www.czasopisma.pan.pl PAN www.journals.pan.pl
INTL JOURNAL OF ELECTRONICS AND TELECOMMUNICATIONS, 2018, VOL. 64, NO. 3, PP. 307–313
Manuscript received May 8, 2018; revised July, 2018. DOI: 10.24425/123523

Comparison of Effective Coverage Calculation Methods for Image Quality Assessment Databases

Mateusz Buczkowski and Ryszard Stasiński

***Abstract*—This article provides a comparison of a three methods that can be used for calculating effective coverage of image quality assessment database. The aim of this metric is to show how well the database is filled with variety of images. For each image in the database the Spatial Information (SI) and Colorfulness (CF) metric is calculated. The area of convex hull containing all the points on SI x CF plane is indication of total coverage of the database, but it does not show how efficiently this area is utilized. For this purpose an effective coverage was introduced. An analysis is performed for 16 databases - 13 publicaly available and 3 artificial created for the purpose of showing advantages of the effective coverage.**

***Keywords*—Image Quality Assessment, IQA database analysis, effective coverage**

I. Introduction

INCREASING computing power, which is available in more and more compact devices allows image processing be used in even more applications. Many of those include a human as an observer of a final image. Due to this we must ensure the high quality of an output image regardless of the used algorithms. Output quality can be measured with basic metrics, such as peak signal to noise ratio (PSNR), but it requires access to the pristine image and do not reflect human perception of the image and the distortion.

The development of the new metrics for image quality assessment (IQA) is still ongoing. Their principle may vary, as some of them estimate quality based on distorted and pristine image (full-reference IQA - FR-IQA), and some are based only on distorted image (no-reference IQA - NR-IQA). There is also a group of algorithms that use partial information from reference image and a distorted image (reduced-reference IQA - RR-IQA).

In order to evaluate a particular method for measuring image quality we must compare it agains a ground truth. This reference is used to determine how close the estimated value is to the expected result. In order to obtain reference values an experiment involving many different human observers is carried out.

During this experiment people rate the quality of the images. Depending on the methodology they can either compare it with the reference image or they have to judge the quality from the single image. Typically observers are asked to assess the image quality by assigning a score from one to five (bad, poor, fair, good, excellent). These scores are gathered for a particular image and distortion level. From this, mean opinion score (MOS) is calculated. This method was first proposed by the International Telecommunication Union (ITU) for evaluation of audio quality, but it was adapted for image quality as well.

This process is costly both in time and resource. Performing such experiment over and over is inefficient, therefore, results are often shared with public as an IQA database and contain:

- reference images
- distorted images
- MOS

and optionally:

- measure of distortion (eg. noise power, compression level)
- individual scores or standard deviation of MOS

II. IQA databases overview

Every available database is described in detail in the paper published by the authors. It covers details on how experiment was conducted, how many images are contained in the database, what kind of distortions were aplied and other relevant parameters.

While evaluating IQA algorithm it is good approach to perform tests on different databases. This allows to avoid overfitting problem for a single database. Therefore it is convinient to have an overview of different databases in one place. There are two research articles worth mentioning in this context. The first one [1] convers 6 monochrome, 8 color and 13 video databases. This work is from 2012 and since then a new, worth noticing databases were published. Author of [1] created a list of publicly available databases for quality assessment which is available online [2].

Another work is [24], where 17 databases with color images are compared. Since the paper is from 2017, it provides a significant update in comparison to [1].

III. Available databases

Table I contains list of databases, which can be used for verifying IQA algorithms and which are compared in this work.

Among these databases, the one that requires longer explanation is the Exploration database. It contains much more images than other databases, but it lacks MOSs. Authors in [20] propose another approach to IQA algorithms validation. It is stated that having huge, diverse database and proposed metrics gives better overview of how good is the solution under evaluation and what are its flaws.

Work funded by the Ministry of Science and Higher Education for the statutory activity of conducting research and development work and related tasks, contributing to the development of young scientists and doctoral students in 2017 under the project number 08/83/DSMK/4716.

M. Buczkowski and R. Stasiński are with the Poznan University of Technology, POLAND, (e-mail: mateusz.buczkowski@put.poznan.pl, ryszard.stasinski@put.poznan.pl).

