|  |  |  |
| --- | --- | --- |
| INTERNATIONAL TELECOMMUNICATION UNION | | **STUDY GROUP 9** |
| **TELECOMMUNICATION STANDARDIZATION SECTOR**  STUDY PERIOD 2009-2014 | |  |
| **English only**  **Original: English** |
| **Question(s):** |  | Stockholm 11 July, 2014 |
|  | | |
| **Source:** | Editors | |
| **Title:** | Output – Multimedia Quality of Experience for Target Recognition Applications | |

----------------P.912

Multimedia Quality of Experience for Target Recognition Applications

Summary

It was once thought that high QoS (Quality of Service) performance solves recurrent problems of low-quality multimedia services. Later, one proposed solutions to ensure a high level of QoE (Quality of Experience). In this document, the authors show the path of work that attempt to understand what really is the quality of multimedia services. By starting from scratch QoS, passing through generalized QoE, ultimately the authors focus on aspects of the subjective and objective quality modelling and optimizing the visual performance for target recognition applications (for example, video surveillance), outlining the path of possible ITU-T standardization in this area. The authors also try to predict at least some of today’s errors of reasoning, which probably will be already evident for the industry in the next decade.

Keywords

QoS, QoE, ITU-T, TRV, P.912, CCTV

Introduction

Decade ago, the telecommunications industry was still convinced that the Quality of Service (QoS) techniques that provide high performance resolve any recurrent problems of low-quality multimedia services. However, a few years later, it became clear that optimization of QoS parameters such as throughput, packet loss, delay, or jitter, it is not the best way to improve the quality experienced by users. The problem of low bandwidth can be compensated by more efficient codecs. The impact of packet loss is strongly dependent on their distribution, and the use of redundancy coding and transmission. In the end, too, for many applications, buffering multimedia data streams can alleviate a large delay and jitter.

Then, the most proposals suggested the possibility of prediction, and consequently: prevention of the above problems of quality multimedia services through the use of generic models to provide a high level of Quality of Experience (QoE) [4] [11]. Concepts of special structures (frameworks) for the integrated assessment of the quality of video sequences began to appear [18]. These structures began to be “filled” with solutions that attempt to model the overall quality, operating at the intersection of QoS and QoE areas [17] or only in the area of QoE. But again, it became obvious that such a general approach, possibly assessing the entertainment media, just does not work at many specific visual applications, for example: target recognition (utility) applications (video surveillance, telemedicine / remote diagnostics, fire safety, backup camera, etc.) [12] [22].

In fact, QoE, thus that how one perceives the quality of multimedia services, depends on a number of contextual parameters, not only objective ones, but also subjective ones [19]. Only a full understanding, usually only possible with strong limiting the area of QoE modelling application, gives a chance to obtain results consistent with the expectations of service users, and consequently, enabling the optimization of quality [13]. Unfortunately, a large number of contextual parameters makes this research problem is still open.

# 1 Target Recognition Video for the Purpose of Use

In many visual applications, the quality of motion picture, in the classic sense, is not as important as the ability of the visual system to perform a specific task, for which they are created, given the processed video sequences. Such sequences are called Target Recognition Video(TRV). Regardless of the different understandings of the concept of TRV quality, its verification is necessary to perform a dedicated quality testing. The basic premise of these tests is to find TRV quality limits, for which the task can be performed with the desired probability or accuracy.

Such tests are usually subjective tests (psychophysical experiments) with a group of subjects. Unfortunately, due to the complexity of the issue and the relatively low level of knowledge of human cognitive mechanisms, besides very limited areas of application, satisfactory results of computer modelling of the quality of TRV has not yet been achieved.

Given the use of TRV, a qualitative test is not focused on the subjective subject satisfaction with the quality of the video sequence, but to measure the performance of the subject, making use of TRV, for the accomplishment of the tasks entrusted to the subject. The purpose of this may be, for example:

* For a video surveillance – recognition of license plate numbers
* For telemedicine / remote diagnosis – correct diagnosis
* For fire safety – fire detection
* For backup camera – parking the car

Naturally, we face an imposed, significant influence of the human factor, so it seems obvious to ask questions on the procedures to be complied with in order to make a subjective assessment of the quality of TRV. In particular, the questions arise on:

