Evaluation of Non-Reference Quality Assessment Algorithms to Curate Born-Digital Video Collections

Maria Esteva, Anne Bowen, Texas Advanced Computing Center; University of Texas at Austin; USA; Todd Richard Goodall, Alan Conrad Bovik; Laboratory for Image and Video Engineering; University of Texas at Austin; USA; Zach Brian Abel; College of Natural Sciences; University of Texas at Austin; USA.

Abstract

As the production, the variety, and the consumption of born-digital video grows, so does the demand for acquiring, curating and preserving large-scale digital video collections. A multidisciplinary team of curators, computer scientists and video engineers we explore the use of Non-Reference Image and Video Quality Algorithms (I/VQA), specifically of BRISQUE in this paper, to automatically derive ranges of video quality. An important characteristic of these algorithms is that they are modeled to human perception. We run the algorithms in a High Performance Computing (HPC) environment to obtain results for many videos at the same time, accelerating time to results and precision in computing per-frame and per-video quality assessment scores. Results, which were evaluated quantitatively and qualitatively, suggest that BRISQUE identifies the distortions in which it was trained, and performs well in videos that have natural scenes and do not have drastic scene changes. While we found that this particular model is not apt for evaluating collections with varied content, the results suggest that research into other I/VQA models is promising, and that their implementation at large scale can narrow the problem of curating very digital video collections and lead to preservation and access decisions based on informed priorities.

Introduction

The use of video has become significant and pervasive in our daily lives, going beyond traditional education and entertainment functions into areas such as personal communications exchange, criminal evidence, surveillance, and marketing. With this functional diversity comes a variety of formats, including advancing compression, and editing mechanisms to facilitate video creation and distribution. The advancements in video technology are important to cultural institutions, responsible for documenting society and of preserving video collections. Over time, these video collections grow without bound, severely encumbering the curation task. Accordingly, collecting institutions realize that individual and manual inspection, a traditional approach to assessing video quality and making subsequent preservation and access decisions, is an insurmountable task. Instead, novel, reliable, and automated methods are required for this purpose.

Motivated by the need to develop curation solutions for large and varied video collections, this project investigates the use of Image and Video Quality Assessment (I/VQA) algorithms to generate data-driven, perceptually relevant indicators of video quality levels for large video collections. I/VQA algorithms are designed to predict the subjective quality of a natural image or video that has been digitally acquired, processed, communicated and displayed as would be perceived and reported by users [1]. Currently, such algorithms are used to assess the quality of images and videos in streaming applications, and to dynamically correct their distortions. In this project we explore if and which I/VQA algorithms can be used to conduct large-scale automated assessment from which the need for more in depth video analysis can be prioritized.

We conducted experiments to understand the adequacy/scope and to refine the I/VQA algorithm BRISQUE using a reference set of videos and a set of artistic videos as testbeds. All the experiments were run using High Performance Computing Resources (HPC). Running parallel computational processes on HPC systems allows generating results for individual frames per video in a collection, promptly and accurately within one workflow. Interpreting these results entailed a qualitative evaluation this is viewing videos with frame-level quality predictions along with a graph indicating a holistic measure of quality over an entire video.

In the context of a digital curation project, experimenting with these algorithms in an HPC environment benefits from an interdisciplinary approach. A collaboration between the Laboratory for Image and Video Engineering (LIVE http://live.ece.utexas.edu), which conducts research in I/VQA, and the Texas Advanced Computing Center (TACC http://www.tacc.utexas.edu), which deploys computational resources for open science research, our team combines the expertise of data curators and computational scientists, with that of video engineers. In this paper we will introduce the I/VQA algorithms, explain how they compare to current methods to estimate video quality in heritage video collections, show the experiments conducted to understand the fitness of the model for video collections' assessment, and discuss the results obtained from testing the model in reference video sets and in a regular video collection.

I/VQA Algorithms

State-of-the-art I/VQA algorithms are based on natural scene statistics (NSS), which function under the premise that scenes have statistical regularities. Because 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. 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 [1].