TABLE I
LIST OF DATABASES AND THEIR REFERENCES

Full name	Short name	No. Ref Imgs	No. Dist Imgs	Year	Reference
Colourlab Image Database: Image Quality	CIDIQ	23	690	2014	[3]
Computational and Subjective Image Quality	CSIQ	30	900	2010	[4]
Single Distortion Imaging and Vision Laboratory	SD-IVL	20	400	2014	[5]
Multi Distortion Imaging and Vision Laboratory	MD-IVL	10	752	2017	[6]
Irvine Valley College	IVC	10	235	2005	[7]
LIVE Image Quality Assessment Database	LIVE2006	29	982	2006	[8] [9] [10]
LIVE In the Wild Image Quality Challenge Database	Challenge	0	1162	2015	[11] [12]
LIVE Multiply Distorted Image Quality Database	MDIQ	15	225	2012	[13]
Media Communication Laboratory	MCL-JCI	50	5000	2016	[14] [15] [16]
Multiply Distorted Image Database	MDID	20	1600	2016	[17]
Tampere Image Database 2008	TID2008	25	1700	2008	[18]
Video Communication Laboratory	VCL@FER	23	575	2012	[19]
Waterloo Exploration Database	Exploration	4774	94880 [1)]	2017	[20]

1) Distorted images are not directly included in the database, but the way of reproducing them is described in the reference.

IV. IQA DATABASES COMPARISON

Comparison of different databases is based on source content charactedistics computed for all pristine images in the particular database.

Formerly used metrics could be insufficient to fully represent database coverage and for that reason in [23] [24] a new metric (*Total effective coverage*) was introduced. This metric address the issue, where database with large uncovered areas could have results in different metrics.

A. Spatial Information (SI)

The measure of SI is based on the edge energy [21]. In order to calculate SI each color channel is filtered with a horizontal and vertical Sobel kernel (s_h and s_v respectively). Let $s_r = \sqrt{s_v^2 + s_h^2}$ denote the edge magnitude at each pixel location. The SI value is a root mean square of s_r normalized with a vertical resolution factor:

$$SI = \sqrt{L/1080}\sqrt{\sum s_r^2/P} \tag{1}$$

where L is a vertical resolution and P is number of pixels in the image.

SI for color images is a weighted average of SI for different RGB channels and is calculated as follows:

$$SI_{RGB} = 0.299SI_R + 0.587SI_G + 0.114SI_B \tag{2}$$

B. Colorfulness

Colorfulness (CF) represents intensity and variety of colors in an image [22]. First step of CF calculation is creating opponent color spaces: $rg = R-G$ and $yb = 0.5(R+G)-B$. From these we calculate standard deviation and mean value along both directions ($\sigma_.$ and $\mu_.$). Formula for CF value based on mean and variance is as follows:

$$CF = \sqrt{\sigma_{rg}^2 + \sigma_{yb}^2} + 0.3\sqrt{\mu_{rg}^2 + \mu_{yb}^2} \tag{3}$$

C. SI × CF convex hulls

Presenting SI vs. CF on a scatter plot shows variety of images creating a database [1].

In order to emphesise advantages and disadvantages of particular methods, three artificial databases were added. This allows to easily understand how different metrics behave for different types of databases. These databases are:

- Uniform - uniformly distributed images along SI and CF space
- Random - randomly distributed (with uniform distribution) images along SI and CF space
- X shaped - images occupying only two perpendicular lines

Besides the points representing particular images, a convex envelope is depicted. As can be seen from Figure 1 more images in the database tend to cover larger area, but this is just a rough estimate of database coverage.

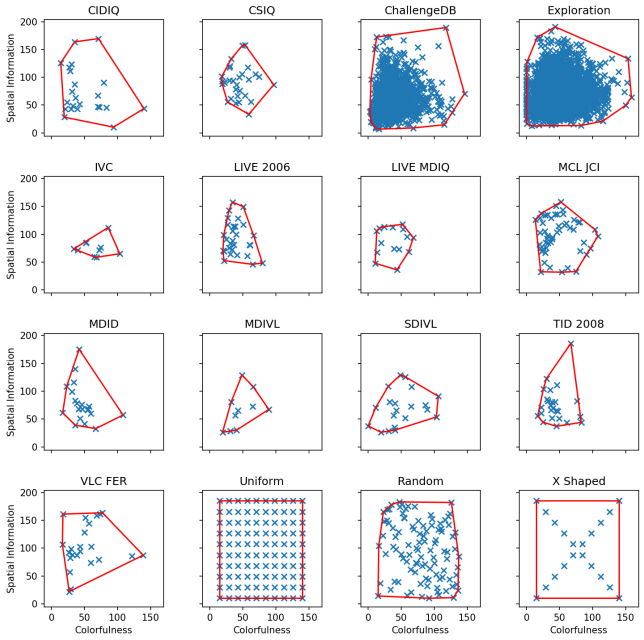

Fig. 1. Spatial Information vs Colorfulness scatter plots for all databases and corresponding convex hull

D. Relative ranges

The first numerical values, that can describe database as a whole is relative range of particular characteristic. It is defined

as follows [1]:

$$R_i = \frac{max(C_i) - min(C_i)}{C_i^{max}} \tag{4}$$

where C_i^{max} is maximum value for given dimension across all databases.

