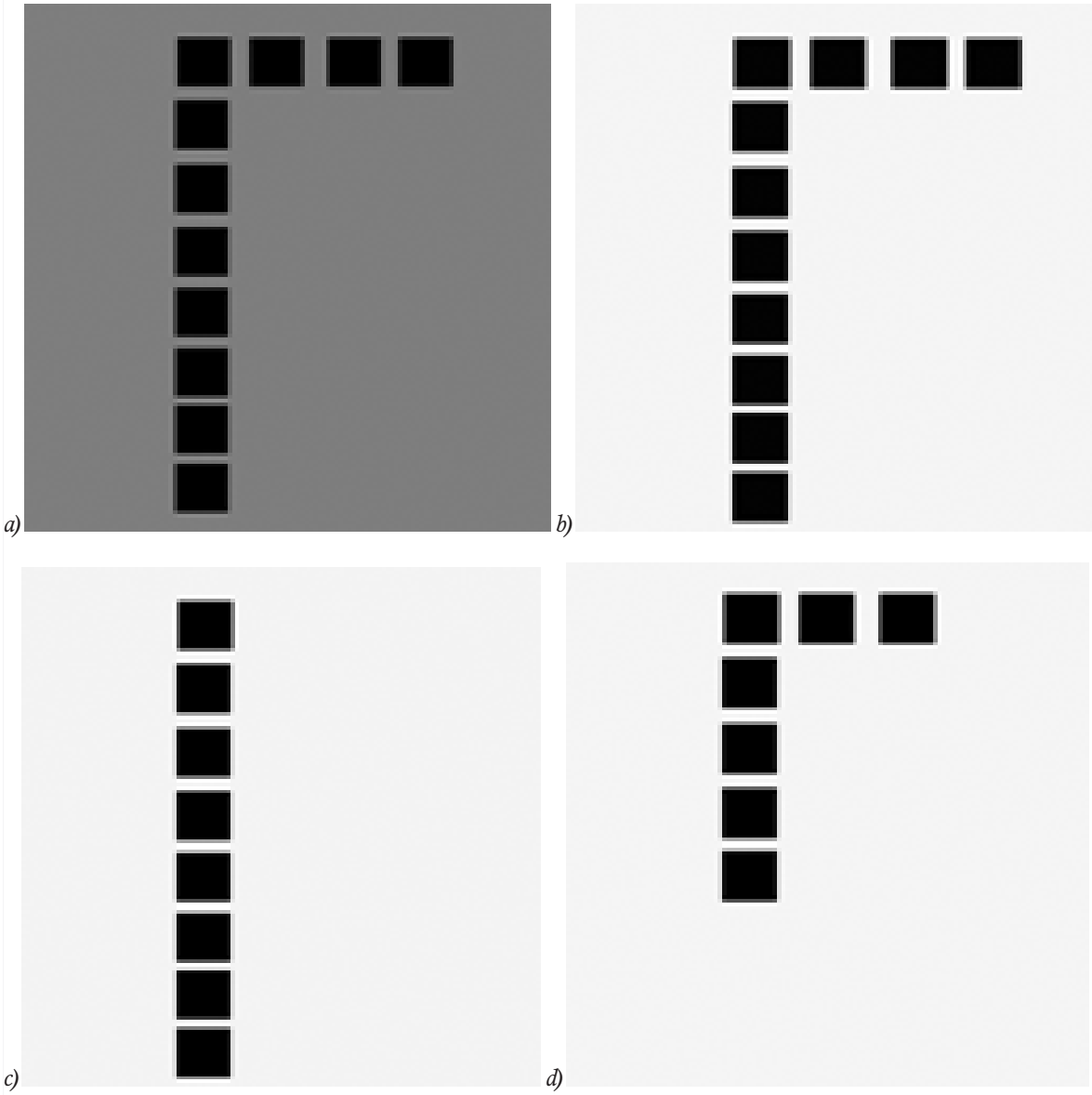


Fig. 1. a) test image, b) corresponding ground truth, c) and d) results of processing a) with two edge detectors

Here, for both algorithms, the values of type I error and Specificity are the same and are equal to 0 and 1, respectively (all of the marked pixels are real edge points). The values of type II error and Sensitivity are equal 0.27 and 0.73 for the first algorithm; for the second one they are equal to 0.36 and 0.64. Thus, if we are guided only by the values of these measures in our evaluation, we must conclude that the performance of the first edge detector on the considered image is better, which contradicts our visual observation (and common sense). Another practical problem for the application of detection performance measures is the problem of threshold selection in order to find the matching pixels on two images. Thus, a small shift in the edge map points of the processed image with respect to the reference edge map, which affects a large number of pixels but does not affect the shape of the pattern (e. g. we see the same apple on the estimated and the ground truth image) can lead to low values of detection performance measures. The above mentioned disadvantages of these measures should be taken into account for practical applications. In particular, this concerns the case, when the preceding edge detection step involves smoothing of the noise.

Among the measures of localization performance, we mention the mean error distance:

$$e(A,B)=\frac{1}{n(B)}\sum_{x\in B}d(x,A)^2,$$

where $d(X,A)=inf\ \rho(x,a)\quad a\in A$, and $\rho(.,.)$ is a shortest path length metric, see [14] and references thereafter; and the popular Pratt's figure of merit:

$$FOM(A,B)=\frac{1}{\max\{n(A),n(B)\}}\sum_{x\in B}\frac{1}{1+kd(x,A)^2},$$

where k is a scaling constant usually set to 1/9, and $\rho(.,.)$ is normalized so that the smallest nonzero distance between pixel neighbors equals 1. One has $0<FOM(A,B)\le 1$ and equals to 1 if and only if $A=B$. The Hausdorff metric can also be attached to this class. It is defined as:

$$H(A,B)=\max\{\sup_{a\in A}d(a,B),\ \sup_{b\in B}d(b,A)\}.$$

Although the classical version of this measure has some desirable topological properties, which are desirable for evaluation of image processing operations, it is rarely used in practice since it is very sensitive to 'noise' and even to changes in a single pixel. The most widely used of the above-mentioned measures is the Pratt's *FOM*.