* Method of selection of source TRV, from which test TRV (with degraded quality) arise
* Subjective testing methods and general manner of conducting the psychophysical experiment
* Method of selecting a group of subjects in the psychophysical experiment
* Instructing and training of subject before the start of the experiment
* Conditions in which the test will be carried out
* Methods of statistical analysis and presentation of results

# 2 Methods for Subjective Evaluation of TRV

Questions articulated in the previous section, are tried to be answered by Recommendation ITU-T P.912 [7]: “Subjective Video Quality Assessment Methods for Recognition Tasks”, published in 2008. In addition, Recommendation P.912 organizes terminology related to subjective TRV testing, introducing appropriate definitions for the methods of testing (psychophysical experiments).

Unfortunately, Recommendation P.912 is only the first step in the standardization of methods of subjective TRV testing. In the opinion of the authors, which is based both on the research results (their own and independent), as well as observation acquired in the course of carrying out numerous experiments with TRV, many claims of Recommendation P.912 is formulated at too high a level of generality. What’s more, the selected statements, not proven by research results, are significantly disputable. In this situation, the authors propose amendments to Recommendation P.912, which detailed scope has been discussed in the following subsections.

**2.1 Source Signal**

In Section 5, Recommendation P.912 says (among others):

*Test sequences should follow the general principles stated in [3] and [6], which specify scenes that should be consistent with the transmission service under test, and should span the full range of spatial and temporal information. It is critical for the nature of these evaluations that the stimuli used actually reflect the true operational parameters of the conditions under which the video material is collected about, and cover the entire range of possible scenarios for the application area identifying that one is.*

Unfortunately, in practice, for some cases, data availability is very limited. Consider the impact of studying quality of images (admittedly still) on the accuracy of X-ray diagnosis of bone fractures. It is obvious that due to the low frequency of certain types of fractures, the availability of the database of corresponding images is very small.

Another example concerns research on the impact of CCTV recordings, the accuracy of license plate recognition [13]. For the purposes of this study, a special video database was created [15]. The recordings have been preparing with the help of the fixed CCTV cameras, recording car entry into the AGH University of Science and Technology in Krakow, Malopolska, Poland (Figure 1). Again, it is obvious that due to the above-mentioned conditions of acquisition, recordings represent a particular CCTV camera, its specific location and direction, a specific distance from the object, as well as specific lighting conditions. What’s more, due to the city in which recordings were made, most of the cars’ license plates that had been filmed have “K” (distinguishing the province of Malopolska) at the first position of the plate, while they have the letter “R” (distinguishing county) in the second position.



Figure 1. Sample frame of a recording from the database [15].

As shown, contrary to Recommendation P.912, it is very difficult to ensure complete coverage of the potential applicability scope for recordings. Any expansion of the record database is either very laborious, time-consuming, or even impossible. This does not mean that the cited studies are useless, but their applicability scope must be explicitly limited to the scope of the database of recordings. Unfortunately, the literature commonly encountered quite unjustified attempt to extrapolate the applicability scope of test results (especially among less experienced researchers), what, according to the authors, may be the effect of the lack of explicit addressing this, it would seem – the obvious, issues in Recommendation P.912, which is often instruction to carry out the tests. Therefore, the authors propose the introduction of appropriate amendments to Section 5 of the Recommendation P.912.

**2.2 Testing Methods**

In Section 6 of Recommendation P.912 we read (e.g.):

*The application of TRV is directly related to the ability of the user that recognizes targets at increasing levels of detail. These levels are referred this as Discrimination Classes (DC). When determining the DC for particular scenarios, they must consider that for a set distance from the camera to the object of interest, the DC directly correlates video is decreasing resolution of the target, and therefore the object is represented by fewer cycles per degree of resolution. Fewer cycles per degree of resolution also means that the object subtends less of the information content of the video, making identification of the target more difficult.*

Unfortunately this subparagraph, in the opinion of the authors, makes it not easy, especially for a novice experimenter, to understand the relationship between the parameters (used partially interchangeably and not quite correctly) such as number of Cycles-Per-Degree (CPD), the resolution of the object and the distance between the camera and the object. In practice, the key parameter is the CPD, on which impact has a resolution of an object, and can be affected by the distance between the camera and the object.

The authors propose changes, mainly editing, involving the ordering relationship between these parameters.

Same Section 6 of Recommendations P.912 defines the multiple-choice method type for testing (Figure 2).