I/VQA algorithms can be full-reference (FR) and no-reference (NR). The former require as input a high quality reference image or video against which a distorted copy can be compared to. In the context of curation, a FR algorithm, the Structural Similarity Index (SSIM), was used to verify if and to what degree the conversion of original video files involved information loss [2]. By contrast, NR algorithms measure the perceived quality in images and videos for which there is no original or pristine version available for comparison [1]. We propose that NR algorithms could be useful to understand a collection's quality without the need for humans to review each video. But, studies have to be conducted to understand which models can be used to assess quality in video collections that are varied in content and distortions. The focus of this paper is evaluating if BRISQUE, a NR algorithm for image quality assessment that can be used to assess video, is appropriate for digital video curation.

Related Work

Collecting institutions have been traditionally focused on digitizing analogue video for preservation and access, and a number of video QC tools have been introduced for purposes of automatic and objective quality assessment of digitized files [3, 4]. This is a great improvement over the traditional approach in which humans reviewed the files to detect both errors originating in the analogue media that was digitized and errors resulting from the digitization process. Indeed, while humans can identify different types of video distortions, manually recording them with precision is extremely time consuming and inconsistent [5]. Aside from individual differences, popular QC tools identify various types of artifacts and noise in individual frames and across frame differences, producing frame-by-frame features [3] or averaged features [4] for each type of detected distortion. In turn, these results have to be interpreted to derive a holistic quality condition per video. Therefore, while these tools assist the curation task by a human, none of them eliminate the need for humans to view the videos. To accurately assess the condition of a video in a perceptually relevant context, these features must be mapped to a quality score which correlates significantly with human-based DMOS scores.

Our work differs in methods and scope from the above, serving a complementary function. As opposed to detecting errors based on distortion-specific filters and corresponding ranges of normalcy, we are introducing perceptual subjective measures based on models of the human visual system to understand the quality of individual digital videos within collections. Importantly, 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. Such consensus can be understood as the collective interpretation of quality. In addition, our project does not focus on detecting analogue distortions or on evaluating the results of the digitization process, but on distortions that are typical of compression algorithms. Because we are interested in processing large video collections, we run the model on a supercomputer allowing us to obtain DMOS predictions both holistically and at the per-frame scale. In addition, we also performed a study without training on rated distortions to remove subjectivity. In the following section we describe the testbed collections used to build and to evaluate our model, and the studies performed to determine its fitness to assess large-scale video collections conditions.

I/VQA Studies

Test Datasets

For building models that correlate with perceptual quality, the CSIQ and LIVE video databases [6, 7], which contain corresponding DMOS scores for each video, were obtained. The CSIQ database contains a total of 12 reference videos and 216 distorted videos. Of these distorted videos, we selected only those distorted videos related to compression and noise. The LIVE VQA database contains 10 reference videos and 150 distorted videos. Of these distorted videos, we selected only distortions related to compression. The final working set, after selecting videos related to compression and noise distortions, contains 22 reference videos with 260 total distorted versions based on these reference videos. Of these reference videos, all are captured from recording the physical world except for one representing an animated scene. The distortions in this final set include H.265, H.264, MPEG-2, MJPEG, Wavelet, and Additive White Noise (AWN).

To evaluate the model within the curation context, we selected a diverse set of twenty-three digital art videos from a museum collection. The videos include natural scenes and artificial elements introduced as part of the artistic intent. All of the videos in this set were sold to the museum as DVDs and thus have MPEG-2 compression [6], but their technical provenance is not documented. Many of the videos could contain mixtures of compression distortions given the circuitous nature of their encoding. Not knowing how and with what tools they were filmed or edited, we did not have precise knowledge of the types of distortions present in the videos.

BRISQUE Model

The following studies illustrate our research involved in identifying suitable I/VQA models for curation. Thus far, we utilize natural scene statistic features developed in the IQA model called Blind Referenceless Image Spatial Quality Evaluator (BRISQUE) [7]. BRISQUE is a general-purpose NR algorithm. Given the high performance of BRISQUE on static images, we use it as a feature extraction model. "Feature extraction" is a common technique used in machine learning and image processing algorithms to derive variables, known as features, based on the initial set of input data. Specifically, the BRISQUE algorithm extracts features by computing a number of variables based on the statistical properties of natural scenes for use in its model. One obvious limitation of the algorithm when applied to video is its lack of modeling distortions related to motion. To partially address this limitation, we incorporate frame differencing, thus each score in the BRISQUE model is produced from a static frame and neighboring frame differences as depicted in Figure 1. For each frame, we use the BRISQUE feature model to compute a total of 144 features. This extension allows the model to capture noise and compression artifacts that appear during inter-coding compression.

