

NISTIR 8187

Evaluation Infrastructure for the Measurement of Content-based Video Quality and Video Analytics Performance

Haiying Guan
Daniel Zhou
Jonathon Fiscus
John Garofolo
James Horan

This publication is available free of charge from:
<https://doi.org/10.6028/NIST.IR.8187>

NIST
**National Institute of
Standards and Technology**
U.S. Department of Commerce

NISTIR 8187

Evaluation Infrastructure for the Measurement of Content-based Video Quality and Video Analytics Performance

Haiying Guan

Daniel Zhou

Jonathon Fiscus

John Garofolo

James Horan

Information Access Division

Information Technology Laboratory

This publication is available free of charge from:

<https://doi.org/10.6028/NIST.IR.8187>

July 2017



U.S. Department of Commerce

Wilbur L. Ross, Jr, Secretary

National Institute of Standards and Technology

Kent Rochford, Acting NIST Director and Under Secretary of Commerce for Standards and Technology

TABLE OF CONTENTS

1	PROJECT GOAL.....	1
1.1	INTRODUCTION.....	1
1.2	WHY CONTENT-BASED VIDEO QUALITY?	1
1.3	PROJECT OBJECTIVE.....	2
2	LITERATURE OVERVIEW	3
2.1	OVERVIEW OF VIDEO QUALITY METRICS RESEARCH.....	3
2.2	NO-REFERENCE VIDEO QUALITY METRIC SURVEY	3
3	VIDEO QUALITY METRICS FOR VIDEO ANALYTICS – OUR APPROACH AND FINDINGS	5
3.1	OUR APPROACH	6
3.2	VIDEO BLIND IMAGE NOTATOR USING DCT STATISTICS	6
3.3	STUDY ON THE 2013 NIST PASC FACE RECOGNITION EVALUATION	6
3.3.1	PaSC Data Subsetting.....	7
3.3.2	Computation cost of Video BLIINDS on PaSC data.....	7
3.3.3	Experimental results of Video BLIINDS on PaSC Data Subset.....	8
3.3.4	Discussion.....	10
3.4	STUDY ON THE 2014 NIST TRECVID MULTIMEDIA EVENT DETECTION (MED) EVALUATION	10
3.4.1	MED Data Subsetting.....	11
3.4.2	Video quality distributions of different datasets	11
3.4.3	Video BLIINDS on MED HAVIC dataset	12
3.4.4	Discussion and Findings.....	14
4	EVALUATION INFRASTRUCTURE.....	15
4.1	EVALUATION FRAMEWORK.....	15
4.1.1	Validation Framework.....	15
4.1.2	Overall Framework.....	16
4.2	NIST DETECTION ANALYSIS PIPELINE RESOURCES (DAPR) INFRASTRUCTURE.....	16
4.3	DAPR FOR VIDEO QUALITY METRICS AND VIDEO ANALYTICS EVALUATION	17
4.3.1	Infrastructure Implementation	17
4.3.2	Cross Evaluation Metrics.....	18
5	FUTURE DIRECTIONS REGARDING EVALUATION FRAMEWORKS AND DATA COLLECTIONS	19
5.1	FUTURE EVALUATION FRAMEWORK REQUIREMENTS	19
5.2	DATA COLLECTION PLAN.....	19
5.3	INITIAL PROOF-OF-CONCEPT VIDEO DATASET DESIGN.....	19
5.4	GENERAL DATA COLLECTION REQUIREMENTS	21
5.5	VIDEO CAPTURE DEVICE.....	21
5.5.1	Type of device	21
5.5.2	Device mount and geolocation	22
5.5.3	Device settings.....	22
5.6	VIDEO SOURCE.....	22
5.7	VIDEO CONTENT	23
5.8	CAPTURE ENVIRONMENT.....	23
5.8.1	Lighting condition	23
5.9	VIDEO PRODUCTION QUALITY	23
5.10	VIDEO ANALYTIC ANNOTATIONS.....	24

5.10.1	Analytic Object Parameters	24
5.10.2	Analytic Event Parameters.....	24
5.11	VIDEO QUALITY ANNOTATIONS	24
5.12	VIDEO ANALYTIC SYSTEM OUTPUT	24
5.13	VIDEO QUALITY METRIC SYSTEM OUTPUT.....	24
6	CONCLUSIONS AND WAY FORWARD	25
7	ACKNOWLEDGEMENT	25
8	DISCLAIMER.....	26
9	REFERENCES	26

LIST OF TABLES AND FIGURES

TABLE 1 VIDEO BLIINDS COMPUTATIONAL COST AT DIFFERENT PIXEL RESOLUTIONS	8
TABLE 2 VIDEO BLIINDS PREDICTED DMOS EXPERIMENT RESULTS ON HAVIC MED BIKE TRICK DATASET	12
FIGURE 1: THE ROLE OF VIDEO QUALITY MEASUREMENT IN FUTURE VIDEO ANALYSIS WORKFLOWS	2
FIGURE 2 VIDEO QUALITY COMPARISON OF PIXEL RESOLUTION VS PREDICTED VIDEO QUALITY SCORE	8
FIGURE 3 VIDEO QUALITY (DIFFERENTIAL MEAN OPINION SCORE) VS. VIDEO RANK WITH A BASELINE FACE RECOGNITION SYSTEM (PCA)	9
FIGURE 4 VIDEO QUALITY (DIFFERENTIAL MEAN OPINION SCORE) VS. VIDEO RANK WITH LRPCA SYSTEM USING LEAST SQUARES FIT	10
FIGURE 5 VIDEO QUALITY (DIFFERENTIAL MEAN OPINION SCORE) VS. VIDEO RANK WITH PITTPATT SYSTEM USING LEAST SQUARES FIT	10
FIGURE 6 HISTOGRAM OF VIDEO QUALITY OF DIFFERENT DATASET	11
FIGURE 7 VIDEO QUALITY SCORE PREDICTED WITH VIDEO BLIINDS OF THE BIKE VIDEOS VS. THEIR RANKS PROVIDED BY CMU'S BASELINE ALGORITHM (DIFFERENTIAL MEAN OPINION SCORE VS SCORE RANK)	13
FIGURE 8 VIDEO QUALITY SCORE PREDICTED WITH VIDEO BLIINDS OF THE BIKE VIDEOS VS. THEIR ANALYTIC CONFIDENCE SCORES PROVIDED BY CMU'S BASELINE ALGORITHM.	14
FIGURE 9: VQM VALIDATION PROCESS.	15
FIGURE 10: THE VIDEO QUALITY METRICS VS. VIDEO ANALYTIC CROSS EVALUATION FRAMEWORK.	16
FIGURE 11: ANALYSIS-CENTRIC EVALUATION INFRASTRUCTURE	17
FIGURE 12: THE DETECTION ANALYSIS PIPELINE RESOURCES (DAPR) DATA MODEL.	18

EXECUTIVE SUMMARY

The use of video cameras is growing at an exponential rate with cell phones, body-worn cameras, drones, all sorts of tactical cameras and low-cost high performance IP surveillance cameras all contributing to rapid growth of video data. A virtual explosion of video from the ground, air, and mobile devices is happening in public safety. Meanwhile, the video data that is being produced by these myriad of devices is driving a data transport and storage explosion. The demand for video will quickly outpace the ability for even the communication networks of the future to carry it all. Human observation and review of every active camera is at best difficult, inefficient and expensive and will eventually become impossible at-scale without some form of automation. Automatic or semi-automatic video analytic technologies are necessary to address both the video demand and analysis challenges.

A key to the efficient and effective use of these technologies are quantitative measures of video quality. The quality of video data significantly affects video analytic system performance as well as optimization of compression and transport and storage techniques and how video can be utilized and interpreted. Objective measurements of video quality are therefore critically important for both real-time applications and forensic applications. Video quality has impact on many application areas, such as public safety, homeland security, video surveillance applications, geographic applications, the video entertainment industry, autonomous vehicles, field robotics, and beyond.

In this report, we discuss the creation of an evaluation framework to measure the predictive power of video quality metrics with regard to the performance of video analytic technologies. The described evaluation framework provides an important future testbed capability for researchers engaged in the development of both quality metrics and video analytic technologies. Ultimately, the measurement framework when combined with reference data can provide the basis for future standards related to video quality. The information gained in this project will be useful in informing future best practices and standards for quality measurement as an objective measurement for improving encoding, collection, compression and storage of video in public safety.

Keywords:

Video Quality Metric (VQM), Video Analytics (VA), Multimedia Event Detection (MED), TREC Video Retrieval Evaluation (TRECVID), Point and Shoot Face Recognition Challenge (PaSC), Face Recognition, Detection Analysis Pipeline Resources (DAPR), Scorer-centric Evaluation, Analysis-Centric Evaluation, National Imagery Interpretability Rating Scale (NIIRS), Video-National Imagery Interpretability Rating Scale (VNIIRS), Mean opinion score (MOS), Difference Mean opinion score (DMOS).

Trademark & Copyright Information

All copyrights, registered trademarks or trademarks are the property of their respective organizations or owners.