Figure 2 presents relative ranges calculated for all databases. Analyzing relative ranges for SI and CF we should mention MCL-JCI, MDID, VCL_FER or CIQID as a very good one and the Exploration, Challenge database as well as all artificial ones as perfect.

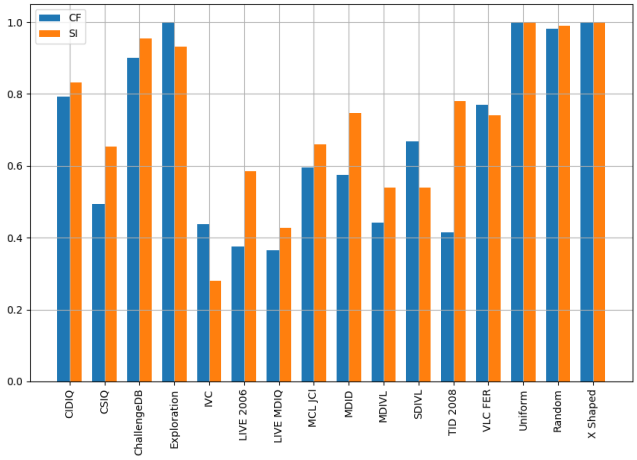

Fig. 2. Colorfulness and Spatial Information relative ranges for all databases

E. Uniformity

Uniformity measure shows how well distributed the images are in the SI×CF space are. This metric is an entropy of the B-bin histogram of particular characteristic:

$$U_i = -\sum_{k=1}^{B} p_k log_B p_k \tag{5}$$

The higher the entropy the more uniform a database is. Values of uniformity for databases are presented in Figure 3. Databases having largest relative ranges might tend to become non-uniform. As an example we can mention CIQID, as it is the most non-uniform database, but its relative ranges were above average.

F. Relative total coverage

Relative total coverage is the square root of area of a convex hull for all points in normalized coordinates C_i/C_i^{max}. Relative total coverage is presented in the Figure 4. This metric was commonly used to evaluate how good is the database in terms of coverage.

As one can see all results shown in Figures 2, 3 and 4 present non-existing databases as the perfect ones. However, from the SI×CF plot (Fig. 1) we can assume that their real coverage

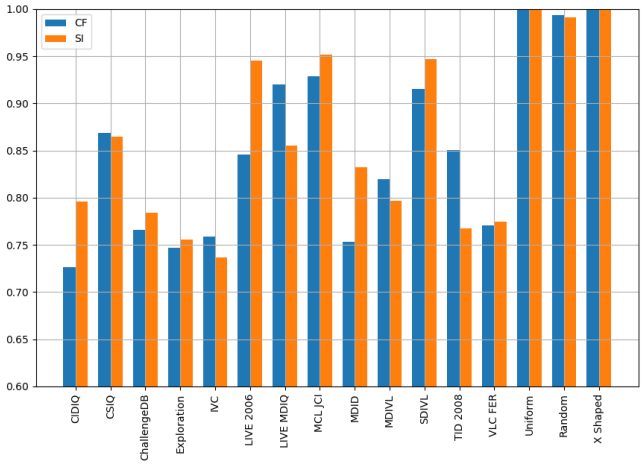

Fig. 3. Colorfulness and Spatial Information uniformity for all databases

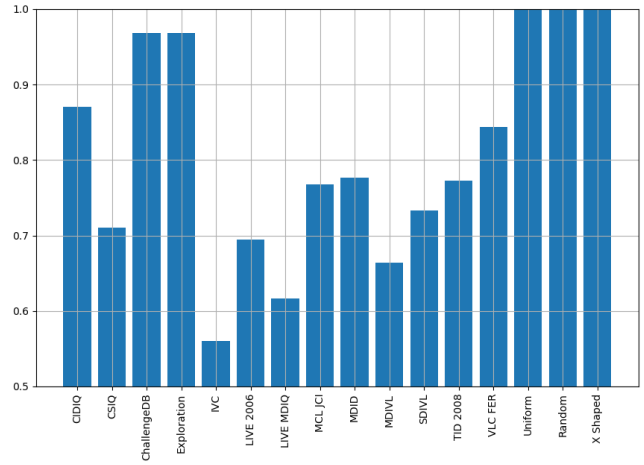

Fig. 4. Total relative coverage for all databases

is very different, and X shaped database is not as good as Uniform and Random. Therefore there was a new metric introduced called *Total effective coverage*, which deals with this weakness.

