Towards Objective Quality Assessment in Digital Collections

Alexander Schindler
Intelligent Vision Systems
Department Safety & Security
Austrian Institute of Technology
alexander.schindler@ait.ac.at

Reinhold Huber-Mörk
Intelligent Vision Systems
Department Safety & Security
Austrian Institute of Technology
reinhold.huber-moerk@ait.ac.at

ABSTRACT

Digital preservation makes high claims concerning the quality of digital artifacts. How to reach or maintain such high quality is currently subject to ongoing research. How to assure this quality is missing sufficient attention. On the one hand, many of the commonly applied quality metrics do not provide an accurate interpretation of information loss and distortions, on the other hand initial quality assessment of digitized collections is in many parts an obstacle yet unsolved. We argue, that perceptual quality metrics should play an essential part in quality assurance of either initial digitization, migration or curation workflows. By identifying tasks critical to subjective quality estimation we outline relevant open research areas in different domains.

Keywords

Digital Preservation, Digital Archives, Digitization, Quality Assurance, Perceptual Quality Estimation, No-Reference Quality Metrics

1. MOTIVATION

Digitization - as one of many techniques in use by the conservation communities - has the goal to fully capture the information of the targeted object [1]. This vaguely defined goal is commonly interpreted from two different points of view - *digitization for public access* and *digitization for preservation* [2]. Quality constraints for providing broad public access to library materials are perceived less severe than those for preservable digital surrogates of real objects. Digitized art-works, document images, audio or video files are made accessible in reduced resolution and converted to compressed media formats. Quality constraints have to conform with bandwidth and copyright limitations and important criteria such as readability of document images or intelligibility of spoken content are treated with less severity. Creating digital artifacts constituting enough value to pursue a digital preservation strategy, adheres more complex quality requirements. Typically, initial acquisitions - also called master files - are of higher quality and consequently claim more storage space. Lossy compression is a contemplated option with a reported reduction of up to 98% for digital image archives [3]. Reducing quality in favor of storage space or bandwidth limitations requires the migration of the master files to different file formats. Quality assurance (QA) being an essential tool for estimating the loss of information during the process concentrates on maintaining the quality after migration or curation actions [4, 5] to verify that the migrated collection adheres to certain quality standards. Quality metrics commonly used or suggested by best practices [6, 7, 8] currently are appropriate to capture deviations in different versions after lossless transformations (e.g. migrating audio documents from mp3 to wav format [9]). Compressing the digital content with an accepted loss of information demands for a subjective interpretation of which amount of degradation is still acceptable. Such relation between low-level quality descriptors and compression artifacts is not available in the context of digital preservation. Acquiring such information by means of user evaluations is required to create significant quality models capable of providing objective quality estimates about subjectively motivated quality attributes such as annoyance of distortions or readability.

Such evaluations should include or focus on already perceptually motivated quality or similarity measures. For instance, Structural Similarity (SSIM) [10] was used to successfully identify near-duplicates within document image collections [11, 12, 13]. Perceptual similarity estimations are more robust against distortions originating from lossy compression. Research towards objective quality measures in digital preservation has to include the adjustment of such similarity measures to evaluations of subjective interpretations of the digital artifact's quality.

Assessing the initial quality of digitized objects or collections lacks the amenity of having reference objects to calculate relative quality estimations. No-Reference quality assessment (also known as blind- or non-intrusive quality assessment) tries to define objective estimates describing distortions of audible or visual stimuli that correlate to subjective mean opinion scores (MOS). Applying the described user-evaluated subjective quality attributes to blind-quality estimates would provide an invaluable objective measure for assessing the quality of newly-digitized objects.