1 PROJECT GOAL

1.1 INTRODUCTION

The measurement of quality of video has traditionally been extremely subjective and focused on human perception of “goodness”, not utility. The questions that this field has historically focused on are along the lines of “Is that a good picture of me?” or “Is this a sharp TV picture?” But the measure of video quality that is meaningful for public safety operations or forensics is related to the successful completion of a specific task. The type of information required from video in the public safety domain is based on the mission space of the organization. Although there is often a significant variance in video quality available, the detection of objects or events such as “Is there a fallen firefighter in the scene?” or “Is there a drowning man in the video field of view?” or “Is the object a gun?” are of key importance. In the research community, these are referred to as high-level semantic features of the video content and are implemented in specific manual video analysis protocols that are used in forensic examination of video and in video analytics for both real-time and forensic applications.

For these kinds of use cases, quality is intrinsically related to the ability to derive useful information from a video feed – either by human interpretation or through automation. In the case of analytics, the quality of this information is directly related to video analytic system performance. For the proper design of a real-time video analytics system, numerous considerations must be factored in. For example, “How should bandwidth be allocated?”, “What level of compression should be employed?”, “Can required analyses be performed?”, and “What is the uncertainty of an interpretation?” are all critically important questions. To evaluate these considerations, an efficient, objective, and repeatable measurement system is necessary. The goal of this effort was to begin to develop the fundamental framework for evaluating the utility of objective video quality metrics (VQM) that are predictive of analytics related to understanding the content in video and to determine the state of the art in current VQM research.

1.2 WHY CONTENT-BASED VIDEO QUALITY?

Fundamental research in the measurement of video data quality as it relates to analytics is critically important for future public safety video systems and applications. Video bandwidth efficiency, video analysis, and forensic video applications can all benefit from this area of research. The future goal is to be able to use quality metrics to objectively ground forensic data standards, optimize compression for public safety, improve communication performance, and reduce persistent storage needs, while improving the performance of operational systems by integrating effective human analyses and analytic technologies.

Future video analysis workflows will integrate automated objective video quality metrics (VQM) in multiple ways. VQM will be utilized to condition the video stream from the point of collection so that it is optimized for downstream analysis, both by humans and automation. Ultimately, VQM metrics will be part of a dynamic feedback loop where downstream analytics are providing quality requirements to upstream processes to support both the optimal conditioning of video and the optimized use of bandwidth and storage (Figure 1). The full integration of VQM at the edge and along the entire

workflow will be essential for large scale video deployments in public safety applications. A quantitative understanding of quality with regard to the utility of the content in the video is key to these future capabilities.

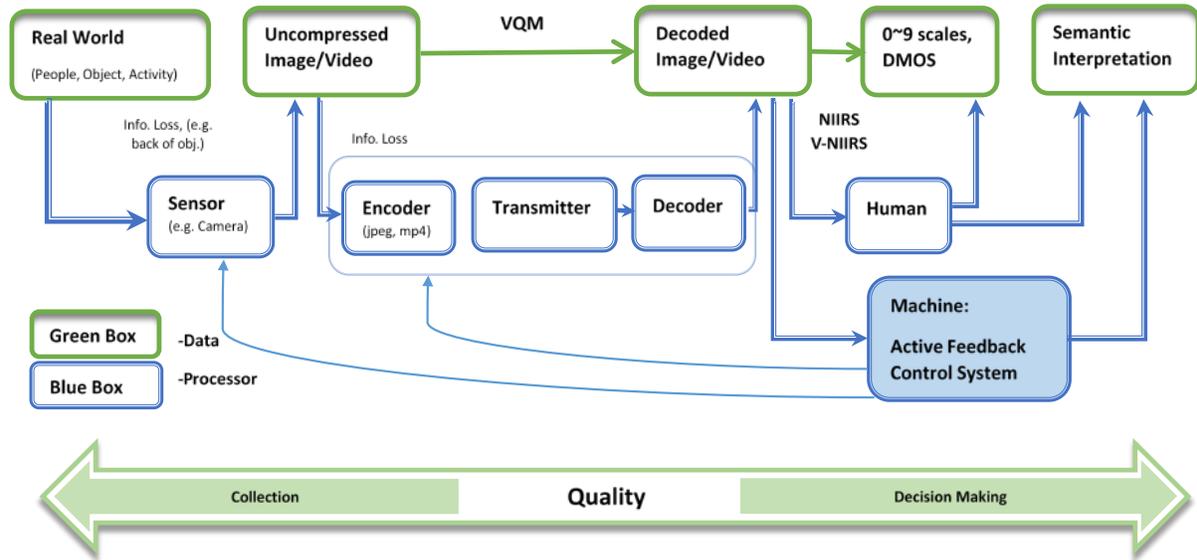


Figure 1: The role of video quality measurement in future video analysis workflows

1.3 PROJECT OBJECTIVE

The overarching objective of the project was to study the relationship between video quality metrics and video analytic algorithm performance and to build a reusable framework to perform experiments in this space. Our implementation goals were to:

1. design and develop a measurement framework that can be used to characterize the predictive power of video quality metrics with regard to the performance of video analytic technologies
2. obtain data from existing video analytics performance evaluations to support pilot experiments to drive the development of the framework
3. identify and obtain existing quality metric algorithms to characterize and assess their utility in predicting analytic performance
4. conduct exploratory experiments with the above
5. define requirements for future reference data to support research and development in this area
6. publish the measurement framework to foster research in this area

The measurement framework would serve several purposes in the project: explore feasibility, define a repeatable methodology, perform an initial assessment of the state-of-the-art in automated video quality metrics, and motivate requirements for the creation of future reference data collections that would support intersectional research in the areas of video quality and video analytics. The information gained

in these measurements would also be useful to drive and inform public safety best practices and standards for quality measurement.

2 LITERATURE OVERVIEW

2.1 OVERVIEW OF VIDEO QUALITY METRICS RESEARCH

The following section describes a survey of existing video quality metrics (VQM).

National Geospatial-Intelligence Agency (NGA) developed the National Imagery Interpretability Rating Scale (NIIRS) for image quality measurement and the Video-National Imagery Interpretability Rating Scale (VNIIRS) for video quality measurement. The metrics that are created are based on human assessment and human-based evaluation. These metrics are focused on interpretation, but are constrained to human perception specifically for the domain of photo reconnaissance and thus require domain-specific metadata.

National Telecommunications and Information Administration (NTIA) and International Telecommunication Union (ITU) developed Video Quality Metric (VQM) Software¹ for signal transformation and communication applications [1].

NIST researchers have performed extensive research and evaluation in the image biometric technology area (fingerprint comparison, face recognition, iris recognition) quality and video quality [2] for decades.

In academic research, video quality metrics can generally be classified as one of three categories: full-reference, reduced-reference and no-reference metrics. Full-reference metrics are used to assess comparative degradations in video that has been put through an encoding, compression, or transmission process and generally impose a precise spatial and temporal alignment of the two videos so that every pixel in every frame can be assigned its counterpart in the reference clip. Aside from the issue of spatiotemporal alignment, full-reference metrics usually do not respond well to global changes in luminance, chrominance or contrast and require a corresponding calibration. For reduced-reference metrics, the restrictions are less severe, as only the extracted features need to be aligned, but they are similar in their comparative use to full reference metrics. No-reference quality metrics are used to determine an absolute measure of quality of a single video source. As such, the focus of this effort was exclusively on no-reference metrics to determine which metrics correlated with semantic features as expressed in video analytics applications relevant to public safety needs.

2.2 NO-REFERENCE VIDEO QUALITY METRIC SURVEY

We performed an informal survey of existing no-reference video quality metrics to determine candidate algorithms for our experiments. In performing this survey, we scanned literature that we could identify on the Web and we reached out to NTIA who has a longstanding program in measuring video quality

¹ <https://www.its.bldrdoc.gov/resources/video-quality-research/guides-and-tutorials/description-of-vqm-tools.aspx>

for transmission protocols. We determined the following set of metrics to have potential applicability for our purposes. We contacted the authors of the metrics to attempt to acquire software implementations of the metrics. Unfortunately, we only received a response from Saad and Bovik [13]. We excluded the metrics from non-responsive authors in our study because it would require a significant effort to code and verify their performance. That was beyond the scope of this work. The following is a complete listing of the no-reference metrics that we surveyed:

1. Video BLind Image Notator using DCT Statistics (BLIINDS) [13] is a no-reference video quality assessment (VQA) algorithm that correlates highly with human judgments of quality. It relies on a spatio-temporal model of video scenes in the discrete cosine transform domain and on a model that characterizes the type of motion occurring in the scenes to predict video quality, and uses the models to define video statistics and perceptual features that are the basis of a no-reference metric. It is shown to perform close to the level of top performing reduced and full-reference VQA algorithms.
2. The Reverse Frame Prediction (RFP) [3] is a no-reference video quality method based on the multi-channel properties of the Human Visual System. Different from most NR methods of its time, no a priori knowledge of impairments is used. Instead, RFP detects frames with degraded quality in a video sequence by making assumptions of the overall quality of frame subsets in the stream. It applies Gabor filtering and is intended for in-service testing and on-line monitoring. The approach is focused on various psycho-perceptual properties of the Human Visual System (HVS). This metric is primarily designed to address transmission impairments for broadcast video and to identify characteristics of the video that would conflict with overall human perception of quality.
3. Oelbaum et al.'s metric [4] is based on assigning a given video to one of four different content classes: very low data rate, sensitive to blocking effects, sensitive to blurring, and a general model for all other types of video sequences. The appropriate class for a given video sequence is selected based on the evaluation of feature values of an additional low quality version of the given video, generated by encoding. The visual quality for a video sequence is estimated using a set of features, which includes measures for the blocking, the blurriness, the spatial activity, and a set of additional continuity features.
4. Farias et al.'s metric in [5] is based on data hiding: it uses a spread-spectrum embedding algorithm to embed a mark (binary image) into video frames. At the receiver, the mark is extracted and a measure of its degradation is used to estimate the quality of the video. Farias and Mitra's metric in [6] is based on individual measurements of three artifacts: blocking, blurriness, and noisiness. These are then tested using a proposed procedure that uses synthetic artifacts and subjective data obtained from previous experiments.
5. Yang et al.'s metric [7] uses information extracted from a compressed bit stream without resorting to complete video decoding. The metric accounts for picture distortion caused by quantization, quality degradation due to packet loss and error propagation, and temporal effects of the human visual system.