G. Total effective coverage

Total effective coverage is an extension of relative total coverage. It is additionally weighed with the fill rate factor. This factor is a measure of how well the convex hull is filled with the images. In other words, large empty spaces lead to decreased fill rate.

In [24] there is described one way of obtaining fill rate, which was calculated for 14 databases. Additional work on this was conducted in [23], where two different methods of fill rate calculation were proposed, but it was limited to only three real databases. This work extends both these works and provides an analysis of three methods of fill rate factor calculation on a large variety of databases.

1) Fixed radius method: The first proposal for fill rate is based on an assumption that single point does not cover only infinitely small area on SI×CF plane, but rather occupies some fixed radius. An approach in this case is as follows:

1) Around each data point in SI×CF space set the $presence$ (p) value to 1 in a circle of radius r. Circle area is cropped by the convex hull boundary.
2) Calculate the total area where p is equal to 1
3) Calculate the ratio of presence area and the total convex hull area

The fill rate factor is then defined as:

$$\eta_{fill} = \frac{\iint_D p\,dx\,dy}{D} \tag{6}$$

The process of calculating fill rate for all databases is depicted in the Figure 5.

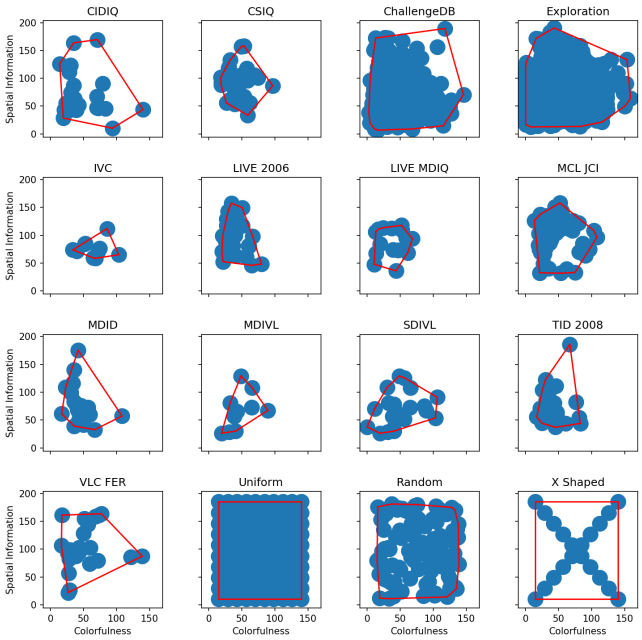

Fig. 5. Databases fill rate calculation - fixed radius method

Instead of points, images are marked with the circles. The difference is best visible between Uniform and X shaped. In case of Uniform database, the whole area is covered, while for X shaped there are large white spots. This is an expected behavior as we want that metric which will compromise a database with large total relative coverage but a lot of unfilled area.

Values of fill rate for all databases are presented in the Figure 6. At this point the weakness of the non-existing database is exposed. The fill rate is the lowest among all databases. The databases, which in fact have good coverage like Exploration or Challenge, still have high value of fill rate.

2) Gaussian radius method: Another approach that is presented in [23] is similar to the fixed radius method. In the previous method presence p is constant and equal to 1. This time presence value reduce with the distance from the original SI×CF point. The method is called gaussian radius method,

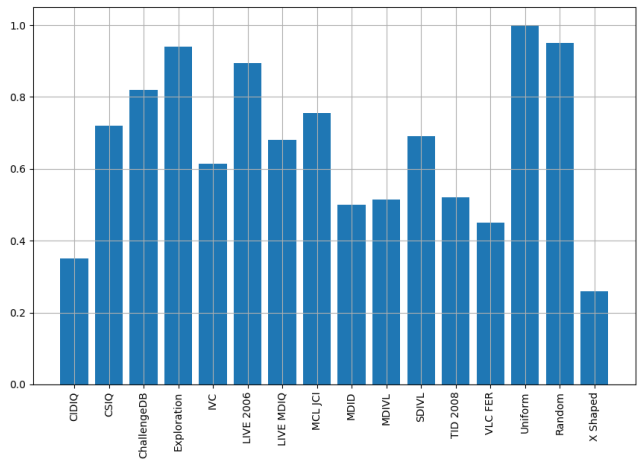

Fig. 6. Fill rate for all databases - fixed radius method

because the presence value lowers according to the Gaussian function. This method is depcicted in the Figure 7.

