Visual Web Archive Quality Assessment

TPDL, September 20th–23rd, 2022
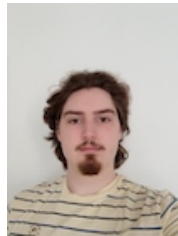
Theresa Elstner[1]   Johannes Kiesel[2]   Lars Meyer[2]   Max Martius[1]   Sebastian Schmidt[1]   Benno Stein[2]   Martin Potthast[1]
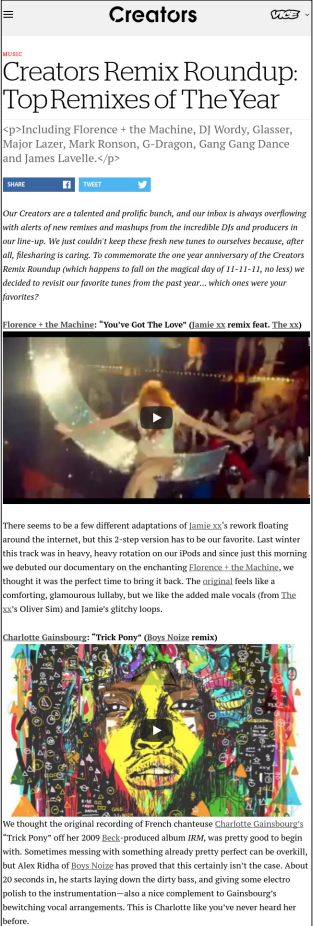
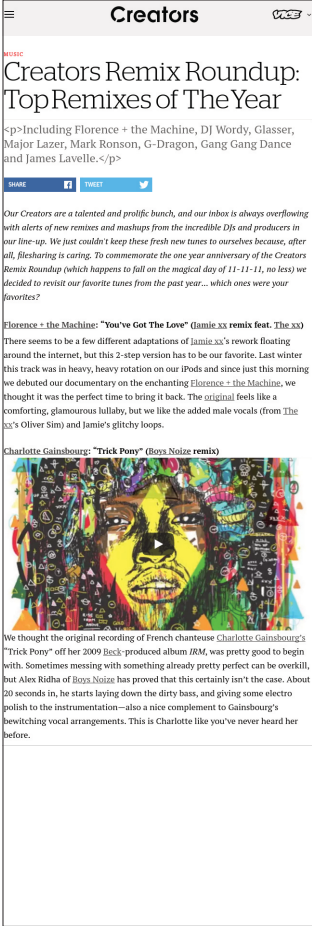[1]Universität Leipzig
first.last@uni-leipzig.de

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

@webis_de   www.webis.de

Visual Web Archive Quality Assessment
What is Web Archive Quality?

Original Screenshot

Archive Screenshot
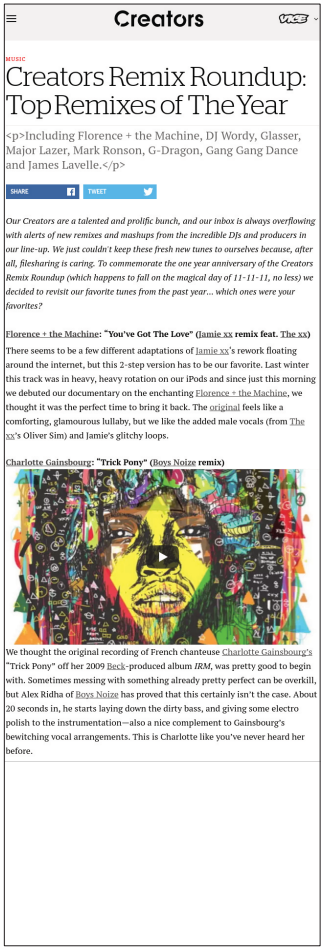
Visual Web Archive Quality Assessment
Pixel Difference

