Automated Quality Assurance for Migration of Born-digital Images

Artur Kulmukhametov and Markus Plangg, Vienna University of Technology, Austria; Christoph Becker, University of Toronto, Canada and Vienna University of Technology, Austria

Abstract

Migration to standardized formats is a common approach for the preservation of digital objects. To ensure the authenticity of the resulting artefacts and the validity of the migration, quality assurance is essential. For large-scale migration, automated quality assurance processes are an essential prerequisite. This paper focuses on the migration processes of born-digital photographs. We describe the particular requirements for successful automation of quality assurance. A key aspect of this is the authenticity of the image, the fidelity of the rendering as it appears to an expert viewer. Automation requires us to substitute the human expert viewer with a software algorithm. The key question is whether existing image comparison mechanisms can be applied. To address it, we introduce a publicly available automated workflow relying on perceptual quality assurance measures and present an experiment testing the correlation of the automated measures to human perception.

Introduction

In ensuring the longevity of digital objects such as images, format conversion plays a central role. By replacing the representation of a digital object according to the needs of a user group, in particular by migrating to well-known and standardized formats, risks can be reduced and costs saved. However, the replacement of the bit stream means that no migration action can be trusted without a full quality assurance mechanism that verifies the authenticity of the resulting object.

Born-digital photographs are increasingly stored in raw formats, as storage size limitations are less of an issue and image quality remains the best achievable out of cameras. A raw format means that there is no image representation that a user can access directly and see the photo. Instead, the data from a camera's sensor is input to a data processing pipeline of considerable complexity that maps the sensor data into a final image when the photograph needs to be shown. Migration of raw images is common and supported by existing tools like Photoshop[1] , CaptureOne[2], Adobe DNG Converter[3] and DigiKam[4] that are widely used for this operation. However, the resulting representation of the photograph is by far not guaranteed to render identically to the original. How do we know if a migration tool performed the conversion correctly according to given criteria? Quality Assurance (QA) answers such questions by measuring properties of interest of objects and calculating similarities between the objects. QA is essential for migration, as it enables the decision maker to assess the trustworthiness of automated conversion processes and software tools. Without QA, proper migration is not possible. Similar considerations of course apply to any other change in the view path [15].

Quality assurance of this conversion is lacking, and there is little assurance as to the reliability of existing tools for quality assurance. Given the limitations of existing tools, QA often focuses on the available measures such as image width, metadata, or format validity. Yet, this is arguably not sufficient to verify that the resulting digital file represents an accurate version of the digital object. Only if we can verify that the significant characteristics of the original object can be reproduced do we have an authentic manifestation. It is this level of Quality Assurance that is the focus of this article.

Manual QA of these aspects will be feasible on small data sets with an expert to judge the correctness of migration output using standardized rendering environments and calibrated equipment. But with limited human resources and rising numbers of photographs, the only feasible approach is automated QA. Here, we need tools that are reliable enough to replace a human expert and provide a judgment on correctness of migration.

This paper addresses the issue of providing automated QA for preservation of born-digital raw photographs. This work is based on the results of earlier research conducted by Bauer et al.[1] .We will briefly describe the requirements needed for successful automation of the processes and a tool for image comparison. Our focus is on an automated, publicly available data processing workflow relying on automated perceptual image QA, and an experiment in which we evaluate the correlation of the tool results to manual annotation carried out in a calibrated environment.

The remainder of this paper is organized as follows: Section 2 gives an overview of related work in image preservation. Section 3 discusses challenges and requirements posed by born-digital photographs. The tool for image comparison and a published, reusable automated workflow is described in Section 4. The results of the experiment are presented in Section 5. Finally, Section 6 provides conclusions and a short outlook on future work.