Like the statistical measures, the localization performance measures show some undesirable features in practice. Namely, they are insensitive to type II errors. For example, if all errors are of type II, $B\in A$, then $e=0$, while $FOM(A,B)=n(B)/n(A)=1-\beta(A,B)$ i. e. the value of $FOM(A,B)$ coincides with the value of specificity and provides no new information. The error distance e and, especially, the Hausdorff metric are highly sensitive to background noise. As to the Pratt's measure, in many situations the *FOM* – optimal images had sections of the true contour missing, or oscillated around the 'true' contour, see [13].

Also, a striking example of behavior of *FOM* has been found by Peli and Malah [14]; it is shown on Figure 2.

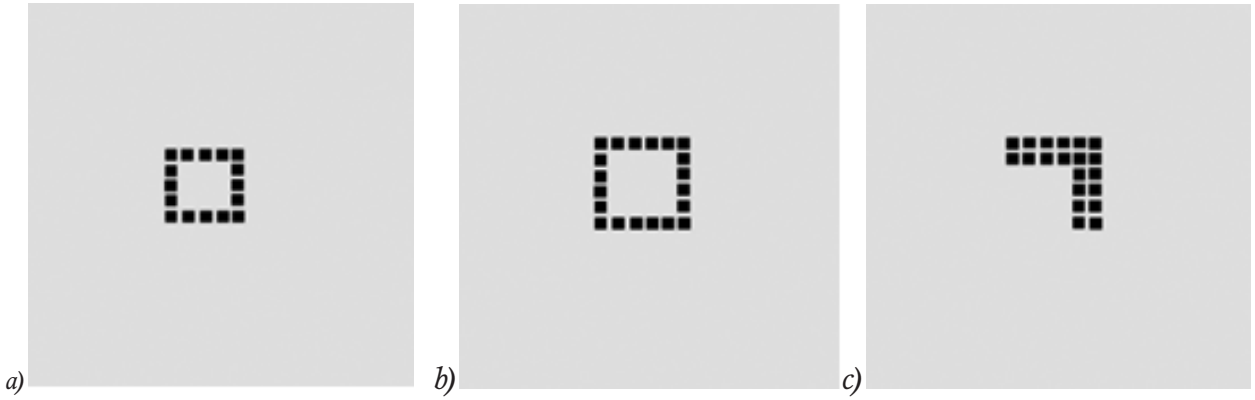


Fig. 2.Peli-Malah example, True picture a) and two estimated pictures b), c) with the same level of FOM =0.941.

If the image shown on Figure 2 a) (boundary of a 5x5 pixel square) is taken as the ground truth image A, then the images B_1 and B_2 shown on Figures 2 b) and 2 c) have the same *FOM* values: $FOM(A,B_1)=FOM(A,B_2)$. Such examples indicate, that the definition of *FOM*, within the selected distance threshold, allows a multiple-to-one correspondence between the detected edge pixels and the ground truth edge pixels. Note that in the Peli-Malah example the values of above detection performance measures are also the same. In the paper [14] a localization performance measure Δ_w^p, which is an L^p modification of the Hausdorf metric, has been introduced:

$$\Delta_w^p=\frac{1}{n(X)}\Big(\sum_{x\in X}|\,w(d(x,A))-w(d(x,B))\,|^p\Big)^{1/p},$$

$1\le p\le\infty,$

where w is a so called cutoff transform: $w(t)=\min\{t,c\}$ for some fixed $c>0$. It is more robust to small pixel changes than the Hausdorf metric. Compared with the latter, it produced more reasonable results in a number of tests. For example, in the Peli-Malah example its values (for c=5) for the images of Figure 2 b)-c) were equal to 0.323 and 0.512 respectively. At the same time, the *FOM* is more robust to small oscillations of edge contour.

In recent years, some new measures of both above classes were offered for edge detector evaluation. They can take into account edge strength, pixel matching on the images as well as the displacement of edge pixel positions in the estimation of similarity [13]. These measures are also not free of disadvantages. Note that unlike the above-considered measures, some of these measures are algorithmically complicated and their calculation is time-consum-

ing. For example, the similarity between two images can be estimated from the cost of the optimal matching between their pixels, where an optimal matching is a matching with a minimal cost among all possible matchings. In order to find it, the tools obtained from graph theory are used. Thus, the development of simple and reliable performance metrics for evaluating edge detection quality remains an actual task.

As to the images used within the empirical supervised evaluation methods, in many papers, simple artificial ground truth images, or real images containing sharp edges are employed. However, in practice, images with hard to detect contours are common, which limits applicability of these methods. Also, the completeness of test sets, used for evaluation, is an open issue for the ground truth based methods.

Note that in some papers (e. g. describing ROC curve evaluation framework, see [13] and references thereafter) the three-valued ground truths were considered. In such images each pixel is marked as either edge, or non-edge or don't-count (the latter are the pixels where the edge status appears ambiguous). It simplifies the generation of ground truths corresponding to real images (e.g. on such ground truths we can attach all pixels forming the texture areas to the third class) and makes the quantitative estimates more informative. Also note such feature of grayscale pictures as uncertainty that exists in locating the exact position of the boundary that separates the object from background. Especially this is typical for blurred images. This uncertainty often makes difficult the generation of two-valued ground truths, containing the reference edge maps, and makes promising the use of fuzzy set theory for evaluation of edge detectors. We consider this issue later in the paper.

3. Quantitative criteria for segmentation quality

Prior to considering some of the currently used quantitative criteria used for segmentation evaluation, some definition of requirements to the segmented image is needed. As mentioned in the section 1, some features of "good" segmentation were formulated by Haralick and Shapiro in [11]:

- regions of an image segmentation should be uniform with respect to some characteristic (e. g. such as grey level intensity and texture);
- adjacent segments should have significantly different values with respect to the characteristic on which they are uniform;
- interiors of the segments should be without many small "holes";
- boundaries of each segment should be simple, not ragged, and must be spatially accurate.

