# Salient Motion Features for Video Quality Assessment

Dubravko Ćulibrk, Milan Mirković, *Student Member, IEEE*, Vladimir Zlokolica, Maja Pokrić, Vladimir Crnojević, *Member, IEEE*, and Dragan Kukolj, *Senior Member, IEEE*

***Abstract*—Design of algorithms that are able to estimate video quality as perceived by human observers is of interest for a number of applications. Depending on the video content, the artifacts introduced by the coding process can be more or less pronounced and diversely affect the quality of videos, as estimated by humans. While it is well understood that motion affects both human attention and coding quality, this relationship has only recently started gaining attention among the research community, when video quality assessment (VQA) is concerned. In this paper, the effect of calculating several objective measure features, related to video coding artifacts, separately for salient motion and other regions of the frames of the sequence is examined. In addition, we propose a new scheme for quality assessment of coded video streams, which takes into account salient motion.**

**Standardized procedure has been used to calculate the Mean Opinion Score (MOS), based on experiments conducted with a group of non-expert observers viewing standard definition (SD) sequences. MOS measurements were taken for nine different SD sequences, coded using MPEG-2 at five different bit-rates. Eighteen different published approaches related to measuring the amount of coding artifacts objectively on a single-frame basis were implemented. Additional features describing the intensity of salient motion in the frames, as well as the intensity of coding artifacts in the salient motion regions were proposed. Automatic feature selection was performed to determine the subset of features most correlated to video quality. The results show that salient-motion-related features enhance prediction and indicate that the presence of blocking effect artifacts and blurring in the salient regions and variance and intensity of temporal changes in non-salient regions influence the perceived video quality.**

***Index Terms*—M5′, motion, no-reference, perceptual quality, regression trees, saliency, video quality assessment.**

## I. Introduction

THERE is an increased need to measure and assess the quality of video sequences, as it is perceived by the multimedia content consumers. The quality greatly 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 quality of the service.

It is well understood that the overall degradation in the quality of the sequence, is a compound effect of different coding artifacts.

A large number of published papers exists, proposing different measures of prominent artifacts appearing in coded images and video sequences [1], [2]. The goal of each no-reference approach is to create an estimator based on the proposed features that would predict the Mean Opinion Score (MOS) [3] of human observers, without using the original (not-degraded) image or sequence data.

While aspects of the Human Visual System (HVS) have been modelled to arrive at an estimate of the perceived level of coding artifacts in video sequences, attention and saliency in videos, due to motion and otherwise, have only recently begun to be considered as a way to enhance VQA [4]. Bottom-up attention can be modelled computationally [5] and has been successfully used in a number of applications such as content-based image retrieval [6], scene classification [7] and vision-based localization [8]. Recently researchers have started looking into using computational models of (motion) attention to enhance the performance of video coding algorithms [9], address the problem of video skimming [10], [11] and improve VQA [4]. Using such a model to enhance no-reference VQA has not been explored before.

When VQA is concerned the motivation for taking attention into account lies in the fact that the HVS sensitivity to motion and texture differs significantly between areas of the stimuli focused upon (attended to) and those in peripheral vision [12]. This leads to different sensitivity to coding artifacts in the two regions of the visual field, which has rarely been taken into account in the MOS estimator design. This paper proposes an approach to VQA which attempts to exploit this effect.

In the following text, we show how a recently proposed multiscale background-subtraction approach can be used to detect salient motion regions in the frames of video sequences efficiently. The saliency information can be calculated in real time for Standard Definition (SD) sequences. Based on saliency information, 17 new features are proposed, which take into account the visual saliency of moving objects in video. We then compare the applicability of 18 published commonly-used features against those proposed, to the problem of MPEG-2 coded no-reference video quality assessment. To achieve this, an approach [13] commonly used in the field of artificial intelligence and data mining [14] is employed to automatically select the features most relevant to the problem at hand.

The selected features are used to train M5′ regression tree estimators. Their performance is compared to existing ap-

Manuscript received November 30, 2009; revised April 16, 2010, July 12, 2010, and September 14, 2010; accepted September 14, 2010. This work was supported in part by Ministry of Science and Technology Development of Republic of Serbia, under Grant 161003 and by the EUREKA E!4160 VICATS project. The associate editor coordinating the review of this manuscript and approving it for publication was .

D. Ćulibrk and M. Mirković are with the Department of Industrial Engineering and Management, Faculty of Technical Sciences, Novi Sad, Serbia (e-mail: http://www.dubravkoculibrk.org; mmirkov@uns.ac.rs).

V. Zlokolica, M. Pokrić, V. Crnojević and D. Kukolj are with the Department of Electrical and Computer Engineering, Faculty of Technical Sciences, Novi Sad, Serbia (e-mail: vladimir.zlokolica@rt-rk.com; maja.pokric@rt-rk.com; crnojevic@uns.ac.rs; dragan.kukolj@rt-rk.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2010.2080279

proaches, in order to gain better insight into the predictive ability of the proposed features. The use of M5′ in this domain is a novelty. The experimental results verify that information regarding salient-motion can be used to improve no-reference VQA significantly and indicate that the presence of blocking effect artifacts and blurring in the salient regions seems to have bearing on human perception, while temporal changes in otherwise non-salient regions influence the perceived video quality.

The rest of the paper is organized as follows: Section II provides an overview of the relevant published work. The methodology used to determine salient motion, extract descriptive features and select the best features for no-reference VQA is described in Section III. Section IV presents the experiments conducted to evaluate the applicability of the proposed features and MOS estimation results obtained. Conclusions and some directions for future work can be found in Section V.

## II. Background and Related Work

### A. Video Quality Assessment

The work presented in this paper relates to no-reference video quality assessment methodologies [15]. No information regarding the original (not-coded) video is used to estimate video quality, as perceived by human observers. A subjective quality measure typically used is the mean opinion score (MOS), which is obtained by averaging scores from a number of human observers [1], [16]. The correct procedure for conducting such experiments was derived from ITU-R BT.500-10 recommendation [3].

Overall degradation in the quality of the sequence is due to encoder/decoder implementations as part of transport stream at various bit rates and is a compound effect of different coding artifacts. Three types of artifacts are typically considered pertinent to DCT block (JPEG and MPEG) coded data: blocking, ringing and blurring. Blocking appears in all block-based compression techniques due to coarse quantization of frequency components [1], [17]. 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 [18]. It is due to improper truncation of high frequency components. This artifact is also known as the Gibbs phenomenon or Gibbs effect. 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 [19].

Even when the reference (not degraded) video is available, objective measures of signal degradation such as Peak Signal-to-Noise Ratio (PSNR) are poorly correlated to MOS [20], leading to significant research effort aimed at the design of measures which will allow computers to determine MOS effectively. The measures typically focus on specific coding artifacts and attempt to take into account the effect of the content of the images (video frames). Thus, when perceived blockiness is concerned, most 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. Spatial activity in images is profoundly related to visual saliency [21]. However, bottom-up attention models have been explicitly considered to enhance VQA in a single published approach [4], dealing with packet loss visibility.