Related Work

Authenticity of digital objects has been studied extensively in the digital preservation domain. A common approach to verify the authenticity of digital objects is to compare their significant characteristics. Given that this significance is contextual, the definition needs to be based on an understanding of the decision context and the stakeholders [2]. As such, it is at the heart of preservation planning [3]. The decision criteria defined to eval-

[1] http://www.adobe.com/products/photoshop.html

[2] http://www.phaseone.com/capture-one

[3] http://www.adobe.com/support/downloads/detail.jsp?ftpID=5645

[4] http://www.digikam.org/

Table 1: The list of raw image properties[1].

Property	Relation	Description
relative AE	content	rAE describes the percentage of pixels that differ when comparing the uninterpolated images
relative MSE	content	rMSE calculates the mean squared error between two uninterpolated images divided by the bit depth
SSIM Value	content	raw data is transformed into fully interpreted 16-bit TIFF, and the brightness of the HSV representation is compared using SSIM
SSIM Saturation	content	SSIM is applied to the S-channel of HSV for comparison
SSIM Hue	content	SSIM is applied to the H-channel of HSV for comparison
Exif (exposure)	context	ExposureTime, FNumber, ExposureProgram, ExposureMode, ISOSpeedRating, SubjectDistance, Flash, FocalLength, ...
Exif (technical)	context	ImageWidth, ImageHeight, CFAPattern, Make, Model, ...
Exif (location)	context	GPSLatitude, GPSLongitudeRef, GPSAltitude, ...
Exif (generated)	context	Metadata added during migration to enable correct rendering: ColorMatrix, AsShotNeutral, BaselineExposure...
IPTC	context	Creator, ContactInfo, SceneCode, Location, City, Province-State, Country, Headline, Description, Keywords, ...
Dublin Core	context	Title, Subject, Description, Source, Coverage, Creator, Rights, Date, Format, Identifier, Audience, ...
Private Tags	context	E.g. ColorTemperature, WhiteBalance, InternalSerialNumber, AFPointsSelected,..
XMP	context	XMP contains Exif and IPTC metadata as well as custom information as used by Adobe Camera RAW and others

uate whether a given preservation action performs well are specified as metrics, each providing a specific measure that can be judged to evaluate whether the significant properties have been adequately preserved. A list of such metrics for raw images in a particular case study is informally presented in Table 1, taken from [1]. While the exact significance of specific elements will vary considerably for comparable scenarios, it is clear that these measures and the properties they refer to have wider relevance in similar environments.

The identified properties may be divided in two groups: content related and context related. The first group relates to any difference between the actual images displayed using the original artefact and environment and any derived artefact. We will focus on content related properties further below. The second group is termed "context" and refers to essential metadata elements of the images. Verifying characteristics of this group is technically straightforward, since they are commonly encoded in a form readily supported by standard tools such as image viewers and metadata extractors. As discussed in [1], these can be verified using metadata extraction[5] and comparison techniques [5].

The content, however, relates to the performance [4] achieved in a given environment and the question whether this performance resembles the original performance. With born-digital raw photographs, the complex processing steps that take place before an image is displayed make this a challenging task: different cameras use a variety of specific presets, the geometric arrangement of the camera sensors do not match the pixels displayed on screen, and the color is a projection created from the digital negative. This paper focuses on the comparison algorithms focused on content elements. Correspondingly, the content rows in Table 1 specify a set of metrics to be evaluated.

In the area of born-digital raw images, there is not a single raw format. Instead, camera vendors each introduce and support their own formats such as Kodak Digital Camera RAW (KDC), Canon RAW 2 (CR2), Epson RAW Format (ERF), Nikon (RAW) Electronic Format (NEF), Olympus RAW Format (ORF), Pentax (RAW) Electronic Format (PEF), Panasonic RAW 2 (RW2) etc. Most of them are based on the TIFF file format, but often include non-standard file headers, additional image data, feature-specific tags etc. A popular alternative for proprietary formats is the open raw format Digital Negative (DNG) by Adobe. The standard is under consideration of being proposed as part of ISO's TIFF/EP[6]. In [6], the authors discuss the suitability of DNG and other raw formats to be used as an image format to store digitized content in archives and libraries.