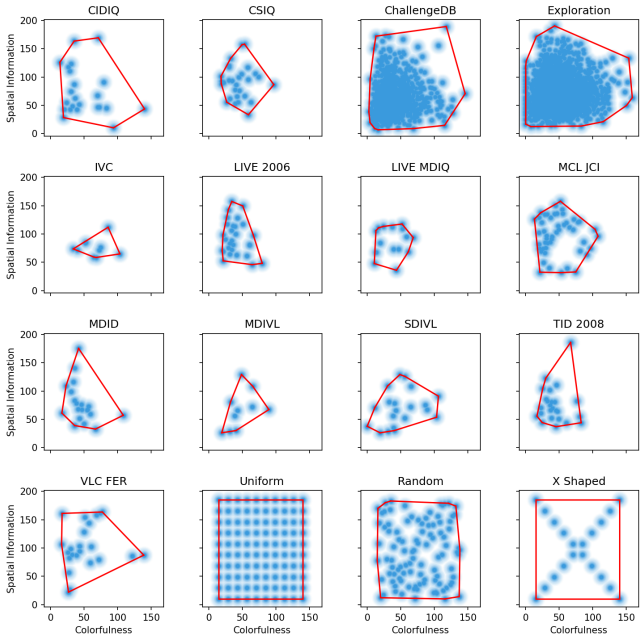

Fig. 7. Databases fill rate calculation - Gaussian radius method

The results for fill rate factor in case of this method are presented in the Figure 8

This method does not provide as hard decision as the fixed radius. In this case obtaining fill rate equal to 1 requires infinite number of images. This method intuitively should provide more reliable results as the further we are from the image, the lower its impact on the database coverage (therefor the Gaussian function). However this method tends to give very low fill rate for databases with low number of images.

3) Delaunay triangulation: Last method, which is improved version of [23] is based on Delaunay triangulation. From set of points P it creates triangulation $DT(P)$ such that

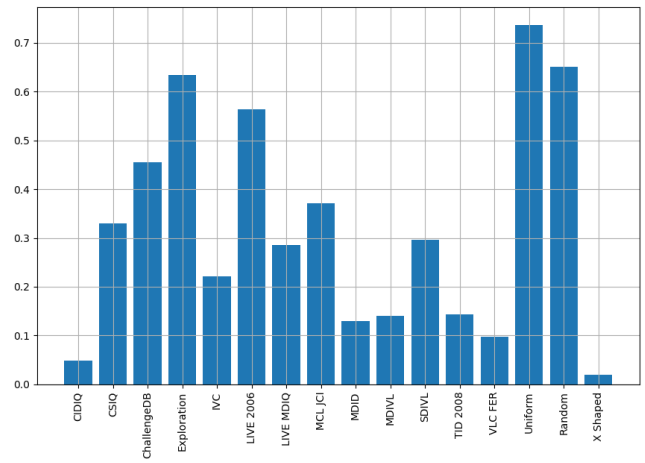

Fig. 8. Fill rate for all databases - Gaussian radius method

no point P is inside the circumcircle of any triangle in $DT(P)$ [25]. Then area of each triangle is calculated and normalized in following way:

$$A_i = \frac{A_i}{\sum_k A_k} \tag{7}$$

where A_i is area if i-th triangle.

The assumtion of this method is that well filled database have equal areas of the triangulation. The higher variation of the areas mean that some points are placed further from the others. Effect of traingulation is depicted in the Figure 9.

