International Journal of Computer Information Systems and Industrial Management Applications.
ISSN 2150-7988 Volume 4 (2012) pp. 392–399

Mining Web Videos for Video Quality Assessment

Dubravko Culibrk[1], **Milan Mirkovic**[2], **Predrag Lugonja**[3] **and Vladimir Crnojevic**[4]

University of Novi Sad, Faculty of Technical Sciences,
Trg Dositeja Obradovica 6, 21000 Novi Sad, Serbia

[1] *dubravko.culibrk@gmail.com*

[2] *mirkovic.milan@gmail.com*

[3] *elpee@gmail.com*

[4] *crnojevic@uns.ac.rs*

Abstract: **Correlating estimates of objective measures related to the presence of different coding artifacts with the quality of video as perceived by human observers is a non-trivial task. There is no shortage of data to learn from, thanks to the Internet and web-sites such as YouTube^{tm}. There has, however, been little done in the research community to try to use such resources to advance our understanding of perceived video quality. The problem is the fact that it is not easy to obtain the Mean Opinion Score (MOS), a standard measure of the perceived video quality, for more than a handful of videos.**

The paper presents an approach to determining the quality of a relatively large number of videos obtained randomly from YouTube^{tm}. Several measures related to motion, saliency and coding artifacts are calculated for the frames of the video. Programmable graphics hardware is used to perform clustering: first, to create an artifacts-related signature of each video; then, to cluster the videos according to their signatures. To obtain an estimate for the video quality, MOS is obtained for representative videos, closest to the cluster centers. This is then used as an estimate of the quality of all other videos in the cluster.

Results based on 2,107 videos containing some 90,000,000 frames are presented in the paper.

Keywords: Video quality assessment, internet data, data mining, YouTube^{tm}

I. Introduction

As the amount of multimedia content generated and consumed grows, there is an increased need to measure and assess the quality of video sequences, as it is perceived by the consumers. The quality depends on the video codec, bit-rates required and the content of video material. User oriented video quality assessment (VQA) research is aimed at providing means to monitor the perceptual service quality.

Overall degradation in the quality of the sequence, is a compound effect of different coding artifacts [1].

A large number of published papers exist that propose different measures of prominent artifacts which appear in coded images and video sequences [2][3]. The goal of each no-reference approach, as is the one proposed in the text below, is to create an estimator based on the proposed features that would predict the Mean Opinion Score (MOS)[4] of human observes, without using the original (not-degraded) image or sequence data.

The procedures for obtaining MOS for a set of degraded sequences are laborious as they involve large number of persons viewing the sequences repeatedly and then averaging the subjective quality score each person gives to the sequence. The size of data sets that can be labelled in such manner makes them more suitable to serve as test sets for approaches relying on designed mathematical models of aspects of the Human Visual System (HVS), than for using machine learning techniques to create an estimate of the MOS. This, of course, is due to the fact that the latter require larger amounts of data, as they first need to undergo training, testing and (often) validation, using different parts of the data set. The quality of data that an algorithm learns from is very important and incomplete and unbalanced data will lead to the learning algorithm being unable to learn the target concept or overfit the data set used for training [5].

Even a carefully selected data set, such as that created by the Video Quality Experts group [6] for the specific purpose of evaluating the quality of video coding and decoding algorithms, leaves something to be desired when machine learning is concerned. In addition, the same study indicates that the results of approaches relying on (HVS) models can be significantly improved upon using machine learning.

The aim of the work described here, is the design of an approach that would use large numbers of multimedia data available on the Internet to determine, using machine learning methods, the appropriate subset of the content that needs to be viewed by the human observers. Doing this using k-means clustering leads to a straightforward VQA approach, which labels the input data according to the obtained MOS of representative sequences. The set of representative sequences contains sequences that fit best into the cluster, for each cluster. To evaluate the quality of the clustering, MOS scores of representative sequences were obtained by subjective assessment and a variance analysis of the opinion scores conducted. The rest of the paper is organized as follows: Section II provides an overview of the relevant published work. The proposed VQA methodology is described in Section III.

Section IV presents the experiments conducted to evaluate the applicability of the proposed method. Conclusions can be found in Section V.

II. Background and related work

When publicly available multimedia content is concerned, videos are usually available only in their coded form. Therefore, the work presented in this paper relates to no-reference video quality assessment methodologies. No information regarding the original (not-coded) video is used to estimate video quality, as perceived by human observers.

In studies such as this, ground truth is usually established in the form of a subjective quality measure mean opinion score (MOS), which is obtained by averaging scores from a number of human observers[1]. The correct procedure for conducting such experiments, in the work presented, was derived from ITU-R BT.500-10 recommendations[4].

The first stage of any no-reference approach is the calculation of metrics designed to quantify the presence of certain predefined artifacts or video features, which can then be related to overall quality for a specific application [1].