6. Kawayoke and Horita's metric [8] estimates quality by a Sobel spatial measurement operator on individual MPEG-2 digital frames first, then adjusts the frame quality using the information from spatial and temporal information respectively. The proposed metric focuses on monitoring the quality of service (QoS) for visual communications.
7. Brandão et al.'s metric in [9] computes quality scores as a linear combination of simple features extracted from the video sequence received at the decoder: the log of the video's bit rate, the log of the mean square error estimate, spatial and temporal activities and their variances. Brandão and Queluz's metric in [10] consists of local error estimation followed by perceptual spatiotemporal error weighting and pooling. Error estimates are computed in the transform domain because the discrete cosine transform (DCT) coefficients are corrupted by quantization noise. The DCT coefficient distributions are modeled using Cauchy or Laplace probability density functions. This form of metric is focused on addressing needs in traditional video encoding applications.
8. Sugimoto et al.'s metric in [11] uses only receiver-side information but extracts quantizer scale information from the bit stream along with two spatiotemporal image features from the baseband signal, which are integrated to express the overall quality using the weighted Minkowski metric. Sugimoto and Naito's metric in [12] is based on parametric analysis of the post-transmission coded bit stream, so it does not have access to the baseband signal (pixel level information) of the decoded video. These parameters are then used to calculate spatiotemporal image features to reflect coding artifacts which have been found to correlate with subjective quality in human perceptual experiments related to entertainment purposes such as IPTV and next generation DVD [11].
9. Lee et al.'s metric [13] uses coding parameters extracted from a bit stream: boundary strengths (BS), quantization parameters (QP), and average bitrates. The metric requires less computation than the other bit stream metrics listed above. So, in theory, it should be easier to implement in realtime.

Given our findings, we chose to implement the VIDEO BLind Image Notator using DCT Statistics (BLIINDS) metric to support our pilot experiments and the development of an evaluation framework. The other metrics we identified might be assessed for their predictive properties for video analysis applications in further work in this area.

3 VIDEO QUALITY METRICS FOR VIDEO ANALYTICS – OUR APPROACH AND FINDINGS

To assess the utility of automated video quality metrics (VQM) in their potential use as a tool in video analytics research and applications related to public safety applications, we created a framework to examine correlations between automatically generated quality rankings and video analytics performance results. Our goal was to create a framework that would permit us to begin to probe the relationship between analytic performance and automatically measurable quality-related factors, an important area affecting performance that has not historically been objectively examined.

3.1 OUR APPROACH

We conducted meta analysis experiments using historical performance results for two video analytics technology evaluations and a baseline quality metric to both show proof of concept and to develop a video quality diagnostic framework to support video analytics evaluation-centered research. To conduct the experiments, we utilized system output scores from the NIST 2013 Point and Shoot Challenge (PaSC) face recognition evaluation (“The challenge of face recognition from digital point-and-shoot cameras” <http://ieeexplore.ieee.org/abstract/document/6712704/>) and the NIST 2014 TRECVID Multimedia Event Detection (MED) challenge (“TRECVID 2014 -- An Overview of the Goals, Tasks, Data, Evaluation Mechanisms, and Metrics” <http://www-nlpir.nist.gov/projects/tvpubs/tv14.papers/tv14overview.pdf>) to provide the analytic metadata for our comparisons. We implemented an automated baseline quality metric that we found in the literature (section 3.2) on the datasets for these evaluations to provide quality data to support the initial correlation experiments with the analytics performance results to drive the framework development. The goal of this study was to both develop the framework to support future research in this area and to determine if an existing automated quality metric would correlate with the performance scores of two different types of analytics. The framework was developed so that it could provide automatically-generated quality-related diagnostic data for future video analytics evaluations as well as support research in automated video quality analysis.

3.2 VIDEO BLIND IMAGE NOTATOR USING DCT STATISTICS

Professor Alan Bovik in University of Texas and his students proposed a blind (no reference or NR) video quality metric in [13]. This metric relies on a spatiotemporal model of video scenes transformed into the discrete cosine transform domain. This also includes a motion model that quantifies motion coherency in video scenes and characterize the type of motion occurring in the scenes, and video statistics and perceptual features that are the basis of a video quality assessment that is used to predict video quality. The Video BLind Image Notator using DCT Statistics (BLIINDS) metric does not require the presence of pristine video to compare against to predict a perceptual quality score. The metric has been successfully tested on the EPFL-PoliMi video dataset. (<http://vqa.como.polimi.it>) Given a video, the Video BLIINDS algorithm will output the predicted Difference Mean Opinion Score (DMOS) (the detail introduction of Mean Opinion Score can be found in [14]), which is a numerical score in [0, 100], where a low score indicates high quality and a high score will indicate low quality.

3.3 STUDY ON THE 2013 NIST PASC FACE RECOGNITION EVALUATION

We designed and conducted an experiment with the performance results data for the 2013 Point-and-Shoot Face Recognition Challenge (PaSC) (“The challenge of face recognition from digital point-and-shoot cameras” <http://ieeexplore.ieee.org/abstract/document/6712704/>). The PaSC challenge is a well-recognized face recognition evaluation and employed a frequently-cited public reference dataset for face recognition research (<http://www.cs.colostate.edu/~vision/pasc/>). In this experiment, we examined the correlation of the Video BLIINDS metric applied to a small subset of this dataset

compared to the performance of three face recognition algorithms in the 2013 evaluation. We measured the correlation between the performance results and the quality metric ranking of the video clips to determine if such a metric could be useful as a diagnostic for face recognition research.

3.3.1 PaSC Data Subsetting

In this experiment, we utilized the system output of three state-of-the-art algorithms in the NIST 2013 PaSC evaluation: PA, local region PCA algorithm (LRPCA), and Pittsburgh Pattern Recognition (SDK 5.2.2). (<http://nvlpubs.nist.gov/nistpubs/ir/2014/NIST.IR.8004.pdf>)

To make the experiments computationally tractable for the Video BLIINDS quality metric (see Section 3.3.2), we selected the system output from the face recognition algorithms for 100 short temporal snippets from the complete dataset. The performance results for each snippet was compared with each of the remaining 99 snippets. We ensured that the system performance distribution for the subset was consistent with the overall distribution of the entire dataset, including accounting for the different cameras that were used in the data collection. We designed and implemented an adapted version of the Greedy Pruned Ordering algorithm based on the methods described in [16] to create the subset so that it was maximally representative of the overall distribution of results for the 3 algorithms.

3.3.2 Computation cost of Video BLIINDS on PaSC data

The first practical issue encountered during this experiment is the impact of high computational cost due to the nature of the open source code of Video BLIINDS. Because the Video BLIINDS algorithm was developed for research purposes, the primary goal is to prove the feasibility of a research concept. Since it is not an industry or commercial-grade development package, this code has not been optimized for computational efficiency and minimum resource impact.

Table 1 shows the computational cost of Video BLIINDS runs on PaSC video data. To mitigate the efficiency issue, we developed two optimized methods to reduce the computation cost of the Video BLIINDS algorithm.

For segmentation, the large videos are segmented into smaller video segments by the fixed length segmentation algorithm and key frame segmentation algorithm respectively. For the fixed length segmentation algorithm, we segment video to a small segment containing 20 frames per segment. We then use the mean or median value of the video qualities of those segments to approximate the overall quality of each video with a reduced computation cost. For the key frame segmentation algorithm, given a large video, we find its key frames, then extract a 20-frame segment around each key frame, and calculate the video quality for each segment. We then obtain the mean or median of all video quality of segments, and use this to approximate the video quality of the whole videos. The table also shows the time used and the values of the original algorithm compared with the approximation algorithm.

