Key words:
Fax over IP, image quality, image distortion, MOS, Mean Opinion Score, Generalized Linear Models

Andrzej GŁOWACZ[*], Michał GREGA[*], Przemysław GWIAZDA[**],
Lucjan JANOWSKI[*], Mikołaj LESZCZUK[*], Piotr ROMANIAK[*]

OBJECTIVE AND SUBJECTIVE EVALUATION METHODS FOR SELECTED ASPECTS OF FAX IMAGE QUALITY

An image transmitted especially with Fax over IP service can be potentially influenced by many mechanical and electronic distortions. This causes the motivation for image quality estimation using objective rules. This paper presents a novel approach to objective multimodal images comparing. Our idea is to determine the visual image quality measure by combining partial measures from component algorithms operating on specified image features. In presented solution, we apply four such algorithms to successfully analyze image contrast, sharpness, granularity, and noise in relation to the original image. To determine statistical influence of these features on perceived image quality, we have conducted a survey of over 5000 Mean Opinion Scores (MOS). The results of objective measure are mapped on these perceived scores to finally obtain overall MOS of the transmitted image, which further enable fax service quality estimation.

1. INTRODUCTION

The „Fax over IP" service [9][10] allows for transmitting facsimile image over an IP network (like the Internet). As the transmission may be carried using the UDP protocol, no guarantee is given on reception of complete image information. Depending on encoding type – this may result in introducing several image distortions and (in consequence) an imperfect reconstruction of the original image.

The main idea of the system is to elaborate a set of the tools to objective image quality assessment. Mapping between objectively obtained values and the Mean Opinion Score – MOS [7], can be assured by the subjective tests.

Subjective tests of the images quality should be conducted on the diverse and numerous groups of experts. Results of the tests should allow one calculating the MOS, for all

[*] AGH University of Science and Technology, Department of Telecommunications.
[**] Telekomunikacja Polska S.A.

the distorted pictures. In order to assess, how strongly few distorted parameters influence the MOS, each test should include evaluation of a single as well as a multi-distorted images (multi-distorted means that more than one parameter was distorted, e.g. contrast and granularity). The images are preferred to belong to the standardized digitized image set [8].

Methodology for the different methods of the subjective tests is described in [6]. The first method, called Double Stimulus Impairment Scale – DSIS, operates on the five-level impairment grading. The reference image is always shown with the distorted one. Assessment of the images quality refers to the distortion level, not the absolute image quality. The second method is called Double Stimulus Continuous Quality Scale – DSCQS. The picture quality is assessed on a continuous quality scale from excellent to bad. Experts are not informed which picture is the reference one, absolute image quality is assessed.

Correct methodology for the subjective tests can differ, depending on the specific test requirements. In order to fulfill the requirements, existing method can be used directly as well as a composition of a few of the existing methods can be applied.

2. OBJECTIVE EVALUATIONS OF DISTORTION ASPECTS

The authors have developed methods for objective evaluating four distortion aspects: contrast, sharpness, granularity and noise. The methodologies used for evaluation are presented in this section.

2.1. CONTRAST

Each single pixel on the image has its own luminance level ranging from ‘0’ to ‘255’. Value ‘0’ means black and ‘255’ means while color. Histogram is the figure that illustrates how many pixels from the whole picture, have specific luminance level. Value ‘0’ at the left side of the histogram indicates the number of the black pixels, and simultaneously, ‘255’ at he right, the number of the while pixels. Hence, histogram is a distribution of the pixels’ brightness for the entire picture.

When a picture has a pure contrast, on its histogram there is only a narrow range of the pixels’ brightness. With the expanding stretch of this range picture’s contrast improves. Maximum range of the values on the histogram means the best possible contrast.

Usually any losses of the contrast have always a global nature and refer to the whole picture, not only to a particular area. Hence, having calculated histogram, it is easy to compare images’ contrast. A decline in the contrast appears as a narrow of the image histogram.