Overall degradation in the quality of the sequence, due to encoder/decoder implementations as part of transport stream at various bit rates, is a compound effect of different coding artifacts. Three types of artifacts are typically considered pertinent to DCT block coded data: blocking, ringing and blurring. Blocking appears in all block-based compression techniques due to coarse quantization of frequency components [2][7]. It can be observed as surface discontinuity (edge) at block boundaries. These edges are perceived as abnormal high frequency components in the spectrum. Ringing is observed as periodic pseudo edges around original edges [8]. It is due to improper truncation of high frequency components. In the worst case, the edges can be shifted far away from the original edge locations, observed as false edge. Blurring, which appears as edge smoothness or texture blur, is due to the loss of high frequency components when compared with the original image. Blurring causes the received image to be smoother than the original one [9].

Measures related to various artifacts are usually evaluated for each frame of the sequence and collapsed temporally to arrive at a quality measure for the whole sequence [10][11][12][3].

Machine learning methods have rarely been used to build MOS estimators. An Multi Layer Perceptron (MLP) neural network estimator was used by Babu and Perkis [13], to estimate MOS for JPEG coded images. In a recent paper Culibrk *et al.* used the same learner to estimate MOS of videos [3]. They evaluated 18 different previously published measures related to image and video quality to determine their suitability to serve as features used in video quality estimation. The final set of measures they selected included no-reference measures proposed by Wang *et al.* for blockiness and blurring [2], as well as a blockiness measure proposed by Babu *et al.* [14].

Most perceived-blockiness measures are based on the notion that the block-edge-related effects can be masked by high spatial activity in the image itself, and that the blockiness cannot be observed in very bright and very dark regions. Wang *et al.*[2] proposed a no-reference approach to quality assessment in JPEG coded images. Their final measure is derived as a non-linear combination of a blockiness, local activity and a so-called zero-crossing measure. The combination is supposed to provide information regarding both blockiness and blurring (via the two latter measures) in JPEG coded images. The blockiness measure of Babu *et al.* [14] takes effects along each edge of the block into account separately. Thus, they derived a measure surpassing the Wang *et al.* approach.

Motion [11] and attention [10] can adversely affect MOS estimation. In [15], authors proposed using a multi-scale background-subtraction approach to detect salient motion in the frame. Based on this information and intra-frame measures proposed by Wang *et al.* [2] and Babu *et al.* [14], they proposed a set of seventeen measures to describe salient motion, blockiness and blurring, for salient and non-salient regions separately. These measures were used to train MOS estimators based on MLP and M5' decision trees [5].

Data sets used to test various metrics and MOS estimators are based on a small number of carefully (and manually) selected short sequences, which are then impaired using different coding algorithms and settings to form the final data set [14][3][16][10]. The sequences are selected to represent different types of content deemed pertinent by the authors. To the best of our knowledge, there have been no attempts to use an automatic procedure to create a reference set of sequences.

Video Quality Experts Group provides a representative set of test sequences, designed specifically for codec testing. These have been used by several authors [11][3][1][15]. The set consists of 8-second scenes comprising both natural and computer-generated scenes with different characteristics (e.g. spatial detail, color, motion) and was selected by independent labs. 10 scenes with a frame rate of 25 Hz and a resolution of 720×576 pixels as well as 10 scenes with a frame rate of 30 Hz and a resolution of 720×486 pixels were created in the format specified by ITU-R Rec. BT.601-5 (1995) for 4:2:2 component video.

In the work presented here, seventeen measures related to salient motion, blockiness and blurring, as proposed in [15], were used to describe frames of the sequence. We then proceed to cluster the frames of the sequence to arrive at a *quality signature* of the sequence, consisting of 100 cluster centroids. The signatures for all the sequences in the data set are, subsequently, clustered to arrive at a set of representative sequences for which subjective MOS is obtained.

III. Video quality assessment

A block diagram of the proposed video quality assessment approach is shown in Fig 1. In *Phase 1*, each video in the data set is processed to extract measures related to motion, salient changes, blurring and blockiness. A total of 17 measures is extracted for half of the frames of video, distributed uniformly - once the measures have been calculated for a frame, the next frame is skipped. This increases the efficiency of the approach without affecting the effectiveness. The values of measures for all frames of a single video are clustered into 100 clusters using k-means clustering [17]. The process yields 100 cluster centroids that represent each video. The set of centroids is a fixed-size representation of a video, regardless of the number of frames it has. This will be referred to as a *Video Quality Signature* (VQS) henceforth.

Phase 2 starts when all the VQSs have been calculated. The values in the all the VQSs undergo another round of clustering, to associate sequences of similar quality. Each VQS is assigned 100 cluster labels - one for each centroid in it, by this final clustering. The cluster a VQS and the corresponding video belong to is determined by majority voting. The cluster that most elements of the VQS belong to, is the cluster the VQS belongs to. Each cluster has one or more videos that are the best representatives of that particular cluster. These are the videos that have the most representatives of that cluster in their VQS. For each cluster, a single of video is selected from those that represent the cluster best. The selected sequences form the data set of representatives videos for which MOS should be measured using human observers. Once MOS values are obtained, the MOS of the representative sequence is propagated to all the videos in the same cluster. This is the estimated MOS, designated in Fig. 1.

The set of features related to video quality was adopted from [15]. Table 1 lists the features used.

The approach proposed in [15] attempts to estimate salient motion in each frame of the sequence and estimate the extent of different coding artifacts in the salient and non-salient parts of the frame separately.

