

MARCH 17, 2016 ECSS SYMPOSIUM

AUTOMATIC QUALITY ASSESSMENT OF DIGITAL VIDEO COLLECTIONS

Texas Advanced Computing Center:

ECSS: Anne Bowen, adb@tacc.utexas.edu

ECSS: John Lockman john@vizias.com

PI Maria Esteva, maria@tacc.utexas.edu

Brian Abel (ICERT-REU intern) UT undergrad

UT Austin Laboratory for Image & Video Engineering :

PI Alan Bovik, bovikac@gmail.com

Todd Goodall, tgoodall@utexas.edu

Jaesong Lee, jason.lee27@utexas.edu (summer contact)

More details:

Maria Esteva, Anne Bowen, TACC; Todd Goodall, Alan Bovik, LIVE; Brian Abel, UT Austin

Evaluation of Non-Reference Quality Assessment Algorithms to Curate Born-Digital Video Collections, Archiving Conference, Volume 2015, Number 1, May 2015, pp. 124-129(6)

OVERVIEW

- ▶ Maria Esteva (Data curation specialist at TACC) approached Alan Bovik (director of [UT Austin's LIVE Laboratory](#) for Image and Video Engineering) to assess the use of I/VQA (Image/Video Quality Assessment) algorithms on museum digital video collections.
- ▶ ECSS support was sought to provide assistance in implementing a large-scale workflow that would use the LIVE algorithms and could be automated and the results easily interpreted (the interpretation of the results ended up being the bulk of the project)

COMPUTATIONAL METHOD AND TEST-BED:

- ▶ **Algorithm:** Alan recommended starting with BRISQUE (Blind/Reference-less Image Spatial Quality Evaluator[1]) a type of Non-Reference Image and Video Quality Algorithms (I/VQA)

1) Mittal, A. K. Moorthy and A. C. Bovik, “ *No-Reference Image Quality Assessment in the Spatial Domain*”, IEEE Transactions on ImageProcessing, 21 (12) Dec. 2012

many other algorithms developed by the LIVE lab: <http://live.ece.utexas.edu/research/Quality/index.htm>

- ▶ **Exemplar Test-bed:** Blanton Museum of Art Archival Video Collection from DVDs (digital versions of the originals). Challenging: Diverse in content and quality. Artistic.

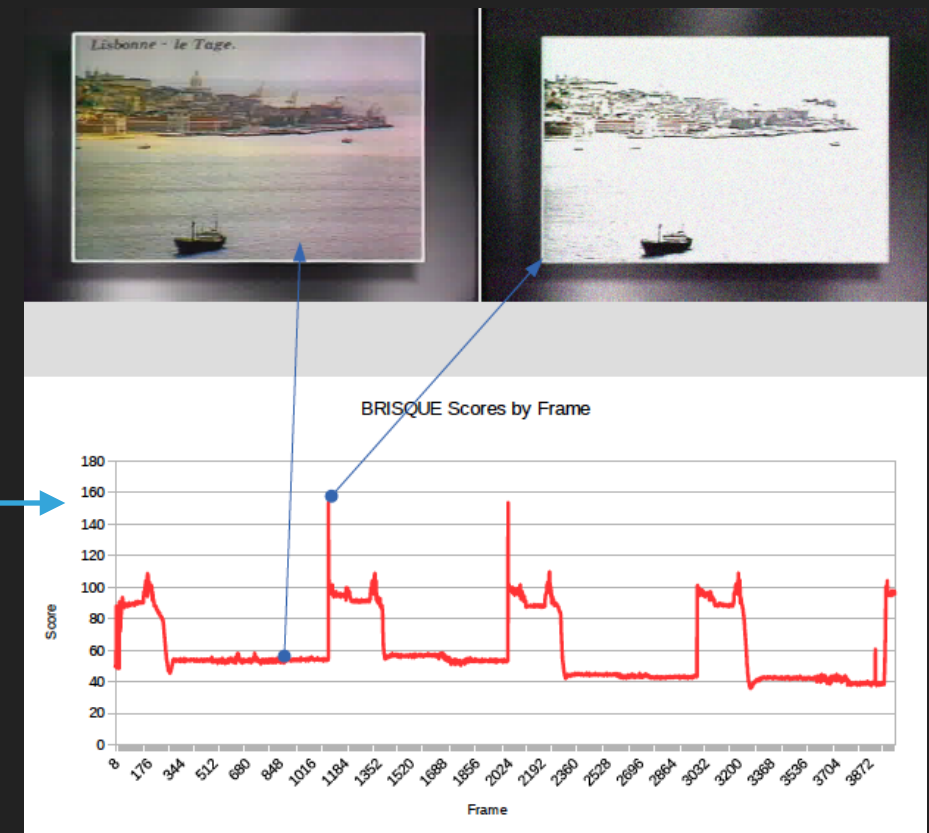
DETAILS OF BRISQUE (AND I/VQA) ALGORITHMS