2. STATE OF THE ART

A set of problems urgent to the document image analysis and retrieval domain is described in [14]. Several factors affecting the quality of the scanned images are categorized and a summary of pre/post-processing steps to enhance the quality is provided. A set of definitions and two models concerning document image quality and degradation are provided in [15]. These models mainly focus on the ability to post-process the scanned content (e.g. in terms of OCR) and are often not generally applicable. A heuristic measure for detecting undesired influences of lossy JPEG 2000 compression on OCR performance is proposed in [5]. A good summary of quality problems, types and causes including suggestions towards appropriate quality assurance methods is provided in [6]. The *IMProving ACcess to Text (IMPACT)* project [16] focused on the development of new approaches to the extraction of text content from historical documents. Thrity-seven characteristics which can affect OCR performance were identified, including bleed-through, stains, page curl, broken characters, low contrast, skew, presence of watermarks. A summary of document digitization from a digital preservation perspective is provided by [2]. Measuring the quality of digitization of newspaper archives is outlined in [17]. It was also demonstrated how to use the *Matchbox Toolset* to categorize defects of document image collections [18, 19]. Different approaches to no-reference/blind image quality assessment have been reported [20] many of them requiring concrete knowlegde of the distortion generating process (e.g. JPEG compression artifacts).
During the *Scalable Preservation Environments (SCAPE)* project a set of tools has been developed to automatically assess the quality of migrated digital collections. *Jpylyzer* [21] can be used to validate that images that have been migrated to the JPEG 2000 format also strictly conform to the JPEG 2000 Part 1 (ISO/IEC 15444-1) specification [22]. The *Matchbox Toolset* [11, 12] can be used to assess the quality within document image collections (e.g. to detect duplicates [13]) or between different collections (e.g. after migration actions). It uses the perceptually motivated Structural Similarity (SSIM) [10] measure to estimate quality deviations between different versions of images.

Summaries and guidelines of audio document preservation, including methodology and software tools are provided in [23, 24]. Good summaries on audio quality assessment techniques including prediction of perceptual quality are summarized in [25, 26]. Particularly two perceptually motivated audio quality estimation models are outlined: the Perceptual Evaluation of Audio Quality (PEAQ) [27] and its predecessor the Perceptual Evaluation of Speech Quality (PESQ) [28] - both audio quality standards BS.1387 and P.862 are provided by the International Telecommunications Union (ITU). Recently the Structural Similarity (SSIM) measure for perceptually motivated comparison of still images [10] has been adapted to estimate audio quality [29, 30]. Extraction of quality parameters from audio fingerprints as a full-reference similarity estimation was suggested in [31]. Such approach seems beneficial for institutions having audio fingerprinting already applied for indexing purposes. As part of the SCAPE project the `xcorrSound` tool package [9] was created for automated audio quality assurance in the context of digital preservation. Blind or no-reference perceptual audio quality measures have not been reported yet.

In the Music Information Retrieval (MIR) domain currently no perceptual quality metrics are reported in literature. A first approach towards investigating the effects of audio degradation on music classification and retrieval is presented in [32].

3. RESEARCH QUESTIONS

In this section we summarize current open research questions concerning quality assurance of digital collections.

3.1 Perceptual Quality

Though it is obviously easy for human assessors to judge the quality of a digital artifact, it is a challenging task to describe objective measures that can be used to automatize quality inspection of digital collections. In textual document images quality is often tightly related to post processing capabilities - especially concerning OCR applicability [5, 16, 15]. Such approaches assert proper performance concerning indexing and retrievability but might affect the overall perceived image quality - especially of mixed content document image pages (e.g. antique books with depictions or pictures). Further problems are: lack of robustness against multilingual or handwritten text, uncommon typefaces and various kinds of image degradation; a main focus on textual content - neglecting multi-modal content; the effects of OCR performance enhancing image pre-processing steps (e.g. noise reduction, contrast enhancement) on the quality non-textual content (e.g. drawings, pictures) has yet not been evaluated. More accurate - the correlations between high OCR performance rates and subjective image quality have yet not been evaluated.
Migrating digital collections to lossy file formats renders commonly used quality estimates (e.g. Peak Signal-to-Noise Ratio (PSNR)) insignificant. It was reported that PSNR poorly correlates with subjective quality and is an unreliable method for assessing image quality [33]. Recent approaches towards perceptually motivated quality estimators and similarity measures (e.g. Structural Similarity (SSIM) [10]) might be more appropriate to objectively assess subjective quality attributes of migrated collections. Concerning current efforts towards migrating digital image collections to the less resource intensive JPEG2000 format [34], approaches using natural scene statistics [35, 36, 37] should be considered and evaluated for their applicability to digital preservation. Especially no-reference image quality estimates [20, 35, 36, 37] have to be analyzed for correlations with subjective opinions concerning the quality of digitized documents. Preferring perceptual quality metrics over common image similarity measures such as Absolute Error (AE) or Mean Squared Error (MSE) was also proposed recently in [8, 38].