In most cases, the researchers are trying to find some matching quantitative criteria for these properties.

In the literature the following two main approaches to image segmentation are indicated:

Separation of the image into regions of similar features by marking their boundaries – edge-based methods [10] (another terms used are boundary-based and contour-based methods);

Clustering the pixels of an image to a set of classes (segments) such that pixels in the same class are having similar quantitative properties (region-based methods [10]).

For quantitative evaluation of the algorithms from the first class, the same criteria as for the edge detectors are mainly applied (see previous section). Below we consider some quantitative criteria used for the segmentation algorithms of the second class.

Perhaps the easiest and one of the most widely used measures of segmentation accuracy is the percent of incorrectly classified pixels in the image. Obviously, this measure is similar to the statistical measures considered in the previous section. However, there are several problems associated with it:

- its value does not always agree with human observation (this is a typical disadvantage of statistical measures, see previous section);
- it does not reflect the spatial information inherent in the pixel misclassification. Obviously, the error on the border of a segment should be penalized differently from the error in the middle of it;
- errors in different pixel classes (segments) are not weighted according to their importance for segmentation accuracy;
- it provides no information about which pixel classes are most responsible for the observed error.

To overcome the last two problems, in [15] two error measures, which generalize the above I and II error rates, were proposed. Both of them are based on the construction of confusion matrix. The columns of the matrix represent the true pixel classes, while the rows represent the chosen classes. Correctly classified pixels appear as entries on the diagonal of the matrix.

The first of these measures is the multiclass type I error for pixel class k:

$$M_1^k = \frac{\left(\sum_{i=1}^{n} C_k\right) - C_k}{\sum_{i=1}^{n} C_k} \times 100,$$

where n – number of classes (dimension of confusion matrix), C_{kk} – number of class k pixels correctly classified (diagonal of confusion matrix),

$\sum_{i=1}^{n} C_{ik}$ – number of pixels truly of class k (column total of confusion matrix). Here the numerator represents the number of pixels of class k not classified as k and the denominator is the total number of pixel of class k. The second measure is the multiclass type II error for class k:

$$M_2^k = \frac{\left(\sum_{i=1}^{n} C_{kk}\right) - C_{kk}}{\left(\sum_{i=1}^{n}\sum_{k=1}^{n} C_{ik}\right) - \sum_{i=1}^{n} C_{ik}} \times 100,$$

where $\sum_{i=1}^{n} C_{ik}$ – number of pixels classified as class k (row total of confusion matrix), $\sum_{i=1}^{n}\sum_{k=1}^{n} C_{ik}$ – total number of pixels or picture size. In this formula the numerator represents the number of pixels of other classes called class k and the denominator – the total number of pixels of other classes.

Thus, for n image segments we get $2n$ evaluation criteria M_1^k, M_2^k, $k=1,2...n$, which allow to analyze the contribution of each segment to the overall error. Besides that, as noted in [15], the confusion matrix may be weighted according to the importance of each type of pixel misclassification. However, in this paper, no concrete weighting procedure has been proposed. Note that the contribution of different segments to the segmentation accuracy can be formalized by using the elements of fuzzy logic. This formalization involves the construction of fuzzy ground truth images and the use of fuzzy performance measures (see the next section).

Another statistical measure of segmentation accuracy based upon the count of misclassified pixels and using the Bayesian approach, was introduced in [16]. In this paper, the probabilities for an arbitrary pixel of segmented image to belong to a foreground object and to the background, are calculated. Using these probabilities, the probability of overall segmentation error is derived:

$$p(err) = p(o)\,p(b\,|\,o) + p(b)\,p(o\,|\,b),$$

where $p(o)$, $p(b)$ – are a priori probabilities for a pixel to be classified as a foreground object, or as a background, respectively. Both are standard geometric probabilities calculated from the ground truth image. Also, $p(o\,|\,b)$ – is the probability of an error for a background pixel to be labeled as an object. It is defined as the ratio of the sum of background pixels labeled as an object to the sum of true background pixels. Finally, $p(b\,|\,o)$ – is the probability of an error for an object pixel to be labeled as the background; it is defined similarly as above. Later, this formula was extended to the case of multi class segmentation.

The above formula based upon the count of misclassified pixels does not take into account the location of such pixels with respect to their "wrong" segments. Obviously, the higher the corresponding distances are, the worse is the segmentation quality and, therefore, the higher it should be penalized.

In the above mentioned paper [15], the following measure, which takes these distance criteria into account, was proposed:

$$\varepsilon = \frac{\sqrt{\sum_{i=1}^{N} d_i^2}}{A} \times 100,$$

where N – the number of misclassified pixels, A – the area of the picture (total number of pixels), d_i – for the i-th misclassified pixel, the Eucledian distance to the nearest point in the "true" picture actually of the misclassified class.

Clearly this measure is similar to the localization performance measures considered in the previous section. Thus in [15], for the evaluation of segmentation algorithms, the measures of both above classes were used: the statistical and the localization performance measures. This approach corresponds with our evaluation method, which we consider in the next section.

The Pratt's measure, considered in the previous section, was utilized for the segmentation quality evaluation task. One of such versions can be written as [17]:

$$FOM_e = \begin{cases} \frac{1}{N_e}\sum_{i=1}^{N_e} \frac{1}{1+\gamma d_i^2}, & N_e > 0 \\ 1, & N_e = 0 \end{cases},$$

where N_e – the number of misclassified pixels, d_i – same as in the above definition of ε, and γ – a scaling constant. It is easily seen that a correct segmentation yields FOM_e=1.