Several published approaches to measuring the different coding effects are of interest for the discussion in the following sections. Wang *et al.* [1] 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. Their approach is usually compared against, when no-reference MOS estimation is concerned. More recently, Babu *et al.* [16] proposed a blockiness measure for use in VQA, which takes effects along each edge of the block into account separately. They report their measure surpassing the Wang *et al.* approach in terms of MOS prediction accuracy. Kusuma and Zepernick [15] described three measures focusing on image-activity and contrast. They propose using two different image activity measures, edge and gradient activity, as a way to detect and measure ringing and lost blocks. Idrissi *et al.* [22] also proposed measures related to texture in an attempt to gain a better description of the spatial activity within the frames of the sequence. Kirenko [18] proposed simple measures for ringing effects detection. Kim and Davis [23] proposed a noise and blur measure, aimed at estimating the quality of video within the framework of automatic surveillance. They show their local-variance-based measure, dubbed fine-structure, to be able to describe video degradation well, in terms of noise and blur.

A recent paper [2] evaluated measures listed in the previous paragraph and additional measures (18 in total) of image and video quality. The additional measures were introduced to account for the temporal dynamics of the sequence. Two motion intensity measures were used: (i) global motion intensity, calculated from the global motion field, and (ii) object motion intensity, calculated by subtracting the global motion from the MPEG motion vectors [17]. The two measures made no attempt to model known aspects of motion-related attention [24]. The full set of measures used in [2] comprises the initial set of features considered for video quality estimation, in the work presented here. In [2], feature selection was performed based on training a simple Multi-layer Perceptron (MLP) estimator with each measure as input, separately. The measures were ranked according to their performance and a subset of 5 measures was selected as input for the final estimator, which was an MLP with 7 nodes in the hidden layer. Since the prediction was done on a single frame basis, median filtering was used to arrive at a single estimate for the whole sequence. The approach achieved better results than any measure considered separately.

### B. Saliency, Motion and Attention

When faced with visual stimuli the human vision system (HVS) does not process the whole scene in parallel. Part of

visual information sensed by the eyes is discarded in a systematic manner to attend to objects of interest. The most important function of selective visual attention is to direct our gaze rapidly towards objects of interest in our visual environment [21], [24]. Objects that are not of interest are still processed, but with reduced spatial resolution and motion sensitivity [12]. Critical fusion frequency, on the other hand, is higher in the peripheral vision, making the HVS more sensitive to sudden changes in illumination in the not-attended region [25].

Such an evolutionary adaptation enables humans to gain insight into the scene quickly, despite the limited processing power of our mind. The ability to orientate rapidly towards salient objects in a cluttered visual scene allows an organism to detect quickly possible prey, mates or predators in the visual world, making it a clear evolutionary advantage [21].

This type of attention is referred to as attention for perception: the selection of a subset of sensory information for further processing by another part of the information processing system [6], [26].

Current research considers attentional deployment as a two-component mechanism [21], [27]. Subjects selectively direct attention to objects in a scene using both bottom-up, image-based saliency cues and top-down, task-dependent cues; an idea that dates back to 19th century work of William James [28].

Bottom-up processing is driven by the stimulus presented [26]. Some stimuli are intrinsically conspicuous or salient (outliers) in a given context. For example, a red dinner jacket among black tuxedos at a sombre state affair, a flickering light in an otherwise static scene or a street sign against gray pavement, automatically and involuntarily attract attention. Saliency is independent of the nature of the particular task, operates very rapidly and is primarily determined in a bottom-up manner. If a stimulus is sufficiently salient, it will pop out of a visual scene. This suggests that saliency is computed in a pre-attentive manner across the entire visual field, most probably in terms of hierarchical *centre-surround mechanisms*. As for the moving stimuli, they are perceived to be moving only if they are undergoing motion different from their wider surround [24]. The speed of saliency-based form of attention is on the order of 25 to 50 ms per item [21]. The second form of attention is a more deliberate affair and depends on the task at hand, memories and even past experience [26]. Such intentional deployment of attention has a price, because the amount of time that it takes (200 ms or more), rivals that needed to move the eyes. Thus, certain features in the visual world automatically attract attention and are experienced as "visually salient". Directing attention to other locations or objects requires voluntary "effort". Both mechanisms can operate in parallel.

Significant progress has been made in terms of computational models of bottom-up visual attention [29]–[32]. While bottom-up factors that influence attention are well understood [33], the integration of top-down knowledge into these models remains an open problem. Because of this, the fact that bottom-up components of a scene influence our attention before top-down knowledge does [27] and that they can hardly be overridden by top-down goals, applications of visual attention commonly rely on bottom-up models [6]–[8], [10]. Although the parallel is seldom made, the centre-surround difference, multi-scale processing, orientation sensitivity and outlier detection properties of HVS, considered within the work on computational modeling of bottom-up visual attention [33], [34], are at the heart of various applications in computer vision such as pattern recognition [35], [36] and texture classification [37].

Full-fledged biologically inspired computational models of attention are too computationally intensive for real world applications such as video skimming [10] and video quality assessment [4]. In the case of VQA this is especially true if a large number of features is calculated based on the output of the visual attention model.

Although the complex saliency model of Itti *et al.* has recently been employed to improve the prediction of packet loss effects, the relatively simple model of Ma *et al.* [10] remains the only attention model that integrates motion cues and has been specifically designed for performance required in real-time video applications. Ma *et al.* distinguish motion and static attention parts of their model, since they rely on previously calculated motion vector field to discern the regions of the frame salient due to motion. They propose measures based on motion intensity, spatial and temporal coherence to detect points salient due to motion and contrast to determine static saliency. It should be noted that the spatial coherency of motion seems to have no bearing on saliency at the lowest levels of attention [24]. Ölveczky *et al.* [24] report that the driving force of the attention at this level is the difference in the speed of motion between a center and the surrounding region.

The approach of Ma *et al.* used general principles of the visual attention in the HVS to drive the design of a lightweight attention model. This is the approach followed in the work presented here. A multi-scale background modeling and foreground segmentation approach proposed in [38] has been employed as an efficient attention model driven by both motion and static cues, which adheres to the principles reported in [24]. The model employs the principles of multi-scale processing, cross-scale motion consistency, outlier detection and temporal coherence. The output of the segmentation has been used to derive features describing the salient motion in the frame, as well as to calculate a number of video quality features separately for regions of the frame observed as salient and the rest of the frame. This enables us to evaluate the influence of the saliency on the predictive ability of the proposed VQA estimators.

### C. Feature Selection and Estimation

Feature selection is a well researched subject within machine intelligence [13], [40]. It relates to selecting the features most relevant to the concept one is trying to learn. The benefit of feature selection is two fold. Not only will a reduced feature set make for more efficient estimation and computational performance once the final estimator is trained, but the presence of features which are not related to the concept will usually increase the error of the learner (estimator), due to its intrinsic bias [14].

Two approaches to feature selection exist [14]. The first is to select the features that are most likely to improve the performance of the specific machine learning approach that will be used to learn the concept. The learner is usually trained using

different subsets of features and its accuracy used to select the features best for the problem at hand. The other approach is to try to evaluate the link between the features and the target concept without a particular learning scheme in mind. Such feature selection is called *filtering* as it filters out the attributes that have no bearing on the target concept.

Systematic feature selection has rarely been applied in the domain of VQA. The work described in Section II-A focuses mainly on proposing new features that describe the perceived intensity of various coding artifacts and fusing several such features to arrive at an estimate of MOS. In [2] an algorithm-specific method of feature selection has been employed to select features best suited for training an MLP neural network to estimate MOS.