Table 1 Video BLINDS computational cost at different pixel resolutions

Reference	Resolution (pixels)	Frame #	File size (MB)	Unzipped frame volume size (3 bytes)	Actual Time (Mins)	Appro. Time (Mins)	Actual Value (Mins)	Appro. Value (Mins)
02463d3613	720*480	115	1.4	720*480*115= 39,744,000	9.1	6.49	83.63	83.04
04385d2834	720*480	242	2,233	720*480*242= 83,635,200	16.79	6.435	71.11	69.28
02463d3464	1920*1080	190	7,915	1920*1080*190= 393,984,000	106.69	37.71	30.81	27.35
04385d2835	1920*1080	267	15,977	1920*1080*267= 553.651,200	233.65 (First 250 frames)	37.88	9.68	6.90

3.3.3 Experimental results of Video BLINDS on PaSC Data Subset

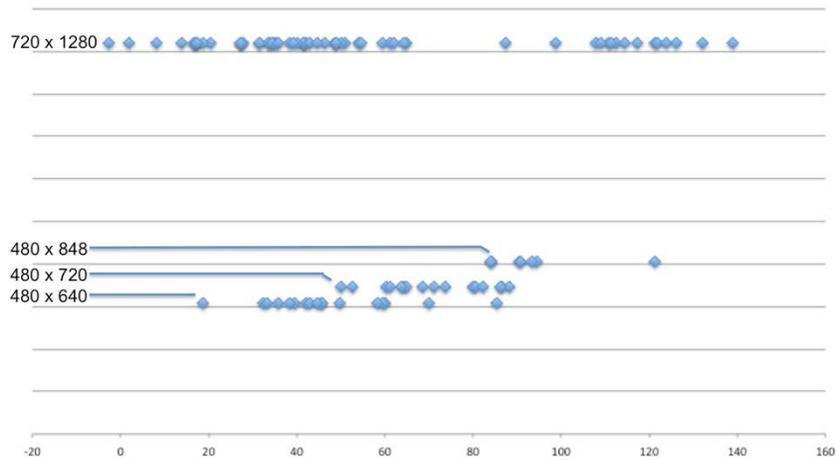


Figure 2 Video quality comparison of pixel resolution vs predicted video quality score

Figure 2 shows that the video resolution is correlated with Video BLIINDS video quality prediction results. The x-axis is the predicted video quality with higher values indicating better quality. The Y axis is the video resolution.

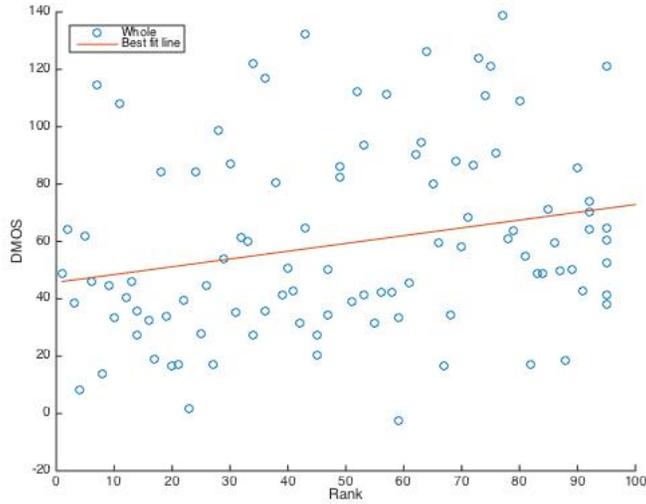


Figure 3 Video quality as estimated by Video BLIINDS (Differential Mean Opinion Score) vs. face recognition performance rank from the PCA system

We analyzed the correlation between Video BLIINDS and the face recognition performance results ranked by Greedy Pruned Ordering (GPO). We took the true positives of the selected datasets with known ranks or scores and compute video quality using Video BLIINDS. Figure 3 shows the correlations between the video qualities with the video rank in a baseline face recognition system (PCA). The correlation coefficient of Principal Component Analysis (PCA) is 0.24.

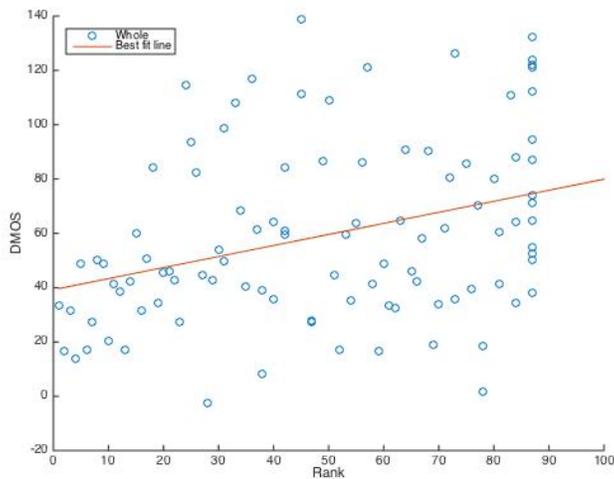


Figure 4 Video quality as estimated by Video BLIINDS (Differential Mean Opinion Score) vs. face recognition performance rank from the LRPCA system using least squares fit

Figure 4 shows the correlations between the video qualities with the video rank in Local Region Principal Component Analysis (LRPCA) system. The correlation of LRPCA is 0.34.

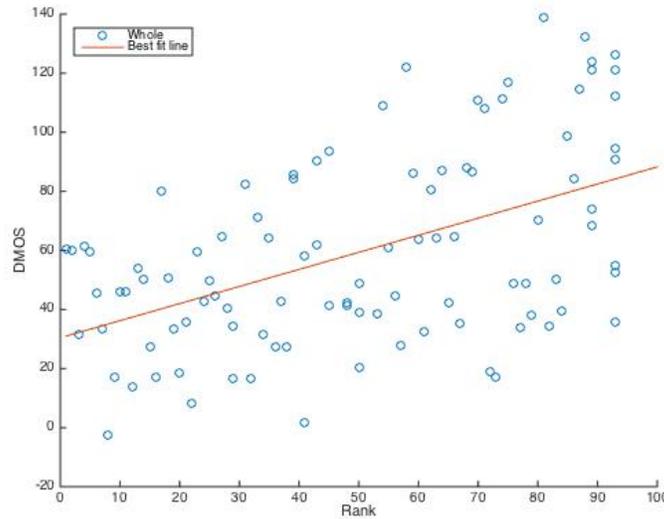


Figure 5 Video quality as estimated by Video BLIINDS (Differential Mean Opinion Score) vs. face recognition performance rank from the Pitt Patt system using least squares fit

Figure 5 shows the correlations between the video qualities with the video rank in Pittsburgh Pattern Recognition (PittPatt) algorithm (Pittsburgh Pattern Recognition SDK 5.2.2) system. The correlation of the PittPatt algorithm is 0.5.

3.3.4 Discussion

Even when utilizing a dataset with relative homogenous recording conditions such as PaSC, we found a correlation between the estimated video quality as measured by Video BLIINDS and the video ranking derived from face recognition system performance, suggesting that there is a measurable relationship between algorithm performance and video quality. Future experiments in this area might further explore this relationship by employing multiple analytics systems in combination with multiple more heterogenous datasets.

3.4 STUDY ON THE 2014 NIST TRECVID MULTIMEDIA EVENT DETECTION (MED) EVALUATION

We performed similar experiments to the PaSC experiment on the algorithm performance data from the 2014 NIST TRECVID Multimedia Event Detection (MED) evaluation.

3.4.1 MED Data Subsetting

The TRECVID 2014 MED video dataset, is a widely cited open video collection developed for video retrieval research which contains a broad variety of internet video and is significantly more varied in recording conditions than the PASC dataset. It is used to evaluate the automatic retrieval of event-related queries (i.e., people interacting with objects and other people). To keep the experiments tractable for the computationally-intensive Video BLIINDS quality metric, video clips from two of the high-motion evaluation events were selected that were likely include significant quality variability: one which contained people performing a trick with a board (skateboard, snowboard, finger board, etc.) and another in which people performed a trick with a bike (motorcycles included). The analytic algorithm system performance scores used for the experiments was developed by Carnegie Mellon University. (http://www.cs.cmu.edu/~yiyang/related_1015.pdf) 18 board trick video clips and 71 bike trick video clips were selected. The Video BLIINDS quality metric was then computed for these clips.

3.4.2 Video quality distributions of different datasets

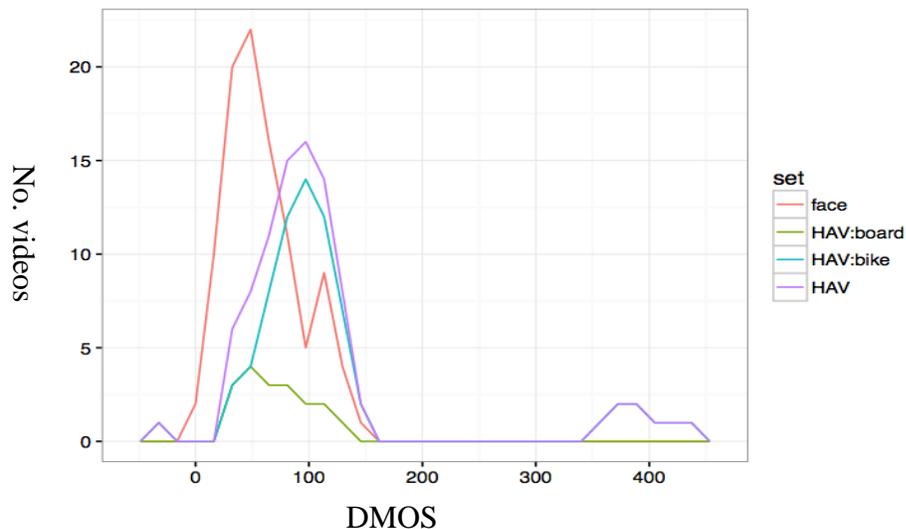


Figure 6 Histogram of video quality of both the PaSC subset and MED subsets

Figure 6 shows the distribution of predicted video quality scores (Video BLIINDS DMOS) of the comparison databases. Notice that the quality scores for the face recognition PaSC subset of 100 videos is, on average, smaller than the quality scores for the MED subset of 89 videos (18 board tricks, 71 bike tricks). The videos in the PaSC dataset are observably of better quality than the MED video clips and were recorded at a higher resolution on average. This difference is measurably depicted in the difference in Video BLIINDS score distributions.