In order to compare contrast, script uses mentioned above method, but not directly. Time and resources consuming method of the image histogram calculation is replaced

with a simpler solution. Histogram of the reference and the distorted image is being normalized by a suitable method. Afterwards, two pairs of the images (one pair consists of the images and its normalized equivalent) are being compared with the use of the PSNR metric. PSNR metric returns similarity level in the dB scale. Result of the PSNR values subtraction stands for the comparison indicator (subtraction in the dB scale is equal to the division in the linear scale). Applied approach assures insensitivity to any other image distortions.

2.2. SHARPNESS

Sharpness is one of the most significant factors that have an influence on the subjective opinion about the picture. It is closely related to the amount of the details that an image can provide.

Sharpness is defined as a distance between the areas having different tones of colors. Fig. 1 illustrates bar pattern of increasing spatial frequency. The top portion represents reference image with a perfect sharpness, bottom represents the same image but with a distorted sharpness. As it is presented on the Fig. 1, higher spatial frequency means lower image sharpness. Clear transitions between black and white stripes appear only for the lowest spatial frequency (bottom portion). Conclusion is simple – edge detector seems to be a good image sharpness indicator. More edges detected on the images means better image sharpness.

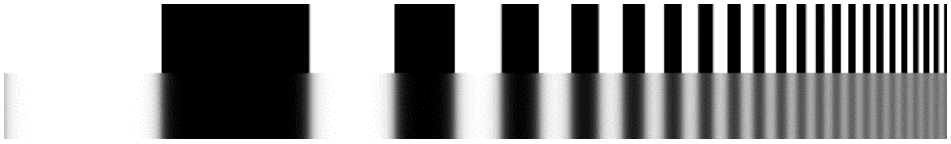

Fig. 1. Bar pattern of increasing spatial frequency (source: [4]).

First step to calculate sharpness is to convert both input images to gray scale. Afterwards, with use of the suitable method all the edges on the images are detected. A filter to detect edges is set on the lowest possible sensitivity, what means that only sharp and clear edges will be detected. Meantime, an auxiliary, entirely black image is created. Next step is to compare both images with the black one, with a use of the PSNR metric (just as mentioned in 2.1, in dB scale). It can be perceives as a absolute edges amount indicator since the images are compared with a solid black one. The result of the subtraction of the PSNR values defines the sharpness comparison value, which is returned as the output of the script. Applied approach assures insensitivity to any other image distortions.

2.3. GRANULARITY

Two images showing the same object, having exactly the same dimensions, can present completely different quality – can provide completely different amount of the details. The reason why it can happen is a fall of the pixels number on the image.

Explaining in other words, the effective size of the single pixel on the image can significantly increase, which will result in the highest granularity of the picture. The type of the distortion being discussed applies to the whole picture, does not have a local nature.

Calculation of the image granularity is performed in a few steps. At the beginning, a number of random points are chosen from the image. Starting from each point, total number of the pixel-changes is calculated for the vertical or horizontal lines (pixel-change appears when at least one of the R, G or B values is different form the previous one). The maximum number of the pixel-changes for all lines is the picture resolution. It is slightly possible, that the real (maximum) value of the image resolution will not be discovered as not the whole image area is being analyzed. However, it is not a problem since the same lines are analyzed on the both images (reference and distorted). As a result of the granularity comparison process, a quotient of the maximum found resolution for reference and distorted image is obtained. Applied approach assures insensitivity to any other image distortions.

2.4. NOISE

Noise distortions can be evaluated using the Hosaka plots [3]. Hosaka plots are objectively calculation comparison metrics, allowing for evaluation of a noise level – *DS*( ) – introduced during image transmission. The noise level is being evaluated in pixel blocks divided into a couple of classes – from 2×2 to 16×16 (see Fig. 2).

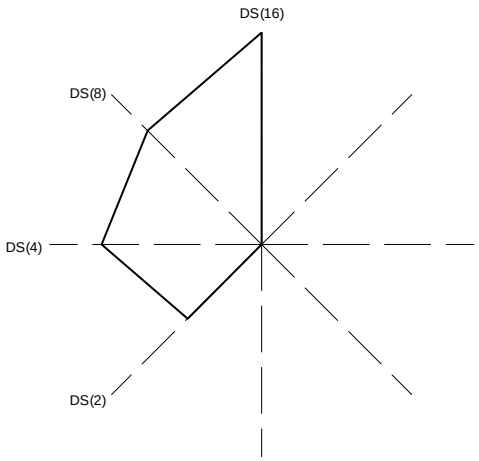

Fig. 2. An example Hosaka plot indicating the noise distortion level

The area of the Hosaka plot is related to the noise level. The shape of the Hosaka plot specifies if the noise distortions are introduced for details (represented by small blocks) or for larger, homogenous areas (large blocks).