For performance evaluation of the image segmentation algorithms, in addition to the statistical measures and the measures of localization performance, some other quantitative criteria are also used. Obviously, one of the features of a correct segmentation is that the evaluated image and its ground truth counterpart should have the same fragmentation level (the number of their segments should coincide). To evaluate this feature, the following measure was introduced in [17]:

$$FRAG = \frac{1}{1+\left|\alpha(n_R - n_I)\right|^{\beta}},$$

where n_R – the actual number of segments (on the estimated image), n_I – the number of segments on the corresponding ground truth, α, β – scaling parameters (in [17] α=0.16, β=2). The parameter α determines the contribution of a deviation of n_I and the parameter β determines the contribution of large deviations relative to small deviation. However, this measure takes no account of the characteristics of the segments. Due to this reason, in the same paper, another measure - *FOC* (figure of certainty) was considered. For its calculations, one such characteristic, namely the grey level intensity, is used:

$$FOC = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{1+\left|\psi(f_i - \mu_j)\right|^{\delta}},$$

where N – the number of pixels, f_i – the grey level of i-th pixel in the test image, μ_j – the representative grey level of the corresponding segment (after segmentation), ψ, δ – scaling parameters (which have the same purpose as α and β in the previous formula). In the simplest cases, then the image consists of several regions of cons tant grey value, the expression $(f_i - \mu_j)$ represents the difference between the grey value of the "true" region of i-th pixel and the value of the corresponding region after segmentation. This measure can be generalized to the case of inhomogeneous (with respect to grey level) regions. Also, instead of the grey level values, another quantitative characteristic of the segments can be taken (e. g. their texture characteristic).

Another class of performance measures concerns the accuracy with which the characteristics of the segments can be determined. As known, image segmentation is one of the initial stages of image analysis. The goal of segmentation is to provide a convenient way to measure the features of image objects. This measurement, in turn, is the ultimate goal of image analysis, and it is heavily dependent on the results of segmentation. Thus, taking into account this ultimate goal, it seems reasonable to evaluate the segmentation quality by comparing the measurement of such features on the segmented and the ground truth images. In the paper [18] two criteria based on this approach were offered: *AUMA* (absolute ultimate measurement accuracy) and *RUMA* (relative ultimate measurement accuracy):

$$AUMA_f = \left|R_f - S_f\right|. \quad RUMA_f = \frac{\left|R_f - S_f\right|}{R_f}\times 100,$$

where R_f – the value of feature f, (geometric, color or texture) obtained from a reference image, S_f – the feature value measured from the segmented image. In fact, we have $2P$ criteria, where P is a number of segments' features. The values of *AUMA* and *RUMA* are inversely proportional to segmentation quality: the smaller the values, the better the results. These criteria can be used to evaluate the importance of different features to the accuracy of segmentation and, therefore, to image analysis.

The above considered criteria of segmentation accuracy are integral part of the empirical supervised evaluation methods and they are based on comparison of the segmented image with the ground truth segmentation. Unlike the latter, the empirical unsupervised evaluation methods are based not on comparison with the reference segmentation, but on some subjective characteristics of "good" segmentation. The criteria used by this methods, are aimed at evaluating such desirable characteristics. Here, the absence of need in the reference images has some potential advantages. For example, it allows one to make the evaluation of segmentation algorithms online. This online evaluation enables one to adjust the algorithm's parameters on the fly, depending on the intermediate results; or to decide about the termination of iterative segmentation after achieving the desired accuracy.

We may mention the following three main groups of unsupervised criteria, aimed to evaluate the following features of the segmented image:

- uniformity of the segments;
- grey level difference between adjacent segments;
- shape of the segments.

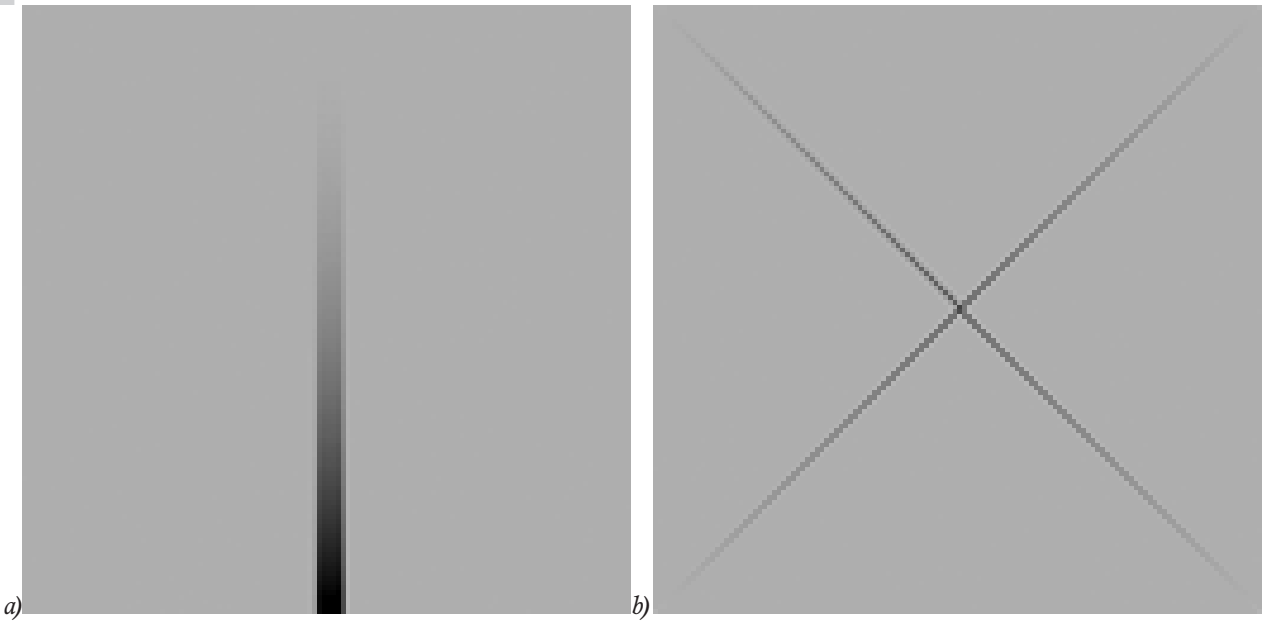
In practice, however, the quantitative performance measures based upon these criteria often produce less adequate results than the ground truth based measures. Besides that, the above criteria are subjective, and therefore their formalization is difficult to carry out. Hence, the empirical supervised evaluation methods are nowadays more reliable than their unsupervised counterparts. In recent years, some complex measures taking into account several of the above criteria (e. g. uniformity of the segments and their number) have been offered (see [19] for details).