These 144 features per F_n are averaged over the entire video and input into an SVR (Support Vector Regressor) to obtain a single score for an input video. The SVR is a supervised machine learning algorithm that is often used for pattern recognition and classification.

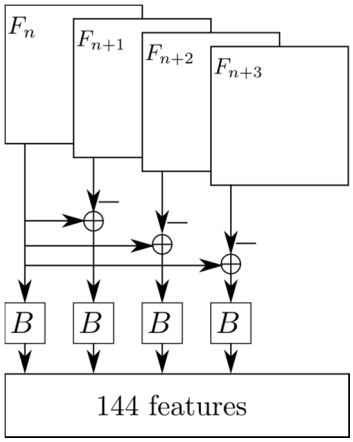

Figure 1. Features used for the BRISQUE feature model. The current frame $\boldsymbol{F_n}$ is passed into the BRISQUE feature model $\boldsymbol{B}$ to extract 36 features. Frame differences produced from $\boldsymbol{F_n - F_{n+1}}$, $\boldsymbol{F_n - F_{n+2}}$, and $\boldsymbol{F_n - F_{n+3}}$ are input separately into the same feature model to produce a 3 sets of frame-difference features. In this model, a total of 144 features is extracted specific to frame $\boldsymbol{F_n}$.

Our first study involved training the proposed model with the corresponding DMOS provided in both CSIQ and LIVE databases for all distortion categories. We used a Support Vector Regressor (SVR) to map averaged (over the entire video) features to these scores. Using leave-one-out cross validation, we trained on 21 content types and tested on the remaining 1 content for the database collection. Figure 2 depicts the scatterplot that results from the 22 resulting tests. The overlaid logistic function, computed using the following remapping:

$$y = \alpha \log(1 + \beta x)$$

where x is the score prediction from the SVR, y is the remapped score, and α and β are computed using least squares minimization between y and ground truth DMOS. This remapping serves to smooth the DMOS predictions for unseen content.

To measure correlation between our model prediction and the ground truth DMOS scores, we computed the median Pearson's linear correlation (LCC) and Spearman's correlation coefficients (SRCC), allowing us to measure how well the relationship between the predicted and ground truth DMOS can be represented by a linear monotonic function. Recall that our reference training set has only 22 types of content. When measuring the correlation for distortions per content type, we observe that the median LCC is 0.80, and the median SRCC is 0.81 (SRCC). However, when measuring correlation regardless of content type, we observe a lower median LCC of 0.55 and SRCC of 0.56. This indicates that within video content, the predictions are highly correlated to the distortion level present in the video. However, these same predictions are weakly correlated with distortion level, across various content. Still, we considered that the results were encouraging enough to pursue the investigation with this model. For example, content is largely similar within each video. Thus the frame-based predictions for a video are well suited for tracking abrupt changes in quality within that video.

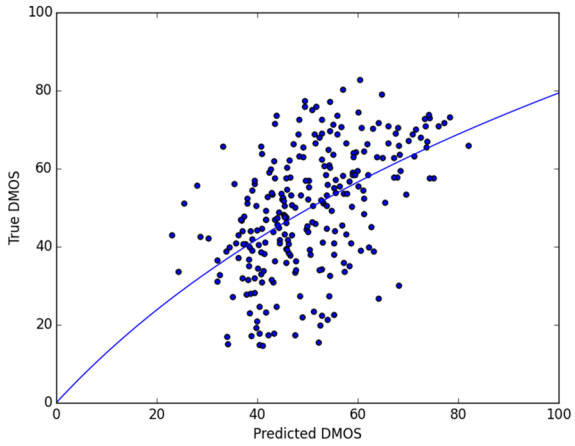

Figure 2. Scatter plot between predicted and actual DMOS. Curve indicates logistic function used for re-mapping predictions.