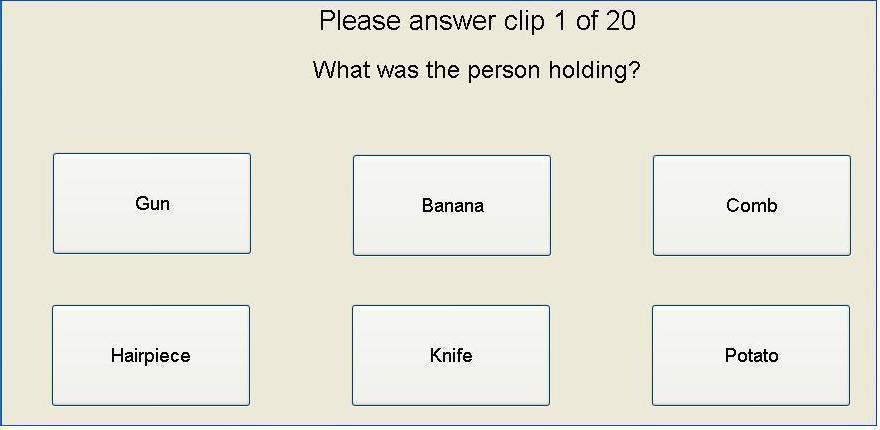


Figure 2. Multiple-choice method [7].

In the description of the multiple-choice method, we can read:

*This method is appropriate for all DC levels and target categories (human, object and alphanumeric). For this method, the video is shown above a letter of verbal labels representing the possible answers. After presenting the video, the viewers must choose the label closest to what they recognized in the clip. The use of fixed multiple choices eliminates any possible ambiguity that could accommodate arise from open questions, and allows for more accurate measurements.*

Unfortunately recommendation says nothing about the potential impact of the order and position of the selection buttons on the preferences of their choice by the subjects. The results of the research (conducted with the authors of this article) have been published [14], which show the existence of the above-mentioned impact (Figure 3).



Figure 3. Selection buttons used in the studies [14].

In this connection, the authors call for the introduction to Section 6 of Recommendation P.912 of an entry proposing a random sequence of buttons.

Next, still in the context of a multiple-choice, in Recommendation P.912 we read:

*The number of choices offered to the viewer will depend on the number of alternative scenes being presented. “Unsure” may be one of the listed choices.*

It should be noted that subjects tend to abuse the “Unsure”. This problem has been observed when applying a Comparison Category Rating (CCR, Table 1), as defined in Recommendation ITU-T P.800 [5], in which method, subjects abuse the response “0” (“About the Same”). A similar trend was observed independently in a TRV study conducted by the authors [14].

Table 1. The scale of the CCR method [5].

|  |  |
| --- | --- |
| 3 | Much Better |
| 2 | Better |
| 1 | Slightly |
| 0 | About the Same |
| -1 | Slightly Worse |
| -2 | Worse |
| -3 | Much Worse |

Unfortunately Recommendation P.912 misses here a clear warning against the prudent use of the “Unsure” (Recommendation P.912 even encourages its use). Therefore, the authors propose the introduction of an appropriate entry into Recommendation P.912.

In Section 6, Recommendation P.912 presents another method of testing – the single choice method (Figure 4).





Figure 4. Single-choice method [7].

About this single choice test method we read the following:

*If there is a non-ambiguous answer is an identification question, the single answer method may be used. This method is appropriate for alphanumeric character recognition scenarios. A viewer is asked what letter(s) or number(s) was present in a specific area of the video, and the answer can be evaluated as either correct or incorrect.*

It should be noted that, contrary to Recommendation P.912, it is possible to also apply fuzzy logic [13]. For scenarios where the result of recognition is an alphanumeric string, such assistance may come by measuring differences between the two strings with the Hamming distance (only for strings of the same length) [21], or its generalization – the Levensthein distance [23] [26]. For example, in practice, to experiment imaged in Figure 4, it can be regarded with success to consider results containing not more than one error as correct ones [13]. It is because even with the wrong result of plate recognition, by correlating it with vehicle database that contains the make and colour of the vehicle, we substantially limit the possibility of ultimate misrecognition.

Consequently, the authors propose to expand the description of the single choice method.