The methodology used in the work presented here, and described in Section III, relies on the attribute selection approach proposed by Hall [13], which selects a subset of features that are most correlated to the target concept. It is applicable to problems with the numeric target variable, as it is the case with MOS. It can be used as a filter method, as well as to determine the subset of features best for prediction using a specific classifier (wrapper method).

Based on the selected features any number of different machine learning algorithms can be trained to serve as estimators of MOS. The VQA approaches described in Section II-A mostly regress simple linear or propose different nonlinear equations to arrive at a MOS estimate. Babu and Perkis [41] proposed using an MLP to estimate the MOS. The same estimator was used with wrapper-based feature selection in [2].

In the VQA approach proposed here an M5′ regression tree, a piece-wise linear algorithm [42], is used as the MOS estimator. Both filtering and wrapper feature selection are performed, based on the approach by Hall.

The work described in this paper bears most similarity with the recently published work of Liu *et al.* [4]. They used a stepwise feature selection approach, adding features one at a time to a general linear model. Such feature selection was used to explore the relevance of saliency-based features in a full-reference packet-loss estimation scenario. The improvement in the prediction accuracy of the model led to a conclusion that the saliency-based metrics are promising in the VQA domain. The saliency model used in that work was that of Itti and Koch [21], which is computationally expensive and cannot be used for real-time applications. The saliency information was used as a weighting function for various existing quality measures, differing from the approach proposed here, which separates each frame into two regions and calculates features for each region independently. Finally, the approach proposed here is a no-reference approach, relying on a different type of estimator and a different set of features.

## III. Features for Video Quality Assessment

The initial set of features evaluated consisted of 18 different features. These features, with their respective references, are listed in Table I. Based on the proposed model of motion-driven visual attention and the results of evaluation of the relevance of each of the 18 initial features performed in [2], this initial set has been extended to include additional 17 salient-motion-related features as listed in Table II.

TABLE I
Initial List of Measures Evaluated With Pertinent References

| # | Feature | Reference |
|---|---|---|
| 1 | Two field difference | [39] |
| 2 | Variance ratio | [23] |
| 3 | Blockiness | [16] |
| 4 | Ringing | [18] |
| 5 | Ringing 2 | [18] |
| 6 | Global motion vector intensity | [17] |
| 7 | Activity | [1] |
| 8 | Blocking effect | [1] |
| 9 | Zero-crossing rate | [1] |
| 10 | Z score | [1] |
| 11 | Gradient activity | [15] |
| 12 | Edge activity | [15] |
| 13 | Contrast | [15] |
| 14 | Correlation | [22] |
| 15 | Energy | [22] |
| 16 | Homogeneity | [22] |
| 17 | Variance | [22] |
| 18 | Contrast | [22] |

TABLE II
List of Proposed Measures With Pertinent References

| # | Feature | Reference |
|---|---|---|
| 19 | Salient reg. count | proposed |
| 20 | Avg. reg. size | proposed |
| 21 | Mean change non-salient | proposed |
| 22 | Change Std.Dev. non-salient | proposed |
| 23 | Mean Change salient | proposed |
| 24 | Change Std.Dev. salient | proposed |
| 25 | Activity non-salient | modified [1] |
| 26 | Blocking effect non-salient | modified [1] |
| 27 | Zero-crossing rate non-salient | modified [1] |
| 28 | Z score non-salient | modified [1] |
| 29 | Activity salient | modified [1] |
| 30 | Blocking effect salient | modified [1] |
| 31 | Zero-crossing rate salient | modified [1] |
| 32 | Z score salient | modified [1] |
| 33 | Blockiness non-salient | modified [16] |
| 34 | Blockiness salient | modified [16] |
| 35 | Blockiness border | modified [16] |

The value of the 35 features has been calculated for sequences the Video Quality Experts Group (VQEG) [43] provided as a benchmark for codec evaluation. MOS was obtained as perceived by human observers for the same sequences. Feature selection based on correlation [13] was performed and most relevant features selected. The selected features have subsequently been used to train an M5′ decision tree, as an estimator for the MOS of new sequences.

### A. Detecting Salient Motion

The proposed approach for the detection of salient motion is derived from the work described in [38]. Specific properties of the background-subtraction algorithm there make it suitable for determining salient motion regions in the frames of a video sequence.

The algorithm employs a multi-scale model of the background in the form of frames which form a Gaussian pyramid, akin to the model employed in the attention model proposed by Itti and Koch [21]. This allows the approach to account for the

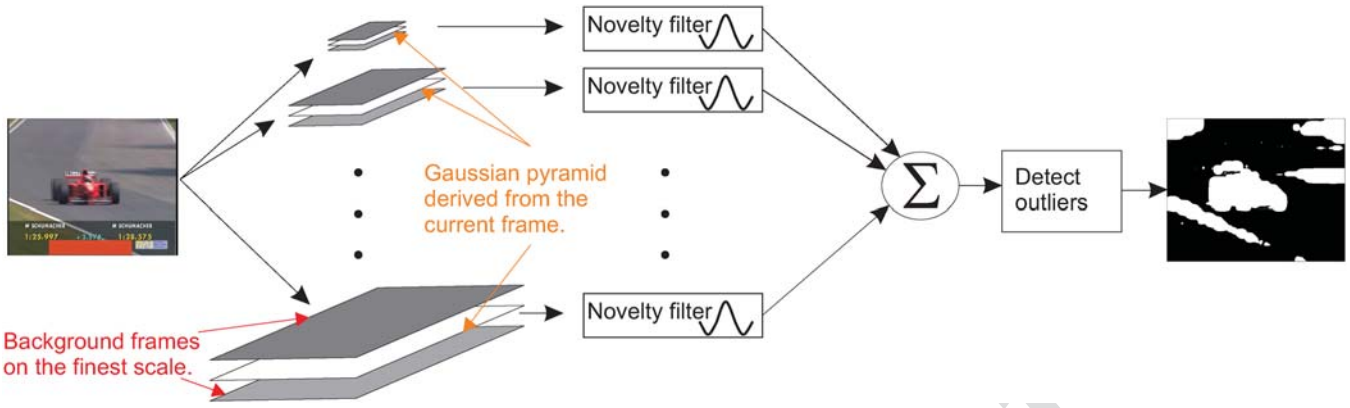


Fig. 1. Salient-motion region segmentation.

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 commonly used in background subtraction [44], [45]. This allows the approach to take into account temporal consistency in the frames. Finally, outlier detection [46] 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.

A block diagram of the approach is shown in Fig. 1. Each frame of the sequence is iteratively passed to a 2-D 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

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

where $\alpha_l$ 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 (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 [38]. 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 30 ms after the viewer is confronted with a visual scene, the value of $\alpha_1$ 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), $\alpha_1$ 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

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

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 [47]. Mean absolute distance (MAD) is calculated using

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

where $\mu$ 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

$$Z_i^{score} = \frac{|fp_i - \mu|}{MAD} \tag{4}$$

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 ($\theta$) with the mean value of the final, normalized set of values ((5)).

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

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 $\mu_{snorm}$ is the mean of the normalized Z-score values. The value of $\theta$ was set to 2.5 in the experiments performed.