Original Screenshot

Pixel Difference

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
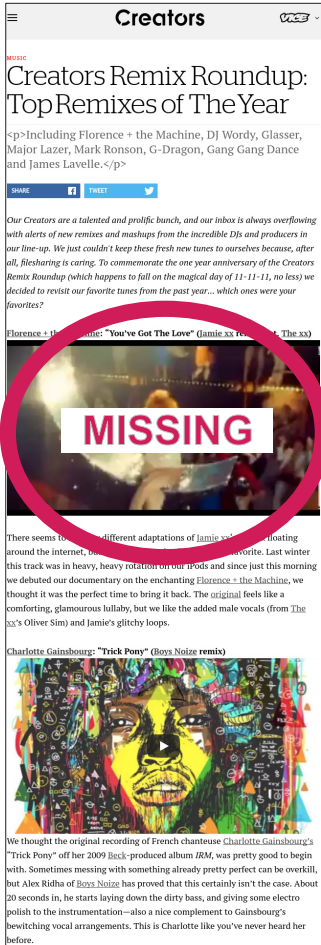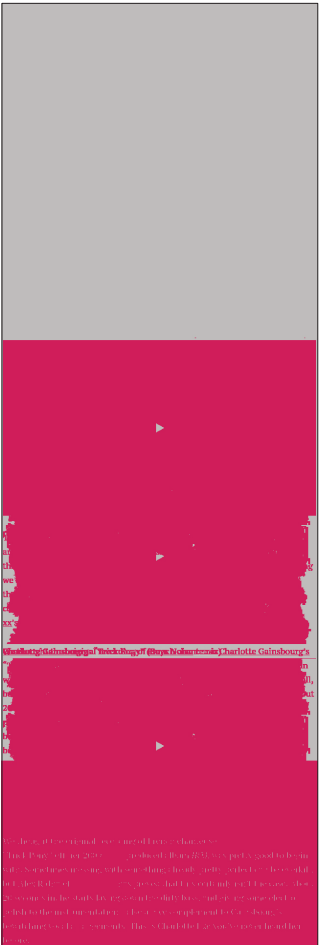

Original Screenshot

Pixel Difference

Archive Screenshot

Visual Web Archive Quality Assessment
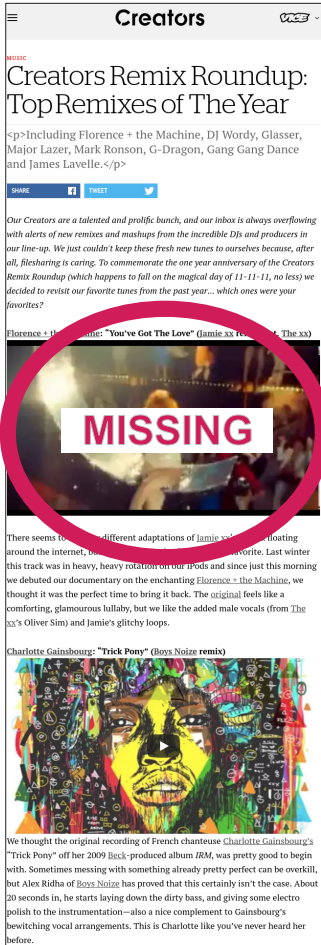## From Pixel Difference to Reproduction Errors