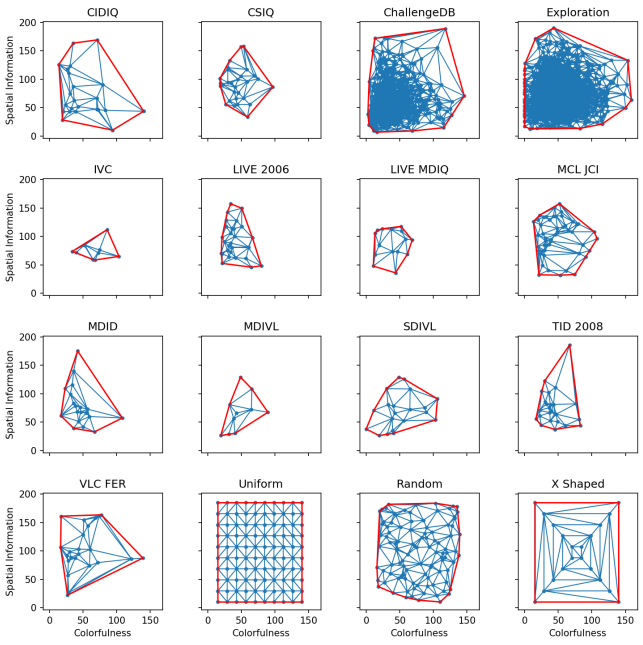

Fig. 9. Delaunay triangulation

Uniformity in this case is calculated differently from [23]. Instead of calculating modified entropy, simply variance of areas is calculated. The higher the variance the lower the uniformity. Calculated values of uniformity are presented in the Figure 10.

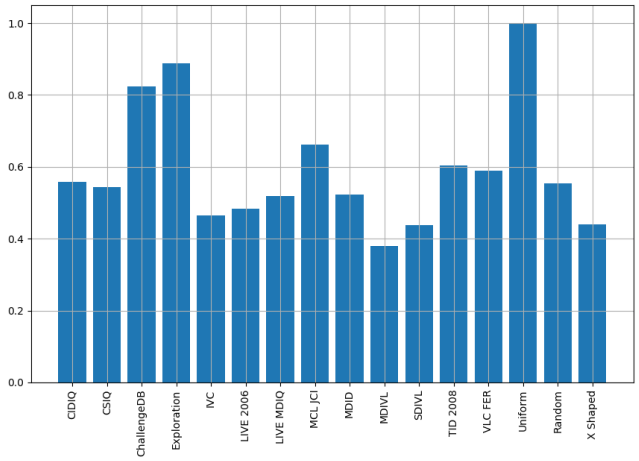

Fig. 10. Fill rate for all databases - Delaunay triangulation method

In this method best result obtain Uniform database, since all areas are equal and variance is zero. Also databases with many images (Exploration and Challenge) have good metric, because only small fraction of areas is different from mean area. However, this approach might not be very reliable, as in case of CIDIQ and CSIQ databases provide similar results. From previous methods and intuitive analysis on SI×CF plane we can see that CIDIQ have lower fill rate (even though it has higher total relative coverage).

Having fill rate η_{fill} we can calculate total effective coverage as:

$$T_{eff} = \eta_{fill} \cdot T_{rel.coverage} \tag{8}$$

Results of total effective coverage are presented in Figures 11, 12 and 13 for fixed radius method, Gaussian radius method and Delaunay trangulation, respectively. As one would expect the quality of X shaped database dropped down in all cases. The Exploration and Challenge databases are boldly dominating over the rest of publicly available ones, which could be expected when inspecting Figure 1.

From provided results one can see that both fixed radius and Gaussian radius methods provide good gradation of databases. In case of fixed radius we can easily point out the perfect database (Uniform) and the bad database (X shaped). In case of Gaussian radius, none of the considered databases reach the perfect score, as the images would have to be infinitely dense in order to provide 1.0 score in this method, but differences between values for different methods are emphesised. While comparing these two methods we can see that decision from Gaussian radius method is softer this method does not priviledge databases easily.

Delaunay triangulation method provides slightly different ordering of the databases. This method tends to favor high regularity or high amount of images in the database. The more uniform distribution of points in both directions, the higher score from Delaunay triangulation is. The weakness