Methods used for image comparison that address content related properties of images use metrics such as Mean Square Error[7] or the Minkowski metric[7]. These metrics have a solid mathematical background and can be used for discovering errors in conversion processes. However, they do not properly describe the human perception[8]. This has the effect that any comparison result but perfect equality is challenging to interpret, since the distance measures do not correspond to human perception of similarity. This is addressed in the Structure Similarity index metric (SSIM)[7]. The difference between rMSE and SSIM is that the first estimates squared intensity difference of original and distorted pixels, while SSIM focuses on the comparison of structural elements that constitute images. It is possible to have two similar images as reported by SSIM even in the case of a high rMSE.

Zauner et al.[10] propose using a so called perceptual image hash function to describe an image. Such a function, at first, decreases an image to resolution of N-by-N pixels (which is usually 8x8) and after binarization encodes the image into a hash string. Using the hashes of two images, it is then possible to calculate an edit distance needed to translate one hash into another.

Since research in [7, 8, 9] showed that SSIM corresponds more closely to human perception, and our goal is to replace a human expert, the SSIM metric is the first candidate.

[5] http://www.sno.phy.queensu.ca/~phil/exiftool/#supported

[6] http://www.dpreview.com/news/2008/5/15/adobeDNG

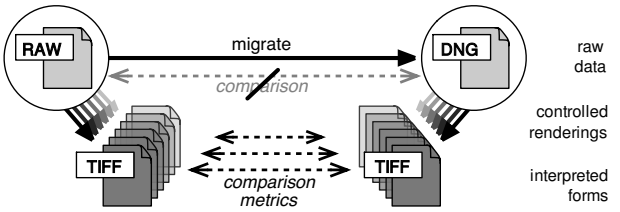

Figure 1: Controlled rendering of raw images enables QA [1].

Automated Measures

The diversity of image formats and the fact that the actual performance of the image is what constitutes the essence of the digital objects, not the encoded file, means that no formal equivalence relation can be defined on the file level. However, as stated in [1], *"two raw files can be called equal if (and only if) they can be transformed into the same interpreted form. In other words, the original and the migrated raw file can be called equal if they are rendered the same way in a reference rendering environment. In the absence of ground truth for verifying conversion and renderings, we need to rely on a standardized and widely used rendering environment"*. This situation is depicted in Figure 1. Before running any property comparison, images must be presented in a unified format with the same data structure. The controlled renderings can potentially introduce additional errors; this can be addressed by relying on multiple renderings. Specific renderings can be used to provide different views on the original objects in order to enable particular comparison metrics.The set of renderings and metrics together needs to be validated for its correspondence to comparing the performance as perceived by a human.

Based on this, we can define a generic method for automated quality-assured conversion:

1. Convert an image from a raw format A to a raw format B;
2. Convert the original image and the converted image into a common format (such as TIFF-6). More than one rendering can be created to compare different properties;
3. Compute distance metrics on the pairs of rendered images;
4. Draw conclusions on the correctness of the raw-to-raw conversion.

It is understood that no absolute certainty can be achieved considering that the intermediate artifacts could be erroneous too. The focus hence must be defensive and focused on falsification.

Image QA tool Photohawk and Taverna workflow

To enable experimentation and usage of these algorithms for photographs, we have developed the QA tool Photohawk[7]. It is able to calculate the set of content comparison metrics from Table 1. For a given pair of images and a preferred comparison metric, the tool produces a value in the continuous range 0 to 1, where 0 stands for completely dissimilar and 1 for identical. To convert raw images to a common TIFF format, we use dcraw[8], an open source tool widely used for raw image processing actions such as enhancements, conversions, cropping etc.