Original Screenshot   Pixel Difference   Archive Screenshot
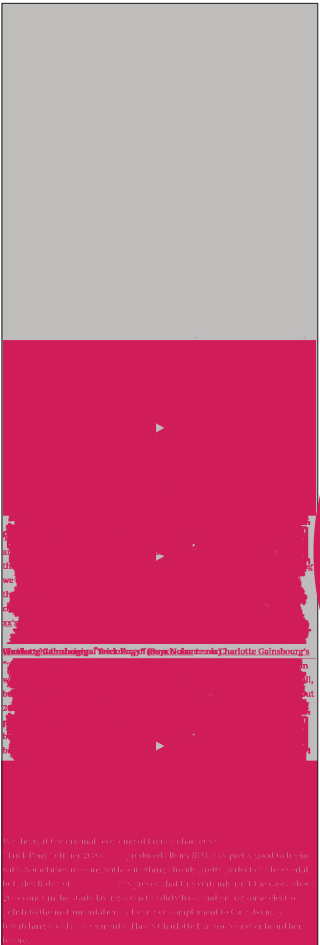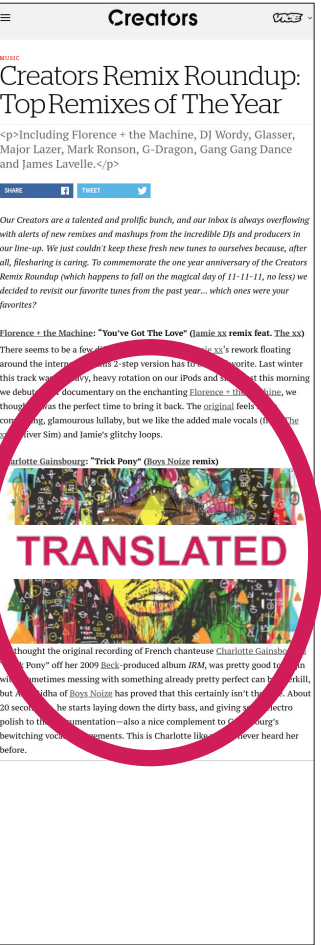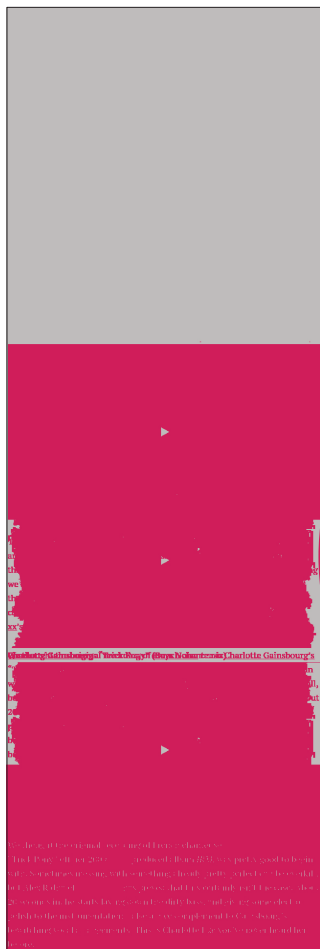
Visual Web Archive Quality Assessment
Reproduction Errors
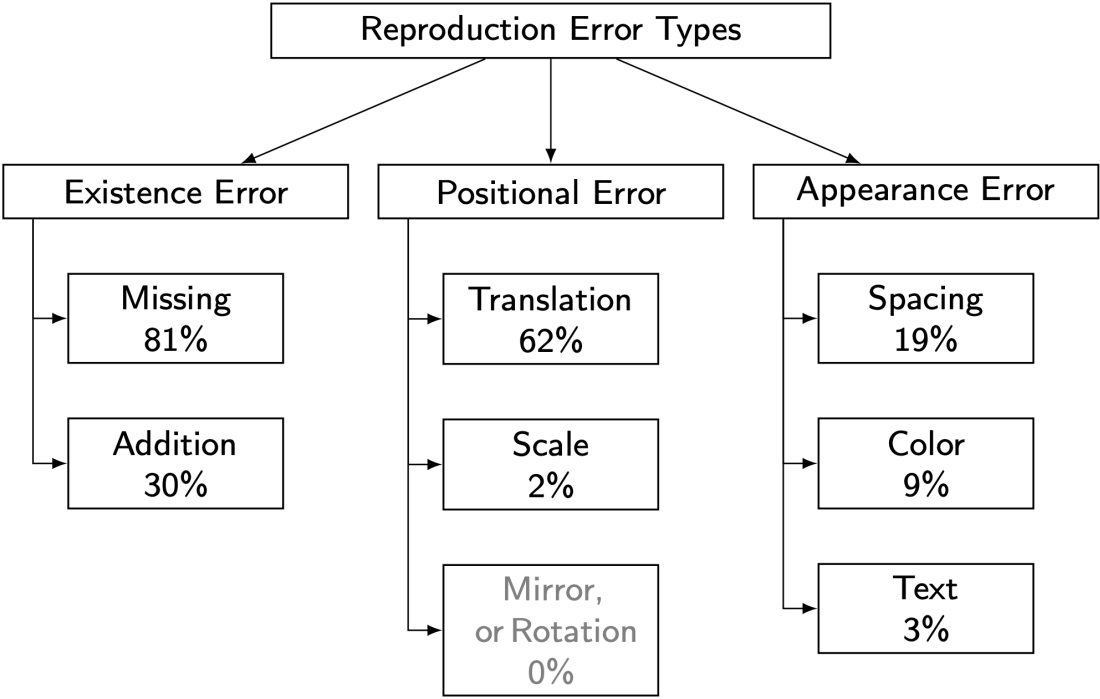
Visual Alignment of Original–Archive Screenshot Pairs