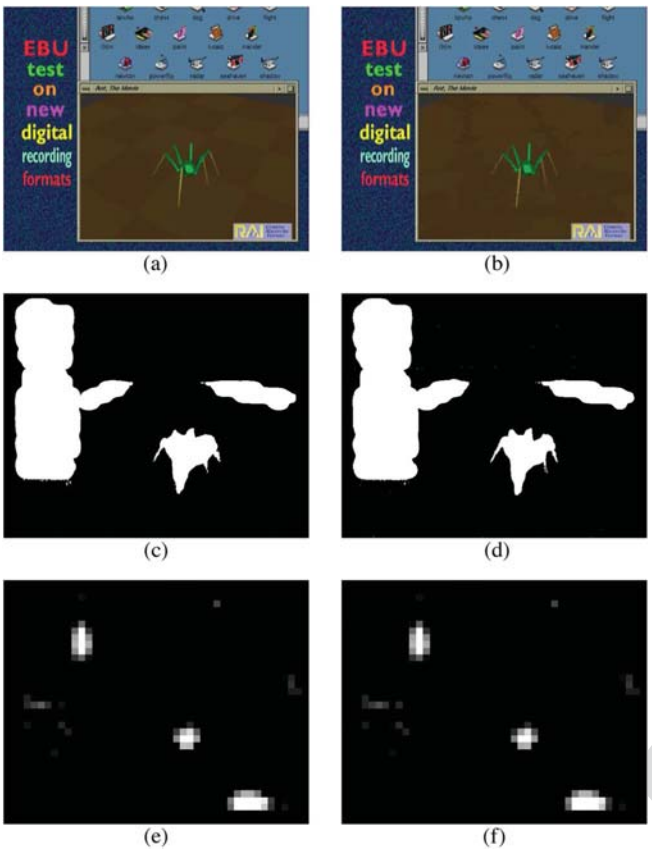


Fig. 2. Salient motion detection for the "Ant" sequence. (a) Sample frame coded at 4 MB/s; (b) sample frame coded at 0.5 MB/s; (c) salient motion detected at 4 MB/s; (d) salient motion detected at 0.5 MB/s; (e) static saliency [34] at 4 MB/s; (f) static saliency [34] at 0.5 MB/s.

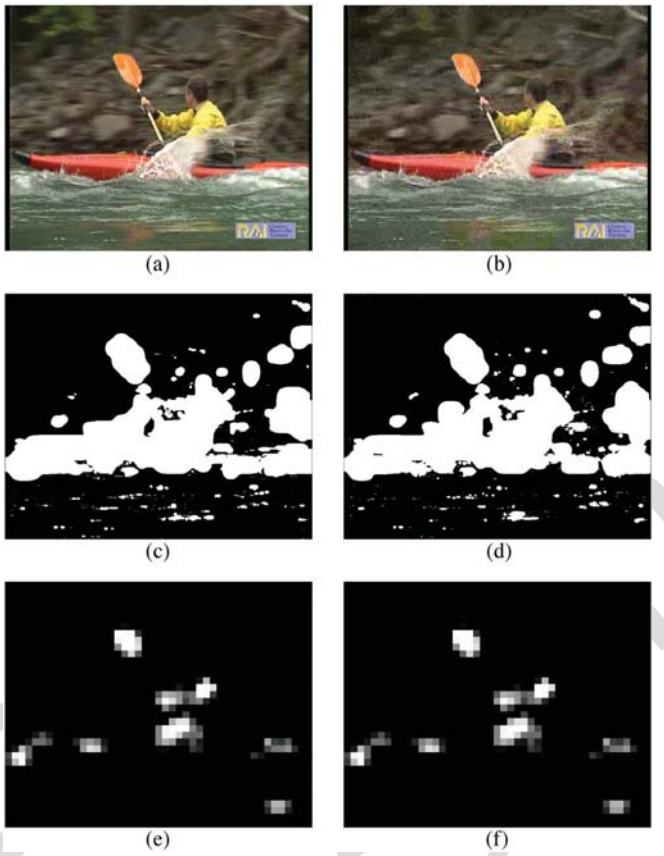


Fig. 3. Salient motion detection for the "Kayak" sequence. (a) Sample frame coded at 4 MB/s; (b) sample frame coded at 0.5 MB/s; (c) salient motion detected at 4 MB/s; (d) salient motion detected at 0.5 MB/s; (e) static saliency [34] at 4 MB/s; (f) static saliency [34] at 0.5 MB/s.

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 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.

Figs. 2 and 3 show sample results of the process, when three scales are used. The salient motion maps remain practically undisturbed by the artifacts introduced through coding the sequence at a lower bit rate. In addition, the saliency maps obtained based on the proposed approach fit better the moving objects of interest in the scene when compared those obtained by the static saliency model of Itti and Koch [21]. The "Ant" sequence is an example of a stationary camera sequence, while the camera in the "Kayak" sequence undergoes motion throughout the sequence as it tracks the man in the kayak. Although the camera is far from stationary in this case, the proposed approach is able to achieve meaningful segmentation of salient moving objects. The limitation of the approach is that it cannot be expected to perform well when the motion of interest forms the bulk of the motion of the frame, i.e., when the scene contains close-ups of objects of interest. This is due to the fact that the motion of such objects cannot be distinguished as outlier motion.

### B. Motion-Related Features

Several features are proposed 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.

To evaluate whether standard static VQA features can benefit from the salient-motion information, features proposed by Wang *et al.* and Babu *et al.* were recalculated based on saliency information, since they scored high in the experiments described in [2]. Wang *et al.* metrics were evaluated separately for salient and non-salient regions of the frames. The blockiness metric proposed by Babu *et al.* was evaluated separately for blocks completely within salient and non-salient regions, as well as for blocks that crossed the border between the two regions of the frame. This extension of the initial feature set brought the final number of features evaluated to 35. The additional features proposed are listed in Table II.

TABLE III
MOS FOR THE TRAINING SEQUENCES

| Test sequence | Bit rate [Mb/s] 0.5 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| "Parade" | 1.85 | 1.80 | 2.85 | 3.5 | 4.55 |
| "Harp" | 2.3 | 3.05 | 3.75 | 4.15 | 4.1 |
| "Ant" | 1.8 | 2.35 | 3.3 | 3.65 | 3.8 |
| "Kayak" | 1.9 | 2.2 | 3.5 | 3.95 | 4.25 |
| "Formula" | 1.95 | 2.7 | 3.6 | 3.9 | 4.2 |
| "Food court" | 1.65 | 2.4 | 3.6 | 3.9 | 4.3 |
| "Scrolling titles" | 2.55 | 3.2 | 3.9 | 3.85 | 3.55 |
| "Football" | 1.6 | 1.8 | 2.9 | 3.35 | 3.8 |
| "Train" | 2.35 | 2.5 | 3.15 | 3.7 | 3.85 |

Salient region count (Salient reg. count) represents the number of connected components corresponding to salient regions detected in the frame. Average region size (Avg. reg. size) is the average size of those regions in pixels. Measures 21–24 are the first moments of the changes in the salient and non-salient regions. Features 25–32 are metrics proposed by Wang *et al.*, calculated separately for non-salient and salient regions. Features 33–35 correspond to the blockiness metric, as proposed by Babu *et al.*, calculated for blocks contained in their entirety in the non-salient regions, salient regions and those crossing the border between the two regions.

