

Visual Web Archive Quality Assessment

TPDL, September 20th–23rd, 2022



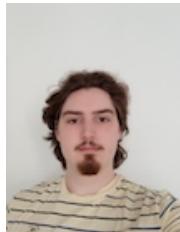
Theresa
Elstner¹



Johannes
Kiesel²



Lars
Meyer²



Max
Martius¹



Sebastian
Schmidt¹



Benno
Stein²



Martin
Potthast¹

¹Universität Leipzig
first.last@uni-leipzig.de

²Bauhaus-Universität Weimar
first.last@uni-weimar.de

Visual Web Archive Quality Assessment

What is Web Archive Quality?

The screenshot shows a news article titled "Creators Remix Roundup: Top Remixes of The Year". It includes a paragraph about the creators' inbox overflowing with alerts, a video thumbnail for "Florence + the Machine: You've Got The Love" (Jamie xx remix feat. The xx), and a second video thumbnail for "Charlotte Gainsbourg: Trick Pony" (Boys Noize remix). The text for the first video discusses its heavy rotation on iPods and debut on their documentary.

Original Screenshot

The screenshot shows the same news article, but the text has been heavily modified. The first video's text now discusses Jamie xx's rework floating around the internet. The second video's text discusses the original recording of "Trick Pony" off her 2009 Beck-produced album *IRM*.

Archive Screenshot

Visual Web Archive Quality Assessment

Pixel Difference

Creators VICE

MUSIC

Creators Remix Roundup: TopRemixes of The Year

<p>Including Florence + the Machine, DJ Wordy, Glasser, Major Lazer, Mark Ronson, G-Dragon, Gang Gang Dance and James Lavelle.</p>

[SHARE](#) [TWEET](#) [TWITTER](#)

Our Creators are a talented and prolific bunch, and our inbox is always overflowing with alerts of new remixes and mashups from the incredible DJs and producers in our line-up. We just couldn't keep these fresh new tunes to ourselves because, after all, filesharing is caring. To commemorate the one year anniversary of the Creators Remix Roundup (which happens to fall on the magical day of 11-11-11, no less) we decided to revisit our favorite tunes from the past year... which ones were your favorites?

Florence + the Machine: "You've Got The Love" (Jamie xx remix feat. The xx)



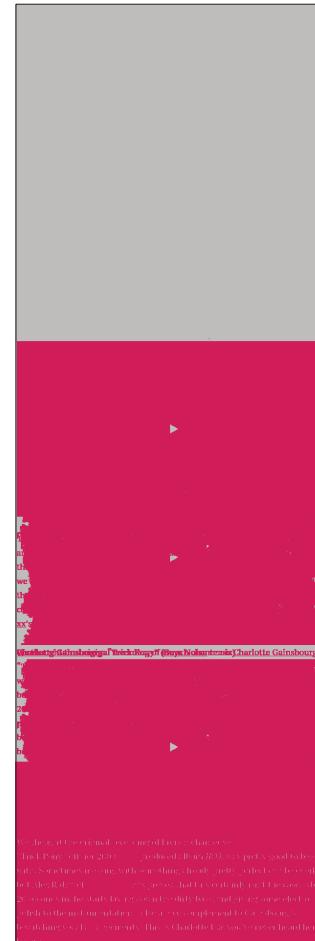
There seems to be a few different adaptations of Jamie xx's rework floating around the internet, but this 2-step version has to be our favorite. Last winter this track was in heavy, heavy rotation on our iPods and since just this morning we debuted our documentary on the enchanting Florence + the Machine, we thought it was the perfect time to bring it back. The original feels like a comforting, glamorous lullaby, but we like the added male vocals (from The xx's Oliver Sim) and Jamie's glitchy loops.