A. Detecting Salient Motion

While the methodology for the detection of salient motion is described in detail in [15], an overview of the algorithm is included here, for the sake of completeness.

The method employs a multi-scale model of the background in the form of frames which form a Gaussian pyramid. This allows the approach to account for the spatial coherence and cross-scale consistency of changes due to motion of both camera and objects. Even with a small number of scales (3-5), the approach is able the achieve good segmentation of interesting moving objects in the scene. Moreover, it is able to do so consistently over a wide range of the amount of coding artifacts present.

The background frames at each level are obtained by infinite impulse response (running average) filtering. This allows the approach to take into account temporal consistency in the frames. Finally, outlier detection [18] is used to detect salient changes in the frame. The assumption is that the salient changes are those that differ significantly from the changes undergone by most of the pixels in the frame.

Each frame of the sequence is iteratively passed to a 2D Gaussian filter and decimated to obtain a pyramid of frame representations at different scales. A background model is maintained in the form of two (background) frames updated in accordance with Eq. 1.

$$b_l(i) = (1 - \alpha_l)b_l(i) + \alpha_l p(i), l \in \{1, 2\} \qquad (1)$$

where: α_l is the learning rate used to filter the l-th background frame, $p(i)$ is the value of pixel at location i in the current frame, $b_l(i)$ is the value of pixel at location i in the l-th background frame.

The initial values for the background frames are copies of the first frame of the sequence. As Equation 1 suggests, the data observed in the frames of the sequence is slowly incorporated into the background. The two background frames are obtained using different learning rates ($\alpha_1 \neq \alpha_2$), allowing for better adjustment of the time taken by the model to adjust to a scene change. Throughout the experiments presented in this paper the relation of $\alpha_2 = \alpha_1/2$ was used, as suggested in [19]. Therefore, the first reference frame incorporated changes twice as fast as the second one. In addition, since the bottom-up saliency of an object dominates the visual search in about $30ms$ after the viewer is confronted with a visual scene, the value of α_1 is set to 0.3 times the reciprocal of the frame rate, i.e. for the sequences with 30 frames per second (as those used in our experiments), α_1 is set to 0.01.

Temporal filtering is then performed to obtain a single image indicating the extent to which the current frame differs from the background frames. This is equivalent to inserting the current frame between the two background frames and employing a temporal filter in the form of Mexican hat function, given by the equation 2.

$$f(x) = -\frac{2}{\sqrt{3}}\pi^{-\frac{1}{4}} \cdot (1 - x^2) \cdot \exp\frac{-x^2}{2} \qquad (2)$$

where x represents the Euclidean distance of the point from the center of the filter.

Once the filter is applied, a modified Z-score test is used to detect the outliers in the frame [20]. Mean absolute distance (MAD) is calculated using Equation 3:

$$MAD = \frac{\sum_{i=1}^{N} |fp_i - \mu|}{N} \qquad (3)$$

where μ is the mean value of the pixels in the filtered image, fp_i is the value of i-th pixel in the filtered frame and N is the number of pixels.

The Z-score values are then calculated using Equation 4:

$$Z_i^{score} = \frac{|fp_i - \mu|}{MAD} \qquad (4)$$

where Z_i^{score} is the Z-score for the i-th pixel.

An additional step is performed once the Z-scores have been calculated, which allows the approach to handle the situations where the outlier detection procedure would be misled by large changes occurring in large parts of the frame. The values are re-normalized to [0,1] range and those smaller than a specified threshold discarded. In the experiments conducted, the threshold was set dynamically by multiplying a threshold coefficient (θ) with the mean value of the final, normalized set of values (Equation 5).

$$out_i = \begin{cases} Z_i^{snorm}, & \text{if } Z_i^{snorm} \geq \theta\mu_{snorm}; \\ 0, & \text{if } Z_i^{snorm} < \theta\mu_{snorm}; \end{cases} \qquad (5)$$

where out_i is the final segmented value of the i-th pixel, Z_i^{snorm} is the normalized Z-score value for the pixel and μ_{snorm} is the mean of the normalized Z-score values. The value of θ was set to 2.5 in the experiments performed.

The result of temporal filtering at each scale is a temporal saliency map containing non-zero real values of the pixels undergoing salient changes.

The saliency maps obtained for different scales are iteratively upsampled and summed to increase the score of pixels scoring high consistently across scales. Thus, a single saliency

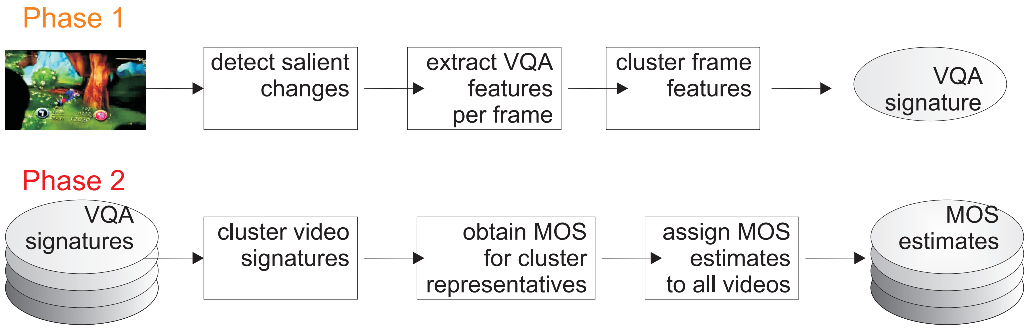


Figure. 1: Proposed VQA approach: in *Phase 1* a VQ signature is extracted for each video, in *Phase 2* the signatures are clustered to determine the set of representative videos, their MOS measured and propagated to other videos

map is obtained per color channel. The value describing the saliency of the pixel is the maximum value across the color channels. The values of the single saliency map obtained in this way are then normalized and compared to a threshold to eliminate the inconspicuous changes. The output saliency map is a binary mask splitting the frame into salient and inconspicuous (non-salient) motion regions.

