Visual Web Archive Quality Assessment

Theresa Elstner[1], Johannes Kiesel[2], Lars Meyer[2], Max Martius[1], Sebastian Schmidt[1], Benno Stein[2], and Martin Potthast[1]

[1] Leipzig University
[2] Bauhaus-Universität Weimar

Abstract. The large size of today's web archives makes it impossible to manually assess the quality of each archived web page, i.e., to check whether a page can be reproduced faithfully from an archive. For automated web archive quality assessment, previous work proposed to measure the pixel difference between a screenshot of the original page and a screenshot of the same page when reproduced from the archive. However, when categorizing types of reproduction errors (we introduce a respective taxonomy in this paper) one finds that some errors cause high pixel differences between the screenshots, but lead to only a negligible degradation in the user experience of the reproduced web page. Therefore, we propose to visually align page segments in such cases before measuring the pixel differences. Since the diversity of reproduction error types precludes a one-size-fits-all solution for visual alignment, we focus on one common type (translated segments) and investigate the usefulness of video compression algorithms for this task.

Keywords: Web archiving · Automatic quality assessment · Visual web page alignment.

1 Introduction

Web archives are created for posterity and they already serve researchers in social sciences, history, linguistics, humanities, and—not least—computer science. The diversity of web technologies and the way users interact with web pages are the main challenges for web archiving. In fact, creating high-quality archives that reliably reproduce layout, content, and behavior of web pages from archived data is a largely unsolved problem: An estimated 44% of archived web pages have minor to major defects due to reproduction errors, while as many as 4% of them are unusable [3]. These are significant portions of the 685 billion web pages archived by the Internet Archive's Wayback Machine at the time of writing [2].

Reproduction errors are parts of a web page that do not look the same when reproduced from a web archive as they did when archived. In this regard, Figure 1 (left) illustrates three common reproduction errors: A missing video causes the translation of underneath content, and the addition of white space or the reduction of screen height. Only the missing video (if relevant to the web page author's intended message) affects the archive quality, not the translation. The example illustrates why it is difficult to identify web pages with low archive quality, even if a screenshot of the original web page is available for reference. Therefore,

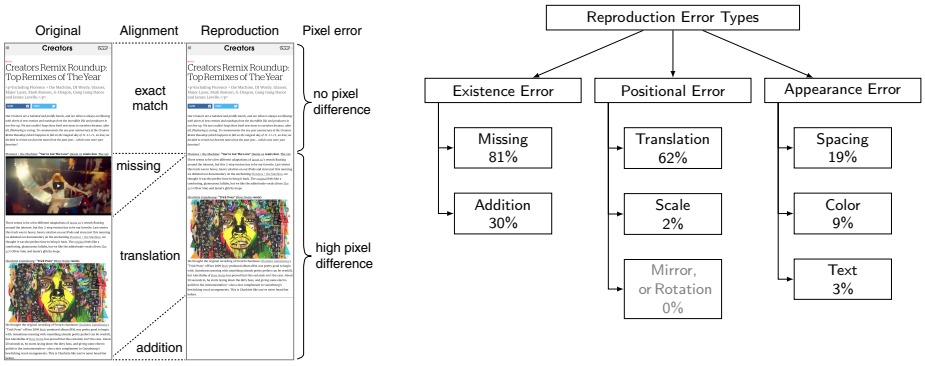

Fig. 1. *Left:* Screenshots of a web page in its original form and when reproduced from an archive with reproduction errors: A missing video causes a vertical translation of content. *Right:* Reproduction error types and the respective percentage of affected pages (i.e., pages with at least one such error) in a web archive sample.

evaluating the quality of web archives requires a user-oriented (semantic) analysis that takes into account the usefulness of an archived page for specific purposes. However, no significant progress has been made in this regard since the task was first proposed along with a benchmark dataset (see Section 2). This paper makes three contributions to take the next steps: (1) a taxonomy of reproduction errors and an empirical assessment of the error frequencies, (2) a proposal for a normalization task (visual segment alignment) to improve automatic web archive quality assessment, and (3) an analysis of the applicability of video compression algorithms for this task. Our code and data is publicly available.[3,4]