3.4.3 Video BLIINDS on MED HAVIC dataset

Table 2 Video BLIINDS predicted DMOS experiment results on HAVIC MED Bike trick dataset

Video	Height	Width	n_Frame	Duration	DMOS	Score	Rank	Decision
HVC055115.mp4	360	480	149	12.329	92.6209	0.70032983	177	y
HVC069983.mp4	360	480	388	12.945	67.9977	0.573108002	831	y
HVC169168.mp4	240	320	399	26.6	121.8692	0.479660504	3031	n
HVC180595.mp4	240	320	86	5.7333	94.6151	0.500019784	2266	n
HVC180729.mp4	360	480	141	7.0983	101.9115	0.610808069	513	y
HVC180794.mp4	240	320	256	16.9333	112.3086	0.614208112	486	y
HVC180945.mp4	144	192	41	6.56	99.3092	0.53671934	1356	n
HVC180946.mp4	360	480	381	25.4	67.0199	0.572380228	839	y
HVC199785.mp4	288	352	371	43.904	55.6548	0.453042375	4475	n
HVC200031.mp4	144	176	314	21.548	418.9927	0.465581272	3690	n
HVC207041.mp4	240	320	133	9.195	115.9876	0.598756709	590	y
HVC207069.mp4	480	640	356	12.074	-35.6465	0.290080663	55922	n
HVC207707.mp4	480	640	379	13.281	58.4406	0.52830402	1533	n
HVC207788.mp4	144	176	181	13.922	402.01	0.560645415	992	y
HVC207831.mp4	480	720	381	25.681	138.2345	0.496259387	2388	n
HVC207860.mp4	240	320	129	13.374	73.1873	0.49795093	2335	n
HVC207871.mp4	144	176	334	14.071	433.9139	0.510583095	1969	n

Table 2 shows the experiment result of sample Video BLIINDS predicted DMOS experiment results on the MED Bike trick subset. The first column (Video) is the full name of the relevant video file; the fourth column (n_Frame) is the number of frames extracted from the video file; the fifth column (Duration) is the duration of the video in seconds; the sixth column (DMOS) is the predicted quality

score Differential Mean Opinion Score (DMOS) using Video BLIINDS, which is a real number, though scores typically range from 0 to 100; a smaller score denotes a higher quality video. The seventh column (Score) is the analytic confidence score as recorded in the original system output. The eighth column (Rank) is the rank of all the videos in the MED dataset as recorded in the system output, ranked by score. The ninth column (Decision) is whether the CMU (baseline) algorithm detected a bike trick or not.

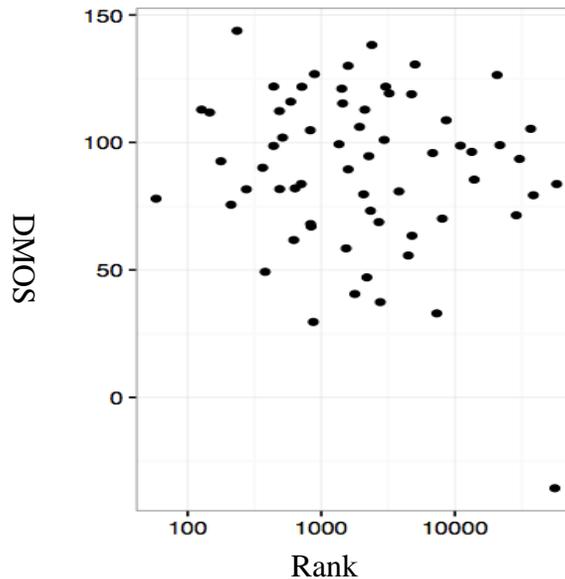


Figure 7 Video quality score predicted with Video BLIINDS of the bike videos vs. their ranks provided by CMU's baseline algorithm (Differential Mean Opinion Score vs score rank)

Figure 7 shows the video quality score predicted with the Video BLIINDS metrics for the bike videos vs. their performance scores derived from the CMU MED system output. It shows that there is almost no correlation between estimated video quality and MED performance rank. Figure 8 shows the video quality score predicted with the Video BLIINDS metrics for the bike videos vs. their analytic confidence scores provided by the CMU system output. It also shows that there is almost no correlation between estimated video quality and MED performance rank.

Although the video quality prediction value DMOS is shown to have very little correlation with either the rank or confidence score, it is premature to conclude that there is no correlation between the quality of the MED video clips and MED algorithm performance. There are a variety of factors that could have swamped the correlation results in the experiment, such as strong signal noise and recording and environmental factors that are not comprehended by the Video BLIINDS algorithm. Of note is that the results showed that the video quality of the MED data is measurably worse than the data that was used to develop the Video BLIINDS model and that the distribution of the MED data is outside of the DMOS

spectrum of performance. Therefore, Video BLIINDS may potentially not be usable with the variety of data that is used in video analytics research.

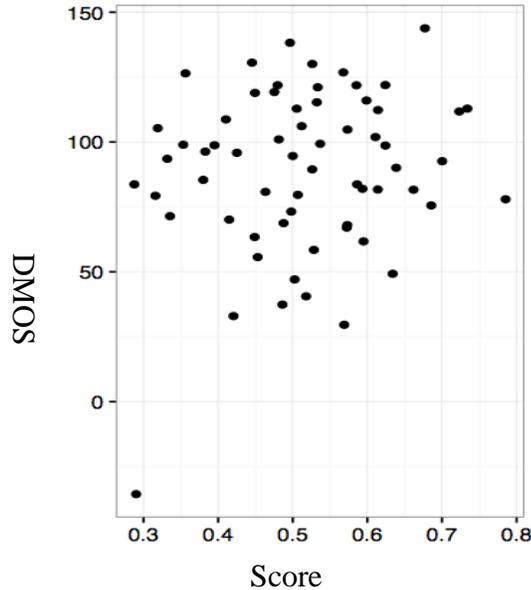


Figure 8 Video quality score predicted by the Video BLIINDS algorithm of the bike videos vs. their analytic confidence scores from the CMU algorithm.

3.4.4 Discussion and Findings

Our study on two different well-known video analytics challenge tasks shows that the video quality metric we studied has a variable correlation to video analytics performance. However, our sample size of both algorithm types and data was extremely small. First, we hypothesize that the quality metric we studied has insufficient expressivity for quality variations outside of the very limited data that was used to originally create and evaluate it. We also hypothesize that existing video analytics datasets do not exhibit the kinds of variability that existing video quality metrics were designed to measure. This finding confirms that datasets need to be designed that more robustly challenge the performance of video analytics with video that varies in quality along important parameters for future public safety applications. Likewise, significant research is needed in quality metrics with regard to video analytics and will require research datasets that can robustly challenge both analytics and video quality metrics. In summary, this effort has shown that significant research is needed both in automated objective video quality measurement and video analytics under varying video quality conditions. With that said, we were able to use existing evaluation data and a baseline video quality metric to develop a framework that could support future robust research in this area.

4 EVALUATION INFRASTRUCTURE

For this effort, we developed an evaluation framework to measure and characterize the performance of content-based video quality metrics. The framework was used to prove feasibility, define a repeatable methodology, evaluate an initial set of automated quality metrics, and motivate requirements for the creation of a future comprehensive framework and reference data collections to support research at the intersection of video quality and video analytics.

4.1 EVALUATION FRAMEWORK

We proposed an evaluation framework for Video Analytics-based Quality Measurement System. The evaluation framework consists of validation framework, VQM framework, video analytics framework, and the overall cross validation framework.

4.1.1 Validation Framework

Both the video analytic and quality metric to be assessed must be validated as part of the import process into the evaluation framework. The video quality metric validation process is shown in Figure 9.

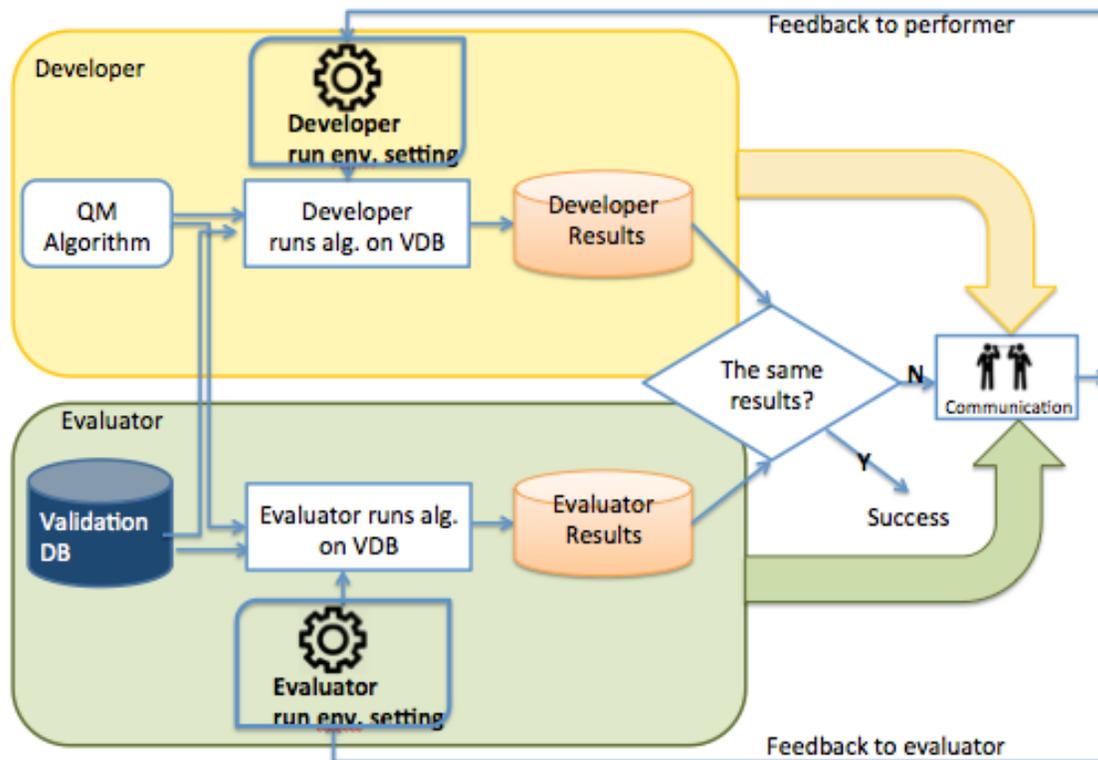


Figure 9: VQM Validation Process.