B. Features for Video Quality Assessment

Several features are used to describe the salient motion in a frame: number of salient regions, their average size, and first moments (mean and standard deviation) of the difference between the current frame and background frames, calculated separately for salient and non-salient regions. Also, to account for blurring and blockiness, Z-score measures proposed by Wang *et al.* [2] and the blockiness measure proposed by Babu *et al.* [14] are calculated separately for salient, non-salient and (int the case of the last feature) border regions.

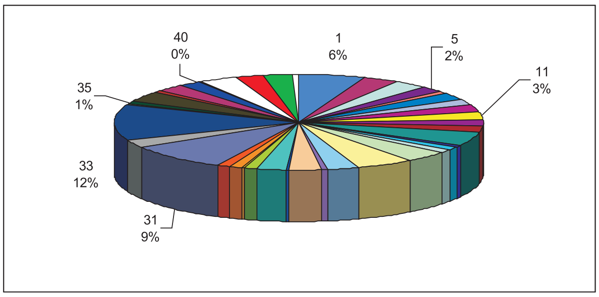


Figure. 2: Portion of the data set corresponding to each of the 45 clusters

The blockiness measures proposed by Wang *et al.* and Babu *et al.* are profoundly different. Babu *et al.* focus on the effects that can be observed along the edges of a single block. Their measure is designed to detect blocks with low spatial activity along the edges, but significant differences across them.

To characterize the activity on the inside of the block edge they calculate the standard deviation of pixel values for 6-pixel long stretches along he border of the block, since they observed that blockiness that spans less than 6 pixels is not perceived as significant. For each edge of the block they try to detect if there is significant activity that could mask the blockiness effect. Let $\{I_{k,j} | k \in [1,4], j \in [1,8]\}$ be the edges of a block and $\{O_{k,j}, k \in [1,4], j \in [1,8]\}$ the corresponding pixels across the edge of the block. We first consider the standard deviation of pixel values on the inside of block edges:

$$\sigma_{k,j} = stddev(I_{k,j}), k \in [1,3], j \in [k, k+5] \qquad (6)$$

Then we compute the gradient across block edges for each subsegment of the edge:

$$\Delta_{k,j} = mean(|I_{k,j} - O_{k,j}|), k \in [1,3], j \in [k, k+5] \qquad (7)$$

If any of $\sigma_{k,i}$ is below an empirically selected threshold ε, than that edge can contribute to the blockiness, but it will do so only if the gradient is larger than a different threshold τ. For a block i of a frame, we define

$$W_i = \begin{cases} 1, & (\exists \sigma_{k,j}, \Delta_{k,j})(\sigma_{k,j} < \varepsilon \wedge \Delta_{k,j} > \tau) \\ 0, & \text{otherwise.} \end{cases} \qquad (8)$$

Finally, we calculate the proportion of blocks that contributes to the blockiness effect as the measure of blockiness:

$$BB = \frac{\sum_{i=1}^{NB} W_i}{NB} \qquad (9)$$

where NB is the number of blocks in the region considered. The authors of the approach [14] suggest $\varepsilon = 0.1$ and $\tau = 2.0$, which are also the values used in the study presented here.

The approach of Wang *et al.* is based on the observation that the artifacts can be detected if the image is transformed to the frequency domain and its power spectrum examined. They design their measures of blurring and blockiness in an attempt to achieve a less computationally intensive approach than that of computing the full power spectrum. Let $x(m,n)$ $m \in [1,M]$ and $n \in [1,N]$, be the pixel values (signal) for a frame. First a differencing signal is calculated along the horizontal lines:

$$d_h(m,n) = x(m,n+1) - x(m,n), n \in [1, N-1] \qquad (10)$$

The blockiness measure proposed by Wang *et al.* tries to take into account the differences between a whole line of blocks, rather than looking at a single block:

$$B_h = \frac{1}{M(\lfloor N/8 \rfloor - 1)} \sum_{i=1}^{M} \sum_{j=1}^{\lfloor N/8 \rfloor - 1} d_h(i, 8j) \qquad (11)$$

Table 1: List of proposed quality features.