- ▶ Rather than detecting errors based on distortion-specific filters, these algorithms use perceptual subjective measures (NSS, "Natural Scene Statistics") based on models of the human visual system to assess quality.
- ▶ The scores produced by the I/VQA algorithms are statistically significant through their correlation with the consensus scores obtained from people that have rated the distortions in reference video sets.

WHAT WE WANTED/HOPED

EXPECTATIONS

- ▶ Overall mean video score could indicate overall video quality.
- ▶ For movies with high deviation, bad frames could be identified.



GOALS

- ▶ BRISQUE (or related I/VQA algorithm) derived scores could be used to show classes of conditions present in a digital video collection (e.g. a digitizing artifact)
- ▶ The workflows could be automated so that curators could use these tools to assist in assessing the quality of large digital video collections.

INITIAL WORKFLOW

Digitized video collection from the Blanton archives (5TB) uploaded to stampede and parametric job “launcher” used to execute scripts. Efficiency tests found that 8 cores/node optimum (bottleneck is ffmpeg frame extraction), each compute node processes two videos at a time and each node can analyze ~40 HD frames per second. BRISQUE was modified slightly allow for parallel execution and easier parsing of output file.

PREPARE/PROCESS DIGITAL VIDEO (FRAMES OR VIDEO)

▶ [curate.py](#)

- ▶ Extract meta-data and header using EXIFtool and/or ffmpeg
- ▶ Frame extraction using ffmpeg (using metadata such as frame-rate).

```
login3.stampede(49)$ cat frames_2003.77.process frames_2003.91.process frames_PG2007.5.1.process
2003.77.iso frames_2003.91.process frames_PG2007.5.1.process
2003.77.mp4 frames_2003.91.process frames_PG2007.5.1.process
2003.91.iso frames_2003.91.process frames_PG2007.5.1.process
2003.91.mp4 frames_2003.91.process frames_PG2007.5.1.process
2004.113.iso frames_2004.113.process frames_PG2007.5.1.process
2004.113.mp4 frames_2004.113.process frames_PG2007.5.1.process
2004.130.iso frames_2004.130.process frames_PG2007.5.1.process
2004.130.mp4 frames_2004.130.process frames_PG2007.5.1.process
2004.92.iso frames_2004.92.process frames_PG2007.5.1.process
2004.92.mp4 frames_2004.92.process frames_PG2007.5.1.process
2005.201.iso frames_2005.201.process frames_PG2007.5.1.process
2005.201.mp4 frames_2005.201.process frames_PG2007.5.1.process
2007.19.iso frames_2007.19.process frames_PG2007.5.1.process
2007.19.mp4 frames_2007.19.process frames_PG2007.5.1.process
2007.28.iso frames_2007.28.process frames_PG2007.5.1.process
2007.28.mp4 frames_2007.28.process frames_PG2007.5.1.process
2007.29.iso frames_2007.29.process frames_PG2007.5.1.process
2007.29.mp4 frames_2007.29.process frames_PG2007.5.1.process
2008.113_orchestra.iso frames_2008.113_orchestra.process frames_PG2007.5.1.process
2008.113_orchestra.mp4 frames_2008.113_orchestra.process frames_PG2007.5.1.process
2008.113.iso frames_2008.113.process frames_PG2007.5.1.process
2008.113.mp4 frames_2008.113.process frames_PG2007.5.1.process
2008.114.iso frames_2008.114.process frames_PG2007.5.1.process
2008.114.mp4 frames_2008.114.process frames_PG2007.5.1.process
2008.115.iso frames_2008.115.process frames_PG2007.5.1.process
2008.115.mp4 frames_2008.115.process frames_PG2007.5.1.process
2008.137.iso frames_2008.137.process frames_PG2007.5.1.process
2008.137.mp4 frames_2008.137.process frames_PG2007.5.1.process
2009.1.iso frames_2009.1.process frames_PG2007.5.1.process
2009.1.mp4 frames_2009.1.process frames_PG2007.5.1.process
2009.5.iso frames_2009.5.process frames_PG2007.5.1.process
2009.5.mp4 frames_2009.5.process frames_PG2007.5.1.process
login3.stampede(49)$
```