It should be noted that until recently, no substantial comparative study of various segmentation evaluation criteria had appeared. In the available papers, either a limited set of test images is considered, or the criteria of the same class, which study the same features of segmented images, are studied. This fact indicates that the development of comparative evaluation methodology remains an actual task.

4. Some features of EDEM method and PICASSO program system

Nowadays, at the Scientific Research Institute of System Analysis (SRISA RAS), the software system PICASSO (PICture Algorithms Study SOftware), aimed for comparative evaluation of image processing and image analysis algorithms, is been developed. The aim of such activity is to create a tool for design of adaptive image analysis systems for a wide range of practical applications. Originally this system was designed to compare various edge detection algorithms (now this is the most advanced part of the system). Its further versions evaluate a wider range of methods including image restoration, texture analysis and image segmentation. The current version of the system contains a large set of images as well as a set of corresponding reference images, a texture database, an image editor, a number of noise generators and filters, filling templates for background and some other components.

The ideological basis for this system is EDEM (Empirical Discrepancy Evaluation Method), which is also under development. Due to variety of application of the algorithms under evaluation and to the fact that some theoretical aspects of these application areas have not yet worked out (for instance, there is no precise definition of the segmentation problem, see [20]), the analytical methods are not used here. In the initial version of PICASSO system intended for testing edge detectors, we worked out a set of synthetic grayscale images. These synthetic images simulate a collection of situations, which are difficult in some sense for edge detection (see Figure 3*a-b* as an example). Here the difficulty is caused by the presence of areas of varying contrast.



a) *b)*

Fig. 3. Examples of images from PICASSO database: a) Degenerating Ridge, b) Degrading Junction

The testing procedure comprised the following steps:

- Selection of an algorithm or a group of similar algorithms for testing.
- Selection of the test images and their corresponding ground truths.
- Selection of the algorithms' parameters. In particular, if the algorithms have a similar input parameter, such selection should enable one to make a joint graphic representation of performance of these algorithms (in terms of a certain measure) with respect to this parameter.
- Choice of distortion methods for the test images. For example, different methods for adding noise to the images can be used. It is desirable that these methods depend on the parameters suitable for graphical representation (e.g. noise deviation).
- Selection of quantitative performance measures.
- Statistical processing of testing results.

This approach was applied for performance testing of several edge detectors on noised and blurred images [21]. Namely, in this paper, the algorithms of Canny, Rothwell, Heitger, Black, Iverson and Smith, different by their nature and designed for solving the same problem, were taken. In all tests, the default values of the algorithms' parameters were used. The variance of Gaussian noise and the width of the blurring window were used as distortion parameters. As the performance evaluation

criteria, the above considered Sensitivity and Specificity were taken. Several meaningful results were revealed after testing. One of them is the possibility in principle to automatically compare the results of testing. Also, the graphical representation of results [21,26] allows one to perform a qualitative analysis of the algorithms.

Besides that, the statistical processing of results (the values of performance measures) revealed that the algorithms of Canny and Rothwell showed the best performance, whereas the algorithm of Iverson was the worst performer of the six algorithms. At the same time, such testing method has several notable disadvantages. As mentioned above, the use of only statistical measures (Sensitivity and Specificity in our case) gives little information about the ability of edge detectors to preserve the shape of contours which separate the objects in an image from background. The use of test images containing only situations which are difficult for edge detection is typical for worst case testing (see section 1 above) and therefore has typical disadvantages of this testing approach. In particular, such approach does not take into account the variety of real-life situation. Finally, this testing methods contains no attempts to separate the flaws of the algorithms, from the problems with their software implementations.

Partly these disadvantages of the initial version of EDEM have been overcome with further development of the PICASSO system. We mention here the paper [22], dealing with the stability study of edge detection under affine transformations (shifts, rotations and scalings) of the objects being tested. This research is essential for identifying objects of in advance unknown size and orientation. In that paper, the same six edge detectors as in [21] were tested. The changes in testing method affected both test images and performance measures. Namely, in addition to the images containing difficult situations for edge detection, their simplified versions were taken. As an example, a simplified version of Degenerated Ridge (Figure *3a*) is presented on Figure *4b*. This simplified picture contains the edges of constant contrast value (average value of its original counterpart).

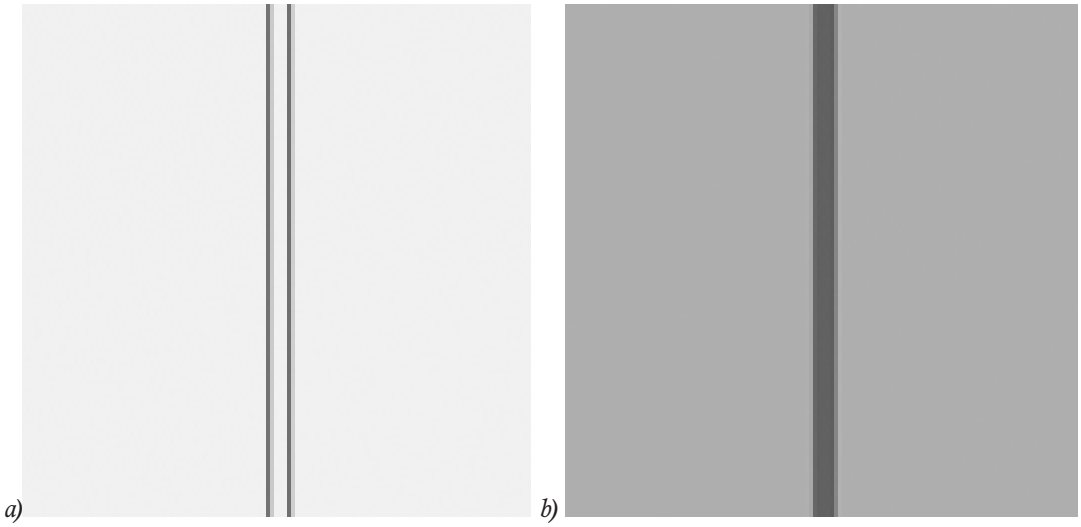


Fig. 4. Pictures corresponding to Figure 3a: a) ground truth image, b) simplified version