#	Feature	#	Feature
1	Salient reg. count	10	Z score non-salient
2	Avg. reg. size	11	Activity salient
3	Mean change non-salient	12	Blocking effect salient
4	Change Std.Dev. non-salient	13	Zero-crossing rate salient
5	Mean Change salient	14	Z score salient
6	Change Std.Dev. salient	15	Blockiness non-salient
7	Activity non-salient	16	Blockiness salient
8	Blocking effect non-salient	17	Blockiness border
9	Zero-crossing rate non-salient		

Thus, the Wang *et al.* provides a more wider-range measure of blockiness, when compared to the basic Babu *et al.* metric. Wang *et al.* proposed two measures in an attempt to characterize the spatial activity of the signal. Their motivation lies in the fact that activity is reduced by blurring. The activity is related to how pronounced the texture is in a particular region of the frame. The first measure is the average absolute difference between in-block image samples:

$$A_h = \frac{1}{7}\left[\frac{8}{M(N-1)}\sum_{i=1}^{M}\sum_{j=1}^{N-1}|d_h(i,j) - B_h|\right] \quad (12)$$

The second measure is the zero-crossing (ZC) rate. They define for $n \in [1, N-2]$:

$$z_h(m,n) = \begin{cases} 1, & \text{horizontal ZC at } d_h(m,n) \\ 0, & \text{otherwise} \end{cases} \quad (13)$$

the horizontal ZC rate can then be estimated as:

$$Z_h = \frac{1}{M(N-2)}\sum_{i=1}^{M}\sum_{j=1}^{N-2} z_h(m,n) \quad (14)$$

The vertical features (B_v,A_v and Z_v) are then calculated in a similar fashion. The overall blockiness, activity and ZC rate are calculated as:

$$B = \frac{B_h + B_v}{2}, A = \frac{A_h + A_v}{2}, Z = \frac{Z_h + Z_v}{2} \quad (15)$$

Finally they formulate an empirical model for the quality score:

$$Z_{score} = \alpha + \beta B^{\gamma_1} A^{\gamma_2} Z^{\gamma_3} \quad (16)$$

They used the non-linear regression routine available in the Matlab statistics toolbox to find the best value of parameters (α,β,γ_1,γ_2,γ_3) for Eq. 16. The values they calculated are used in the study presented here: $\alpha = -245.9$,$\beta = 261.9$,$\gamma_1 = -0.0024$,$\gamma_2 = 0.016$,$\gamma_3 = 0.0064$.

Blockiness is masked by the texture(spatial activity) in the region for which it is calculated. Activity measures are directly related to texture properties within blocks. The final Z-score is a nonlinear combination of these measures, that emulates the properties of the human visual system.

All the quality related features used are listed in Table 1. Z score, activity, blocking effect and zero-crossing rate were originally proposed by Wang *et al.*. Blockiness designates the measure proposed by Babu *et al.*.

IV. Experiments and Results

A. Data

The experimental data used consists of 2,107 videos downloaded from YouTube(tm). The videos were downloaded automatically in a random fashion using a download tool customized at our lab.

The tool, which is an extension of the TubeKit toolkit designed by Shah [21], relies on a set of random keywords (phrases) to achieve randomness of the downloaded material. The core components of the TubeKit were used to create a system that takes a set of keywords and the number of query results desired for each keyword, and acquires videos based on that data. Keywords were obtained using a shuffling function to get a word from the Ubuntu Linux 9.10 built-in English dictionary. Fig. 5 shows some frames from the material.

B. Objective Video Quality

To calculate the values of features in Table 1 the algorithm was implemented in C++, using OpenCV[22] to read the videos.

On an Intel Core2Duo processor the algorithm was able to extract features in near real time, taking environ 25 ms per frame. Once the features were calculated a programmable graphics hardware implementation of k-means clustering [17] was used to cluster the feature values for a single sequence into 100 clusters, whose centroids comprise the VQS for that sequence. This took between 0.025 and 1.74 additional seconds per sequence, when running on an MSI 9500GT graphics card.

Once all the VQS-es were collected a final round of clustering was conducted to form 45 clusters. The number of clusters was selected to correspond to the number of impaired sequences used by Culibrk *et al.* in [3] to train and evaluate the approach proposed there. Fig. 2 shows the data set contribution of each cluster. The contributions of the largest clusters and some clusters of interest for the discussion in further text are given in detail.

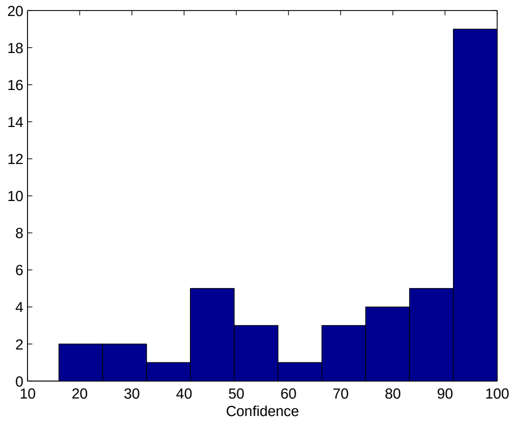


Figure. 3: Histogram of the number of representative sequences' VQS elements that correspond to pertaining clusters

Once the clusters are obtained, sequences with the largest portion of the VQS corresponding to a cluster are selected as representative of that cluster. Fig. 3 shows the histogram of *confidence values* indicating the portions of the VQS-es of representative sequences that correspond to the cluster they represent. As the figure shows, 29 of the clusters have representatives that have a confidence value over 70%. Representative sequences for these clusters were selected for subjective assessment. Five of these clusters have more than one sequence that is representative of them, the duplicates were included in the subjective assessment test, bringing the total number of sequences to 34.