4.1.2 Overall Framework

The overall framework for Video Analytics-based Quality Measurement System is shown in Figure 10. As previously depicted, there are two major components in the evaluation framework: a quality metrics sub framework and a video analytics sub framework. Based on these two sub-frameworks, the evaluation tool can perform cross-evaluation and study the relationship between video analytic performance and the video quality score.

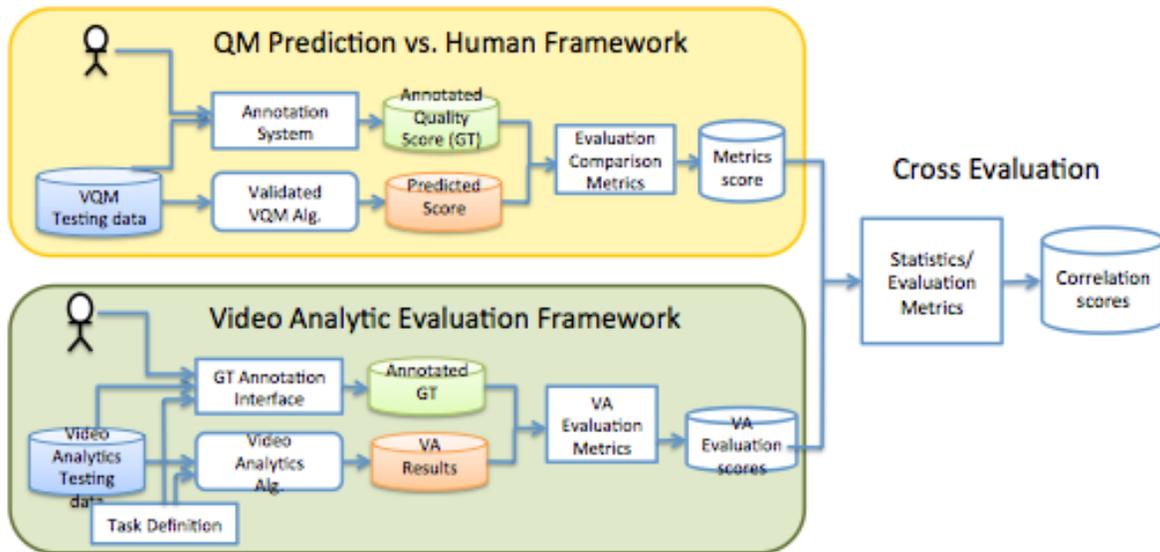


Figure 10: The Video Quality Metrics vs. Video Analytic Cross Evaluation Framework.

Using the proposed infrastructure, we can report:

- (1) correlations among video quality metrics and the human annotated video quality values;
- (2) correlations among video analytics system performance and the human annotated video analytics ground-truth;
- (3) evaluations of emerging automatic video quality measurement technologies;
- (4) evaluations of emerging video analytic technologies;
- (5) cross evaluation of video quality and video analytics system performance.

4.2 NIST DETECTION ANALYSIS PIPELINE RESOURCES (DAPR) INFRASTRUCTURE

The evaluation framework was implemented using a MySQL database with the NIST Detection Analysis Pipeline Resources (DAPR) Data Model developed at NIST to support a variety of speech and video analytic performance evaluations. The benefits of the DAPR analysis-centric infrastructure are: it separates different types of technology specific data from evaluation mechanics and it can organize and manage different kinds of metadata easily. DAPR can easily support multiple metrics

(embedded or plugin). It supports the controls of complexity and a generalized nomenclature. The overview of DAPR structure is shown in Figure 11.

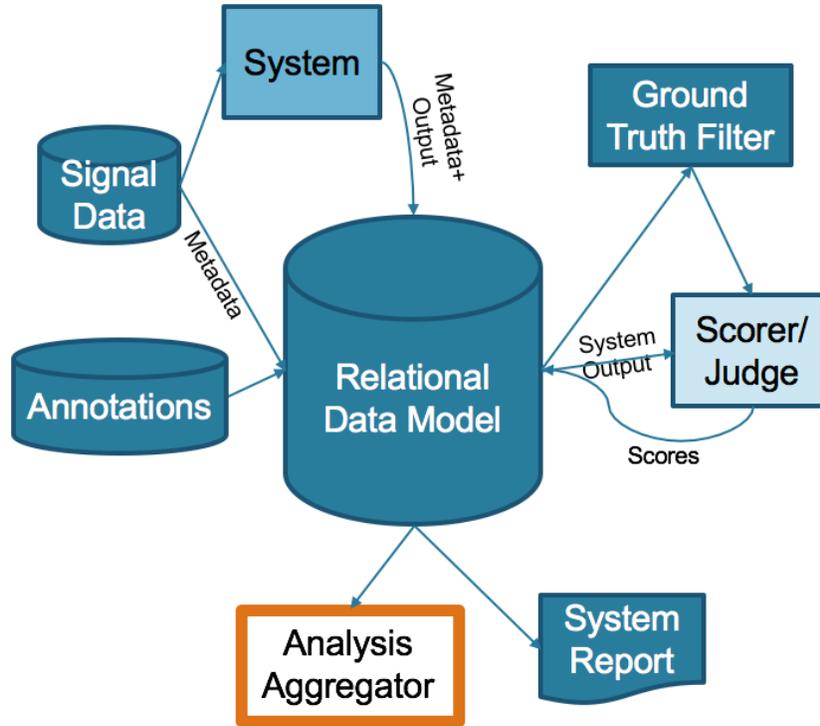


Figure 11: Analysis-Centric Evaluation Infrastructure

4.3 DAPR FOR VIDEO QUALITY METRICS AND VIDEO ANALYTICS EVALUATION

We developed an infrastructure and used a baseline automated quality testing framework to incorporate a representative subset of automated quality metrics and analytics.

4.3.1 Infrastructure Implementation

Figure 12 shows the major modules and tables associated with them in the DAPR infrastructure. This model is used for the organization and analysis of the data. DAPR supports as reusable API design. It consists of the following modules: data import module, ground-truth generation module, task definition, annotation data and related files. The DAPR output is ground-truth table, scoring module, factor analysis module, and reporting module.

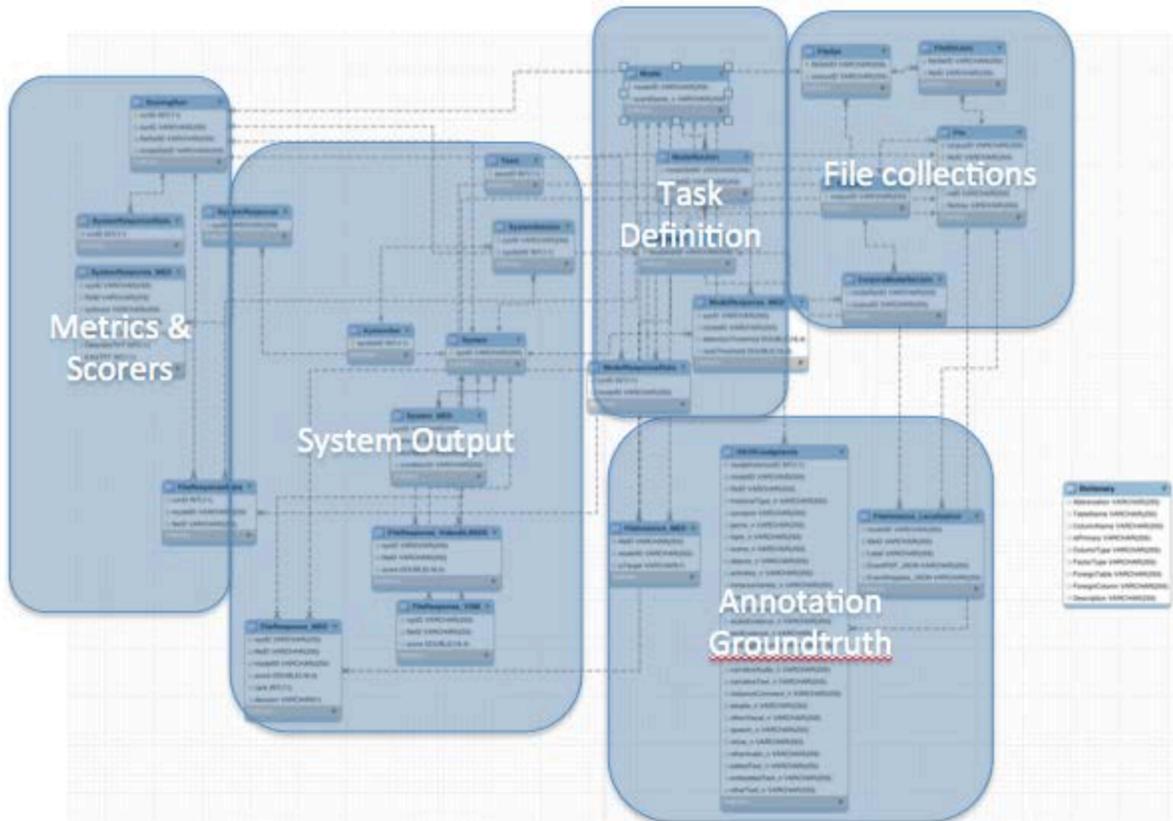


Figure 12: The Detection Analysis Pipeline Resources (DAPR) Data Model.