Quality assessment for music and audio in general is more complex and less matured. The low number of available audio file formats provides relative stability concerning their probability of deprecation. Consequently, preservation actions focus on identity verification which can be analyzed through already reported approaches (e.g. `xcorrSound` [9]). Automatically evaluating the quality of digital audio collections after migration requires identifying the amount of lost information. Due to semantic discrepancies between low-level audio descriptors (e.g. Mean Squared Error (MSE)) and perceived quality degradation (e.g. unnoticeable, an-

noying), similarity measures estimating the perceived quality migrated audio files are required. Declared standards (e.g. PEAQ [27]) should be considered and evaluated. Recently approaches analyzing the structural similarity from audio files [29, 30] - analogous to the structural similarity of images [10] seem highly promising for assessing the migration quality in digital preservation. Though, a significant interpretation of their estimation is yet missing. Typically a respond value of 1 corresponds to identity, but how is a similarity value of 0.9 perceived? Will this be an unnoticeable difference or already an annoying degradation? The applicability of these approaches to the specific domain of digital preservation has to be evaluated thoroughly.

3.2 Perceptions of Quality

In this section we give examples of quality criteria that are highly subjective and require perceptually motivated algorithms for their estimation.

3.2.1 Readability / Intelligibility

OCR based quality estimation [5, 16, 15] uses text-based distance or similarity measures to compare text extracted from document images with manually annotated ground truth. Such measures only describe deviations between two instances but give no semantic interpretation of their values. Conclusions about readability or completeness of the digitized content cannot be drawn. Chunks might be missing or even parts of a page could be cut off. Further analysis of the content - often requiring concrete domain knowledge - might be necessary.
Accordingly, the quality of spoken content depends on its intelligibility. To which extend PESQ [28] aligns to digital preservation needs has yet to be evaluated.

3.2.2 Detectability

To which notion are distortions within digital documents detectable. A severity scale needs to be defined to classify corruptions occurring during acquisition or migration. Existing quality metrics have to be evaluated how they align to such severity definitions and if they are robust against certain levels.

3.3 Assessing the Initial Quality

The main focus of quality assurance in digital preservation is on migration actions. Based on the premise that the existing quality has to be remained after migration, quality estimates are based on full-reference similarity or quality measures.

3.3.1 No-Reference Quality Assessment

Initial quality assessments are lacking references to be compared to. Thus, no-reference or blind quality assurance algorithm should be considered [20, 35, 36, 37]. Though there is little substantive work on this topic yet - and in some modalities no work at all has been reported in literature - a no-reference approach seems most promising to solve initial quality assurance tasks.
Most of the work on no-reference quality estimation so far reported relies on the use of prior knowledge of quality degrading processes (e.g. permanent stretching of the polyester of magnetic tapes causing noticeable pitch dropping). Overviews of problems occurring during digitization of document image collections are provided in [15, 17]. Such processes have to be identified and analyzed in order to define proper models that can be used to formulate adequate no-reference quality estimates for digital collections.

3.3.2 Completeness

Currently it is not possible to automatically assess the completeness of a digital object or collection. For instance, displacements of original documents during the digitization process may result in cropped images. Without a reference image the loss of valuable information is hardly assessable. If sequential post-processing (e.g. layout detection) confirms the validity of the digitized image, it is not possible to determine which part of the content is absent and to which notion the missing information degrades the overall quality of the document. On a broader scale, detecting missing pages in document image collection is still not generally solved.

3.3.3 Duplicates

Solutions concerning the detection of duplicates in document image collections are described in [4, 11, 12, 13]. The approach described is based on perceptually motivated similarity estimates to detect duplicated content in near duplicate images. Similar approaches are required for the curation of audio archives.

3.4 Sacrificing Quality

When opting for reduced quality in favor of reducing storage space and costs, primitive quality estimators (e.g. PSNR, MSR) give no conclusive description of the resulting quality. Also the `xcorrSound` tool package [9] is inappropriate in such cases. Although being developed for audio quality assurance in migration actions, it focuses on mp3 to wav conversion to verify identical content of both versions. Perceptual similarity measures (e.g. PEAQ [27]) are required to estimate the degradation of sound quality related to the amount of annoyance to the listener. The applicability of such measures proposed by literature to digital preservation scenarios has yet not been evaluated. Especially music is lacking proper definitions of quality issues and degradation models as well as attempts towards comprehensive objective perceptual quality metrics.

3.4.1 Adaptive Resolutions / Sampling Rates

A common approach to compression in digital collections is to apply the same configuration to all items in the collection. It has been demonstrated that the perceived quality of compressed audio depends on the dynamic range of the encoded track [39]. Similar observations have been reported for document images [3]. Taking this into account the rate of degradation of the original artifacts has to be estimated in advance.

3.5 Acceptable for Preservation

The quality metrics so far mentioned describe intrinsic properties of digitized objects. The most pressing question is: is this digital object worth of being preserved for long time? This question includes the previous quality characteristics - is it perceived as good, is it readable, are distortions acceptable or annoying?