Charlotte Gainsbourg: "Trick Pony" (Boys Noize remix)



We thought the original recording of French chanteuse Charlotte Gainsbourg's "Trick Pony" off her 2009 Beck-produced album *JRM*, was pretty good to begin with. Sometimes messing with something already pretty perfect can be overkill, but Alex Ridha of Boys Noize has proved that this certainly isn't the case. About 20 seconds in, he starts laying down the dirty bass, and giving some electro polish to the instrumentation—also a nice complement to Gainsbourg's bewitching vocal arrangements. This is Charlotte like you've never heard her before.

Original Screenshot



Pixel Difference

Creators VICE

MUSIC

Creators Remix Roundup: TopRemixes of The Year

<p>Including Florence + the Machine, DJ Wordy, Glasser, Major Lazer, Mark Ronson, G-Dragon, Gang Gang Dance and James Lavelle.</p>

[SHARE](#) [TWEET](#) [TWITTER](#)

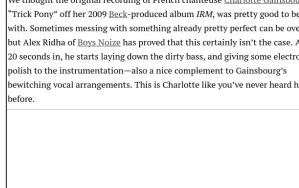
Our Creators are a talented and prolific bunch, and our inbox is always overflowing with alerts of new remixes and mashups from the incredible DJs and producers in our line-up. We just couldn't keep these fresh new tunes to ourselves because, after all, filesharing is caring. To commemorate the one year anniversary of the Creators Remix Roundup (which happens to fall on the magical day of 11-11-11, no less) we decided to revisit our favorite tunes from the past year... which ones were your favorites?

Florence + the Machine: "You've Got The Love" (Jamie xx remix feat. The xx)



There seems to be a few different adaptations of Jamie xx's rework floating around the internet, but this 2-step version has to be our favorite. Last winter this track was in heavy, heavy rotation on our iPods and since just this morning we debuted our documentary on the enchanting Florence + the Machine, we thought it was the perfect time to bring it back. The original feels like a comforting, glamorous lullaby, but we like the added male vocals (from The xx's Oliver Sim) and Jamie's glitchy loops.

Charlotte Gainsbourg: "Trick Pony" (Boys Noize remix)



We thought the original recording of French chanteuse Charlotte Gainsbourg's "Trick Pony" off her 2009 Beck-produced album *JRM*, was pretty good to begin with. Sometimes messing with something already pretty perfect can be overkill, but Alex Ridha of Boys Noize has proved that this certainly isn't the case. About 20 seconds in, he starts laying down the dirty bass, and giving some electro polish to the instrumentation—also a nice complement to Gainsbourg's bewitching vocal arrangements. This is Charlotte like you've never heard her before.

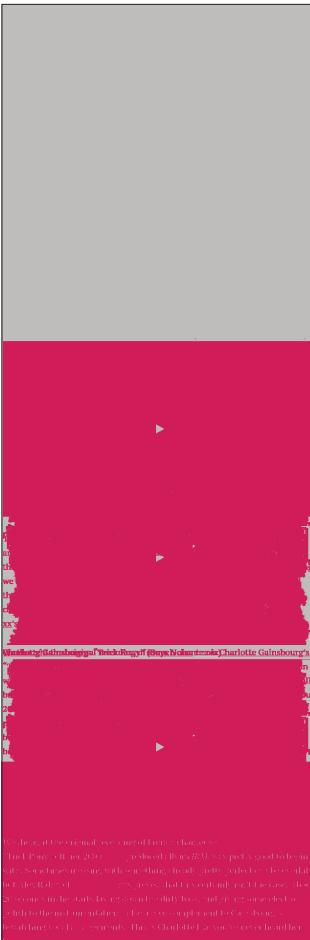
Archive Screenshot

Visual Web Archive Quality Assessment

From Pixel Difference to Reproduction Errors



Original Screenshot



Pixel Difference



Archive Screenshot

Visual Web Archive Quality Assessment

From Pixel Difference to Reproduction Errors



Original Screenshot



Pixel Difference



Archive Screenshot

Creators Remix Roundup: Top Remixes of The Year

<sp>Including Florence + the Machine, DJ Wordy, Glasser, Major Lazer, Mark Ronson, G-Dragon, Gang Gang Dance and James Lavelle.</p>