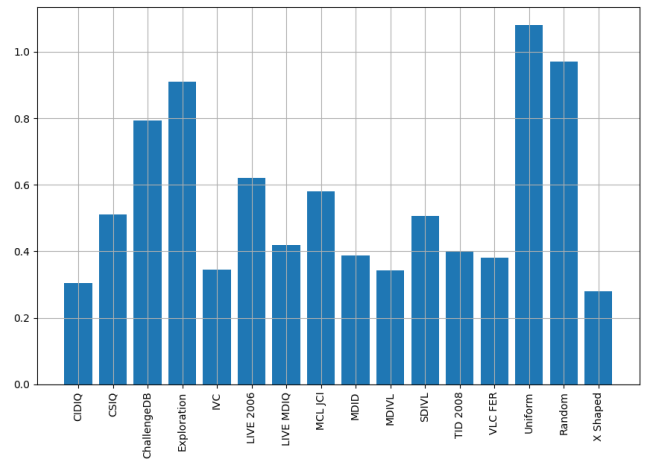

Fig. 11. Total effective coverage for all databases - fixed radius method

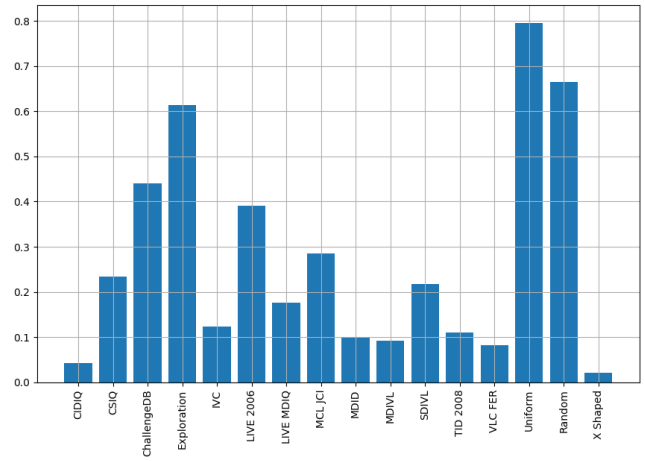

Fig. 12. Total effective coverage for all databases - Gaussian radius method

of this method is particularly well visible in case of Random database. Its coverage is very good area is completely filled with images, yet the uniformity is slightly above 0.5. This is due to the fact that some areas have higher density of images than other regions. Because of this, the database is punished with lower score. Such phenomenon does not occur in case of first two methods. If additional images appear in the database, causing uniformity, the score can only increase while calculating completeness by fixed or Gaussian radius. For Delaunay triangulation adding new images can decrease the score, which can be considered as counterintuitive. This behavior can interpreted in another way. If we would consider this metric not as a how well images cover the particular convex hull, but rather how big area this amount of images cover. One would expect that adding new images to the database increase its variety. If it is not the case and the convex hull is not changed after the new image is added we might expect the score to be lower.

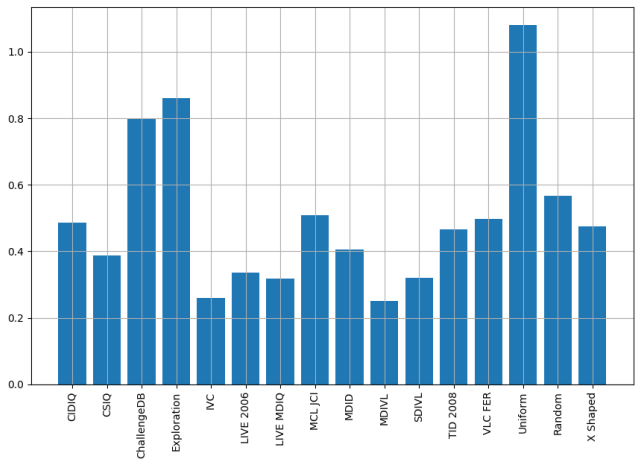

Fig. 13. Total effective coverage for all databases - Delaunay traingulation

V. Conculsions

In the article the measures for calculating IQA databases coverage were confronted for large number of databases. Three possible solutions are presented and discussed, namely fixed radius coverage, Gaussian radius coverage and Delaunay triangulation uniformity.

Two of the proposed metrics can be successfully used in the intended purpose of measuring effective coverage, while the third one does not fully meet the criterion of grading databases in order of perceptual coverage. This metric, however, can be used in different aspect. It shows how uniform and regular the database is.

References

[1] S. Winkler, Analysis of Public Image and Video Databases for Quality Assessment, Sel. Top. Signal Process. IEEE J., vol. 6, no. 6, pp. 616625, 2012.
[2] S. Winkler, ”Image and Video Quality Resources”, http://stefan.winkler.site/resources.html, [Mar, 2017]
[3] Liu X., Pedersen M., Hardeberg J.Y. (2014) CID:IQ A New Image Quality Database. In: Elmoataz A., Lezoray O., Nouboud F., Mammass D. (eds) Image and Signal Processing. ICISP 2014. Lecture Notes in Computer Science, vol 8509. Springer, Cham
[4] E. C. Larson and D. M. Chandler, ”Most Apparent Distortion: Full-Reference Image Quality Assessment and the Role of Strategy,” Journal of Electronic Imaging, 19 (1), March 2010.
[5] Silvia Corchs, Francesca Gasparini, Raimondo Schettini, No Reference Image Quality classification for JPEG-Distorted Images, In Digital Signal Processing, volume 30, pp. 86-100, Elsevier, 2014.
[6] Silvia Corchs, Francesca Gasparini, Raimondo Schettini , Noisy Images-JPEG Compressed: Subjective and Objective Image Quality Evaluation, In Image Quality and System Performance XI, volume 9016, pp. 90160-, SPIE, 2014
[7] Patrick Le Callet, Florent Autrusseau, ”Subjective quality assessment IRCCyN/IVC database”, http://www.irccyn.ec-nantes.fr/ivcdb/ [Mar, 2017]
[8] H.R. Sheikh, Z.Wang, L. Cormack and A.C. Bovik, ”LIVE Image Quality Assessment Database Release 2”, http://live.ece.utexas.edu/research/quality [Mar, 2017]
[9] H.R. Sheikh, M.F. Sabir and A.C. Bovik, ”A statistical evaluation of recent full reference image quality assessment algorithms”, IEEE Transactions on Image Processing, vol. 15, no. 11, pp. 3440-3451, Nov. 2006.
[10] Z. Wang, A.C. Bovik, H.R. Sheikh and E.P. Simoncelli, ”Image quality assessment: from error visibility to structural similarity,” IEEE Transactions on Image Processing , vol.13, no.4, pp. 600- 612, April 2004.