4. CONCLUSION AND OUTLOOK

By elaborating the state-of-the-art of quality assurance in digital preservation, we outlined some shortcomings concerning the objective quality estimation of digital artifacts. We argue that for processes that alter the content of a digital item (e.g. lossy compression, noise reduction) conventionally chosen quality measures (e.g. PSNR) are insignificant and perceptually motivated quality estimates are indispensable. To apply such measures reported in literature to the domain of digital preservation, they have to be thoroughly evaluated according their relative compliance to subjectively interpreted quality attributes.

5. REFERENCES

[1] P. Conway, "Preservation in the age of google: Digitization, digital preservation, and dilemmas," *Preservation*, vol. 80, no. 1, 2010.

[2] A. Kavčič-Čolić, "Approaching digitisation through a digital preservation perspective," in *VIII SEEDI Conf.*, (Ljubljana, Slovenia), pp. 93–103, May 2012.

[3] R. Gillesse, J. Rog, and A. Verheusen, "Alternative File Formats for Storing Master Images of Digitisation Projects," tech. rep., National Library of the Netherlands, Mar. 2008.

[4] S. Schlarb, P. Cliff, P. May, W. Palmer, M. Hahn, R. Huber-Moerk, A. Schindler, R. Schmidt, and J. van der Knijff, "Quality assured image file format migration in large digital object repositories," in *Proc. 10th Int. Conf. Digital Preservation (IPres2013) to appear*, (Lisbon, Portugal), September 2-5 2013.

[5] S. Schlarb and C. Neudecker, "A heuristic measure for detecting influence of lossy jp2 compression on optical character recognition in the absence of ground truth," in *Archiving 2012*, (Copenhagen, Denmark), June 12-15 2012.

[6] X. Lin, "Quality assurance in high volume document digitization: A survey," *Document Image Analysis for Libraries, Int. Workshop on*, vol. 0, pp. 312–319, 2006.

[7] M. Casey and B. Gordon, *Sound directions: Best practices for audio preservation.* Indiana University, 2007.

[8] R. Buckley, "Using lossy jpeg 2000 compression for archival master files," tech. rep., Office of Strategic Initiatives, Library of Congress, March 2013.

[9] B. A. Jurik and J. A. S. Nielsen, "Audio quality assurance: An application of cross correlation," in *Proc. 9th Int. Conf. Digital Preservation (IPres2012)*, (Toronto, Canada), Oct 1-5 2012.

[10] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *Image Processing, IEEE Transactions on*, vol. 13, no. 4, pp. 600–612, 2004.

[11] R. Graf, R. Huber-Moerk, and A. Schindler, "An expert system for quality assurance of document image collections," in *Proc. Int. Conf. Cultural Heritage (EuroMed2012)*, Lecture Notes in Computer Science, (Lemesos, Cyprus), Springer, Oct 29 - Nov 3 2012.

[12] R. Huber-Moerk and A. Schindler, "Quality assurance for document image collections in digital preservation," in *Proc. 14th Int. Conf. Advanced Concepts for Intelligent Vision Systems (ACIVS 2012)*, Lecture Notes in Computer Science, (Brno, Czech Republic), Springer, Sept 4-7 2012.

[13] R. Huber-Moerk, A. Schindler, and S. Schlarb, "Duplicate detection for quality assurance of document image collections," in *Proc. 9th Int. Conf. Digital Preservation (IPres2012)*, (Toronto, Canada), Oct 1-5 2012.

[14] H. S. Baird, "Difficult and urgent open problems in document image analysis for libraries," in *Proc. 1st Int. Workshop on Document Image Analysis for Libraries (DIAL'04)*, (Washington, DC, USA), pp. 25–, 2004.

[15] H. S. Baird, "The state of the art of document image degradation modelling," in *Digital Document Processing*, pp. 261–279, Springer, 2007.

[16] H. Balk and A. Conteh, "Impact: centre of competence in text digitisation," in *Proc. 2011 Workshop on Historical Document Imaging and Processing*, HIP '11, (New York, NY, USA), pp. 155–160, ACM, 2011.

[17] S. Tanner, T. Muñoz, and P. H. Ros, "Measuring mass text digitization quality and usefulness," *D-Lib Magazine*, vol. 15, no. 7/8, pp. 1082–9873, 2009.

[18] R. Huber-Moerk and A. Schindler, "Automatic classification of defect page content in scanned document collections," in *Proc. 8th Int. Symposium on Image and Signal Processing and Analysis (ISPA 2013) to appear*, (Trieste, Italy), Sept 4-6 2013.