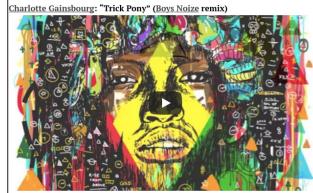
SHARE TWEET

*Our Creators are a talented and prolific bunch, and our inbox is always overflowing with alerts of new remixes and mashups from the incredible DJs and producers in our line-up. We just couldn't keep these fresh new tunes to ourselves because, after all, **filesharing is caring**. To commemorate the one year anniversary of the Creators Remix Roundup (which happens to fall on the magical day of 11-11-11, no less) we decided to revisit our favorite tunes from the past year... which ones were your favorites?*



MISSING

There seems to be two different adaptations of Jamie xx's "You've Got The Love" floating around the internet, but this is our favorite. Last winter this track was in heavy, heavy rotation on our iPods and since just this morning we debuted our documentary on the enchanting Florence + the Machine, we thought it was the perfect time to bring it back. The original feels like a comforting, glamourous lullaby, but we like the added male vocals (from The xx's Oliver Sim) and Jamie's glitchy loops.



*We thought the original recording of French chanteuse Charlotte Gainsbourg's "Trick Pony" off her 2009 Beck-produced album *IRM*, was pretty good to begin with. Sometimes messing with something already pretty perfect can be overkill, but Alex Ridha of Boys Noize has proved that this certainly isn't the case. About 20 seconds in, he starts laying down the dirty bass, and giving some electro polish to the instrumentation—also a nice complement to Gainsbourg's bewitching vocal arrangements. This is Charlotte like you've never heard her before.*

Creators Remix Roundup: Top Remixes of The Year

<sp>Including Florence + the Machine, DJ Wordy, Glasser, Major Lazer, Mark Ronson, G-Dragon, Gang Gang Dance and James Lavelle.</p>

SHARE TWEET

*Our Creators are a talented and prolific bunch, and our inbox is always overflowing with alerts of new remixes and mashups from the incredible DJs and producers in our line-up. We just couldn't keep these fresh new tunes to ourselves because, after all, **filesharing is caring**. To commemorate the one year anniversary of the Creators Remix Roundup (which happens to fall on the magical day of 11-11-11, no less) we decided to revisit our favorite tunes from the past year... which ones were your favorites?*

Florence + the Machine: "You've Got The Love" (Jamie xx remix feat. The xx)

There seems to be a few different versions of Jamie xx's rework floating around the internet, this 2-step version has to be our favorite. Last winter this track was in heavy, heavy rotation on our iPods and since just this morning we debuted our documentary on the enchanting Florence + the Machine, we thought it was the perfect time to bring it back. The original feels like a comforting, glamourous lullaby, but we like the added male vocals (from The xx's Oliver Sim) and Jamie's glitchy loops.

Charlotte Gainsbourg: "Trick Pony" (Boys Noize remix)

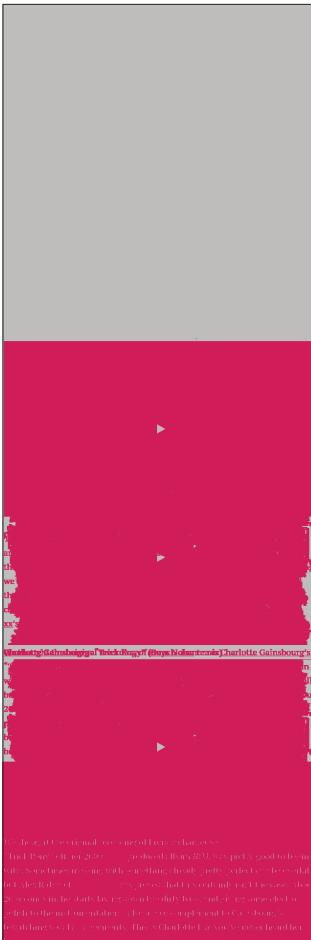
*We thought the original recording of French chanteuse Charlotte Gainsbourg's "Trick Pony" off her 2009 Beck-produced album *IRM*, was pretty good to begin with. Sometimes messing with something already pretty perfect can be overkill, but Alex Ridha of Boys Noize has proved that this certainly isn't the case. About 20 seconds in, he starts laying down the dirty bass, and giving some electro polish to the instrumentation—also a nice complement to Gainsbourg's bewitching vocal arrangements. This is Charlotte like you've never heard her before.*