3. SUBJECTIVE MEAN OPINION SCORE MAPPINGS

In the research described in this paper a DSIS [6] operating on the five-level impairment grading has been used. A scale described by words is easy to understand for people (we all use words, not numbers, to describe things) but it has no mathematical meaning. The way of mapping the verbal answer on numbers is just a matter of convention. The only mathematical property of the verbal answers is that they are ordinal i.e. “bad” is better than “unaccepted”, but worse than “average”. Since a distance between “bad” and “unaccepted” or “bad” and “average” cannot be found (everyone has their own measure of these differences), verbal answers modeling in the same way as length or speed is a common mistake [1].

In order to model such values properly, a more general model than a simple regression model has to be used [1]. A model used in this paper is a GLZ (Generalized Linear Model) with an ordinal multinomial distribution and a logit link function. An idea of GLZ models is described in [5], where a difference between categorical, categorical ordinal and continuous variables is given.

An important property of the GLZ model used in this paper is that it does not describe MOS directly but a probability that a particular distortion results in a particular answer. Note that more than one answer is possible since people are different and the only thing that can be found is a probability of a particular answer. Knowing the probability of a particular answer makes possible to compute a MOS value.

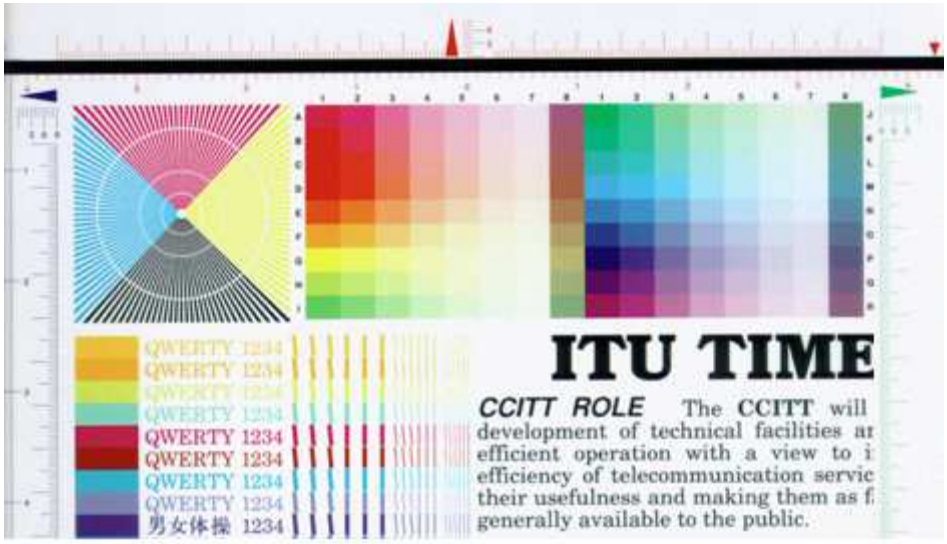

Fig. 3. An example test picture with noise distortion.

An interesting observation made during this research is that a variation of possible answers given for the same picture is very high. The only reason is that different people understand words “bad”, “average” etc differently. For example Fig. 3 presents a picture with noise distortion for which 2 from 15 answers were “unaccepted” and 7 “very good”.

The analysis of data with a high variation has to be repeated and carefully tested. Therefore, the analysis followed an algorithm:

1. Obtained data have been cleaned i.e. if a tester has given an answer a for better picture and b for a worse one and $b-a>1$, all tester’s answers have been removed.