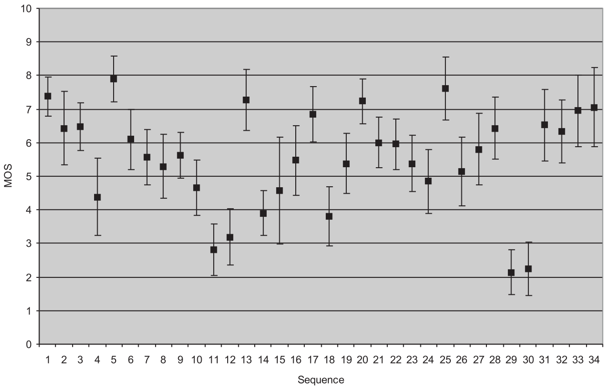


Figure. 4: MOS and standard deviation of opinion score

C. Subjective Assessment

The subjective video quality assessment method used was Absolute Category Rating (ACR), described in detail in ITU Recommendations [4, 23], and successfully implemented for similar applications[24]. In this method, test clips are presented to assessors one at a time, and rated independently on a discrete 9-level scale, ranging from "Bad" to "Excellent". The ratings for each test clip are then averaged over all subjects to obtain a Mean Opinion Score (MOS).

The subjective test consisted of two sessions of about 15 minutes, including training. Two sessions were conducted to allow for unreliable observers to be eliminated from the final MOS scores, using a paired t-test [25]. Before the actual test, written instructions were given to subjects and a test session was executed. The test session consisted of five videos demonstrating the extremes of expected video quality ranges. The actual test comprised thirty four segments cut from original videos, each around 10 seconds long. A sequence was made, comprising all of the videos (with 8s pause between them for voting). Subjects were grading the sequence twice, but at different points in time (once per session). In that way intra-subject reliability as well as inter-subject variability could be measured. 21 subjects - 13 male and 8 female - participated in the test, their age ranging from 19 to 30. None of them were familiar with video processing nor had previously participated in similar tests. All of the subjects reported normal or corrected vision prior to testing.

Fig. 4 shows a plot MOS values and standard deviation of opinion scores, for the 34 sequences assessed.

Table 2: Opinion score statistics for the representative sequences used in the evaluation

Sequence	Min OS	MOS	Max OS	Std. Dev.	Cluster
clip 20	5.88	6.48	7.08	1.40	1
clip 15	5.32	6.1	6.87	1.81	
clip 9	4.87	5.57	6.27	1.63	
clip 14	3.96	4.67	5.37	1.65	11
clip 7	5.30	5.95	6.6	1.53	
clip 25	6.51	7.28	8.06	1.82	35
clip 29	6.04	7.05	8.06	2.36	
clip 11	4.6	5.48	6.36	2.06	40
clip 3	5.64	6.43	7.21	1.83	
clip 23	4.62	5.38	6.14	1.78	5
clip 5	4.27	5.14	6.01	2.03	

Table 3: T-test results for pairs of sequences representative of a single cluster:t values, degrees of freedom (df) and p-values

t	df	p-value	Cluster
-0.2502	19	0.8051	1
-1.7119	19	0.1032	
-1.3966	19	0.1786	
-3.3804	19	0.003	11
0.2239	19	0.8252	35
0.1724	19	0.865	40
-0.4774	19	0.6385	5

D. Evaluation

To evaluate the ability of the proposed approach to cluster the videos according to their perceived quality and detect reliable representative sequences, statistical tests were performed on representative sequences of clusters which had more than one representative in the subjective assessment test. These sequences are listed along with their opinion score statistics in Table 2.

If the proposed approach is able to fulfill its intended purpose, one expects the representative sequences within the same cluster to have similar MOS, while there should be significant differences in OS between different clusters. Rather than simply comparing MOS values, we test a more stringent requirement: the differences in the opinion scores (OS) of representative sequences gathered from the human observers should not be statistically significant for most clusters, while there should be statistically significant differences between them.

To verify the first part of the hypothesis, t-tests were conducted between representative sequences within the same cluster. The null hypothesis was that there are no significant differences between the pairs of representative sequences in the clusters. Table 3 shows the results for the OS of ten reliable observers.

The p-values for all but one cluster are well over 0.05, meaning the first part of the hypothesis cannot be rejected based on the experimental data. For cluster 11, however, the t-test rejected the hypothesis that the opinion scores of the two sequences show no statistically significant difference. While this means that the observers found the two sequences different, their MOS is sufficiently close to allow them to be placed in the same cluster. To test the second part of the hypothesis, Analysis of Variance (ANOVA) technique was used [25], with a null hypothesis that there are no differences in the OS values between the clusters. ANOVA returned an f-measure of 11.1 and significance value below 0.001, rejecting the null hypothesis.

V. Conclusion

The paper presents a novel approach to Video Quality Assessment, which utilizes large amounts of publicly available data from web-sites such as YouTube^{tm}.