Visual Web Archive Quality Assessment

From Pixel Difference to Reproduction Errors



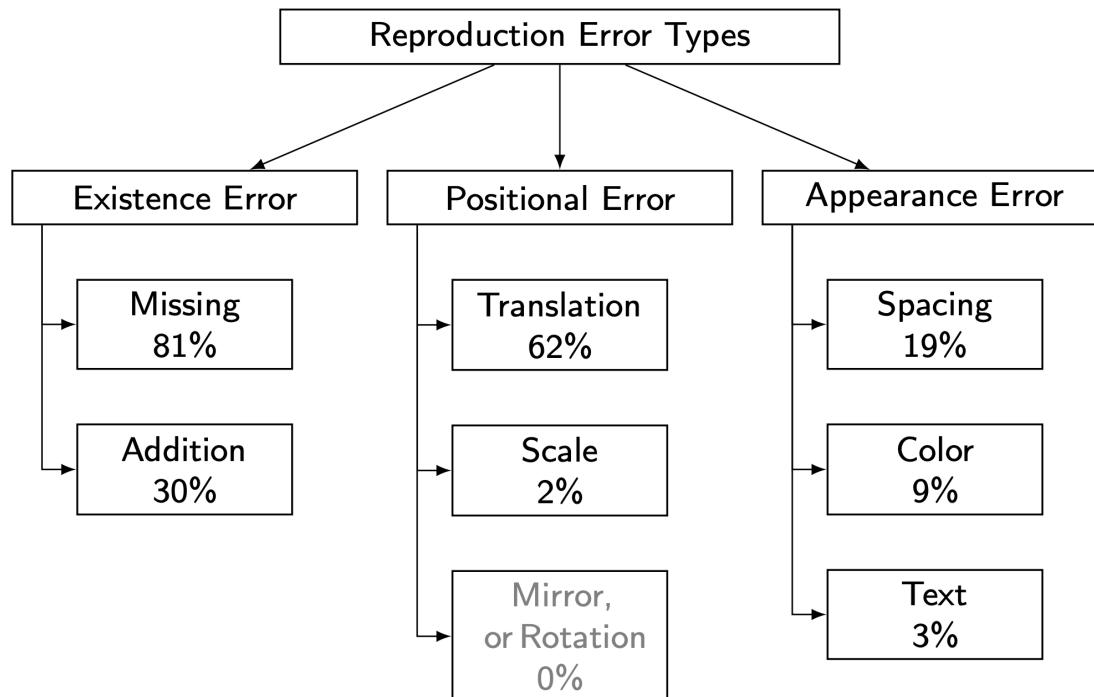
Original Screenshot



Archive Screenshot

Visual Web Archive Quality Assessment

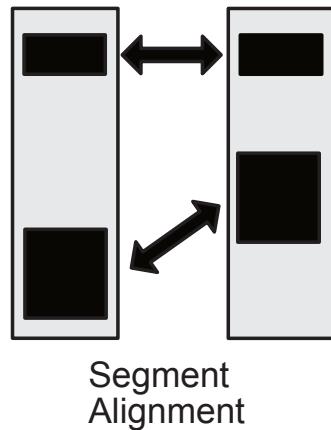
Reproduction Errors



Visual Alignment of Original–Archive Screenshot Pairs

Generic Framework for Quality Assessment

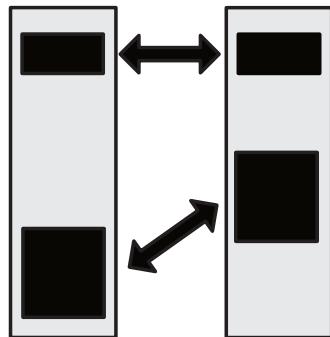
- Step 1: Visual segment alignment



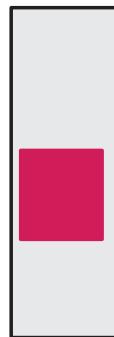
Visual Alignment of Original–Archive Screenshot Pairs