As to performance criteria, alongside with Sensitivity and Specificity the Pratt's *FOM* (localization performance measure) was used. Also, the metric of Hausdorff was taken as auxiliary measure. The test results showed that two of the six algorithms displayed an unstable behavior on the simplified test images. Another four algorithms showed acceptable results (confirmed by visual inspection). At the same time, based on these performance criteria, it is impossible to find the "best" algorithm. Also, it was noted that in some cases for two algorithms applied to the same test image, the values of all the measures were practically identical except for the Hausdorff metrics (and visually the performance results were identical). Here the Hausdorff metric played a role of magnifying glass which allows one to see the difference between two pictures undetectable by other means. Thus, the use of this metric in combination with other performance measures can be useful (in particular, it rejects the claim made in [14] that the Hausdorff metric is "practically unusable" for the performance evaluation task).

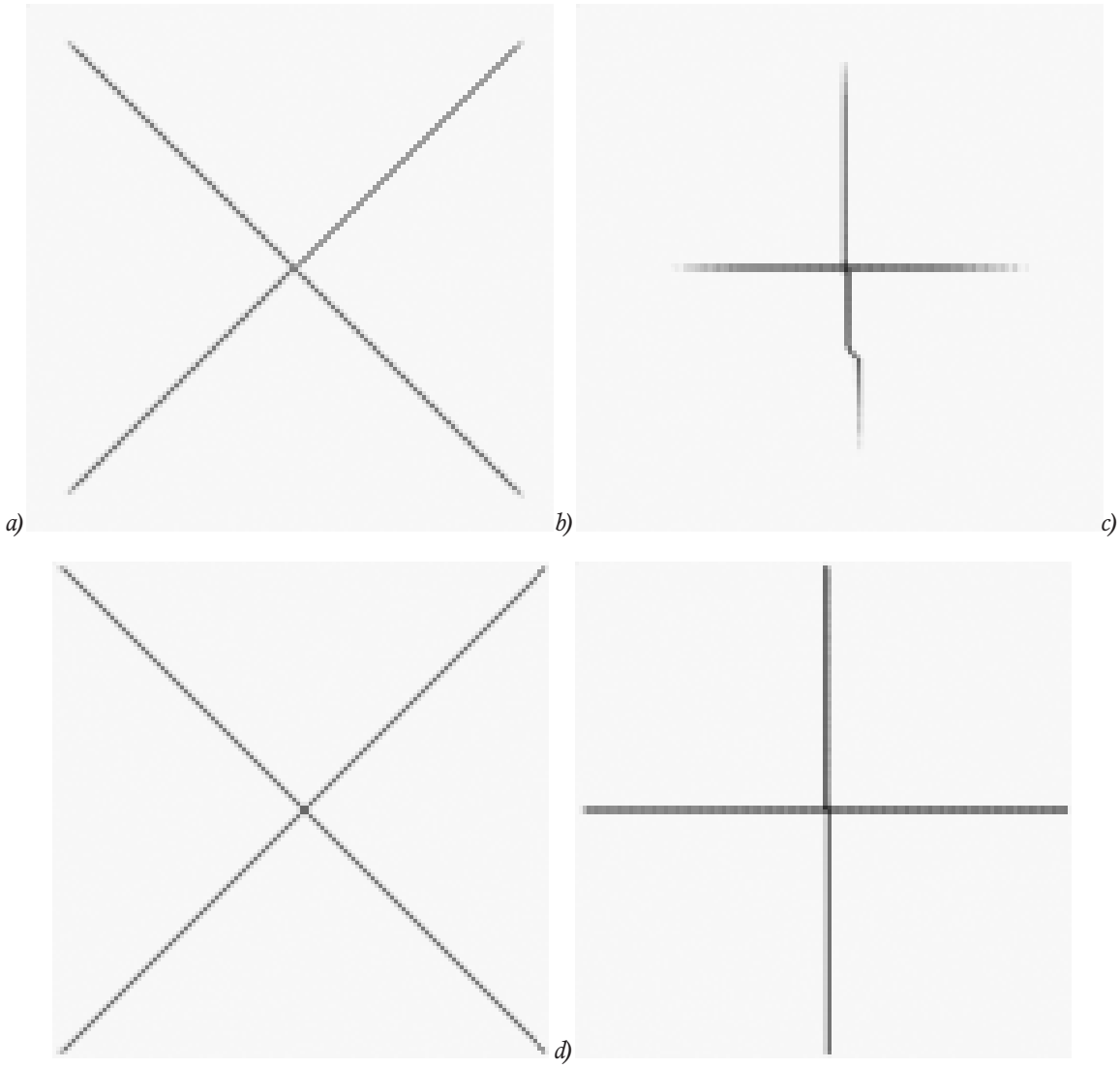


Fig. 5. Results of processing Degrading Junction (Figure 3b) and its 45 degree rotation by Smith detector:
a) b) original picture
c) - d) simplified picture.

On the initial images containing edges of varying contrast, the performance of all six algorithms was significantly worse (as also confirmed by visual assessment). A corresponding example is given on Figure *5*.

The paper [22] also contains some simple tests aimed at finding problems of software implementations of the tested algorithms. In one of such tests, the algorithm's performance results on a given image and on its rotation by 180 degrees were compared (obviously, for reliable software implementations, these results should be close to identical). For the Canny algorithm, the results of this test were worst among the tested edge detectors. Taking into account the high popularity of the latter (or in other words using the previous testing experience), a natural conjecture was made that the source code of this algorithm contains some flaws. To prove that, instead of this software realization, the MATLAB implementation of this edge detector was taken, and for this version the results of such tests were one of the best among the tested algorithms.