Generic Framework for Quality Assessment

❏ Step 1: Visual segment alignment
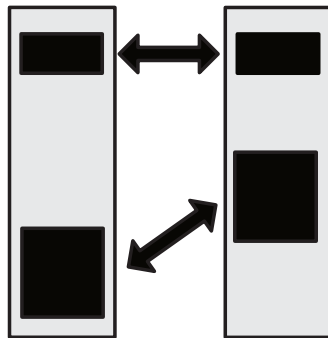
Segment
Alignment

Visual Alignment of Original–Archive Screenshot Pairs

Generic Framework for Quality Assessment

- ❑ Step 1: Visual segment alignment
- ❑ Step 2: Visual Edit Distance calculation
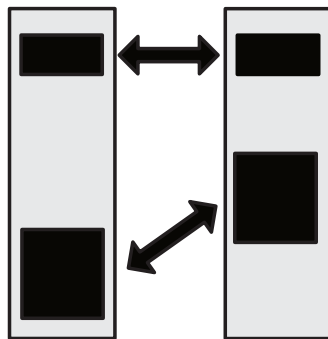
Segment
Alignment

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
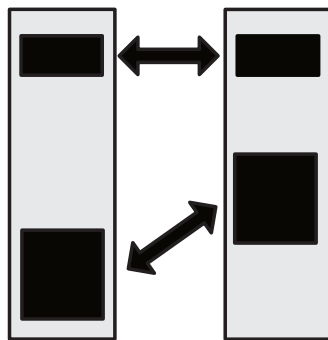
Segment
Alignment

Visual Edit Distance
Calculation
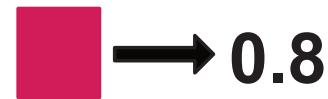
Quality
Prediction

Visual Alignment of Original–Archive Screenshot Pairs
Data

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

Total: 6531 web pages[1]

---

[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
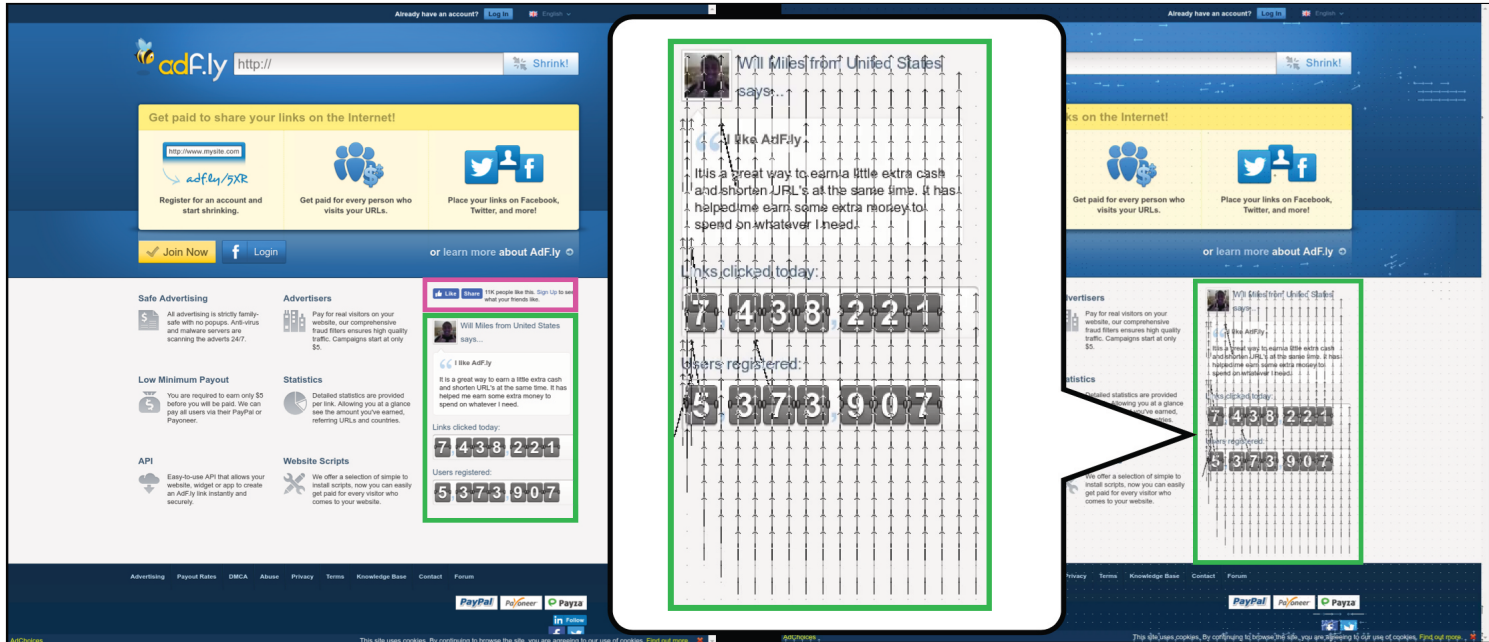
Frame 1: Original Screenshot   Frame 2: Archive Screenshot