The problem of obtaining the Mean Opinion Score (MOS), a standard measure of the perceived video quality, for large numbers of videos is addressed by determining a small set of representative examples using machine learning techniques. We show that the several recently proposed measures related to motion, saliency and coding artifacts can be used to cluster the videos reliably using k-means. To obtain an estimate for the video quality of all videos, subjective assessment can be conducted for a few representative videos within each cluster. This is then used as an estimate of the quality of all other videos in the cluster. Results based on 2,107 videos containing some 90,000,000 frames are presented in the paper, along with the statistical analysis of the properties of representative videos obtained via the proposed approach, within and between clusters.

The results suggest that the approach is viable. The potential implication is that the vast amounts of publicly available multimedia content could be exploited to create a relatively small set of sequences representative of all that data. This data could then be used to estimate video quality based on the approach proposed here, testing other conventional approaches and serve as a data set allowing the application of other machine learning techniques to the problem of VQA.

Further and more in depth testing using an even larger database of videos should be conducted. This will probably require moving the feature extraction process to programable graphics hardware. Other features and distance metrics guiding the clustering process may be explored.

References

[1] S. Winkler, *Digital video quality: vision models and metrics*. John Wiley & Sons, 2005.

[2] Z. Wang, H. R. Sheikh, and A. C. Bovik, “No-reference perceptual quality assessment of jpeg compressed images,” in *Proceedings of IEEE 2002 International Conferencing on Image Processing*, 2002, pp. 477–480.

[3] D. Culibrk, D. Kukolj, P. Vasiljevic, M. Pokric, and V. Zlokolica, “Feature selection for neural-network based no-reference video quality assessment,” in *Proceedings of the International Conference on Neural Networks (ICANN)*, 2009, pp. 633–642.

[4] ITU-R BT.500, “Methodology for the subjective assessment of the quality of television pictures,” International Telecommunication Union, Tech. Rep., 2002.

[5] I. H. Witten and E. Frank, *Data Mining: Practical machine learning tools and techniques, 2nd Edition*. Morgan Kaufmann, San Francisco, 2005.

[6] VQEG, “The video quality experts group,” http://www.vqeg.org.

[7] G. Warwick and N. Thong, *Signal Processing for Telecommunications and Multimedia, Chapter 6: Classification of Video Sequences in MPEG Domain*. Springer, 2004.

[8] I. Kirenko, “Reduction of coding artifacts using chrominance and luminance spatial analysis,” *Consumer Electronics, 2006. ICCE ’06. 2006 Digest of Technical Papers. International Conference on*, pp. 209–210, Jan. 2006.

[9] R. Ferzli and L. Karam, “A no-reference objective image sharpness metric based on just-noticeable blur and probability summation,” *Image Processing, 2007. ICIP 2007. IEEE International Conference on*, vol. 3, pp. III –445–III –448, 16 2007-Oct. 19 2007.

[10] S. Wolf and M. Pinson, “Spatial-temporal distortion metrics for in-service quality monitoring of any digital video system,” in *Proc. SPIE*, vol. 3845, 1999, pp. 266–277.

[11] Z. Wang, L. Lu, and A. Bovik, “Video quality assessment based on structural distortion measurement,” *Signal processing: Image communication*, vol. 19, no. 2, pp. 121–132, 2004.

[12] K. Kim and L. Davis, “A fine-structure image/video quality measure using local statistics,” 2004, pp. V: 3535–3538.

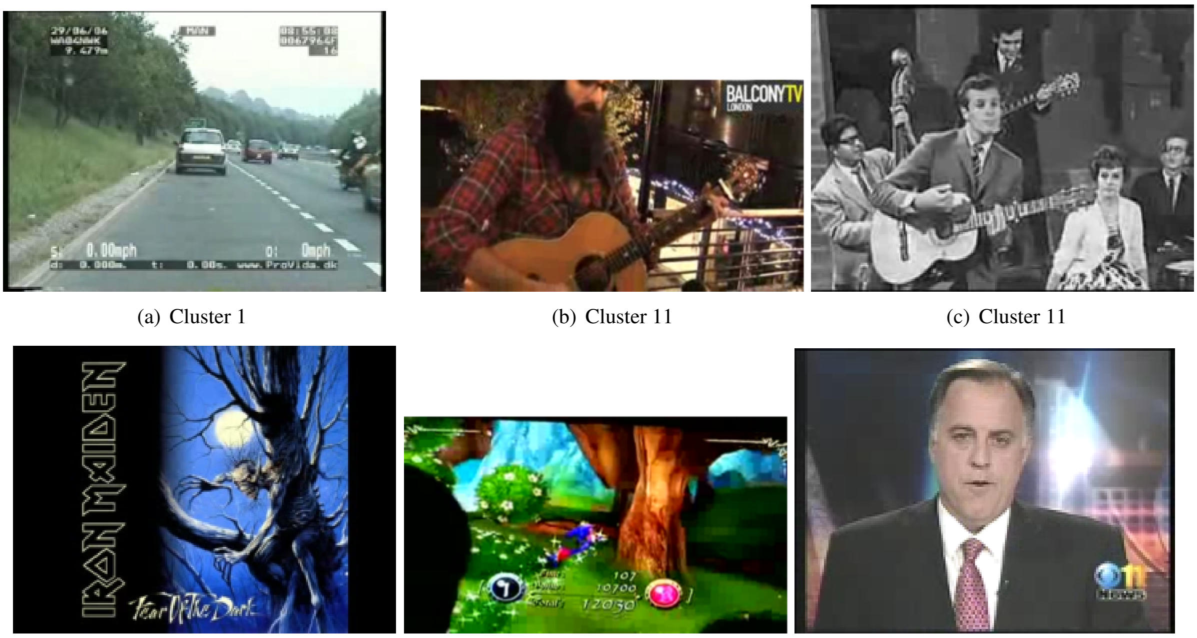
[13] R. Babu and A. Perkis, “An hvs-based no-reference perceptual quality assessment of jpeg coded images using neural networks,” in *Image Processing, 2005. ICIP 2005. IEEE International Conference on*, vol. 1. IEEE, 2005, pp. I–433.