Thus, changes in the testing method such as matching two types of test images and two types of performance measures led to improvement of quality of evaluation and allowed to obtain new practical re-

sults. For example, if the location and size of a certain object in an image are not known in advance, and the contrast level along its boundaries is constant, then for finding the contours of such object, four of the six edge detectors with default values of their parameters can be applied. In the case of varying contrast level, the use of all six algorithms is not recommended (at least, a procedure for tuning the algorithm's parameters is required).

As noted in section 2, grayscale images are inherently fuzzy in nature due to the uncertainty which exists in locating the exact position of the boundary which separates the object from background. Also, in processing of remote sensing images, due to insufficient resolution of the sensor, often it is difficult to assign some pixels to one pure class (e. g. to the "forest", "water", or "urban land"). This uncertainty leads to the thought of using elements of fuzzy set theory in image processing and analysis. In particular, it concerns such tasks as edge detection and image segmentation. In recent years, there is a growing number of algorithms handling these tasks as well as image restoration, boundary improvement and texture analysis, which rely on fuzzy logic. Evaluation of these algorithms required a modification of EDEM (see [13], [23] for details). This modification improved the quality of testing for traditional algorithms which do not use fuzzy logic. Moreover, it enables to compare simultaneously both "fuzzy" and "non-fuzzy" algorithms.

It should be noted that methods of comparative evaluation of these two types of algorithms have not been properly developed until now. In some papers about the analysis of remote sensing images, some statistical performance measures (similar to the measures considered in the previous section) were generalized for fuzzy case. In particular, these measures allow one to compare fuzzy sets with respect to their crisp counterparts. Such measures were called in [22] fuzzy similarity measures. Also, in [22] the concept of a fuzzy ground truth image was considered. Although these papers were directed at aerial image classification, some of their results can be applied as well to comparative evaluation of edge detectors. This observation resulted in a paper [23].

In this paper, the fuzzy ground truth images used for testing edge detection algorithms were identified with membership functions of edge class, which for each pixel of a given test image takes the values from 0 to 1. The ordinary ground truths containing the reference edge maps were identified with the characteristic functions of the pixels forming these edge maps (taking the value 1 for edge pixels and 0 otherwise). A fuzzy ground truth used for image segmentation quality evaluation can be identified with a set of membership functions of the reference segments.

One of the new features of EDEM offered in [23] is that depending on the specific feature being tested, different fuzzy ground truths corresponding to the same test image can be used. For example, one fuzzy reference image can be used to test the detection of weak edges, and another one - to test the ability to generate continuous edge contours (its membership function is sensitive to the gaps in the edge map).

Another promising application of these fuzzy ground truths is that they can be used to study the ability of edge detectors to find image feature points (for example, the corner points of a rectangular). In particular, the knowledge of these points is important for the edge linking procedure. Assigning the higher values of the membership functions to these pixels compared to the other pixels on the ground truth edge map, we get the higher values of fuzzy similarity measures when such pixels were marked as edges by the tested algorithm. Note that the use of fuzzy ground truth images can be also useful for testing image segmentation algorithms. At present, we are developing a method of generating various fuzzy ground truth images for their use in evaluation of image processing algorithms, and we study the applicability for this purpose of different fuzzy similarity measures.

5. Conclusions

In last two decades, the problem of performance evaluation of image processing algorithms has received a growing interest in the literature. Since the general theory of image processing and analysis is not completed so far, the analytical evaluation methods have a limited application restricted to some special cases. Nowadays the priority is given to the empirical methods which use the ground truth images for evaluation (supervised methods), as well as to the empirical methods where such images are not required (unsupervised methods). The first of these methods provide more accurate evaluation than the second ones, whereas the latter can be applied to real-time software for evaluating the algorithms online.

The current version of our method EDEM, aimed at empirical evaluation of various computer vision algorithms, has the following main features:

- Use of "difficult" test images for the evaluated algorithms.
- Matching these images with their simplified versions.
- Use of performance measures from different classes for quantitative evaluation of the algorithms (e. g. the

use of statistical and localization performance measures for comparative evaluation of edge detectors).

- Possibility to evaluate qualitatively the results of tests (e. g. their representation in graphical form).
- Application of elements of fuzzy logic, including the concept of fuzzy ground truths. The use of several fuzzy ground truth images for the same test image for the purpose of more profound evaluation. Utilizing existing fuzzy similarity measures for the evaluation of computer vision algorithms.

Our method found several practical applications (its latest application is the design of a software system for automated blood cell image segmentation, [24]-[25]). At the same time, there are some open issues concerning the general methodology of comparative evaluation, as well as the EDEM method itself (see [26] for details). For example, a significant difficulty appearing during the evaluation process is how to select a proper performance measure giving a reliable assessment of the performance of the algorithms. Most of the measures used in practice evaluate reliably only a certain feature of the tested algorithms, and no all-inclusive evaluation criteria exist. Thus, the design of reliable performance measures and their matching within one testing method remain actual issues. Especially it concerns the matching of the measures of the same class (e. g. localization performance measures). The use of several such measures inevitably leads to the question of their proper ranking (which one is most reliable). During our tests within EDEM, we found that the Hausdorff metric can be used as a second (auxiliary) localization performance measure for evaluation of edge detectors and image segmentation algorithms [20], [22].