To enable reuse and experimentation with Photohawk and ensure it is accesible to the public, easy to install and use, and

[7] http://datascience.github.io/photohawk/
[8] http://www.cybercom.net/~dcoffin/dcraw/

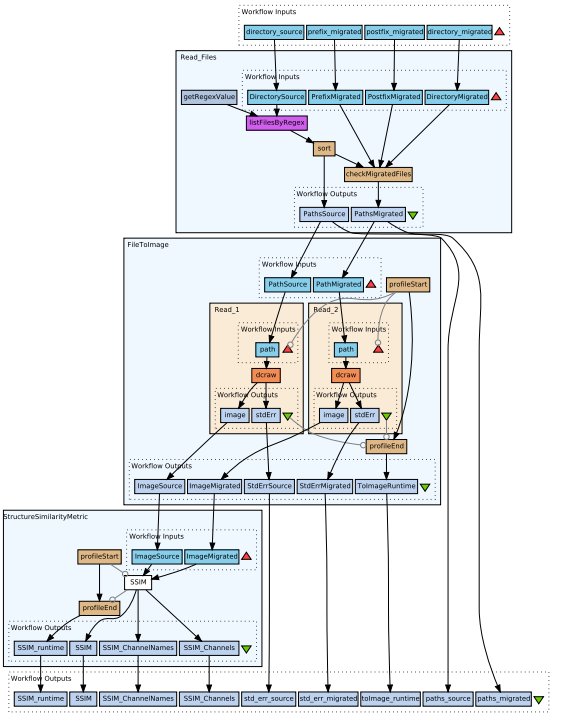

Figure 2: A Taverna workflow based on Photohawk.

modular and reusable, a Taverna workflow based on Photohawk has been created. Taverna is a workflow management system originally developed for bioinformatics[11]. It is open-source and platform-independent, which means that this workflow may be run on any popular OS. Taverna is used to design and execute workflows. The workflow is illustrated in Figure 2. It consists of several smaller nested workflows.

Originally intended primarily for e-science experimentation, the digital preservation community has discovered the benefits of sharing and reusing workflows in recent years[9]. A key advantage of such approach is the integration with myExperiment[10][12], a social networking and workflow sharing environment hosting a growing pool of workflows that can be used to conduct experiments, share expertise and avoid unnecessary reinvention. The workflows created for Photohawk are published on myExperiment[11].

Experimentation

To enable the QA software mechanism capable to identify similarity of images to reliably replace the judgment efforts of experts, we need to verify whether the QA mechanism is able to provide results that can be used as reliable indicator verifying the quality of conversion results. Any finding regarding our hypothesis is of crucial importance for automated image QA, but also for automated QA in general. Once an annotated dataset for a certain type of content is presented and a QA tool under consideration is run on that dataset, that QA tool may be tested in much the same

[9] http://www.myexperiment.org/groups/490.html
[10] http://www.myexperiment.org/
[11] http://www.myexperiment.org/packs/576.html

way as described below.

First, a test dataset was generated containing 230 raw images that were sampled from a collection of about 12000 raw images. Each image was converted to DNG format using the following converters with default configuration: CaptureOne, Adobe DNG Converter, and DxO[12].

Second, the updated dataset (the original images and the images obtained from the converters) was annotated. For this, we set up hardware and software for 2 independent groups of reviewers. Monitors of the same model and vendor were calibrated using professional calibration equipment to ensure they display images in the same manner. OS and rendering software (Adobe Lightroom[13]) was identical. Each group provided the following information (factors) based on their perceived experience for each pair of images:

Framing refers here to any differences in image dimensions and distortion as perceived in the viewer. These differences can typically and intentionally be caused by post-processing instructions correcting lens distortions. If these lens corrections are intended, they rely on both the encoded lens information and specific computing steps. If either are lost, the perceived image will change. On the other hand, conversion software can introduce these corrections, even if this is not inteded. Additionally, border pixels on sensors are often meant to be left out when creating an interpreted image, but some conversion processes may in fact include them, increasing the size of the image that is produced.