We then wanted to evaluate how close the predicted scores aligned with the ground truth scores for the distortions for which it had been specifically trained. Using the same leave-one-out cross validation approach, we tested the 22 individual content categories in our training set. Our best results, quantified by the minimum average error between predicted scores and ground truth DMOS, exemplify the model's ability to distinguish the level of distortion in a perceptually relevant way for the content types "Riverbed" and "ParkScene." The prediction results for "Riverbed," obtained in the LIVE database, are depicted in Figure 3. Additional results for "ParkScene," obtained in the CSIQ database, are depicted in Figure 4. When taking into account the standard deviation of the ground truth DMOS, we observed that the currently proposed model captures the monotonic relationship across distortion levels (i.e. increasing the level of distortion increases the predicted DMOS score and vice versa). This is certainly an important relationship to capture for ranking video collections from best to worst quality.

Knowing that all of the museum collection had MPEG-2 type compression, we particularly focused on the behavior of the model with this distortion. Analysis over the LIVE collection provides insight into the performance on MPEG-2. We observed that the BRISQUE-based model produced DMOS predictions which when compared to the ground truth DMOS scores, achieved 39% SRCC and 36% LCC correlations when considering all content together. On a particular video content, we find 50% correlation for both SRCC and LCC. The correlation is low, but is still useful for determining the worse and the best videos in a collection. This proposed model does not capture MPEG-2 distortions in isolation, but it is a general model that applies to many types of compression distortions. Given the circuitous nature of the encoding within the museum collection, this generalization is an important property.

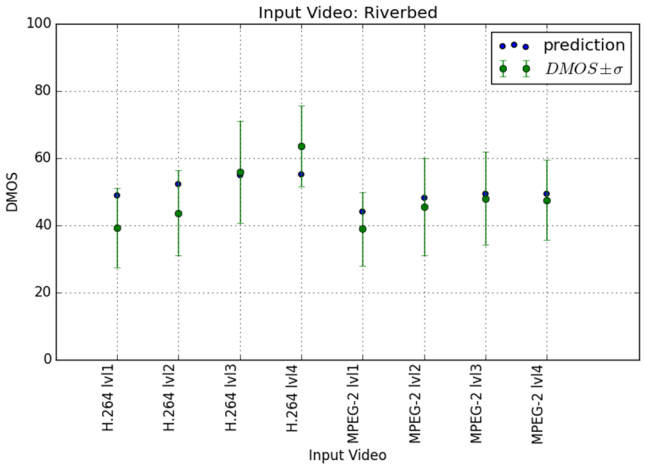

Figure 3. Predicted results for the video "Riverbed" from the LIVE database. The x -axis shows the different distortions and their levels.

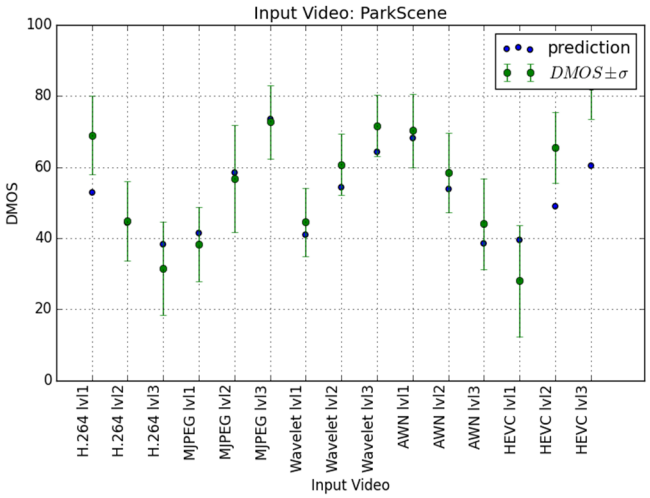

Figure 4. Predicted results for the video "ParkScene" from the CSIQ database. The x-axis shows the different distortions and their levels.