As for the methodology for application of test images and the corresponding ground truths, one of the main open questions here is the completeness of the test set. Within our method, the selection of test images is practical task-based. For example, for testing edge detection algorithms, the edge density in the test images should correspond to such density in the real images, to which the tested algorithms are supposed to be applied. Accordingly, we are developing a technology for generation of such test images in the current version of PICASSO system. Also, we are developing a method for generation of various fuzzy ground truth images as well as a method of using several fuzzy similarity measures to make practical evaluation. Finally, our experience in the comparative evaluation shows that in some cases it is impossible to choose the absolute best algorithm on the basis of the values of performance measures; and at the same time there are several algorithms with similar (acceptable) results. In these situations, to choose most suitable algorithm for practical applications, one should take into account such its properties as processing complexity, resource efficiency, parallelizability, etc. These properties are studied by analytical evaluation methods. We are planning to work out a method for the evaluation of such features in the future versions of our system.

References

1. Kirsch RA, Kahn L, Ray C, Urban GH. Experiments in processing pictorial information with a digital computer. Proceedings of the Easterm Joint Computer conference 1957; 221-9.

2. Zhang YJ. Evaluation and comparison of different segmentation algorithms. Pattern Recognition Letters 1997; 18(10): 963-74.

3. Heath MD, Sarkar S, Sanocki T, Bowyer K. Robust visual method for assessing the relative performance of edge detection algorithms. IEEE Transactions on Pattern Analysis and Machine Intelligence 1997; 19(12): 1338-59.

4. Wirth MA. Performance evaluation of image processing algorithms in CADe. Technology in Cancer Research and Treatment 2005; 4(2): 159-72.

5. Zhang H, Fritts JE, Goldman SA. Image segmentation evaluation: A survey of unsupervised methods. Computer Vision and Image Understanding 2008; 110(2): 260-80.

6. Shin MC, Goldgof D, Bowyer K. Comparison of edge detector performance through use in an object recognition task. Computer Vision and Image Understanding 2001; 84(1): 160-78.

7. Cardoso JS, Corte-Real L. Toward a Generic Evaluation of Image Segmentation. IEEE Transactions on Image Processing 2005; 14(11): 1773-82.

8. Thomas GA, Grau O. 3d image sequence acquisition for tv and film production. Proceedings of 1st International Symposium on 3D Data Processing, Visualisation and Transmission 2002; 320-6.

9. Zhang YJ. A survey on evaluation methods for image segmentation. Pattern Recognition 1996; 29(8): 1335-46.

10. Zhang YJ. Image segmentation evaluation in this century. Encyclopedia of Information Science and Technology. 2nd edition. IGI Global; 2009: 1812-17.

11. Haralick RM, Shapiro LG. Image segmentation techniques Computer Vision, Graphics, and Image Processing 1985; 29(1): 100-32.

12. Canny J. A computational approach to edge detection. IEEE Transactions on Pattern Analysis and Machine Intelligence 1986; 8(6): 679-98.

13. Gribkov IV, Koltsov PP, Kotovich NV, Kravchenko AA, Kutsaev AS, Osipov AS, Zakharov AV. On some issues of the quantitative performance evaluation of edge detectors [In Russian]. Programmnye Produkty I Sistemy 2011; 2: 13-9.

14. Baddeley AJ. Errors in binary images and Lp version of the Hausdorff Metric. Nieuw Archief voor Wiskunde 1992; 10: 157-83.

15. Yasnoff WA, Mui JK, Bacus JW. Error measures for scene segmentation. Pattern Recognition 1977; 9(4): 217-31.

16. Van Droogenbroeck M, Barnich O. Design of Statistical Measures for the Assessment of Image Segmentation Schemes. Proceedings of 11th International Conference on Computer Analysis of Images and Patterns (CAIP2005), Lecture Notes in Computer Science 2005; 3691: 280-7.

17. Strasters KS, Gerbrands JJ. Three-dimensional image segmentation using a split, merge and group approach. Pattern Recognition Letters 1991; 12(5): 307-25.

18. Zhang YJ, Gerbrands JJ. Objective and quantitative segmentation evaluation and comparison. Signal Processing 1994; 39(1-2): 43-54.

19. Koltsov PP, Kotovich NV, Kravchenko AA, Kutsaev AS, Osipov AS, Zakharov AV. Criteria for evaluating image segmentation [In Russian]. Trudy NIISI RAN 2012; 2(2): 87-99.

20. Gribkov IV, Koltsov PP, Kotovich NV, Kravchenko AA, Kutsaev AS, Osipov AS, Zakharov AV. Testing of image segmentation methods in PICASSO system [In Russian]. Moscow: NIISI RAN; 2007.

21. Gribkov IV, Koltsov PP, Kotovich NV, Kravchenko AA, Kutsaev AS, Nikolaev VK, Zakharov AV. PICASSO – A System for Evaluating Edge Detection Algorithms. Pattern Recognition and Image Analysis 2003; 13(4): 617

22. Gribkov IV, Koltsov PP, Kotovich NV, Kravchenko AA, Kutsaev AS, Osipov AS, Zakharov AV. Edge Detection under Affine Transformations: Comparative Study by PICASSO 2 System. WSEAS Transactions on Signal Processing 2006; 2(9): 1215-21.

23. Osipov A. A fuzzy approach to performance evaluation of edge detectors. Lecture Notes in Signal Science, Internet and Education. WSEAS Press; 2007: 94-9.

24. Koltsov PP, Kotovich NV, Kravchenko AA, Kutsaev AS, Kuznetsov AB, Osipov AS, Sukhenko EP, Zakharov AV. On one approach to blood cell image segmentation. The 11th International Conference "Pattern Recognition and Image Analysis" (PRIA-11-2013), Conference Proceedings 2013; 2: 615-18.

25. Belyakov VK, Koltsov PP, Kotovich NV, Kravchenko AA, Kutsaev AS, Kuznetsov AB, Osipov AS, Sukhenko EP, Zakharov AV. On one method of blood cell classification and its software implementation [In Russian] Programmnye Produkty I Sistemy 2014; 2: 46-56.

26. Koltsov PP, Kotovich NV, Kravchenko AA, Kutsaev AS, Osipov AS, Zakharov AV. Direct assessment of software quality. Criteria and materials for testing [In Russian]. Programmnye Produkty, Sistemy I Algoritmy 2014; 3: P. 1-8.