2. Cleaned data have been split up in order to obtain training set and test set.

3. On the basis of Schwarz criterion the best GLZ model has been chosen.

4. The comparison of the obtained GLZ model distribution and test set according to Pearson χ^2 test has been done. If the obtained model has not passed the test, other model has been analyzed or a different function of distortion has been used.

The final model describes a probability as a function of particular distortion. Note that the probability is different for each answer (1, 2, 3, 4 or 5), therefore five different probability functions represent the final result. MOS value is given by

$$MOS(d) = p_1(d) + 2p_2(d) + 3p_3(d) + 4p_4(d) + 5p_5(d), \quad (1)$$

where d is a distortion value.

A different result that can be given by knowing the probability functions is the most probable answer MPA i.e. the answer that is given by more testers than any other answer. Note that extreme values have influence on a mean value, such as MOS

is, and therefore MOS for Fig. 3 equals 4.2 (little more than good), even if 7 of 15 answers were 5 (very good). MPA can be used as an alternative for MOS since it is not under the influence of extreme values. Of course, MPA is not a perfect measure since most users can mean only 21% of them.

MOS (blue line) and MPA (red line) have obtained four different distortions; the values have been presented in Fig. 4.

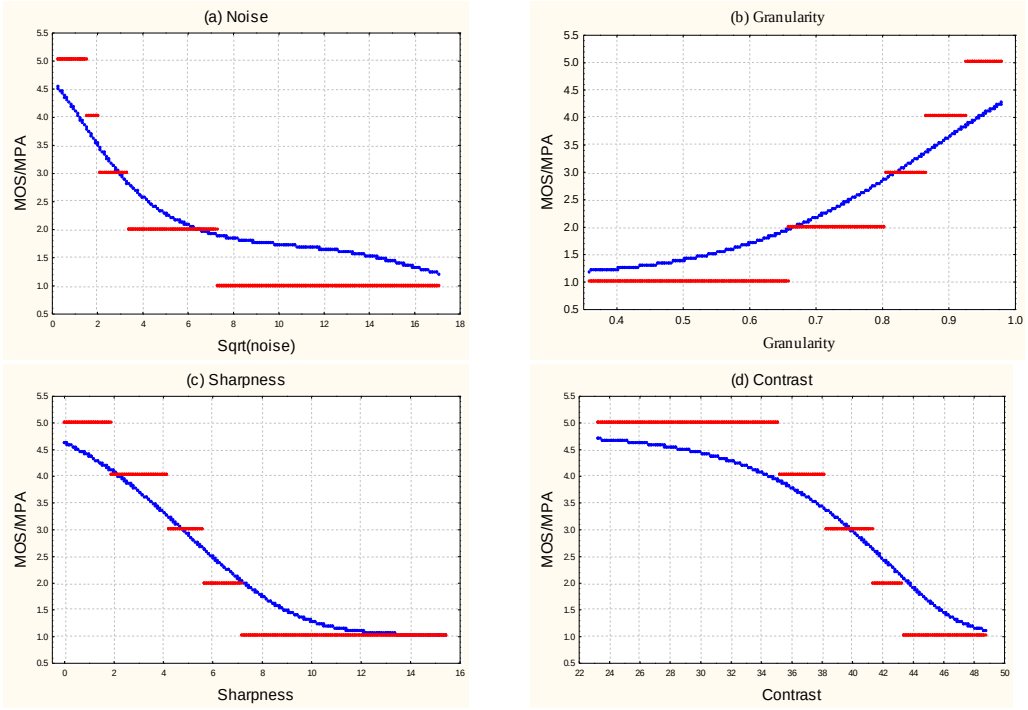

Fig. 4. MOS and MPA obtained for different distortions.

All MOS functions are decreasing functions of increasing distortion (for granularity a higher value of x axe is obtained for less distorted picture). The most interesting is the range for which MPA values are different from 1 or 5, since for this range a picture was distorted strongly enough that most people noticed some differences but still could accept the picture. A deep analysis of MPA function shows that noise and granularity distortion is less linear than these obtained for sharpness and contrast. ”Less linear” means that an interval for which MPA equals 3 is not the same as an interval obtained for MPA equals 2.