RUN I/VQA ALGORITHM

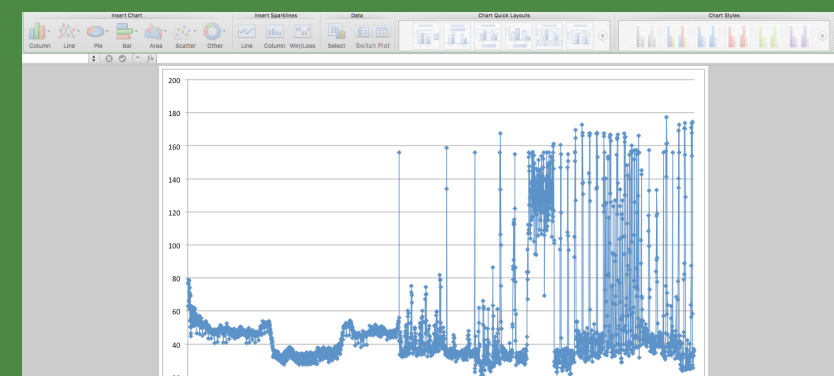
▶ [analyze.py](#)

- ▶ Run BRISQUE analysis on every frame
- ▶ Compute BRISQUE statistics per movie

```
login3.stampede(49)$ cat frames_2003.77.process frames_2003.91.process frames_PG2007.5.1.process
2003.77.iso frames_2003.91.process frames_PG2007.5.1.process
2003.77.mp4 frames_2003.91.process frames_PG2007.5.1.process
2003.91.iso frames_2003.91.process frames_PG2007.5.1.process
2003.91.mp4 frames_2003.91.process frames_PG2007.5.1.process
2004.113.iso frames_2004.113.process frames_PG2007.5.1.process
2004.113.mp4 frames_2004.113.process frames_PG2007.5.1.process
2004.130.iso frames_2004.130.process frames_PG2007.5.1.process
2004.130.mp4 frames_2004.130.process frames_PG2007.5.1.process
2004.92.iso frames_2004.92.process frames_PG2007.5.1.process
2004.92.mp4 frames_2004.92.process frames_PG2007.5.1.process
2005.201.iso frames_2005.201.process frames_PG2007.5.1.process
2005.201.mp4 frames_2005.201.process frames_PG2007.5.1.process
2007.19.iso frames_2007.19.process frames_PG2007.5.1.process
2007.19.mp4 frames_2007.19.process frames_PG2007.5.1.process
2007.28.iso frames_2007.28.process frames_PG2007.5.1.process
2007.28.mp4 frames_2007.28.process frames_PG2007.5.1.process
2007.29.iso frames_2007.29.process frames_PG2007.5.1.process
2007.29.mp4 frames_2007.29.process frames_PG2007.5.1.process
2008.113_orchestra.iso frames_2008.113_orchestra.process frames_PG2007.5.1.process
2008.113_orchestra.mp4 frames_2008.113_orchestra.process frames_PG2007.5.1.process
2008.113.iso frames_2008.113.process frames_PG2007.5.1.process
2008.113.mp4 frames_2008.113.process frames_PG2007.5.1.process
2008.114.iso frames_2008.114.process frames_PG2007.5.1.process
2008.114.mp4 frames_2008.114.process frames_PG2007.5.1.process
2008.115.iso frames_2008.115.process frames_PG2007.5.1.process
2008.115.mp4 frames_2008.115.process frames_PG2007.5.1.process
2008.137.iso frames_2008.137.process frames_PG2007.5.1.process
2008.137.mp4 frames_2008.137.process frames_PG2007.5.1.process
2009.1.iso frames_2009.1.process frames_PG2007.5.1.process
2009.1.mp4 frames_2009.1.process frames_PG2007.5.1.process
2009.5.iso frames_2009.5.process frames_PG2007.5.1.process
2009.5.mp4 frames_2009.5.process frames_PG2007.5.1.process
login3.stampede(49)$
```

▶ collect data, plot (gnuplot not good for diagnostics)

COMPUTE STATISTICS ANALYZE RESULTS



The workflow ran smoothly, but the results were not as expected. Sitting down to analyze the results with Alan was insightful, and made it obvious we needed a better human-centered way to analyze the results

CHALLENGES OF ART COLLECTIONS:

- ▶ Multiple conversions from original to archived version introduced compounded compression artifacts (e.g. a digitized VHS recording), algorithms were not trained for this.
- ▶ Artistic videos might intentionally use compression artifacts for effect (e.g. low-fi storytelling)
- ▶ BRISQUE designed for evaluation of “natural” scenes (e.g. black borders, text overlays, animations), BRISQUE designed of still frames and lacks modeling of distortion related to motion.
- ▶ For human-opinion scoring tests, how do you keep the subjective nature of evaluations interfere with the score given.