**2.3 Subjects**

In Section 7.3, Recommendation P.912 states:

*Subjects who are experts in the application field of the target video recognition should be used.*

In order to verify this finding, experiments were carried out (of which, the authors, were co-organizers), asking subjects for recognition of objects (mobile phone, torch-flashlight, gun, mug, radio, aluminium soda can, electric stun gun – "Taser"), presented in video sequences. First, the experiment hired subjects who are experts – law enforcement department’s officers [24] [25]. Then the experiment was repeated with non-experts to give very similar results, as long as the work of non-experts has been previously paid [14].

Consequently, the authors call for the introduction to the Recommendation of the entry, which allows the use of non-expert subjects, if they will be motivated in other than professional, manner (for example, will be paid up). Naturally, this will be possible only for certain areas of application testing – because it seems impossible to use of the non-experts in tests associated with (for example) medical diagnostics.

**2.4 Instructing and Training of Subjects**

Recommendation P.912 relates to the issue of instructing and training of testers, or more broadly, any interaction with the testers, mainly in Section 7.4:

*The subject should be given the context of the task before the video clip is played, and told what they are looking for or trying to accomplish. If questions are to be answered about the content of the video, the questions should be posed before the video is shown, so the viewer knows that what the task is.*

This is not however the only place where the subject is touched, and so, a reference to this issue can be found also in Section 6.2:

*Care must also be taken to avoid terminology that may differ from participant to participant.*

Consequently, given the breakdown of this subject for many sections in the Recommendation P.912 document, the authors call for assembling it only in one (dedicated) Section 7.4.

Changes (extensions) also seem to be required in the same descriptions of interactions with the subjects. In particular, it is important to inform the subjects as to how to act in situations of doubt as to the result of recognition (which, moreover, in a properly designed experiment should happen very often). Experience from previous experiments (e.g., [10], [16], license plate recognition, Figure 5) shows that in doubtful cases, some subjects attempt by all means to give the answer (guessing, often correctly), and some other ones – resign from reply.

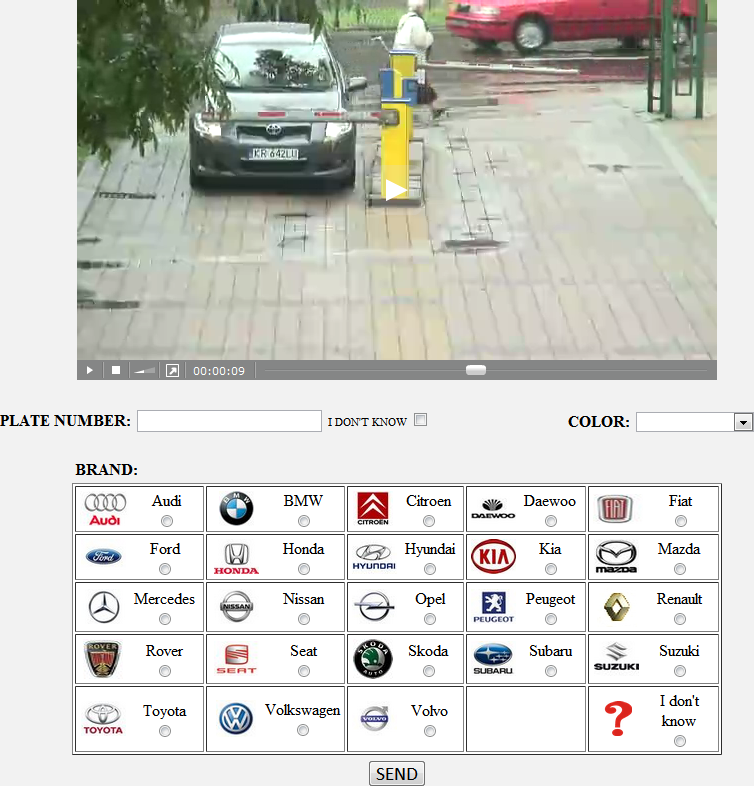


Figure 5. License plate recognition [10] [16].

Without proper instruction and / or training, the default behaviour seems to be an individual feature, so a random selection of testers can make a significant and undesirable ultimate affect in the results obtained.

Naturally, we can not be sure to what extent the better instruction and / or training would alter the results, due to the known observed effect of diversified engagement of subjects into the experiment. However, clear instruction, combined with the training session, containing the video sequence of license plate: difficult but possible to read, would allow subjects to explain that in cases of doubt, they should still try to read the characters.