2 Related Work

While a considerable amount of research addresses web archive quality, little work has been done on its automatic assessment using visual information. The Internet Archive's Archive-It suggests to "Browse through your archived site(s), clicking links and activating dynamic media players in order to make sure that they were archived in accordance with your expectations" [1]. Reyes Ayala et al. [5, 6, 4] outline the need for an automated solution using screenshots based on an analysis of Archive-It's support tickets. The degree of similarity between screenshots of the original and archived versions of a page contributes significantly to manual evaluation. Reyes Ayala et al. [6] study image similarity measures in terms of their ability to detect such differences. Kiesel et al. [3] have created a large benchmark dataset of 10,000 archived pages for this task. They introduce the root-mean-square error (RMSE), which is calculated from the pixel difference of original and archive screenshots, as a well-performing measure, yet conclude that "[...] a single missing advertisement banner at the top of the page can shift up the entire web page, causing the RMSE to become unjustifiably high."

[3] Code https://github.com/webis-de/TPDL-22
[4] Data https://zenodo.org/record/6881335

3 A Taxonomy of Web Archive Reproduction Errors

Figure 1 (right) shows a taxonomy of reproduction error types identified when comparing screenshots of an original web page with its archived version (original–archive screenshots). Relative frequencies were determined manually using a random sample of 100 original–archive screenshots with pixel differences from our dataset.[5] Each frequency estimates the proportion of archived pages for which a given reproduction error is expected, i.e., a pair may contain multiple errors. Each of the three branches in the taxonomy describes the way in which a segment in the archive screenshot may look different compared to the corresponding segment in the original screenshot: (1) existence errors indicate elements that are present in the original but not in the archive screenshot or vice versa; (2) position errors indicate where a segment appears in the archive screenshot and how its position differs from the corresponding segment in the original screenshot; and (3) appearance errors indicate differences in the content or layout of an archived item compared to its original, including color differences.

A reproduction error can affect either the whole or a part of a segment. One segment may also be affected by multiple reproduction errors. Existence errors occur most frequently, followed by translation and spacing errors, and they can even cause a number of other errors. For example, a missing segment may cause translation errors affecting segments further down the page (as in Figure 1), which may result in additional segments (content) at the bottom.

Beyond visual reproduction error types, a web page's interactive content (audio, animation, or any functional elements triggered by interaction, including hyperlinks) may also contain errors. These errors cannot be identified using screenshots. Furthermore, some seemingly erroneous differences may only affect its screenshot, not the archived web page itself: 41% of the 100 manually annotated web pages were affected by this, with many of them being timing issues, e.g. showing different parts of a GIF animation.

4 Visual Alignment of Original–Archive Screenshot Pairs

We present a generic framework for web archive quality assessment with three steps: (1) visual segment alignment, where original–archive screenshots with reproduction errors that do not affect the quality of an archived web page are aligned (in our prototype translated DOM elements are returned to their original position), (2) visual edit distance calculation, and (3) quality prediction, where machine learning is used to predict human quality annotations based on the visual edit distance. Our experiments are based on a sample of original screenshots from the Webis-Web-Archive-17 [3], supplemented by new reproductions and a corresponding description of structure. To ensure the validity of the human annotations from Webis-Web-Archive-17 for this dataset, it does not contain web pages whose reproductions differ by more than 5% of the corresponding pixels.

[5] The fraction of screenshots without pixel differences (no reproduction errors) in our dataset is 12.9% (845/6531).

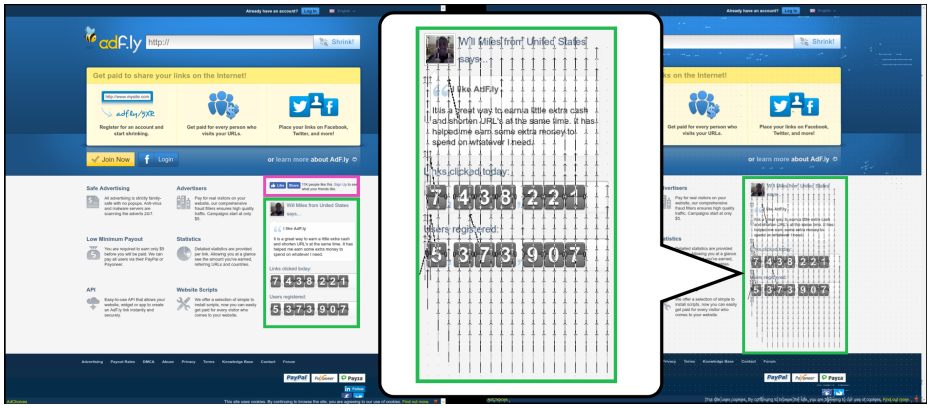

Fig. 2. Original and archive screenshot as two frames of a H.264-encoded video. The element marked by a pink border in the original screenshot is missing in the archive screenshot, so the element below (marked by a green border) is translated upwards as indicated by the motion vectors (black arrows).