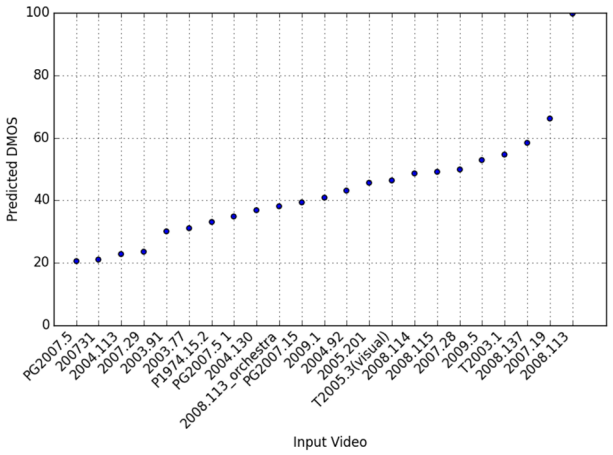

Figure 5. DMOS scores predicted over 23 video art pieces going from 0 excellent to 100 worse.

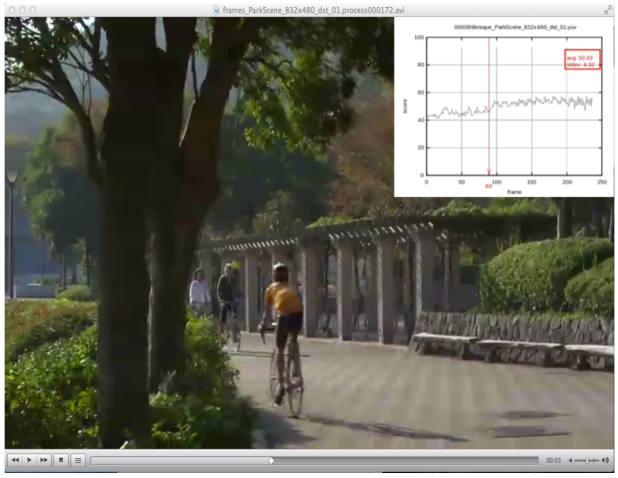

Figure 6. Screenshot from the visual diagnostic tool showing the current frame along with its score and the scores for all other frames.

Having the model trained for DMOS, we next studied how well the predictions served to assist curation on a collection of twenty-three video art pieces. Since this model makes predictions relative to the distortions it was trained on, it is crucial to identify where prediction accuracy suffers, both at the per-frame level and over an entire video. Figure 5 shows the predictions made on the test-bed collection in which the highest scores indicate lower video quality. Note that in general, the collection ranges from good to poor. We comment on the qualitative evaluation of these results in the Visual Evaluation Section.

Parallel Implementation of the Algorithms

To improve precision and time-to-results, the computational process for the quality assessment was implemented as a workflow using High Performance Computing (HPC) resources that allow multiple tasks to be executed in parallel. Parallel processing is essential to analyze large video collections in a timely fashion. The "Stampede" supercomputer at the Texas Advanced Computing Center (TACC) was used as a test-bed for the development and testing of the parallel quality assessment workflow. Stampede is one of the most powerful supercomputers in the U.S. for open science research with 522,080 processing cores spread over 6400 nodes and 260 TB of memory. To ensure that the supercomputer resources are used efficiently, our parallel quality assessment workflow was designed to be scalable and use the appropriate number of processes for the analysis at hand.

The steps in the workflow consist of a frame and metadata extraction of the video by ffmpeg, followed by the execution of the BRISQUE-based feature extraction algorithm on each frame; the final frame-by-frame prediction for each movie are then plotted and embedded into the original movie frame to allow for live viewing of the BRISQUE results for diagnostic purposes. This final stage in the workflow creates the visual diagnostic tool which superimposes the graph of the BRISQUE results along with the score of the current frame onto the original source video. As shown in Figure 6, the score of the current frame (indicated by the red line) can be viewed in the context of the overall plot of the

scores from the entire video. The scripts that control the workflow allow for each computational process to be run in parallel using the 'launcher' job execution framework at TACC. We studied the efficiency of ffmpeg and our proposed model running on the Stampede supercomputer and found that an optimal run uses 8 cores per node. In our workflow implementation, each compute node works on two videos at once and when finished starts again. Each node can analyze ~40 high definition frames per second.