Brightness refers to differences in brightness and contrast levels of images. This factor is more difficult to detect for the human eye for numerous contextual reasons including the sequence of contrasts perceived before looking at the image. It is also highly dependent on hardware and software, rendering settings and monitor calibration.

Hue The perceived color fidelity is of course a crucial requirement, and at the same time one that poses challenges, in particular when camera settings such as white balance are not correctly encoded.

Each factor was graded by a number from 1 to 3, where

- 1 denotes poor performance (the images are not identical, there are easily detectable differences;
- 2 denotes problematic performance, the images are almost identical, but there are slight differences; and
- 3 denotes excellent performance, images appear identical, there is no noticeable difference.

The experiment checks whether automated measures are capable of corresponding adequately to changes in values for these factors. The scaling of 1 to 3 was intentionally used to ensure a robust experiment setting and avoid a false sense of precision on the side of human judgment.

Thirdly, we used Photohawk to compute distance metrics on the dataset. To calculate similarity in an automated fashion, the Taverna workflow for Photohawk is used to produce SSIM and MSE measures. In addition to the content properties from

[12] http://www.dxo.com/

[13] http://www.adobe.com/products/photoshop-lightroom.html

Table 2: Correspondence between image metrics and reviewers' factors.

Metric	Framing	Brightness	Hue
rAE	+	+	
rMSE	+	+	
rMAE	+	+	
rPAE	+	+	
rAE*	+	+	+
rMSE*	+	+	+
SSIM Hue			+
SSIM Saturation			+
SSIM Value		+	

Table 1, Photohawk is able to calculate Relative Mean Absolute Error (rMAE) and Relative Peak Absolute Error (rPAE). Table 2 shows which of the calculated metrics are seen as candidates for the approximation of reviewers' judgments. For example, to consider automation of the Hue factor, the most appropriate properties to choose from are SSIM Hue, SSIM Saturation, rAE* and rMSE*. The introduced metrics rAE* and rMSE* are rAE and rMSE as calculated on the fully interpreted images. These are included to evluate if these simple metrics provide additional useful indicators on image similarity.

Finally, we analysed the experiment results and assessed the correspondence between the automated and human judgement. Some result samples are presented in Figure 3. Size limitations prevent an in-depth discussion of the entire experiment, but the following section will attempt a discussion of selected aspects of interest.

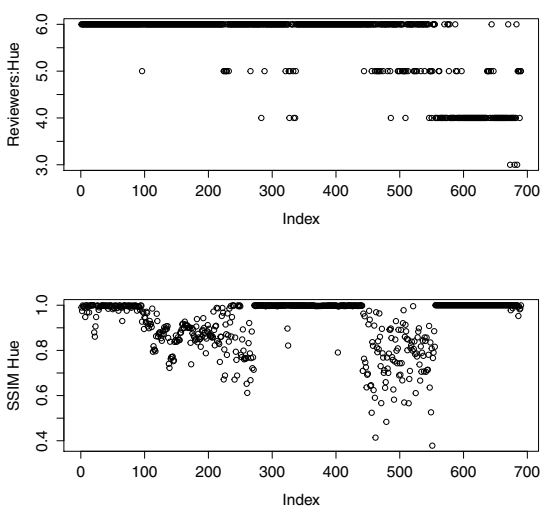

Figure 3: Distributions of reviewers' Hue factor scoring and SSIM Hue property of the dataset.