### C. Creating the Data Set

The data set used is based on nine SD sequences made available by Video Quality Experts Group (VQEG) for purposes of testing the quality of video codecs. Each sequence has been encoded using five different bit-rate settings (0.5 Mb, 1 Mb, 2 Mb, 3 Mb, 4 Mb), using MPEG-2 codec. Values of the features have been calculated for 110 frames of the sequences, i.e., half of the frames of the sequence, distributed uniformly.

The mean opinion score (MOS), which is a subjective quality measure obtained by averaging scores from a number of human observers (assessors), was derived from tests created according to ITU-R BT.500-11 recommendations.

Double Stimulus Impairment Scale (DSIS) Variant I method was used, where the assessor is first presented with an unimpaired, reference sequence, and then with the same sequence impaired. He is then asked to vote on the second sequence, keeping in mind the first. Voting is done on a 1 to 5 scale, 1 being the lowest score where perceived impairments are very annoying, and 5 being the highest, where impairments cannot be perceived. Series of sequences with random levels of impairments are presented and, for control purposes, reference sequences are also included in the assessment set, but assessors are not informed about it. The final MOS value for a sequence is the average score over all assessors for that sequence. These values are listed in Table III.

Laboratory where the tests were conducted was setup as proposed in [3]. Sequences were presented to assessors on a 17″ Philips monitor (170S7FB), which was operated at its native resolution of 1280 × 1024 pixels, and a specialized software was used for their playback and voting [48].

After presenting the assessor with the reference sequence and an impaired one, the software displays a dialog-box for voting, where the assessor is asked to rate the impaired sequence on a 1–5 scale (labelled as described in [3]) by using a slider.

The test consisted of one session of about 30 minutes, including training. Before the actual test, written instructions were given to subjects, and a test session was run that consisted of videos demonstrating the extremes of expected video quality ranges. The actual test comprised 9 series of 6 sequences, each 10 seconds long. 20 subjects—11 male and 9 female—participated in the test, their age ranging from 22 to 33. 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.

TABLE IV
FEATURES IN THE BEST SUBSET BASED ON [13], WITH PERTINENT REFERENCES

| # | Feature | Reference |
|---|---|---|
| 22 | Change Std.Dev. non-salient | proposed |
| 34 | Blockiness salient | modified [16] |
| 31 | Zero-crossing rate salient | modified [1] |
| 31 | Zero-crossing rate | [1] |
| 10 | Z score | [1] |

### D. Feature Selection

To evaluate the predictive capability of each feature (measure), when MOS estimation is concerned, filter methodology for attribute selection has been used [14], relying on the correlation based approach of Hall [13]. A genetic algorithm [49] was employed to search the solution space, in order to escape the sensitivity of forward selection to local minima. Starting from the initial population of 10 random subsets, over 20 generations, it converged to a subset of 5 features listed in Table IV. Three of the five features selected are determined using the saliency information: the standard deviation of the difference (change) between the current frame and the background in the non-salient region and Zero-crossing rate and blockiness in the salient-motion part of the frame. An additional two features have been selected by the procedure are Zero-crossing rate calculated over the whole frame and Z-score.

To arrive at a subset of features specifically tailored for the type of estimator evaluated, wrapper method of correlation-based feature subset selection proposed by [13] was performed. The selection was based on the results of prediction of an M5′ algorithm trained using a specific feature subset, rather than the features themselves. Again, a genetic algorithm was used to search solution space. This yielded a significantly larger subset of 16 features listed in Table V. As the table shows, half of the features selected using this approach are calculated using saliency information.

The features selected suggest that the intensity of the blurring and blocking effects in the salient regions have most bearing on the perceived video quality. On the other hand, variance and intensity of temporal changes in the part of the frame not attended to, i.e., non-salient, seem to influence the judgment of video quality. Both Table IV and Table V indicate that the variance and intensity of temporal changes (respectively) in the part of the frame that is not attended to, impact perceived video quality. This corroborates past results in the literature showing the HVS being more sensitive to temporal changes in peripheral vision

TABLE V
WRAPPER-BASED SELECTED FEATURES

| M5' | |
|---|---|
| Feature | Reference |
| Mean Change non-salient | proposed |
| Mean Change salient | proposed |
| Change Std.Dev. salient | proposed |
| Activity non-salient | modified [1] |
| Zero-crossing rate non-salient | modified [1] |
| Activity salient | modified [1] |
| Zero-crossing rate salient | modified [1] |
| Two field difference | [39] |
| Variance ratio | [23] |
| Blockiness | [16] |
| Activity | [1] |
| Zero-crossing rate | [1] |
| Z score | [1] |
| Edge activity | [15] |
| Homogeneity | [22] |
| Contrast | [22] |

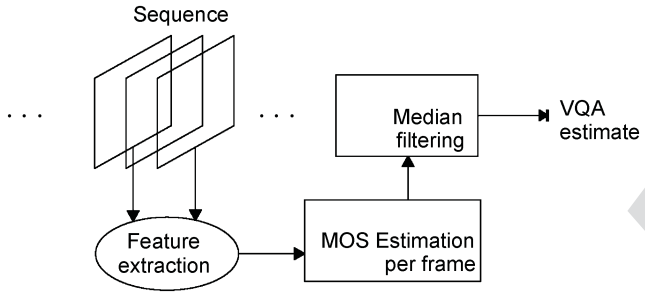


Fig. 4. Block diagram of the proposed MOS estimation approach.

than in the area focused on [25]. The effect of such changes is not enough to cause the HVS to attend to their location, so they remain in the peripheral vision, but they have a definite effect on our perception of the video quality. They are in fact, for the lack of a better expression, very annoying and observable.

### *E. Salient-Motion Driven VQA Estimation*

A block diagram of the proposed video-quality estimator is shown in Fig. 4. Based on the selected set of features an M5′ decision tree is used for frame MOS estimation.

The video quality assessment is conducted by calculating the selected features for half of the frames of the sequence, uniformly distributed (i.e., the frame rate was halved to make the approach more efficient). The processing of a single SD frame took around 170 ms on a single core of 3 GHz Intel Core2Duo processor, enabling the whole approach to run in real time.

The features obtained for each evaluated frame were fed into the estimator and the measure of the quality for that frame obtained.

Since the standard deviation of the estimator's prediction error over the frames of a single sequence is relatively high, robust statistics should be used to arrive at the final single measure of sequence video quality. Kim and Davis [23] suggest using the median of the quality values across the frames to achieve this. We followed their recommendation and adopted the median of values across the evaluated frames of the sequence as the final measure of sequence quality. Median is known to be a measure robust to the outliers, which commonly occurred in the experiments performed.

TABLE VI
CROSS-VALIDATION RESULTS FOR DIFFERENT FEATURE SUBSETS: FEATURES LISTED IN TABLE IV (A), FEATURES LISTED IN TABLE V (B), FEATURES USED IN [2](C)

| Estimator | Features | CC | MAE | RMSE | RAE | RRSE |
|---|---|---|---|---|---|---|
| M5' | A | 0.92 | 0.22 | 0.33 | 30.43 % | 40.06 % |
| | B | 0.93 | 0.20 | 0.31 | 27.8 % | 37.77 % |
| | C | 0.88 | 0.26 | 0.39 | 35.23 % | 46.93 % |
| MLP | C | 0.82 | 0.37 | 0.49 | 49.78 % | 58.98 % |

## IV. RESULTS AND DISCUSSION

To evaluate the descriptive capability of the features selected in Section III 10 fold cross-validation of the M5′ model, trained using a subset of features, has been performed, similar to the validation performed in [4].