(images removed)

VISUAL ANALYSIS TOOL FOR DIAGNOSTICS

- ▶ Needed a better tool to assess the results/scores frame by frame with QA experts (like Alan Bovik and Todd)
- ▶ This post processing tool regenerated a movie from the original source movie and gnuplot generated graph embedded in the upper right along with a “ticker” line to show the current frame and score.



HOW WELL CAN BRISQUE PERFORM UNDER IDEAL CONDITIONS?

- ▶ Enlisted help of Todd Goodall (LIVE PhD student) to identify good test cases for BRISQUE and also to help us to understand/interpret/make sense of the results and fine-tune BRISQUE.
- ▶ Tried the analysis again with movies from public QA assessment databases [CSIQ](#) (OSU Vision lab) and [LIVE VQA](#) which have movies with associated quality ratings (Difference Mean Opinion Scores)

VISUAL EVALUATION

- ▶ We (Anne, Maria, Todd) watched each video using the visual diagnostic tool developed for this project. We did not know what types of distortions were present in the museum collection (aside from MPEG2 present in all), and how the distortions affected the scores. We incorporated the visual identification of distortions in the analysis. and noted the presence of distortions including those in which the algorithm was not trained on (e.g. interlacing, VHS blips, sensor noise and lens flair), and if the video had non-natural scenes such as animations or other special effects, to compare our scores with BRISQUE and anomalies.
- ▶ Conclusions: BRISQUE is not appropriate for videos that have non-natural scenes, and it does not perform well with movies in which there are drastic scene changes or that have noise, interlacing and other distortions that are not captured by the algorithm. The model performed well in videos had distortions on which the algorithm was trained.

CONCLUSIONS AND FUTURE WORK

- ▶ To try to obtain a more objective result from our subjective visual analysis, Todd Goodall tried to correct the BRISQUE results to account for the presence of distortion types it wasn't trained for (process described in the paper) and we repeated the analysis on a smaller subset of the videos.
 - ▶ Conclusion: BRISQUE is very sensitive to noise, it isn't expected to perform well on varied collections with high accuracy, but is expected to assess relative quality within individual videos and across videos that comply with certain characteristics (described on previous slide).
- ▶ A diverse collection would need to be filtered first to provide BRISQUE with appropriate videos (e.g first using NIQE (Naturalness Image Quality Evaluator) algorithm).
- ▶ We are continuing this study with a related algorithm "VideoBLINDS" (using (Blind Image Integrity Notator using DCT Statistics) with Todd Goodall and Jaesong Lee (graduate student at LIVE) and is currently conducting a user-study to collect human opinion scores on a more comprehensive (compression) movie collection that better represents a diverse museum collection and training set for VideoBLINDS.

EXTRA DETAIL: MORE BACKGROUND ON I/VQA ALGORITHMS

- ▶ I/VQA algorithms are based on natural scene statistics (NSS). NSS function under the premise that scenes have statistical regularities and that the human visual system is tuned to note regularities from irregularities. The statistics sensitive to these variations in regularity have been shown to correlate well with difference mean opinion scores (DMOS) of images and video.
- ▶ **Must be trained:** To successfully map these statistics to a single perceptual quality score, these algorithms train on both images and videos that have corresponding opinion scores. These DMOS scores are computed from a set of subjective evaluations obtained from humans watching sets of videos that have specific types and degrees of distortions. These videos are rated using a continuous sliding scale with the labels "Worst," "Poor," "Fair," "Good," and "Excellent."
- ▶ The user scores are combined to compute the DMOS score on the range of [0-100], where 0 is "Excellent" and 100 is "Worst." These human scores are necessary for measuring the impact that different distortions have on perceptual quality.

A. C. Bovik, Automatic Prediction of Perceptual Image and Video Quality. Proceedings of the IEEE, 101(9), pp. 2008-2024, September 2013.