Visual Segment Alignment. Our visual segment alignment method moves translated DOM elements in archive screenshots to their respective positions in the original screenshot. The positions of elements in the archive screenshots are derived from the available page structure descriptions. As these are unavailable for the screenshots of the original web pages, we derive the original positions using motion estimation algorithms from video encoding software as follows.

We use the media conversion framework ffmpeg [7] to encode original–archive screenshots into an H.264 video. From this video, we obtain motion vectors containing information about the translation of pixel blocks in the archive screenshot compared to the original screenshot. Figure 2 shows a corresponding example. Together with the position information of the elements from the page structure description, we can recover the positions of the moved elements according to the motion vectors. It should be noted that motion vectors for an element do not always indicate the same original position, so we use majority voting to decide whether to perform a move. This may lead to errors (e.g., large translated child elements cause translations of their parents).

Calculating the Visual Edit Distance. The second component is implemented as a visual edit distance. It compares each pixel in the aligned and archive screenshot with the pixel at the same position in the original screenshot. This results in two visual edit distances: *aligned* and *baseline*. Both of them count the minimum number of applications of the substitution, insertion, and deletion operations.

In a pilot study, we evaluated three methods for coloring gaps created when translating elements: (1) using duplicates of translated elements, which leaves no gaps; as well as coloring the gap in the most frequent color (2) around the edges of the translated element; and (3) in the archive screenshot. Method 3 achieved the least visual edit distance. Hence it was considered in the following step.

Table 1. *Left:* Confusion matrix Δ from baseline to aligned regression, with correct classifications (i.e., on the main diagonal) highlighted. *Right:* Precision, recall, and F_1-score for detecting pages with low reproduction quality (per targeted "low quality") per aligned regression, including delta from baseline to aligned regression.

Truth	Predicted quality				
	1	2	3	4	5
1	+1	+4	−6	+1	±0
2	−108	+115	−8	±0	+1
3	+2	−7	+4	±0	+1
4	−10	+7	+3	−3	+3
5	−7	+7	+3	−3	±0

Target	Precision (Δ)	Recall (Δ)	F_1 (Δ)
2–5	0.850 (+0.008)	0.770 (+0.045)	0.808 (+0.029)
3–5	0.644 (+0.053)	0.186 (+0.012)	0.289 (+0.020)
4–5	0.518 (−0.035)	0.101 (−0.007)	0.169 (−0.011)
5	0.185 (−0.019)	0.049 (±0.000)	0.077 (−0.002)

Predicting Web Archive Quality. We implement a linear regression model evaluated with a 10-fold cross-validation to predict the quality of archived web pages on a 5-point scale as per the used benchmark dataset's human annotation [3]. As an input to this model, we test features from three categories. *Unaligned features* are: (u1) the size difference of the unaligned archive and original screenshots in pixels due to different screenshot heights; (u2) the baseline visual edit distance as described above, which includes u1; (u3) the baseline visual edit distance normalized by the original screenshot size; and (u4) the pixel size of the original screenshot to put u1 and u2 into context. *Aligned features* (a1–a3) correspond to u1–u3 for the aligned archive screenshot. *Translation features* count the DOM elements translated between original and archive as per the visual alignment: (t1) all translated elements; (t2–t9) elements translated into a particular direction as per the motion vectors, namely towards the top, top right, right, bottom right, etc.; (t10, t11) elements translated by a large or small distance, the latter being too small to indicate a missing content element;[6] (t12–t27) the same as t2–t9, but once counting only large and once counting only small translations like for t10 and t11; and (t28) all elements from the archived page, again for context.