During training, the M5′ algorithm was required to produce a tree with leaves covering no less than 55 instances, corresponding to half of the frames in each coded sequence. The algorithm created a tree of 69 leaf nodes (regression equations), when the 5 features in Table IV were used. For the 5 features used in [2] the tree contained 59 leaf nodes. Finally, when all the features listed in Table V are used, the tree contained 77 leaf nodes. Table VI shows the different performance measures obtained for different estimators: Correlation Coefficient (CC), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Relative Absolute Error (RAE), Root Relative Squared Error (RRSE). The same measures were obtained for the approach described in [38]. The M5′ estimators in general performed better than the Multi Layer Perceptron (MLP) neural network approach, which in turn did significantly better than the best single quality measure evaluated (Z-score of Wang *et al.*, for which a RMSE of 1.0264 has been reported in [2]).

The RRSE of the M5′ estimator decreases by 6.9% when saliency based features are considered and the overall number of features used is limited to 5, as suggested in [2]. This indicates that the proposed saliency-based features are indeed relevant to the problem of MOS estimation. In addition, the combination of saliency-related features and M5′ decision trees represents an improvement of 18.9% in terms of RRSE, when compared to the MLP based approach described in [2].

Introduction of additional features, selected using the wrapper methodology, leads to an additional 2.3% decrease in the estimation RMSE, but increases the number of features to be calculated significantly.

The results of the frame-by-frame MOS estimation need to be integrated into a single measure of video quality per sequence. To do this we use median filtering. To evaluate the impact of this procedure, median filter was applied to prediction results of the proposed M5′ approach, based on 5 selected features listed in Table IV and the MLP approach proposed in [2], based on the same number of non-saliency related features. Both M5′ and MLP performed significantly better when final results after median filtering are concerned, achieving an RMSE of 0.1413 and 0.4647, respectively. For reference, the average RMSE of Wang

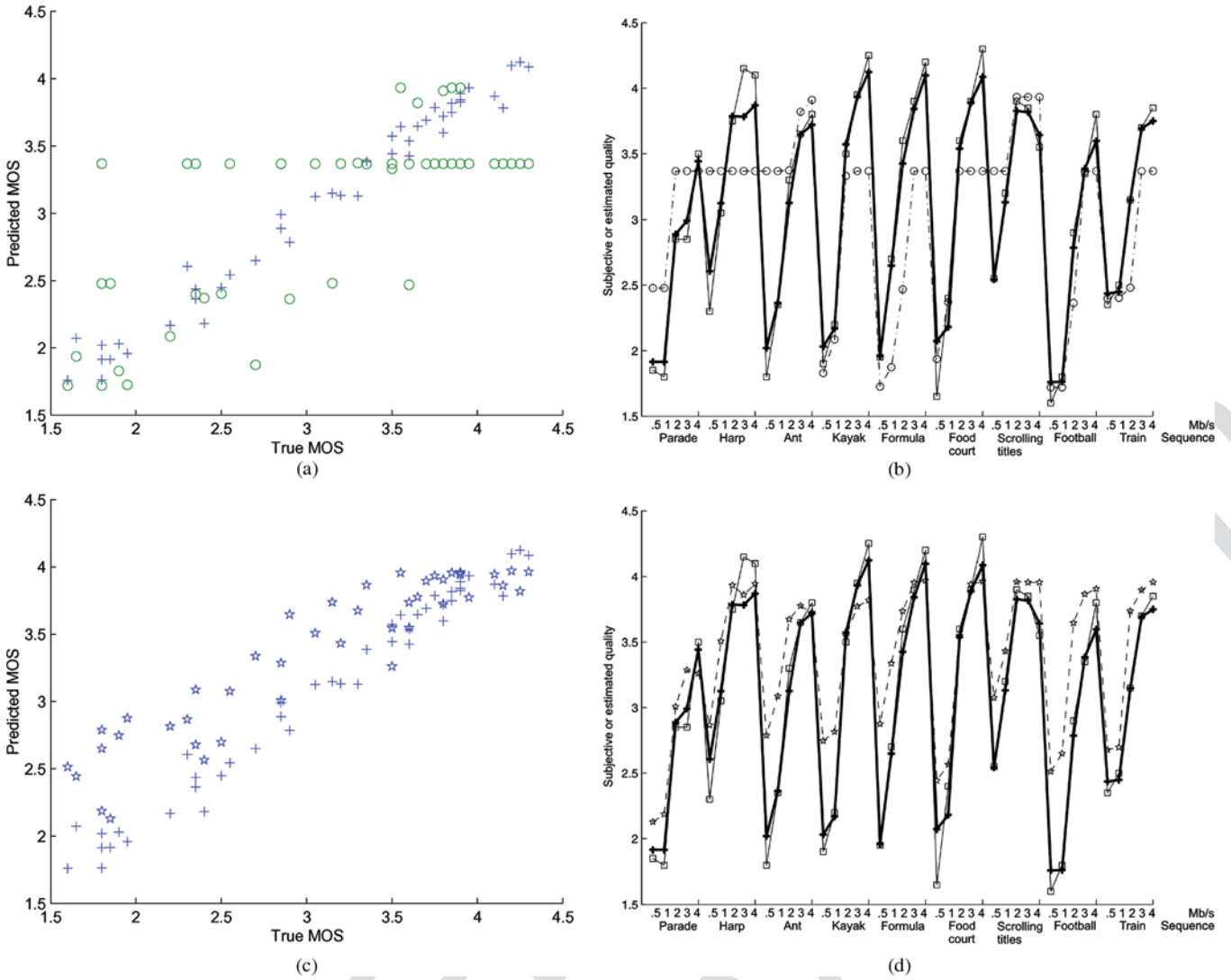


Fig. 5. Test results for the test set containing 10% of data. (a) Estimate scatter plot: proposed M5′ (crosses) and Wang *et al.* (circles). (b) Estimate over the sequences: proposed M5′ (bold solid line, cross markers), Wang *et al.* (dashed line, circle markers) and true subjective quality (MOS) values (thin solid line, square markers). (c) Estimate scatter plot: proposed M5′ (crosses) and [2] approach (stars). (d) Estimate over the sequences: proposed M5′ (bold solid line, cross markers), [2] approach (dashed line, star markers) and true subjective quality (MOS) values (thin solid line, square markers).

*et al.* with median filtering applied (0.5789) is 4 times higher than that of M5′.

Fig. 5 shows the plots of the quality (MOS) estimates obtained for each sequence, when median filtering is applied to per-frame estimates obtained using the proposed approach (crosses), the MLP approach proposed in [2] and the quality metric proposed by Wang *et al.* [1]. It should be that the Wang *et al.* metric is an excellent way of determining the quality of coded still images or single frames. It does not take motion into account and this is, probably, the main reason why it fails on some sequences. We have selected it to demonstrate that quality metrics designed for still images cannot be readily extended to video. The combined effect of the median filtering and insensitivity to motion-related artifacts is that the Wang *et al.* metric is not capable of discerning the changes in quality between the same sequence coded at several bit rates, leading to a number of data points aligned horizontally in Fig. 5(b).