Figure 3 contains distributions of experiment results for Hue factor and SSIM Hue metric. We combine outputs of 2 groups of the reviewers by summing their scoring. It could be noticed that most of the highest results (6.0) for Hue factor correspond to values over 0.8 in SSIM Hue. We have to keep in mind limitations of human visual perception, which could be the case in detecting

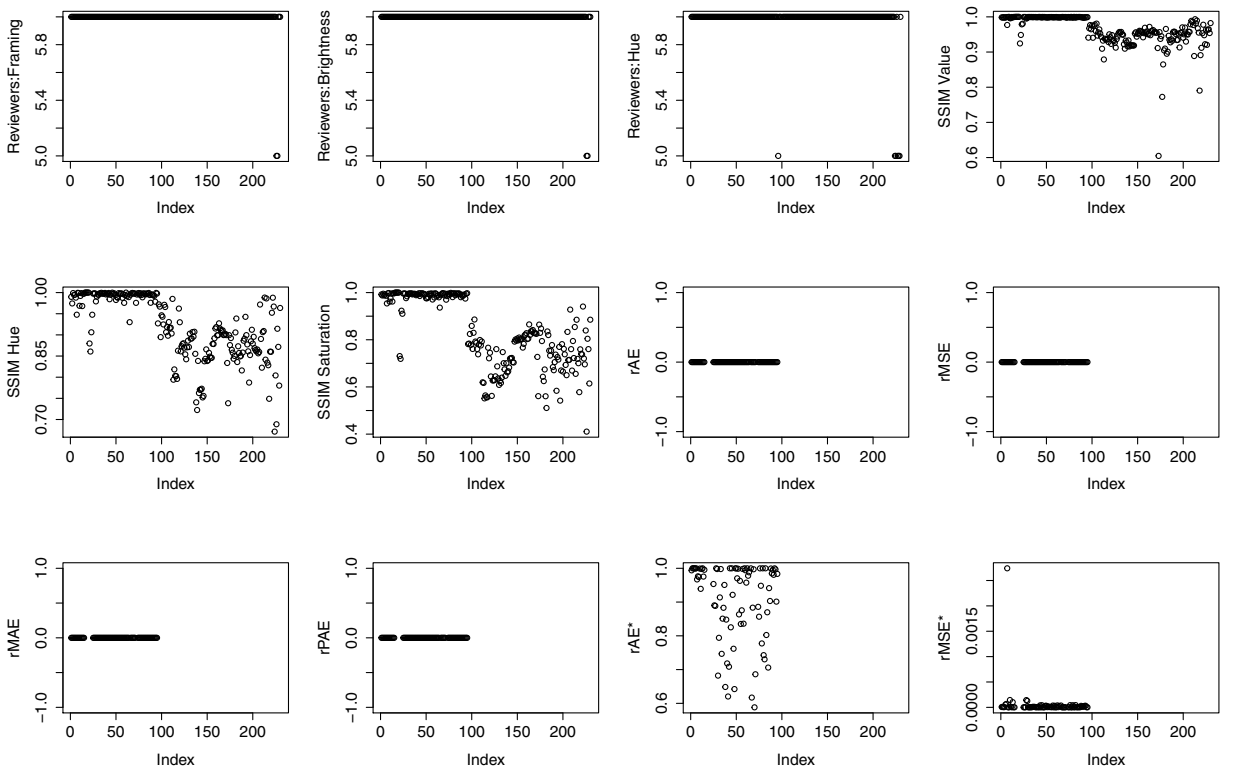

Figure 4: Experiment results for first 230 paired image samples.

difference in color. To a first approximation, this value (0.8) could be considered as an empirical threshold for decisions made on the similarity of image pairs. If that was the case, images that were evaluated by Photohawk with a SSIM Hue metric value over 0.8 could be considered similar. However, the last 150 images perform differently: reviewers and Photohawk disagree on their similarity. Photohawk reports them completely identical, while reviewers give them the lowest scores. We will come return to these images below.

As Figure 3 shows, there are clusters of values that can be visually distinguished. For example, in SSIM Hue there are 3 explicit groups of pairs with a score of 1.0. These groups correspond to the 3 conversion tools described previously. The problematic last 150 images are obtained from one source and converted by DxO. The same images converted by other tools are treated identically well both by SSIM Hue and reviewers' Hue factor.