[14] V. Babu, P. Andrew, and H. O. Inge, “Evaluation and monitoring of video quality for uma enabled video streaming systems,” *Multimedia Tools Appl.*, vol. 37, no. 2, pp. 211–231, 2008.

[15] D. Culibrk, M. Mirkovic, V. Zlokolica, M. Pokric, V. Crnojevic, and D. Kukolj, “Salient Motion Features for Video Quality Assessment.” *IEEE Trans. on Image Processing*, vol. 20, pp. 948 – 958, 2010.

[16] S. Winkler and R. Campos, “Video quality evaluation for Internet streaming applications,” in *Proc. IS&T/SPIE Electronic Imaging 2003: Human Vision and Electronic Imaging VIII*, vol. 5007, 2003, pp. 104–115.

[17] M. Zechner and M. Granitzer, “Accelerating k-means on the graphics processor via cuda,” in *Intensive Applications and Services, 2009. INTENSIVE’09. First International Conference on*. IEEE, 2009, pp. 7–15.



(a) Cluster 1 (b) Cluster 11 (c) Cluster 11

(d) Cluster 35 (e) Cluster 5 (f) Cluster 40

Figure. 5: Sample frames from representative sequences

[18] V. J. Hodge and J. Austin, "A survey of outlier detection methodologies," *Artificial Intelligence Review*, vol. 22, pp. 85–126, 2004.

[19] D. Culibrk, V. Crnojevic, and B. Antic, "Multiscale background modelling and segmentation," in *Digital Signal Processing, 2009 16th International Conference on*. IEEE, 2009, pp. 1–6.

[20] E. McBean and F. Rovers, *Statistical procedures of environmental monitoring data and risk assessment*. Prentice Hall PTR, 1998.

[21] C. Shah, "Tubekit: a query-based youtube crawling toolkit," ACM, pp. 433–433, 2008.

[22] G. Bradski and A. Kaehler, *Learning OpenCV: Computer Vision with the OpenCV Library*. O'Reilly Media, Inc., 2008.

[23] I.-T. R. P.910, *Subjective video quality assessment methods for multimedia applications.*, Sept. 1999.

[24] S. Winkler and C. Faller, "Audiovisual quality evaluation of low-bitrate video," *SPIE/IS&T Human Vision and Electronic Imaging*, vol. 5666, pp. 139–148, 2005.

[25] H. R. Lindman, *Analysis of variance in complex experimental designs*. W. H. Freeman & Co., 1974.

Author Biographies

Dubravko Ćulibrk is an Assistant Professor in the Department of Industrial Engineering and Management at Faculty of Technical Sciences in Novi Sad, Serbia. He received his B.Eng. degree in automation and system control as well as a M.Sc. degree in computer engineering from the University of Novi Sad, Novi Sad, Serbia, in 2000 and 2003. In 2006 he received a Ph.D. degree in computer engineering from Florida Atlantic University, Boca Raton, Florida, USA. His research interests include video and image processing, computer vision, neural networks and their applications, cryptography and evolutionary computing.

Milan Mirković is a Teaching Assistant at Faculty of Technical Sciences in Novi Sad, Department of Industrial Engineering and Management. He received his B.Eng. degree in industrial engineering and management, and his M.Sc. degree in business processes automation from the University of Novi Sad. He is pursuing a Ph.D. degree in information and communication systems at University of Novi Sad, his current interests focusing on image processing, data mining, web and persuasive technologies, and their application.

Lugonja Predrag was born in Novi Sad, Serbia in July 24, 1983. He received M.S. degree in electrical engineering form Novi Sad University, Serbia, in 2002. He is currently working towards the Ph.D. degree at Faculty of Technical Science on University of Novi Sad. His research interests include remote sensing and video quality assessment.

Vladimir Crnojević received the Diploma degree and the M.Sc. and Ph.D. degrees in electrical engineering from the University of Novi Sad, Serbia, in 1995, 1999, and 2004, respectively. In 1995, he joined the Communications and Signal Processing Group, Department of Electrical Engineering, University of Novi Sad, where he was a teaching and research assistant; in 2004, he became an Assistant Professor in the same department. He is a coordinator of several projects from the EU program (FP7, EUREKA!) and national research programs. His research interests include image processing, computer vision, and evolutionary algorithms.