The DAPR evaluation infrastructure supports multiple databases to include reference DMOS from a subset of the PaSC dataset, and the subset of MED dataset: bike trick and board trick datasets. As public safety research evolves, we expect to have access to public safety video datasets in the future.

4.3.2 Cross Evaluation Metrics

There are many correlation metrics that could be adapted in our evaluation infrastructure. For example, correlations used in MetricsMaTr program provides an approach to develop correlated machine translation automatic metric vs. human assessment output. Three correlation measures are used in Machine translation metrics matters (MetricsMaTr): Spearman's Rho, Kendall's Tau, and Pearson's R at different levels: segment, document, and system level. The datasets and results can be found at: <http://www.itl.nist.gov/iad/mig//tests/metricsmatr/2008/results/correlationResults.html>.

We adopt the same approach to implement two additional correlations, Spearman's Rho, and Pearson's R in our infrastructure using MySQL.

5 FUTURE DIRECTIONS REGARDING EVALUATION FRAMEWORKS AND DATA COLLECTIONS

5.1 FUTURE EVALUATION FRAMEWORK REQUIREMENTS

A framework that supports future robust evaluation would require the following components:

- (1) Datasets with ground-truth human annotation for both VQM and video analytics (VA)

An initial small collection with approximately 100 video clips with a constrained diversity of people, scenes, and events would provide great utility in beginning to foster research in this area. This will service as an initial proof-of-concept video dataset. Subsequent collection for the benchmark dataset will require the creation of significantly larger collections of annotated videos that support the diversity of quality conditions and video content necessary to develop, evaluate, and field applications with a high confidence in performance and measured understanding of uncertainty.

- (2) Annotation tools and best practices

The development of best practices and tools for efficient ground truth annotation of both quality and content for analytics are needed to support a future evaluation program.

- (3) Evaluation infrastructure (DAPR)

The evaluation framework tool would need to be expanded to support a wide variety of analytics as well as diagnostic metrics for measuring quality with regard to analytic performance. Development of a rich set of evaluation-focused research tools will support agile research in both quality and analytics. Further refining of the existing framework would support its growth in the open source community.

- (4) Baseline algorithms and their performance

In order to perform further research in this area, the research community needs access to both baseline algorithms for quality measurement as well as video analytics so that agile collaborative research is enabled. There is a significant repository of video analytic algorithms that are now available in open source. However, there are few quality analytics in the open domain that would support research. Fostering the development of open source quality tools would support significant potential growth in this area.

5.2 DATA COLLECTION PLAN

We are beginning interagency discussions to develop design requirements for a future dataset and challenge problem set that would support research in both video quality measurement and video analytics. The following subsections describe guidance we have currently compiled.

5.3 INITIAL PROOF-OF-CONCEPT VIDEO DATASET DESIGN

There are numerous factors involved in video dataset collection for both quality and analytics research and evaluation. Traditionally, such datasets have focused on a very narrow set of challenges and factors

to ensure that there is sufficient data to support the evaluation of significance in the research. We must necessarily prioritize a set of factors to keep both the collection effort and size of the dataset for research tractable while strategically driving research in necessary technology gaps. Based on lessons learned from our study and experiments on video quality metrics for video analytics, we proposed to prioritize the following factors for a future video collection that would support intersectional research on video quality and video analytics:

- (1) Video capture device: the video quality heavily depends on capture devices. Capture devices cover a wide range, from Hollywood-specific digital and analog capture devices to low quality, low resolution cell phone cameras. In addition, police body camera's video quality varies greatly. The sensor size, the year of camera release and the type of camera such as smartphone, compact, Digital Single-Lens Reflex (DSLR), Closed-circuit television (CCTV), and studio quality cameras are all factors that contribute to the initial dataset design and should be considered. It's important that a representative set of collection devices are included so that research addresses the broad challenges.

The mounting position of the video capture device is another important factor. The video analytics algorithms for videos captured from a fixed position device mounted on a solid object, with pan/tilt/zoom, or a tripod with no vibration may greatly differ from videos captured using body-worn cameras. We may classify the videos by mount factors, camera motion and setting, which may greatly affect the analytic system's performance.

The video capture device setting is another factor for consideration. Frame rate, image resolution, scan setting (interlace or progressive) are other factors that may affect video quality and performance of video analytics.

- (2) Video source: the initial database should cover a large variety of sources. For example, commercial advertisement samples, home security video data samples, airport surveillance video data samples, body camera video samples, Flickr videos, traffic management video, drone videos, etc.
- (3) Video content: the video content will have a strong impact on the analytic system performance. For example, simple scenes that include just a face would be simpler to detect than video from a complex marathon scene for face detection. The initial dataset should cover the video scene complexity. Pinson *et al.*[1][2] described important video characteristics, such as degree of motion, level of detail, color strength, contrast and sharpness, that have a major impact on VQM development. Many video analytics applications are related to the detection of people. The aspect and number of people in a video is another factor that can be collected: no people, one person head & shoulders, multiple people, crowd, etc. A description of video with free text annotation used to describe what happens is also helpful, and structuring the categorization of data based on similar content is even better.

The video realism is another factor that may affect the real system's performance. The video analytics algorithms trained by video collected simulations designed from real situations or artificial scenarios (e.g., man walks past camera holding a cup) may not work well or accurately represent the same performance of videos collected in real situations.

- (4) Capture environment parameters: lighting condition (full sun, partial shade, dusk or dawn, night, artificial lighting, outdoor day time, outdoor evening, indoor diffuse light, spot light) is the major challenge of video analytics. Indoor environment with controlled lighting and outdoor environment also may have great impact on analytic algorithm performance. Other capture factors like the physical distance of the object and camera, light capture mode (normal YCbCr / RGB, infrared) are all factors that need to be collected.
- (5) Video production quality (professional, consumer enthusiast or amateur) is another factor for collection. Compression format, parameters, and video duration are also interesting factors for the database collection. Once we have original videos, the metadata and annotation data are also needed.

For the initial data collection, we need to pre-define the analytic tasks first to ensure that the collection is populated with the necessary instances of the phenomena to be researched. Face and object detection and recognition are good analytics to focus on since they are well-studied domains and many algorithms are readily available. The measurement of quality in these areas is also extremely important for the public safety and forensic domains. Studying the effect of human annotation is also important since humans play a large role in these analytic tasks.

5.4 GENERAL DATA COLLECTION REQUIREMENTS

A critical requirement for data collection is usage rights. If the end users do not have rights to use and freely distribute for research and development purposes, then the data set has very limited value. Many datasets do not clearly specify usage rights, and so people cannot use it. Many organizations host a dataset for a long time, but this is problematic since funding may cease and datasets may no longer be available. One possible solution is Consumer Digital Video Library (CDVL)[3]. CDVL also hosts quite a few PSCR videos.

Based on the current literature survey results regarding the research work on video quality analytics and video quality metrics domain, we would propose a conceptual database collection design that specifies the key factors that should be considered when collecting data for capture into the database. Both the metadata related with collection conditions (subject, lighting, etc.) and the details of how the video is collected and processed may impact the video quality and video analytics algorithm performance. The factors/parameters for each category suggested in previous section are listed in each subsection below.

5.5 VIDEO CAPTURE DEVICE

5.5.1 Type of device

- Device model: which may cover other factors like sensor type and size, manufacture year, etc.

- Device type: smartphone, body-worn, compact, Digital Single-Lens Reflex (DSLR), Closed-circuit television (CCTV), studio quality professional capture device.

5.5.2 Device mount and geolocation

- Geolocation of devices
- Device registration
- Still mount without vibration: on tripod or still object.
- Still mount with vibration: fixed on vibration object such as traffic light arm on a windy day.
- Smooth motion: fixed on object with motion (e.g. vehicle such as an automobile).
- Moderate motion: (e.g., hand held device with image stabilizer)
- Severe motion: bodycam or moving camera with bouncy video without image stabilizer.
- Pan/Tilt/Zoom (PTZ): with or without.