Visual Evaluation

In projects involving the use of models to automate assessment processes, it is important to obtain reality checks by manually reviewing results in relation to data. In this case, a visual, qualitative evaluation of each video was required to understand the factors that contributed to the proposed model prediction scores, and the extent to which they can substitute human judgment of quality. Given that, except for the MPEG-2 compression present in all the movies, we did not know a priori what other types of distortions were present in the museum collection, and how the presence or absence of distortions affected the scores, we incorporated the visual identification of distortions in the analysis. Three team members watched each video using the visual diagnostic tool developed for this project (See Figure 6). During the viewing of each video, we 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. The three of us also scored each video according to the same scale in which human subjects are asked to score reference videos, from 1 to 5 in ranges that go from excellent to worse. We used this scores to evaluate if our assessment coincided or not with the algorithm.

Judging from the agreement or disagreement between our scores and those of the model, we concluded that our proposed model is not appropriate for videos that have non-natural scenes, and that it does not perform as well with movies in which there are drastic scene changes or that have interlacing and other distortions that are not captured by the algorithm. Instead, the model performed well in videos that do not have too many scene changes, and that have distortions on which the algorithm was trained. This agrees with the results noted in Figures 3 and 4 produced from our training set.

Objective Evaluation

On a last study we attempted to remove the subjectivity from the quality assessment process. For this we extracted the BRISQUE-based features, as depicted in Figure 1, from both static image frames and frame differences, from both the LIVE and CISQ databases, to evaluate the ability of these natural BRISQUE features to detect degree of distortion without training on rated distortion, i.e. completely blind. From the distorted and pristine reference videos, we extracted the features and trained a one-class Support Vector Machine on the pristine feature set. We then input the features from the distorted set to this model to produce distances from natural. We define this distance from natural as the distance from the separating hyperplane surrounding the pristine features. Essentially, natural distortion-free videos should produce

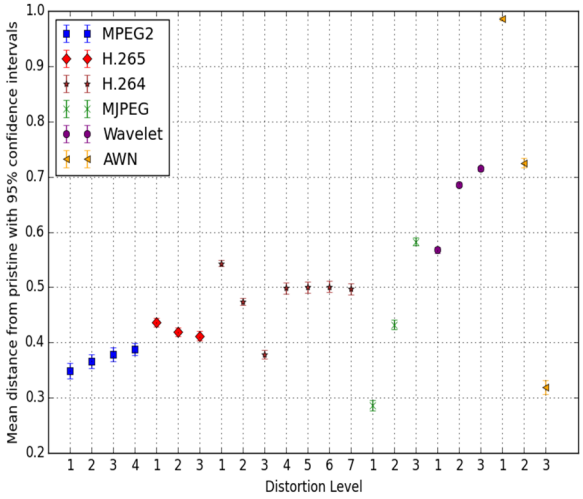

Figure 7. Distances from "natural" across distortion categories and corresponding levels.

features that are statistically well-behaved, only varying within some tolerance. Measuring this distance from the normal tolerance can also be seen as a distance from "naturalness." Figure 7 depicts the capability of this model at predicting relative magnitudes of degradation across several distortion categories and levels when averaged across content. Here, distance is only meaningful in a relative sense and distance of 0 would indicate excellent quality. In this figure, all distortions have at least some distance from 0 indicating the distortion presence. Distances for the AWN distortion category appear to differ from each other more than distances for other distortion categories. This is a direct result of BRISQUE feature sensitivity to noise. This distance depends directly on the learned statistics and only appears to have meaning in a relative sense.

We then correlated the results between the distance measures and DMOS to determine if the results are statistically significant and found that the correlation is too weak. Videos under curation may have unknown and sometimes novel distortion patterns for which a distortion-blind model would be theoretically ideal. However, it is difficult to produce a completely distortion-blind model without a more complete and accurate model of the human visual system in place. Human visual systems can be modeled, and if a model were good, it would not be necessary to use the human opinion for training the algorithms. Also worth noting is that rating quality is a highly subjective experience even with an ideal model of human vision, and thus the proposed model works well because it uses opinions for the training process.