[19] R. Huber-Moerk and A. Schindler, "A keypoint based approach for content characterization in document collections," in *Proc. 9th Int. Symposium on Visual Computing (ISVC'13) to appear*, (Rethymnon, Crete, Greece), July 29-31 2013.

[20] A. K. Moorthy and A. C. Bovik, "Blind image quality assessment: From natural scene statistics to perceptual quality," *Image Processing, IEEE Transactions on*, vol. 20, no. 12, pp. 3350–3364, 2011.

[21] D. Tarrant and J. V. D. Knijff, "Jpylyzer: Analysing jp2000 files with a community supported tool," in *Proc. 9th Int. Conf. Digital Preservation (IPres2012)*, (Toronto, Canada), October 1-5 2012.

[22] "JPEG 2000 image coding system: Core coding system," 2004.

[23] K. Bradley, I. A. of Sound, and A. A. T. Committee, *Guidelines on the Production and Preservation of Digital Audio Objects: Standards, Recommended Practices and Strategies*. Aarhus, Denmark: International Association of Sound and Audiovisual Archives, second ed., 2009.

[24] F. Bressan and S. Canazza, "A systemic approach to the preservation of audio documents: Methodology and software tools," *Journal of Electrical and Computer Engineering*, vol. 2013, 2013.

[25] D. Campbell, E. Jones, and M. Glavin, "Audio quality assessment techniques: A review, and recent developments," *Signal Processing*, vol. 89, no. 8, pp. 1489–1500, 2009.

[26] K. Seshadrinathan and A. C. Bovik, "Automatic prediction of perceptual quality of multimedia signalsŮa survey," *Multimedia Tools and Applications*, vol. 51, no. 1, pp. 163–186, 2011.

[27] R. Itu, "Method for objective measurements of perceived audio quality," in *ITU-R Recommendation BS.1387*, (International Telecommunications Union, Geneva), 1998.

[28] J. G. Beerends, A. P. Hekstra, A. W. Rix, and M. P. Hollier, "Perceptual evaluation of speech quality (pesq) the new itu standard for end-to-end speech quality assessment part ii: psychoacoustic model," *Journal of the Audio Engineering Society*, vol. 50, no. 10, pp. 765–778, 2002.

[29] S. Kandadai, J. Hardin, and C. Creusere, "Audio quality assessment using the mean structural similarity measure," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pp. 221–224, 2008.

[30] J. P. Bello, "Measuring structural similarity in music," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 7, pp. 2013–2025, 2011.

[31] P. J. O. Doets and R. L. Lagendijk, "Extracting quality parameters for compressed audio from fingerprints.," in *ISMIR*, pp. 498–503, Citeseer, 2005.

[32] M. Mauch and S. Ewert, "The audio degradation toolbox and its application to robustness evaluation," in *Proc. 14th Int. Soc. for Music Information Retrieval Conf. (ISMIR2013) to appear*, (Curitiba, PR, Brazil), Nov 4-8 2013.

[33] Q. Huynh-Thu and M. Ghanbari, "Scope of validity of psnr in image/video quality assessment," *Electronics Letters*, vol. 44, no. 13, pp. 800–801, 2008.

[34] H. Kulovits, A. Rauber, M. Brantl, A. Schoger, T. Beinert, and A. Kugler, "From tiff to jpeg2000? preservation planning at the bavarian state library using a collection of digitized 16th century printings," *D-Lib Magazine*, vol. 15, no. 11/12, 2009.

[35] Z. Wang, H. R. Sheikh, and A. C. Bovik, "No-reference perceptual quality assessment of jpeg compressed images," in *Proc. Int. Conf. Image Processing*, vol. 1, pp. I–477, IEEE, 2002.

[36] H. R. Sheikh, A. C. Bovik, and L. Cormack, "No-reference quality assessment using natural scene statistics: Jpeg2000," *Image Processing, IEEE Transactions on*, vol. 14, no. 11, pp. 1918–1927, 2005.

[37] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind image quality assessment: A natural scene statistics approach in the dct domain," *Image Processing, IEEE Transactions on*, vol. 21, no. 8, pp. 3339–3352, 2012.

[38] S. Bauer and C. Becker, "Automated preservation: The case of digital raw photographs," in *Proc. 13th Int. Conf. Asia-Pacific Digital Libraries (ICADL 2011)*, (Beijing, China), p. 39, 2011.

[39] E. Ruzanski, "Effects of mp3 encoding on the sounds of music," *Potentials, IEEE*, vol. 25, no. 2, pp. 43–45, 2006.