Fig. 4.d shows that even for small distortions MOS is less than 5. The reason is that some testers give answer 4 or even 3 for a picture identical to the original one (sic!). Therefore it is almost impossible to obtain MOS=5. One can conclude that some people do not give answer “very good”. On the other hand almost all people

answer “unacceptable” for highly distorted pictures since for almost all presented distortions MOS value for high distortions is 1 or very close to 1.

4. CONCLUSIONS

Driven by a need of automatic image quality assessment a set of image quality analysis methods was designed. Algorithms for objective testing of contrast, sharpness, granularity and noise were developed. Contrast is measured using the histogram analysis and PSNR metric. Sharpness is measured using edge detection. Granularity is assessed by analysis of resolution and noise is measured with use of Hosaka Plot. In order to calibrate the objective assessment method a series of subjective tests was run. The subjective tests were performed according to the ITU-T recommendations. The results of objective were analyzed and a process of model finding was applied with aid of GLZ data analysis method. Finally, a set of software tools was developed, which is capable of evaluating the quality of an image in a scale which is complimentary with human reception.

5. FURTHER WORK

As the further work, methods for evaluating three more image distortion aspects (geometric deformations, colour accuracy and greyscale accuracy) are planed to be worked out. Then, the next step will be development of a software tool for comparing the transmitted (original) and the received (reconstructed) images. The tool will specify distortions at the reconstructed image and map them in the subjective MOS scale. It is also expected for the tool to be able to execute selected tests described in [2].

6. ACKNOWLEDGEMENTS

The work presented in this paper was fully financed by Telekomunikacja Polska S.A.

REFERENCES

[1] AGRESTI A., *Categorical Data Analysis*. Wiley, 2nd edition, 2002.

[2] ETSI ETS 300 242, *Terminal Equipment (TE); Group 3 facsimile equipment*. Sophia Antipolis, ETSI, 1997.

[3] HOSAKA K., *A new picture quality evaluation method*. In: Proc. International Picture Coding Symposium, Tokyo, Japan, 1986, 17-18.
[4] IMATEST LLC, *Imatest – Sharpness*. http://www.imatest.com/docs/sharpness.html.
[5] McCULLAGH P. and NELDER frs J.A., *Generalized Linear Models*. Chapman&Hall, 2nd edition, 1990.
[6] RECOMMENDATION ITU-R BT.500-11, *Methodology for the subjective assessment of the quality of television pictures*. Geneva, Switzerland, 2002.
[7] RECOMMENDATION ITU-T P.800, *Methods for subjective determination of transmission quality*. Geneva, Switzerland, 1996.
[8] RECOMMENDATION ITU-T T.24. *Standardized digitized image set*. Geneva, Switzerland, 1998.
[9] RECOMMENDATION ITU-T T.37, *Procedures for the Transfer of Facsimile Data Via Store-and-forward on the Internet*. Geneva, Switzerland, 1998.
[10]RECOMMENDATION ITU-T T.38, *Procedures for Real-time Group 3 Facsimile Communication over IP Networks*. Geneva, Switzerland, 1999.

METODY OBIEKTYWNEJ I SUBIEKTYWNEJ OCENY WYBRANYCH ASPEKTÓW OBRAZU FAKSOWEGO

Obrazy transmitowane przy użyciu usługi Fax over IP mogą zostać zniekształcone wskutek działania czynników mechanicznych lub elektronicznych. Zapewnienie odpowiedniej jakości transmisji wymaga więc obiektywnej miary jakości obrazów na etapie wdrażania usługi. Artykuł przedstawia nowe podejście do zagadnienia obiektywnej wielokryterialnej oceny jakości obrazu. W przedstawionym rozwiązaniu zastosowano cztery algorytmy składowe dla celów analizy kontrastu, ostrości, ziarnistości oraz szumu. Dla statystycznego oszacowania wpływu tych składowych na postrzeganą jakość obrazu przeprowadzono serię 5000 badań subiektywnej oceny MOS. Obiektywne wyniki algorytmów zostały zmapowane na oceny subiektywne dla oszacowania końcowej oceny całego obrazu, co pozwoli na ocenę jakości usługi Fax over IP.