Generic Framework for Quality Assessment

- Step 1: Visual segment alignment
- Step 2: Visual Edit Distance calculation



Segment
Alignment

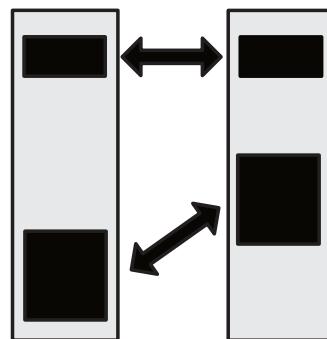


Visual Edit Distance
Calculation

Visual Alignment of Original–Archive Screenshot Pairs

Generic Framework for Quality Assessment

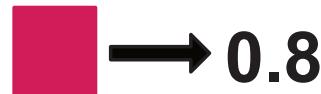
- Step 1: Visual segment alignment
- Step 2: Visual Edit Distance calculation
- Step 3: Quality prediction



Segment
Alignment



Visual Edit Distance
Calculation



Quality
Prediction

Visual Alignment of Original–Archive Screenshot Pairs

Data

- Original*–Archive screenshot pairs
- Human annotations*
- Description of structure

Total: 6531 web pages¹

¹(*) from Webis-Web-Archive-17: Johannes Kiesel, Florian Kneist, Milad Alshomary, Benno Stein, Matthias Hagen, and Martin Potthast. Reproducible Web Corpora: Interactive Archiving with Automatic Quality Assessment. Journal of Data and Information Quality (JDIQ), 10 (4) : 17:1-17:25, October 2018

Visual Alignment of Original–Archive Screenshot Pairs

Prototype Step 1: Visual Segment Alignment

Our Prototype:

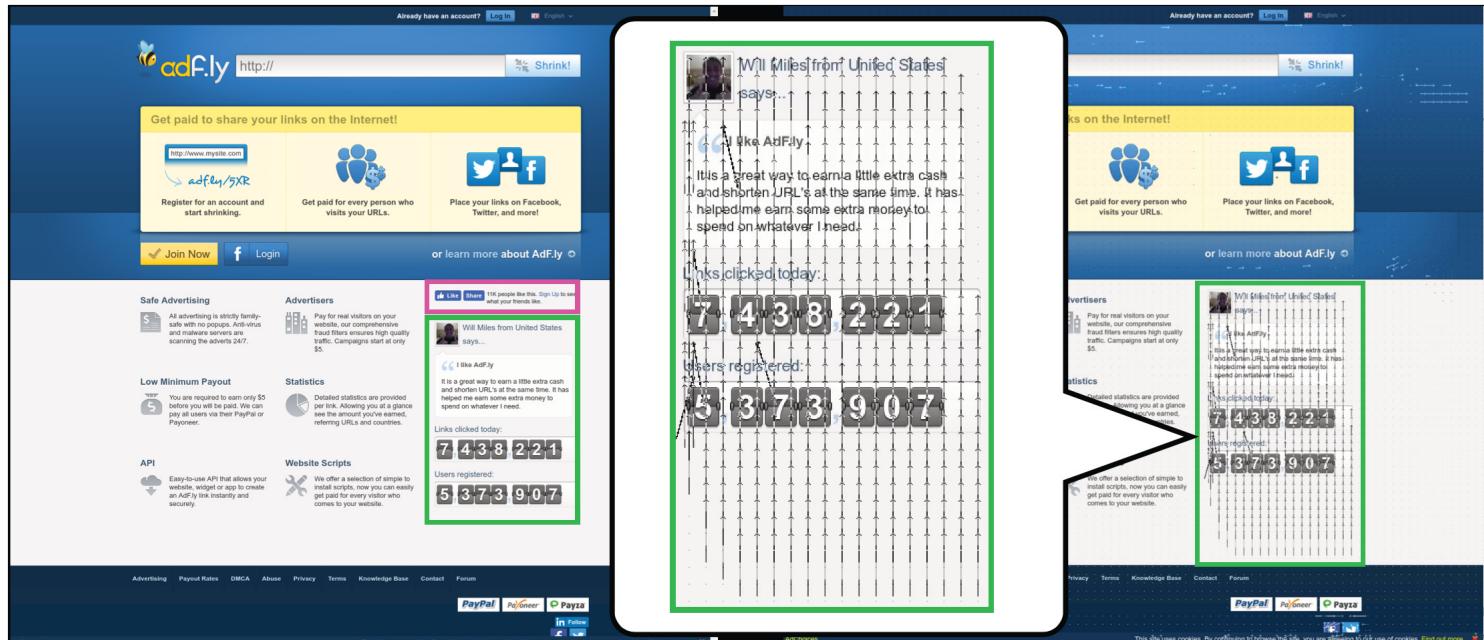
- ❑ Move translated DOM elements in archive screenshots to their respective positions in the original screenshot:
 - Derive positions of elements in the archive screenshots from the page structure descriptions
 - Derive positions of elements in the original screenshots using video encoding