Conclusions

Anticipating the growing and increasingly varied videos cared for by collecting institutions, this project aims to implement a large-scale curation system to categorize videos into ranges of quality. We started our research investigating the referenceless BRISQUE-based features, and we have learned both its limitations and potential advantages. We learned that we could not expect this model to assess varied collections with high accuracy, but that we could expect it to assess relative quality within individual videos and across videos that comply with certain characteristics. Overall, after going through the process of evaluating and testing

we consider that the utilization of I/VQA algorithms for curation is promising, and more studies need to follow. To further this goal, we are analyzing techniques within the top-performing Video BLINDS (Blind Image Integrity Notator using DCT Statistics) [8] and NIQE (Naturalness Image Quality Evaluator) algorithms [9].

We have also identified the need to conduct studies addressing the issue of diversity in content across videos, as is the case with collections belonging to cultural and educational institutions. We found that current models appear not to consider distortions due to artistic intent, and that are often proven using a relatively small collection of pristine videos, which are then distorted artificially. We are amassing a large and varied testbed collection to provide a thorough examination of the statistical variations. Analyzing larger statistics will provide new insights about video quality by removing the aforementioned limitations. The results from these algorithms can inform about the types of distortions that are present in digital video collections, provide insights about the videos technical provenance, and help make curatorial decisions.

References

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

[2] M. Esteva, K. Vega, B Scott, S. Gunnels, K. Kumar, Automated Workflow for Archiving Video Art in DVD format. *Proceedings of the Archiving 2013 Conference*, pp. 19-24,April 2013.

[3] Bay Area Video Coalition.Quality Control Tool for Video Preservation. Retrieved December 8, 2014 http://bavc.org/qctools

[4] National Archives Video Frame Analyzer. Retrieved December 8, 2014 https://github.com/usnationalarchives/Video-Frame-Analyzer

[5] J Gfeller, A Jarczyk, J. Phillips. Compendium of Image Errors in Analogue Video. University of Chicago Press. 2012.

[6] A. Mittal, A. K. Moorthy, and A. C. Bovik, No Reference Image Quality Assessment in the Spatial Domain. *IEEE Transactions on Image Processing*, 21(12), pp. 4695-4708, December 2012.

[7] K. Seshadrinathan, R. Soundararajan, A. C. Bovik and L. K. Cormack, Study of Subjective and Objective Quality Assessment of Video, *IEEE Transactions on Image Processing*, 19(6), pp. 1427-1441, June 2010.

[8] P. V. Vu and D. M. Chandler, ViS3: An Algorithm for Video Quality Assessment via Analysis of Spatial and Spatiotemporal Slices, *Journal of Electronic Imaging*, 23(1), 013016, February 2014.

[9] M.A.Saad, A. C.Bovik, and C. Charrier.Blind Prediction of Natural Video Quality. *IEEE Transactions on Image Processing*, 23(3), No. 3, pp. 1362-1365, March 2014.

[10] A. Mittal, R. Soundararajan and A. C. Bovik, Making a Completely Blind Image Quality Analyzer, IEEE Signal processing Letters, 22(3), pp. 209-212, March 2013.

Author Biography

Maria Esteva has a PhD in Information Sciences. She is a Data Curator at TACC where she designs architectures for scientific and humanities collections. Her research includes the design of large-scale data curation strategies.

Anne Bowen has a PhD in Computational Chemistry from the University of Zurich. She is a visualization scientist at TACC where she develops tools for analysis and visualization of large datasets.

Todd Goodall received his B.S. in Computer Engineering from Clemson University followed by his M.S.E. in Electrical and Computer Engineering from the University of Texas at Austin. He joined the Laboratory for Image and Video Engineering (LIVE) in August 2013, and is pursuing his Ph.D. in Electrical and Computer Engineering.

Alan Conrad Bovik is the Cockrell Family Regents Chair Professor with the Department of Electrical and Computer Engineering and the Institute for Neuroscience, UT Austin. His research interests include image and video processing, computational vision, and visual perception. He has published over 700 scientific articles and holds several US patents.

Brian Zach Abel is a senior studying Physics and Computer Science at UT Austin. Previously an ICERT REU intern, he is currently an Industry Partner sponsored undergraduate research assistant at TACC.