5.5.3 Device settings

- Frame rate: the number of frames captured per second.
- Image/video resolution: more pixels generally (but not necessarily) improve the video analytic algorithm performance.
- Interlace or progressive scan: Interlace refers to even / odd number of lines in image; progressive means all lines are captured and processed in sequence.
- Aspect ratio

Other capture characteristics listed below are preferred. These are optional based on availability:

- Contrast: the difference between the light and dark parts of an image.
- Exposure: Video that is overexposed or underexposed makes analytic tasks difficult and at times impossible.
- Aperture: one of three parameters that determines proper exposure.
- Shutter speed: one of three parameters that determines proper exposure.
- ISO: one of three parameters that determines proper exposure.
- White/color balance: the parameter that adjust the color balance on a digital camera.
- Dynamic range: when you cannot find a single exposure that can capture both the brightest and darkest areas, dynamic range helps.
- Focal length: focal length and sensor size determine field of view.
- Optical aberration, and distortion: it affects the video boundary video analytics performance.

5.6 VIDEO SOURCE

The factors/parameters that should be considered includes but is not limited to the video purpose, application, and usage etc.

- Commercial advertisement
- Home security video
- Airport surveillance video

- Body camera video
- Flickr videos
- Traffic management video
- Drone videos
- Warehouse or shopping mall product display video
- Public safety video
- Military video
- Geographic video
- Daily life video etc.

5.7 VIDEO CONTENT

Pinson *et al.*[1][2] listed some video content characteristics that have a major impact on VQM development, such as:

- Motion and motion blur,
- Details: repetitious or indistinguishable fine details (e.g., gravel, grass, hair, ...)
- Number of moving objects
- In-focus foreground or blurred background
- Analog noise
- Night scene or dimly lit scene

5.8 CAPTURE ENVIRONMENT

5.8.1 Lighting condition

- Indoor/outdoor
- Outdoor:
 - shading, clouds, transient shadows
 - time of day
- Indoor: light type –
 - artificial lighting
 - indoor diffuse light
 - spot light

5.9 VIDEO PRODUCTION QUALITY

- Video length or duration: the volume of information captured.
- Video coding format: MPEG-2, H.264, and H.265 (HEVC)
- Compression parameters

Other factors listed below are preferred:

- Bitrate: Constant bitrate (CBR) or variable bitrate (VBR).
- Color subsampling: 4:2:0 vs. full color resolution (4:4:4).

5.10 VIDEO ANALYTIC ANNOTATIONS

5.10.1 Analytic Object Parameters

- Object pixel size: having more pixels on the subject is likely to make it easier to identify the object.
- Object physical size: the physical size of the object.
- Object category: ImageNet object category definition.
- Object characteristics: rigid/non-rigid; texture (with/without reflection);
- Object position related factors
 - Object position within the scene
 - Object View angle
 - Camera-object distance
- Object motion related
 - Object moving speed (absolute and relative speed to the camera)
 - Motion parameters (absolute vs. relative to the camera)
 - Motion blur

5.10.2 Analytic Event Parameters

- The event descriptions: e.g. wedding or political speech.
- Human annotated video analytics ground-truth: the ground-truth of the video analytics, for example, person A's face is in the video for the face detection and recognition; this is wedding event for event detection video analytic system.

5.11 VIDEO QUALITY ANNOTATIONS

The following annotation parameters are needed:

- Human annotation on the video quality: Human annotated Mean Opinion Score (MOS).

5.12 VIDEO ANALYTIC SYSTEM OUTPUT

The following video analytic system output or video metrics prediction system output are needed:

- Video analytics algorithm detection outputs from different video analytics systems. The content elements related with video analytics system are required to be annotated. For example, the face should be labeled with a ID and located in the video with a bounding box for face detection and recognition system.

5.13 VIDEO QUALITY METRIC SYSTEM OUTPUT

- The video quality metrics predicted output from different video quality metrics systems.
- The statistical density distributions of key factors should be satisfied. Thus, we can ensure a measurable level down to a specific level of uncertainty. For example, the statistical distribution of human annotated quality should cover the whole quality spectrum.

In summary, data collection design is a very critical step for a successful benchmark process involved in a dataset collection. In general, before we could explore the salient characteristics of an unknown domain, we prefer a breadth-first approach instead of a depth-first approach. When we have gained a more detailed understanding about the described domain, we can fine-tune our design and customize it leveraging current state-of-the-art research.

Generally, data collection is tedious, time-consuming, and costly. We can mitigate issues arising from use of an existing dataset, annotating factors and parameters that are missing. We also continue to advocate and pursue a public safety collection.

6 CONCLUSIONS AND WAY FORWARD

In this effort, we developed a framework and toolset to evaluate the predictive value of objective video quality metrics with regard to the performance of video analytics algorithms. We performed an informal survey of the state-of-the-art in video quality metrics and found a dearth of research in this area. We utilized the framework to perform initial experiments using a baseline academic metric we found with face recognition and event detection performance results data that we obtained from previous NIST evaluations. We learned that there was only a weak correlation of the academic metric with the performance of the video analytic algorithms, likely due to its very limited training on non-representative data as well as limited variability in the quality of the data that has historically been used to evaluate video analytics performance.

We hypothesize that reference data is needed that exercises both the range of challenges in video quality along with range of challenges in key video analytic tasks to properly explore this area. Moreover, it's clear that a robust research effort is needed to develop the video quality measurement technologies needed to support future needs in public safety video workflows. Such an effort will also require such data and measurement tools.

We have begun the specification process for future expansion of the framework and dataset that would be necessary to properly support robust R&D and evaluation efforts for future automated objective video quality metrics. We shall make the framework and tools developed in this work available to the public and we'll make the framework and data requirements available to Federal research organizations to support future critical research in this area.

7 ACKNOWLEDGEMENT

We would like to thank Dr. Jonathon Phillips and his team for their assistance in providing the PaSC dataset, baseline PASC system performance results and input into future requirements for video data. We would also like to thank Margaret Pinson for her extensive input on applied needs and optical variability requirements for future video research data to support video quality metrics.

This research was supported by Office for Interoperability and Compatibility (OIC), Department of Homeland Security (DHS), Public Safety Communications Research (PSCR) funding.

8 DISCLAIMER

Any mention of commercial products or reference to commercial organizations in this report is for information only; it does not imply recommendation or endorsement by NIST nor does it imply that the products mentioned are necessarily the best available for the purpose.

9 REFERENCES

- [1] M. H. Pinson, "Video Quality Measurement User's Manual," Feb. 2002.
- [2] C. Fenimore, J. Libert, and S. Wolf, "Perceptual Effects of Noise in Digital Video Compression," *SMPTE J.*, vol. 109, no. 3, pp. 178–187, Mar. 2000.
- [3] J. Guo, M. V. Dyke-Lewis, and H. R. Myler, "Gabor difference analysis of digital video quality," *IEEE Trans. Broadcast.*, vol. 50, no. 3, pp. 302–311, Sep. 2004.
- [4] T. Oelbaum, C. Keimel, and K. Diepold, "Rule-Based No-Reference Video Quality Evaluation Using Additionally Coded Videos," *IEEE J. Sel. Top. Signal Process.*, vol. 3, no. 2, pp. 294–303, Apr. 2009.
- [5] M. C. Q. Farias, M. Carli, and S. K. Mitra, "Objective video quality metric based on data hiding," *IEEE Trans. Consum. Electron.*, vol. 51, no. 3, pp. 983–992, Aug. 2005.
- [6] M. C. Q. Farias and S. K. Mitra, "No-reference video quality metric based on artifact measurements," in *IEEE International Conference on Image Processing 2005*, 2005, vol. 3, p. III-141-4.
- [7] F. Yang, S. Wan, Q. Xie, and H. R. Wu, "No-Reference Quality Assessment for Networked Video via Primary Analysis of Bit Stream," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 11, pp. 1544–1554, Nov. 2010.
- [8] Y. Kawayoke and Y. Horita, "NR objective continuous video quality assessment model based on frame quality measure," in *2008 15th IEEE International Conference on Image Processing*, 2008, pp. 385–388.
- [9] T. Brandao, L. Roque, and M. P. Queluz, "Quality assessment of H. 264/AVC encoded video," in *Proc of conference on telecommunications-ConfTele, Sta. Maria da Feira, Portugal*, 2009.
- [10] T. Brandao and M. P. Queluz, "No-Reference Quality Assessment of H.264/AVC Encoded Video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 11, pp. 1437–1447, Nov. 2010.
- [11] Osamu, S. Naito, S. Sakazawa, and A. Koike, "Objective perceptual video quality measurement method based on hybrid no reference framework," in *2009 16th IEEE International Conference on Image Processing (ICIP)*, 2009, pp. 2237–2240.
- [12] O. Sugimoto and S. Naito, "No reference metric of video coding quality based on parametric analysis of video bitstream," in *2011 18th IEEE International Conference on Image Processing*, 2011, pp. 3333–3336.
- [13] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind Prediction of Natural Video Quality," *IEEE Trans. Image Process.*, vol. 23, no. 3, pp. 1352–1365, Mar. 2014.
- [14] R. C. Streijl, S. Winkler, and D. S. Hands, "Mean opinion score (MOS) revisited: methods and applications, limitations and alternatives," *Multimed. Syst.*, vol. 22, no. 2, pp. 213–227, 2016.
- [15] J. R. Beveridge *et al.*, "The challenge of face recognition from digital point-and-shoot cameras," in *2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, 2013, pp. 1–8.

- [16]P. J. Phillips *et al.*, “On the existence of face quality measures,” in *2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, 2013, pp. 1–8.