Median filtering reduces the RMSE of the proposed estimator by a factor of 2. This is consistent with the hypothesis that the need for robust statistics is even greater when the salient-motion-related features are used, since the proposed salient-motion detection approach cannot be expected to produce stable results in the first several frames of the sequence, as well as in this interval after the shot changes radically. The effect is due to the need to update the background model.

The RMSE of the proposed VQA approach (0.1413) is significantly lower than the mean standard deviation of the opinion scores (0.7610) of the human observers in the subjective tests conducted.

The full data set used in our experiments, as well as videos and other materials can be found at http://www.dubravko-culibrk.org/VQA.

## V. Conclusion

The effect of bottom-up saliency due to motion on the problem of video quality assessment has been explored in the paper. Specifically, we compare the applicability of a large set of published commonly-used features against that of a number of new features designed to take into account the motion-based visual saliency of objects in video, to the problem of MPEG-2 coded video quality assessment. To derive the new features we use a multi-scale background-subtraction approach and show that it can efficiently be used to determine objects undergoing salient motion even when the camera is not stationary (a restriction commonly imposed on background subtraction methods). Features used to describe coding artifacts were then calculated separately for salient and non-salient regions.

Insights from the field of artificial intelligence and data mining [14] were used to automatically select the features most relevant for the problem at hand. An M5′ decision tree estimator was trained based on the selected features and its performance compared to existing approaches.

The proposed estimator achieved superior results when able to use features taking into account salient motion.

Such results suggest that bottom-up saliency due to motion can be used to enhance the performance of video quality assessment approaches. The improvement can be achieved even when computationally inexpensive approaches, such as that proposed, are used to determine salient regions in the frame.

Finally, the best features selected suggest that the intensity of the blurring and blocking effects in the salient regions have most bearing on the perceived video quality. On the other hand, variance and intensity of temporal changes in the part of the frame not attended to, i.e., non-salient, seem to influence the judgment of video quality.

## Acknowledgment

The authors would like to thank J. Filipović and P. Lugonja for their help in implementing the various quality metrics.

## References

[1] Z. Wang, H. R. Sheikh, and A. C. Bovik, "No-reference perceptual quality assessment of jpeg compressed images," in *Proc. IEEE Int. Conf. Image Process.*, 2002, pp. 477–480.

[2] D. Culibrk, D. Kukolj, P. Vasiljevic, M. Pokric, and V. Zlokolica, "Feature selection for neural-network based no-reference video quality assessment," in *Proc. ICANN (2)*, 2009, pp. 633–642.

[3] *Methodology for the Subjective Assessment of the Quality of Television Pictures*, ITU-R BT.500, Video Quality Experts Group, 2002.

[4] T. Liu, X. Feng, A. Reibman, and Y. Wang, "Saliency inspired modeling of packet-loss visibility in decoded videos," in *Proc. Int. Workshop VPQM*, 2009, pp. 1–4.

[5] L. Itti and P. F. Baldi, "Bayesian surprise attracts human attention," *Vis. Res.*, vol. 49, no. 10, pp. 1295–1306, May 2009.

[6] O. Marques, L. M. Mayron, G. B. Borba, and H. R. Gamba, "An attention-driven model for grouping similar images with image retrieval applications," *EURASIP J. Adv. Signal Process.*, vol. 2007, 2007.

[7] C. Siagian and L. Itti, "Rapid biologically-inspired scene classification using features shared with visual attention," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 2, pp. 300–312, Feb. 2007.

[8] C. Siagian and L. Itti, "Biologically inspired mobile robot vision localization," *IEEE Trans. Robotics*, vol. 25, no. 4, pp. 861–873, Jul. 2009.

[9] Z. Liu, H. Yan, L. Shen, Y. Wang, and Z. Zhang, "A motion attention model based rate control algorithm for h.264/avc," in *Proc. 8th IEEE/ACIS Int. Conf. Comput. Inf. Sci.*, 2009, pp. 568–573.

[10] Y.-F. Ma, X.-S. Hua, L. Lu, and H.-J. Zhang, "A generic framework of user attention model and its application in video summarization," *IEEE Trans. Multimedia*, vol. 7, no. 5, pp. 907–919, Oct. 2005.

[11] Z. Longfei, C. Yuanda, D. Gangyi, and W. Yong, "A computable visual attention model for video skimming," in *Proc. 10th Int. ISM '08*, Washington, DC, 2008, pp. 667–672.

[12] J. Solomon and G. Sperling, "1st-and 2nd-order motion and texture resolution in central and peripheral vision," *Vis. Res.*, vol. 35, no. 1, pp. 59–64, 1995.

[13] M. Hall, "Correlation-based feature selection for discrete and numeric class machine learning," in *Proc. Mach. Learning Int. Workshop*, 2000, pp. 359–366.

[14] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques, 2nd Edition*. San Francisco, CA: Morgan Kaufmann, 2005.

[15] T. Kusuma, M. Caldera, and H. Zepernick, "Utilising objective perceptual image quality metrics for implicit link adaptation," in *Proc. IEEE 2004 Int. Conf. Image Process.*, 2004, pp. IV: 2319–IV: 2322.

[16] R. Venkatesh Babu, A. Perkis, and O. Hillestad, "Evaluation and monitoring of video quality for UMA enabled video streaming systems," *Multimedia Tools Appl.*, vol. 37, no. 2, pp. 211–231, 2008.

[17] G. Warwick and N. Thong, *Signal Processing for Telecommunications and Multimedia,*. New York: Springer, 2004, ch. 6.

[18] I. Kirenko, "Reduction of coding artifacts using chrominance and luminance spatial analysis," in *Proc. ICCE*, Jan. 2006, pp. 209–210.

[19] R. Ferzli and L. Karam, "A no-reference objective image sharpness metric based on just-noticeable blur and probability summation," in *Proc. IEEE ICIP*, 16 2007–Oct. 19 2007, vol. 3, pp. III-445–III-448.

[20] L. L. Zhou Wang and A. C. Bovik, "Video quality assessment based on structural distortion measurement," *Signal Process.: Image Commun.*, vol. 19, no. 2, pp. 121–132, 2004.

[21] L. Itti and C. Koch, "Computational modelling of visual attention," *Nature Reviews Neurosci.*, vol. 2, no. 3, pp. 194–203, Mar. 2001.

[22] N. Idrissi, J. Martinez, and D. Aboutajdine, "Selecting a discriminant subset of co-occurrence matrix features for texture-based image retrieval," in *Proc. ISVC05*, 2005, pp. 696–703.

[23] K. Kim and L. Davis, "A fine-structure image/video quality measure using local statistics," in *Proc. IEEE 2004 Int. Conf. Image Process.*, 2004, pp. V: 3535–V: 3538.

[24] B. P. Olveczky, S. A. Baccus, and M. Meister, "Segregation of object and background motion in the retina," *Nature*, vol. 423, pp. 401–408, 2003.

[25] R. Brooke, "The variation of critical fusion frequency with brightness at various retinal locations," *J. Opt. Soc. Amer.*, vol. 41, no. 12, pp. 1010–1016, 1951.