How?

- ❑ Encode original–archive screenshots into an H.264 encoded video
- ❑ Obtain motion vectors with information on translated pixel blocks

Visual Alignment of Original–Archive Screenshot Pairs

Prototype Step 1: Visual Segment Alignment



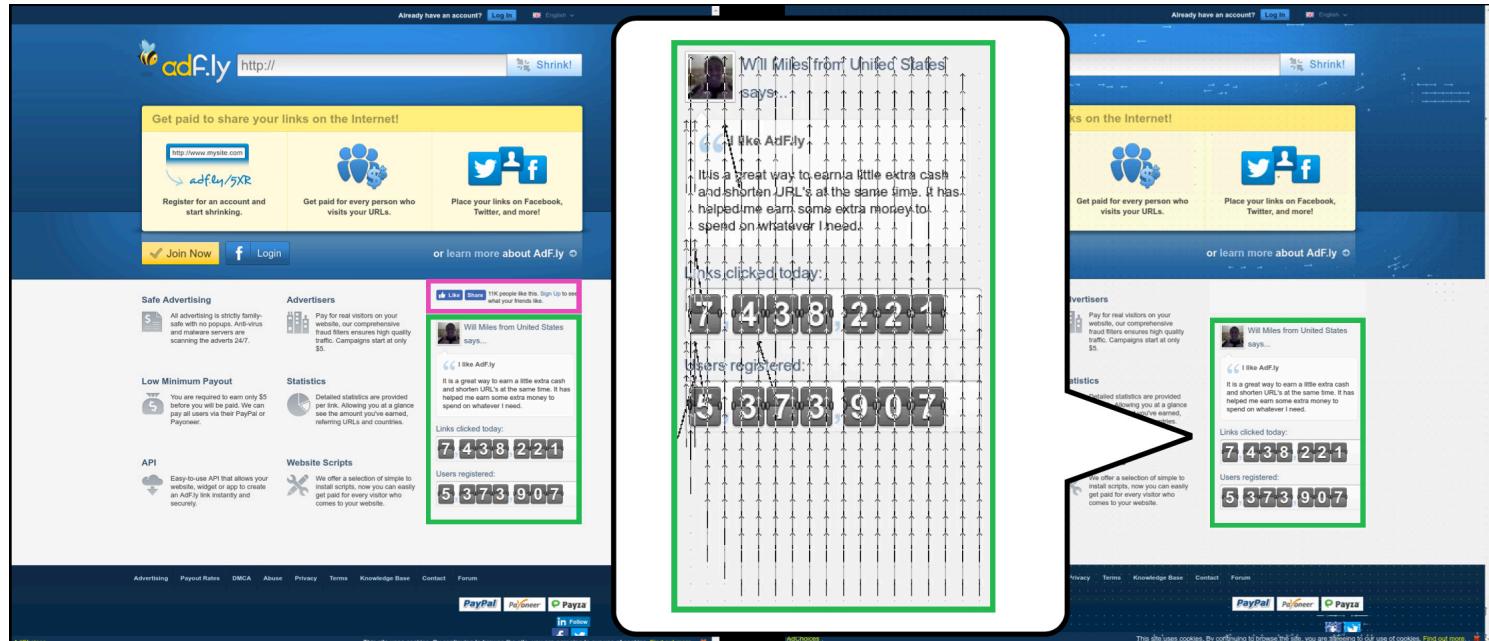
Frame 1: Original Screenshot

Frame 2: Archive Screenshot

- Original and archive screenshot as two frames of H.264 encoded video
- Missing element (pink) causes shift of elements below (green)
- Motion vectors indicate the original position of the shifted element

Visual Alignment of Original–Archive Screenshot Pairs

Prototype Step 1: Visual Segment Alignment



Aligned Screenshot

- Aligned screenshot without shifted element

Visual Alignment of Original–Archive Screenshot Pairs

Prototype Step 2: Visual Edit Distance Calculation

- Count the number of changed pixels on
 - Archive–Original screenshot pair (Unaligned Visual Edit Distance)
 - Aligned–Original screenshot pair (Aligned Visual Edit Distance)

Visual Alignment of Original–Archive Screenshot Pairs

Prototype Step 3: Quality Prediction

- Linear regression, 10-fold cross validation
- Prediction target: human annotations on 5-point scale (1 perfect reproduction, 5 unusable)
- Input features:
 - Unaligned Regression:
 - Size difference of unaligned archive and original screenshots
 - Unaligned visual edit distance (normalized)
 - Aligned Regression:
 - Aligned visual edit distance (normalized)

Evaluation

Improvement of Aligned Regression over Unaligned Regression:

Truth	Predicted quality				
	1	2	3	4	5
1 (<i>not affecting the visitor</i>)	+1	+4	-6	+1	±0