5 Evaluation

Table 1 shows the improvement of an aligned regression over a baseline regression restricted to unaligned features (u1–u4). We tested each feature combination for the baseline regression and multiple inter- and intra-group combinations of the three groups of features (u,a,t) for the aligned regression. Δ describes the change in the respective value from the best-performing baseline regression system (which uses the size difference u1 and visual edit distance u3) to the best-performing alignment regression system (which uses only the aligned visual edit distance a3). For the confusion matrix on the left of Table 1, we find that the number of correct classifications increases, indicating small classification improvements despite the very prototypical nature of our approach and features. As the table shows, the main improvement of the aligned regression lies in a

[6] Based on a frequency analysis, we set the threshold for small to be 5 or fewer pixels vertically and 8 or fewer pixels horizontally.

better differentiation between quality 1 ("not affecting the visitor" per [3]) and 2 ("small effect on a few visitors"). Following Kiesel et al. [3], the right hand side of Table 1 shows the achieved precision, recall, and F_1 score for detecting pages of low reproduction quality. Across all three scores, our prototypical approach yields some improvements for identifying reproductions of quality 2–5 and 3–5. If the targeted low quality includes only 4 ("affect, but page can still be used") and 5 ("unusable page"), the prototypical alignment slightly reduces the scores. However, we assume that a quality of 3 ("small effect on many or all visitors") is already a cause for concern for web archivists worth detecting.

Limitations. Our approach considers translation errors only. However, also other types of reproduction errors lead to a higher visual edit distance than warranted for the associated decrease in reproduction quality. While the selected translation features (t) could not improve the regression in our prototypical study, more elaborate features that go beyond counting translated DOM elements may be able do so. Similarly, other algorithms than linear regression might improve the quality assessment. For example, one could also integrate our alignment into a convolutional neural network [3].

Our study employs only a single approach for the critical task of detecting translated elements. But other approaches, not based on video encoding, might also be suited for web pages. Indeed, we found that the video encoding misses some translations. We manually annotated translations in 80 randomly sampled web pages from the dataset: Our annotation identified 148 million translated pixels, but only 28 million of them (about 1/5) were found automatically. Although many translations were detected correctly, we noticed that especially translations of large areas are sometimes missed. This suggests that more work on tuning the video encoding as well as alternative approaches are needed.

6 Conclusion

We explore a novel approach to automatic quality assessment of archived web pages based on visual alignment. To this end, we categorized visually perceivable reproduction error types, proposed a three-step framework for automatic quality assessment, and implemented and tested a prototypical implementation based on detecting translated elements from motion vectors. Our implementation shifts translated elements in the archive screenshot back to their original position, thereby bringing the pixel difference between original and archive screenshot closer to the human perception of reproduction quality. A comparison of two linear regression models, using features from unaligned and aligned screenshots respectively, shows small improvements over the baseline when using aligned screenshots even for our prototypical implementation. However, the design space for archive quality assessment algorithms is vast, and a more thorough exploration is necessary to develop reliable assessment technology—potentially based on our generalizable framework of normalization through visual segment alignment.

References

1. Internet Archive: Quality Assurance Overview. https://support.archive-it.org/hc/en-us/articles/208333833-Quality-Assurance-Overview (2022)
2. Internet Archive: Wayback Machine size as displayed on its front page. https://web.archive.org/web/20220531094827/https://web.archive.org/ (2022)
3. Kiesel, J., Kneist, F., Alshomary, M., Stein, B., Hagen, M., Potthast, M.: Reproducible Web Corpora: Interactive Archiving with Automatic Quality Assessment. Journal of Data and Information Quality (JDIQ) **10**(4), 17:1–17:25 (Oct 2018). https://doi.org/10.1145/3239574, https://dl.acm.org/doi/10.1145/3239574
4. Reyes Ayala, B., Phillips, M., Ko, L.: Current quality assurance practices in web archiving. UNT Digital Library pp. 1–34 (August 2014)
5. Reyes Ayala, B.: Correspondence as the primary measure of quality for web archives: A grounded theory study. In: Hall, M., Merčun, T., Risse, T., Duchateau, F. (eds.) Digital Libraries for Open Knowledge. pp. 73–86. Springer International Publishing, Cham (2020)
6. Reyes Ayala, B., Hitchcock, E., Sun, J.: Using image similarity metrics to measure visual quality in web archives. In: JCDL 2019: Web Archiving and Digital Libraries (WADL) workshop. pp. 11–13. ACM (2019). https://doi.org/10.7939/r3-yh2n-rx10
7. Tomar, S.: Converting video formats with ffmpeg. Linux Journal **2006**(146), 10 (2006)