- 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
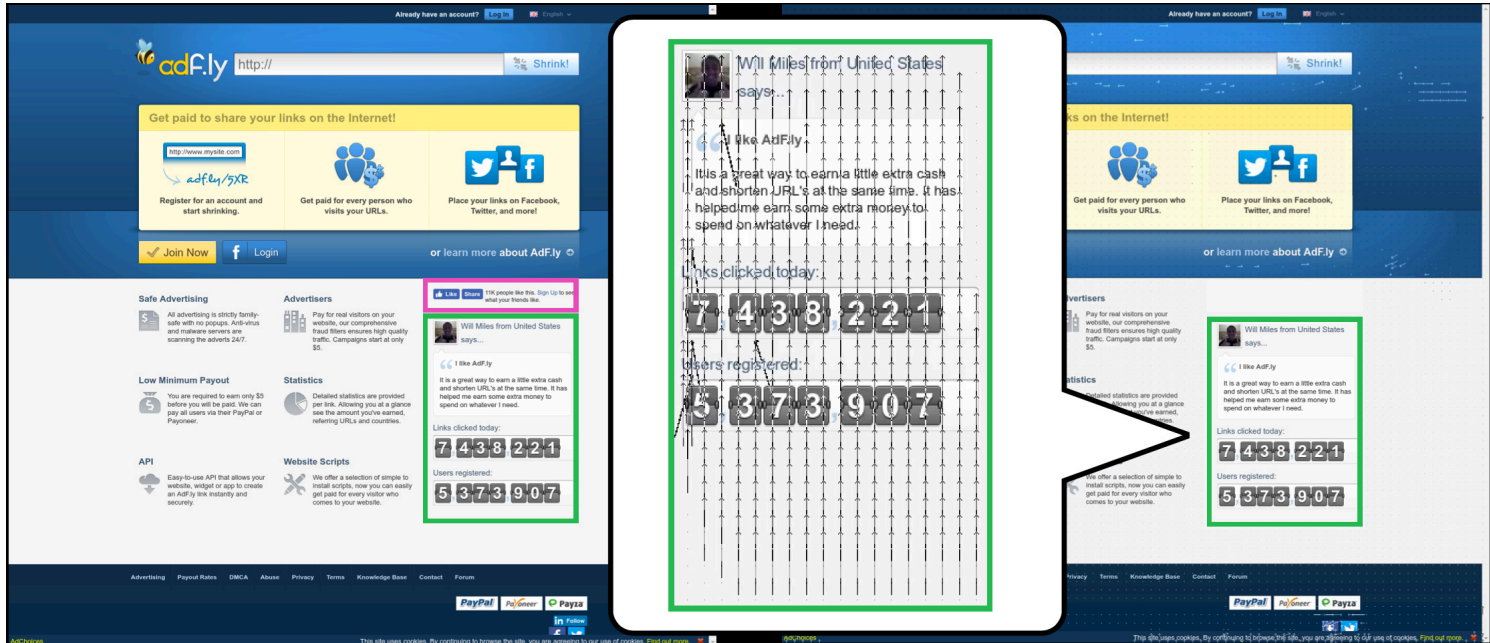
Aligned Screenshot

❏ 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 *(not affecting the visitor)*	+1	+4	−6	+1	±0
2 *(small effect on a few visitors)*	−108	+115	−8	±0	+1
3 *(small effect on many or all visitors)*	+2	−7	+4	±0	+1
4 *(affect, but page can still be used)*	−10	+7	+3	−3	+3
5 *(unusable page)*	−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 ($\triangle$)	Recall ($\triangle$)	F_1 ($\triangle$)
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!