[11] D. Ghadiyaram and A.C. Bovik, "Massive Online Crowdsourced Study of Subjective and Objective Picture Quality," IEEE Transactions on Image Processing, accepted arXiv 2015 [arXiv]
[12] D. Ghadiyaram and A.C. Bovik, "LIVE In the Wild Image Quality Challenge Database," Online: http://live.ece.utexas.edu/research/ChallengeDB/index.html [Mar, 2017]
[13] Dinesh Jayaraman, Anish Mittal, Anush K. Moorthy and Alan C. Bovik, Objective Quality Assessment of Multiply Distorted Images, Proceedings of Asilomar Conference on Signals, Systems and Computers, 2012.
[14] Lina Jin, Joe Yuchieh Lin, Sudeng Hu, Haiqiang Wang, Ping Wang, Ioannis Katsavounidis, Anne Aaron and C.-C. Jay Kuo. Statistical Study on Perceived JPEG Image Quality via MCL-JCI Dataset Construction and Analysis. Electronic Imaging (2016), the Society for Imaging Science and Technology (IS&T).
[15] Sudeng Hu, Haiqiang Wang and C.-C. Jay Kuo, A GMM-based stair quality model for human perceived JPEG images, IEEE International Conference on Acoustic, Speech and Signal Processing, Shanghai, China, March 20-25, 2016
[16] Joe Yuchieh Lin, Lina Jin, Sudeng Hu, Ioannis Katsavounidis, Anne Aaron and C.-C. Jay Kuo. Experimental Design and Analysis of JND Test on Coded Image/Video. SPIE Optical Engineering+ Applications. International Society for Optics and Photonics, 2015
[17] W. Sun, F. Zhou, Q. M. Liao. MDID: a multiply distorted image database for image quality assessment, Pattern Recognit. 61C (2017) pp. 153-168.
[18] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, M. Carli, F. Battisti, "TID2008 - A Database for Evaluation of Full-Reference Visual Quality Assessment Metrics", Advances of Modern Radioelectronics, Vol. 10, pp. 30-45, 2009.
[19] A.Zaric, N.Tatalovic, N.Brajkovic, H.Hlevnjak, M.Loncaric, E.Dumic, S.Grgic, "VCL@FER Image Quality Assessment Database", AUTOMATIKA Vol. 53, No. 4, pp. 344354, 2012
[20] K. Ma et al., "Waterloo Exploration Database: New Challenges for Image Quality Assessment Models," in IEEE Transactions on Image Processing, vol. 26, no. 2, pp. 1004-1016, Feb. 2017.
[21] ANSI T1.801.03 "Digital transport of one-way video signals - parameters for objective performance assessment", American National Standards Institute, New York, 1996
[22] D. Hasler, S. Susstrunk, "Measuring colourfulness in natural images", Proc. SPIE Human Vision and Electronic Imaging vol. 5007, Santa Clara, CA, January 21-24, 2003, pp.87-95
[23] M. Buczkowski, "Measuring the effective coverage of the image databases", Measurement Automation Monitoring, vol 63, 2017
[24] M. Buczkowski, R. Stasiski, "Effective coverage as a new metric for image quality assessment databases comparison," International Conference on Systems, Signals and Image Processing (IWSSIP), Poznan, 2017
[25] B. Delaunay, Sur la spheere vide. A la meemoire de Georges Voronoi, Bulletin de lAcademie des Sciences de lURSS. Classe des sciences mathematiques et na, no. 6, pp. 793800, 1934