We focus on the first 230 pairs of the experiment. Figure 4 plots measures corresponding to the choices outlined in Table 2. For the selected samples, all factors demonstrate high scores. For Framing, the corresponding values in the error-based diagrams (rAE, rMAE, rMSE, rPAE, rAE*, rMSE*) show values only for the first half of samples. Absence of values in the diagrams is due to difference in the size of images. The missing values correspond to the set of problematic images. While all error metrics agree on incorrect image sizes, the reviewers still give these images the highest scores for Framing factor based on the performance in the reference rendering environment. These particularities point to errors in the rendering dealing with particular properties in the converted files, the cause of which needs to be verified in detail.

The remaining error-metric values behave as expected (lower values mean higher similarity), except rAE* with a mean value close to 1.0. This can be the explained by the fact that during interpolation, color noise is introduced. The metric rAE* is very sensitive to differences in each pixel of images, which leads to a high error reported. Regarding the Brightness factor, SSIM Brightness demonstrates almost the same distribution with little deviation in the second half of samples. The difference may be explained by the fact that for a human, it is not a trivial task to capture changes in brightness levels of two images. SSIM shows a more differentiated picture of this factor. As for Hue, SSIM Hue and SSIM Brightness are closely corresponding to it. Most of the values are above 0.8 in SSIM Hue and 0.7 in SSIM Saturation. The obtained results point to the possibility of using SSIM Hue and SSIM Saturation as reliable indicators for the factor Hue.

The images under consideration in the previous paragraph were converted by AdobeDNC. These images demonstrate the highest performance regarding both human judgment and Photohawk metrics. Other tools produce results that are more varied. In general, we find a lot of noise in results obtained through such experimentation. Partial results are promising, such as the possible suitability of using SSIM Hue indicator for comparing color fidelity. This is confirmed by statistical testing with a correlation of 0.72. However, the results are in no way robust enough to warrant excessive confidence. Some insight is obtained on the feasibility of the approach in this scenario, more robust experimentation and analysis is required before the quality assurance mechanisms can be declared trustworthy and thresholds values can be determined with certainty.

Conclusions

Without a reliable verification of the tests and measures that we use for conducting quality assurance on digital content, preservation actions will never be both trustworthy and scalable [14, 16]. While this article focused the discussion on a migration scenario, it is really any change in the viewpath that requires us to validate whether the performance of digital objects is in fact resembling the elusive 'original' performance [15]. This article presented an open source approach and publicly available workflow for computing perceptual QA measures on pairs of images, applied to born-digital raw photographs.

The main purpose of this experiment is to investigate approaches to determine the equivalence of automated measures to human judgments and the degree of confidence we can obtain in substituting expert judgment with automated QA mechanisms. While the results point to the feasibility of doing so, it is clear that careful validation needs to be carried out before being able to rely on such automated measures. The controlled rendering is an inevitable intermediary required to provide the means of comparison. At the same time, this artefact needs to be robust enough to assure us that it does not introduce additional errors. The behavior observed in some of the clusters points to potential errors in the renderings that need to be investigated in detail. The data set should be statistically representative of the entire content set, which is a challenging subject. Sophisticated content analysis and aggregation mechanisms can address this challenge, but need further development [13]. Combinations of metrics seem to be required to detect differences reliable, based both on the uninterpreted data and on (multiple) renderings. Finally, larger studies are needed to evaluate the initial hypotheses generated through such exploratory testing. This should lead to robust thresholds that can be used as guidance for validating QA results. In addition, a crucial question arises: For distance metrics that are close to 1, but not 1, is it possible to define a threshold that will reliably identify problematic renderings? Can this be done generally? Current and future work includes conducting larger tests to identify such thresholds in the QA outputs linked to human perception levels.

The tools described are fully open source and can be applied to conduct QA on any digital photograph collection. It is hoped that other experiments will be conducted and published to enable wider comparison and benchmarking of QA mechanisms

Acknowledgments

Part of this work was supported by the Vienna Science and Technology Fund (WWTF) through project ICT12-046 (BenchmarkDP) and by the 7th Framework Program, IST, through the SCAPE project, Contract 270137.