More information on the errors made by the subjects in this, and other related experiments sensing target recognition tasks (utility applications), is included in [9].

**2.4 Conditions for Testing**

In the range of conditions under which testing subjective (psychophysical experiment) is performed, Recommendation P.912 refers (several times: Section 5, Section 6, Section 6.6, Section 6.7, Section 7.1, Section 7.2 and Section 7.3) to the ITU-T Recommendation P.910 [6]:

*The Experimenter should follow the guidelines outlined in [6].*

Recommendation P.910 is from 1998 and at the time of approval Recommendation P.912 (reminder: 2008) was probably the most recent Recommendation on testing conditions to which one could then refer. As a result, the vast majority of subjective tests (psychophysical experiments) were performed previously under strictly controlled conditions (Figure 6), defined in Recommendation P.910.



Figure 6. Strictly controlled test conditions [14].

By 2014; however, ITU-T Recommendation P.913 was approved [8], which largely displaces Recommendation P.910, including defining a much smoother (the easier to meet) requirements for the testing. By that way of work on the upgrade of Recommendation P.912, the authors call for the introduction of references to Recommendation P.913, replacing references to Recommendation P.910.

**2.5 Statistical Analysis and Reporting**

In Section 8, with respect to the single choice method, Recommendation P.912 states:

*For single answer conditions, where the answers are correct or incorrect, a statistical metric to determine if the subject is performing above the level of chance for answering correctly should be implemented. “Unsure” answers should be pooled with the incorrect answers.*

At the same Section 8, in relation to the multiple-choice method, it is written:

*For multiple-choice answers, the probability of an incorrect answer needs to be balanced against the ability to answer the questions correctly. The statistic metric in this situation will require an examination of the stability of the answers within and between subject performance metrics. “Unsure” answers should be pooled with the incorrect answers.*

Statements definitely do not take into account the possibility of using more sophisticated statistical techniques. In the study [13], the authors are showing the possibility of using a logit function of statistical analysis of results. In other studies, they are showing also the possibility of using a Generalized Linear Model (GLZ) [2]. In the end, too, despite the existence of already published proposals [9] – there is no discussion about removing outlier’s responses from the pool of results, what is standard procedure in the case of other studies in the field of subjective QoE.

# 2 Summary – Standardization

The presented discussion of the statements contained in Recommendation P.912, shows that some of the findings and the observations, require verification of certain provisions of the Recommendation. Therefore, the authors have taken steps to introduce significant modifications (amendments) Recommendations. For this purpose they established working cooperation with the U.S. Agency of National Telecommunications and Information Administration (NTIA, whence comes the original version of the Recommendation), and, in order to formalize the procedures, with the Polish Ministry of Administration and Digitization. The authors have already received from the Ministry formal nomination as delegates of Polish government. The procedure for submitting amendments is likely to start even in 2014.

Ultimately, the amendment recommendations are to have a broader scope and, for example, have the extended applicability of the techniques in the field of crowdsourcing for the subjective assessment of the quality of TRV. In cooperation with NTIA, they are also plans to introduce (to the Recommendation) metrics of Video Acuity, created at NASA Vision Group [1].

Acknowledgements

The Polish National Centre for Research and Development funds this work, Contract No. C2013/1-5/MITSU/2/2014 under the international EUREKA programme: MITSU – Next Generation Multimedia Efficient, Scalable and Robust Delivery.