2 (<i>small effect on a few visitors</i>)	-108	+115	-8	±0	+1
3 (<i>small effect on many or all visitors</i>)	+2	-7	+4	±0	+1
4 (<i>affect, but page can still be used</i>)	-10	+7	+3	-3	+3
5 (<i>unusable page</i>)	-7	+7	+3	-3	±0

- Correct classifications increased in the aligned regression
- Main improvement of the aligned regression lies in a better differentiation between quality 1 and 2

Evaluation

Improvement of Aligned Regression over Unaligned Regression:

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)

- Improvements for identifying reproductions of quality scores 2–5 and 3–5
- Reduced scores for identifying reproductions of quality scores 4–5 and 5

Conclusion

- ❑ Explored an approach to automatic quality assessment of archived web pages using their screenshots
- ❑ Categorized visually perceivable reproduction errors
- ❑ Presented a 3-step framework for automatic quality assessment
- ❑ Implemented a prototype
- ❑ Comparison of aligned and unaligned linear regression models yields small improvement for our prototype

Conclusion

- ❑ Explored an approach to automatic quality assessment of archived web pages using their screenshots
- ❑ Categorized visually perceivable reproduction errors
- ❑ Presented a 3-step framework for automatic quality assessment
- ❑ Implemented a prototype
- ❑ Comparison of aligned and unaligned linear regression models yields small improvement for our prototype

Limitations

- ❑ Prototype considers translations only
- ❑ Linear regression might be outperformed by more sophisticated systems like CNNs
- ❑ Video encoding does not detect all translations

THANK YOU!