References

[1] Bauer, S., Becker, C., "Automated preservation: the case of digital raw photographs." In Proc. ICADL, Springer Berlin Heidelberg, pp. 39-49, 2011.

[2] Dappert, Angela, and Adam Farquhar. "Significance is in the eye of the stakeholder." In Proc. ECDL. Springer Berlin Heidelberg, 2009. 297-308.

[3] Becker, Christoph, et al. "Systematic planning for digital preservation: evaluating potential strategies and building preservation plans." International Journal on Digital Libraries 10.4 (2009): 133-157.

[4] Heslop, Helen, Simon Davis, and Andrew Wilson. "An approach to the preservation of digital records." Canberra: National Archives of Australia, 2002.

[5] Becker, Christoph, and Andreas Rauber. "Decision criteria in digital preservation: What to measure and how." JASIST 62.6 (2011): 1009-1028.

[6] Bennett, Michael J., and F. Barry Wheeler. "Raw as archival still image format: A consideration." Archiving Conference. Vol. 2010. No. 1. Society for Imaging Science and Technology, 2010.

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

[8] Wang, Zhou, and Alan C. Bovik. "Mean squared error: love it or leave it? A new look at signal fidelity measures." Signal Processing Magazine, IEEE 26.1 (2009): 98-117.

[9] Wang, Z., Simoncelli, E.P., Bovik, A.C., "Multiscale structural similarity for image quality assessment," Signals, Systems and Computers, 2004. Conference Record of the Thirty-Seventh Asilomar Conference on, vol.2, no., pp.1398-1402 Vol.2, Nov. 2003.

[10] Zauner, Christoph. "Implementation and benchmarking of perceptual image hash functions." Master's thesis, Upper Austria University of Applied Sciences, Hagenberg Campus 43 (2010).

[11] Oinn, Tom, et al. "Taverna: a tool for the composition and enactment of bioinformatics workflows." Bioinformatics 20.17 (2004): 3045-3054.

[12] Goble, Carole Anne, and David Charles De Roure. "myExperiment: social networking for workflow-using e-scientists." Proceedings of the 2nd workshop on Workflows in support of large-scale science. ACM, 2007.

[13] Petrov, Petar, and Christoph Becker. "Large-scale content profiling for preservation analysis." Proc. iPRES (2012).

[14] Becker, Christoph, and Kresimir Duretec. "Free benchmark corpora for preservation experiments: using model-driven engineering to generate data sets." ACM/IEEE JCDL, 2013.

[15] Guttenbrunner, Mark, and Andreas Rauber. "A measurement framework for evaluating emulators for digital preservation." ACM Transactions on Information Systems (TOIS) 30.2 (2012): 14.

[16] Rosenthal, David SH. "Format obsolescence: assessing the threat and the defenses." Library hi tech 28.2 (2010): 195-210.

Author Biography

Artur Kulmukhametov is project assistant and PhD student at the Software and Information Engineering Group, Vienna University of Technology and involved in the projects SCAPE and BenchmarkDP. The focus of his research is systematic analysis and evaluation of scalable digital preservation environments and their longevity.

Markus Plangg is project assistant at the Information and Software Engineering Group, Vienna University of Technology. He is involved in the SCAPE project and focuses on scalable digital preservation.

Christoph Becker is Assistant Professor at the Faculty of Information and Associate Director of the Digital Curation Institute, University of Toronto, and Senior Scientist at the Software and Information Engineering Group, Vienna University of Technology. He was involved in the European research projects DELOS, PLANETS, DPE, and SHAMAN. He was leading the sub-project Scalable Planning and Watch of the SCAPE project and he is Principle Investigator of the new project BenchmarkDP. His research takes an information systems and design perspective on questions of digital longevity, curation, and archiving.