Bibliography

1. AB Watson, "Video Acuity: A Metric is a Quantify the Effective Performance of Video Systems", *Optical Society of America, Imaging Systems Applications Topical Meeting*, 2011
2. A. Głowacz, M. Grega, P. Gwiazda, L. Janowski, M. Leszczuk, P. Romaniak, "Automated qualitative assessment of multi-modal distortions in digital images based on GLZ", *Annals of Telecommunications-Annales des Télécommunications*, vol 65, p 3-17, 2010
3. ANSI *T1.801.01-1995: Digital Transport of Video Teleconferencing / Video Telephony Signals – Video Test Scenes for Subjective and Objective Performance Assessment*, 1995
4. E. Cerquiera, S. Zeadally, M. Leszczuk, M. Curado, A. Mauthe, "Recent advances in multimedia networking", *Multimedia Tools and Applications*, vol 54, pp. 635-647, 2011
5. ITU-T *P.800: Methods for Subjective determination of transmission quality*, Geneva 1998
6. ITU-T *P.910: Subjective video quality assessment methods for multimedia applications*, Geneva 2008
7. ITU-T *P.912: Subjective video quality assessment methods for recognition tasks*, Geneva 2008
8. ITU -T *P.913: Methods for the Subjective assessment of video quality, audio quality and audiovisual quality of Internet video and television distribution quality in any environment*, Geneva 2014
9. L. Janowski, "Task-based validation subject: more reliability metrics", *Fourth International Workshop Quality of Multimedia Experience*(QoMEX), 2012
10. L. Janowski, P. Kozlowski, R. Baran, P. Romaniak, A. Głowacz, T. Rusc, "Quality assessment for a visual and automatic license plate recognition", *Multimedia Tools and Applications*, vol 68, pp. 23-40, 2014
11. M. Grega, L. Janowski, M. Leszczuk, P. Romaniak, Z. Papir, "Estimation of the perceived quality of service (QoE) for multimedia communication", *Review of Telecommunications, Telecommunication News*, vol 81, p 142-153, 2008
12. M. Leszczuk, "Image quality for utility applications: definitions, testing, standardization and current trends", *Review of Telecommunications, Communications News*, vol 83, pp. 242-246, 2010
13. M. Leszczuk, "Optimizing task-based video quality: a journey from psychophysical experiments Subjective quality is objective optimization", *Multimedia Tools and Applications*, vol 68, pp. 41-58, 2014
14. M. Leszczuk, A. Koń, J. Dumke, L. Janowski, "Redefining the ITU-T recommendation P.912 for requirements of the subjects of quality assessments in recognition tasks", *Multimedia Communications, Services and Security, 5th International Conference* (MCSS), pp. 188-199, 2012
15. M. Leszczuk, L. Janowski, "Database for video quality assessment in license plate *recognition*", *Signal Processing: Algorithms, Architectures, Arrangements, and Applications* (SPA)*,* pp. 51-55, 2013
16. M. Leszczuk, L. Janowski, P. Romaniak, A. Głowacz, R. Mirek, "Quality assessment for a license plate recognition task based on a video streamed in limited networking conditions", *Multimedia Communications, Services and Security* (MCSS),pp. 10-18, 2011.
17. M. Leszczuk, L. Janowski, P. Romaniak, Z. Papir, "Assessing quality of experience for high-definition video streaming packet loss under diverse patterns", *Signal Processing-Image Communication*, vol 28, pp. 903-916, 2013
18. M. Mu, P. Romaniak, A. Mauthe, M. Leszczuk, L. Janowski, E. Cerquiera, "Framework for the integrated video quality assessment", *Multimedia Tools and Applications*, vol 61, pp. 787-817, 2012
19. P. Le Callet, S. Moeller, A. Perkis (ed.), *QUALINET White Paper on QoE Definitions*, European Network on QoE in Multimedia Systems and Services (COST Action IC 1003), Dagstuhl 2012
20. P. Romaniak, L. Janowski, M. Leszczuk, Z. Papir, "Perceptual quality assessment for H.264/AVC compression", *IEEE Consumer Communications and Networking Conference* (CCNC), pp. 597-602, 2012
21. RW Hamming, "Error detecting and error correcting codes", *Bell System Technical Journal*, vol 29, p 147-160, 1950
22. S. Moeller, A. Raake (ed.), *Quality of Experience: Advanced Concepts, Applications and Methods*, Springer, Berlin 2013
23. VI Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals", *Soviet Physics Doklady*, vol 10, pp. 707-710, 1966
24. VQiPS, "Public Safety Communications DHS-TR-PSC-10-09: Video quality tests for object recognition applications", *U.S. Department of Homeland Security's Office for Interoperability and Compatibility*, 2010
25. VQiPS, "Public Safety Communications DHS-TR-PSC-11-01: Recorded and video quality tests for object recognition tasks ", *U.S. Department of Homeland Security's Office for Interoperability and Compatibility*, 2011
26. В. И. Левенштейн, "Двоичные коды с исправлением выпадений, вставок и замещений символов", *Доклады Академий Наук СCCP*, vol 163, p. 845, 1965

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_