[26] E. A. Styles, *Attention, Perception, and Memory: An Integrated Introduction*. New York: Taylor & Francis Routledge, 2005.

[27] C. Connor, H. Egeth, and S. Yantis, "Visual attention: Bottom-up versus top-down," *Current Biol.*, vol. 14, no. 19, pp. R850–R852, 2004.

[28] W. James, *The Principles of Psychology, Vol. 1*. New York: Dover Publications, Jun. 1950 [Online]. Available: http://www.worldcat.org/isbn/0486203816

[29] L. Itti, "Automatic foveation for video compression using a neurobiological model of visual attention," *IEEE Trans. Image Process.*, vol. 13, no. 10, pp. 1304–1318, Oct. 2004.

[30] ***[Page numbers?]***F. W. Stentiford, "An attention based similarity measure with application to content-based information retrieval," in *Proc. SPIE Electronic Imag.*, 2003.

[31] J. K. Tsotsos, S. M. Culhane, W. Y. K. Winky, Y. Lai, N. Davis, and F. Nuflo, "Modeling visual attention via selective tuning," *Artif. Intell.* vol. 78, no. 1–2, pp. 507–545, Oct. 1995 [Online]. Available: http://dx.doi.org/10.1016/0004-3702(95)00025-9

[32] O. Le Meur, P. Le Callet, D. Barba, and D. Thoreau, "A coherent computational approach to model bottom-up visual attention," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 5, pp. 802–817, 2006.

[33] J. M. Wolfe, "Visual attention," in *Seeing*. New York: Academic, 2000, pp. 335–386.

[34] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[35] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. ICCV*, 1999, pp. 1150–1157 [Online]. Available: http://citeseer.ist.psu.edu/lowe99object.html

[36] P. Viola and M. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, May 2004.

[37] M. Varma and A. Zisserman, "A statistical approach to texture classification from single images," *Int. J. Comput. Vis.*, vol. 62, no. 1–2, pp. 61–81, Apr. 2005.
[38] D. Culibrk, V. Crnojevic, and B. Antic, "Multiscale background modelling and segmentation," in *Proc. 16th Int. Conf. Digit. Signal Process.*, 2009, pp. 922–927.
[39] S. Wolf and M. Pinson, "Ntia Report 02-392: Video Quality Measurement Techniques," Institute for Telecommunication Sciences, 2002 [Online]. Available: http://www.its.bldrdoc.gov/pub/ntia-rpt/02-392/, Tech. Rep.
[40] A. L. Blum and P. Langley, "Selection of relevant features and examples in machine learning," *Artif. Intell.*, vol. 97, no. 1–2, pp. 245–271, 1997.
[41] ***[Page numbers?]*** R. Babu and A. Perkis, "An hvs-based no-reference perceptual quality assessment of jpeg coded images using neural networks," in *Proc. ICIP*, 2005, vol. 1.
[42] R. J. Quinlan, "Learning with continuous classes," in *Proc. 5th Australian Joint Conf. Artif. Intell.*, 1992, pp. 343–348.
[43] [Online]. Available: ftp://ftp.crc.ca/crc/vqeg/TestSequences/Reference/
[44] L. Li, W. Huang, I. Gu, and Q. Tian, "Statistical modeling of complex backgrounds for foreground object detection," *IEEE Trans. Image Process.*, vol. 13, pp. 1459–1472, 2004.
[45] ***[Page numbers?]*** L. Rosin, "Thresholding for change detection," in *Proc. 6th Int. Conf. Computer Vision (ICCV'98)*, 1998.
[46] V. J. Hodge and J. Austin, "A survey of outlier detection methodologies," *Artif. Intell. Rev.*, vol. 22, pp. 85–126, 2004.
[47] E. McBean and F. Rovers, *Statistical Procedures of Environmental Monitoring Data and Risk Assessment*. Englewood Cliffs, NJ: Prentice Hall, 1998.
[48] [Online]. Available: http://www.compression.ru/video/
[49] D. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*. Reading, MA: Addison-Wesley, 1989.

**Dubravko Ćulibrk** received the B.Eng. degree in automation and system control and the M.Sc. degree in computer engineering from the University of Novi Sad, Novi Sad, Serbia, in 2000 and 2003, respectively. In 2006, he received the Ph.D. degree in computer engineering from Florida Atlantic University, Boca Raton.

He is an Assistant Professor in the Department of Industrial Engineering and Management at Faculty of Technical Sciences in Novi Sad, Serbia. His research interests include video and image processing, computer vision, neural networks and their applications, cryptography, hardware design and evolutionary computing.

**Milan Mirković** received the B.Eng. degree in industrial engineering and management and the M.Sc. degree in business processes automation from the University of Novi Sad, Serbia, where he is pursuing the Ph.D. degree in information and communication systems.

He is a Teaching Assistant at Faculty of Technical Sciences in the University of Novi Sad, Department of Industrial Engineering and Management. He his current interests focusing on image processing, data mining, web and persuasive technologies, and their application.

**Vladimir Zlokolica** received the M.Sc. degree in electrical engineering from the University of Novi Sad, Serbia, in 2001 and the Ph.D. degree in 2006 from the Ghent University, Belgium.

He worked as Doctoral Researcher (2001–2006) and Postdoctoral Researcher (2006–2007) at the Department of Telecommunications and Information Processing, Ghent University, Belgium. In March 2007 he joined MicronasNIT, where he worked as a Video Engineer in the R&D Department for TV Systems until April 2009. Since 2008 he has been an Assistant Professor at Computer Engineering and Computer Communications Department, University of Novi Sad, Serbia. Since April 2009 he has been working in RT-RK Computer Based Systems, as a Video Processing System Architect. His expertise includes video processing and medical imaging, with special accent on video enhancement, image and video objective quality assessment, motion estimation and segmentation, and video conversion.

**Maja Pokrić** received the B.S. degree in electronic and electrical systems engineering in 1994, and the Ph.D. degree in video inspection system and data modeling in 1999, from from Leeds Metropolitan University, Leeds, U.K.

She worked on various multi-national projects as Post-doctorate Research Associate at Department of Electrical Engineering and Electronics, Imaging and Neural Computation, UMIST, and the University of Manchester, Computing and Imaging Science and Biomedical Engineering. Currently she is employed at University of Novi Sad, Faculty of Technical Sciences as lecturer and collaborative engineer. Her projects were related to scientific and medical imaging including X-ray solid-state imaging devices and systems, virtual reality (VR) simulators for surgery, MR, MRI and CT data modeling and, most recently, video quality assessment. Her main areas of research interest are video and image processing, video system modeling and medical imaging.

**Vladimir Crnojević** (M'97) 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.

**Dragan Kukolj** (M'97–SM'06) received the Diploma degree in control engineering in 1982, the M.Sc. degree in computer engineering in 1988, and the Ph.D. degree in control engineering in 1993, all from the University of Novi Sad, Serbia.

He is currently a Professor of computer-based systems with Department of Computing and Control, Faculty of Engineering, University of Novi Sad. His main research interests include soft computing techniques and their applications in signal processing and process control. He has published over 100 papers in journals and conference proceedings. He is the consultant of R&D companies, where he is involved in the areas of soft computing and computer-based systems integration with application in high-tech sensors, audio and video processing.