Quantifying the amount of spatial and temporal information in video test sequences

A. Ostaszewska, R. Kłoda

Warsaw University of Technology, Faculty of Mechatronics,
Sw. Andrzeja Boboli 8 Str., Warsaw, 02-525, Poland

Abstract

In case of compressed video quality assessment, the selection of test scenes is an important issue. So far there was only one conception for evaluation the level of scene complication. It was given in International Telecommunication Union recommendations and was broadly used. Authors investigated features of recommended parameters. The paper presents the incompatibility of those parameters with human perception that was discovered and gives a proposal of modification in algorithm, which improves accordance of parameters with observers' opinion.

1. Introduction

The rapid growth of digital television, DVD editions and video transmission over the Internet has increased the demand for effective image compression techniques and the methods of coding/decoding systems evaluation. There are two alternative ways of compressed video quality evaluation: perceptual (sometimes called subjective) and computational (also referred to as objective). No matter what the method is, the crucial role in results of a coder evaluation is played by the scene selection. The algorithm (or the whole system) performance is strictly dependant on the amount of perceptual information that the picture contains. In case of a video, the perceptual information can be divided into spatial and temporal. Test sequences must span the full range of spatial and temporal information of interest to users of the system under test. Considering test sequence

selection, the need to quantify the amount of this information seems to be obvious.

2. SI and TI according to ITU Recomendations

The spatial and temporal information measures proposed by International Telecommunication Union [1] are represented by single values for the whole test sequence.

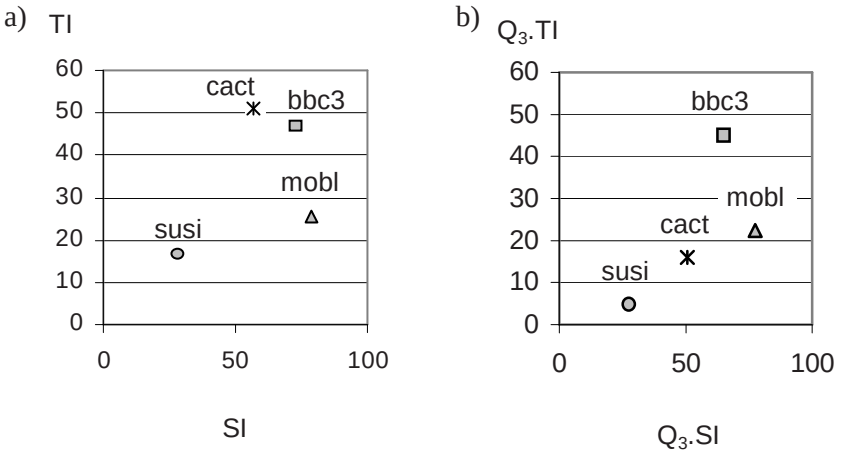

Fig.1. The comparison of SI(TI) plot (a) and Q_3.SI(Q_3.TI) (b)

The SI (Spatial perceptual Information) takes into consideration the luminance plane only and is computed on the base of Sobel filter. Each video frame at time n (F_n) is transformed with the Sobel filter [Sobel(F_n)]. Then the standard deviation over the pixels (std_{space}) in each Sobel-filtered frame is computed. This operation is repeated for each frame in the video sequence and afterwards the maximum value in the time series (max_{time}) is chosen:

$$SI = \max_{time} \left\{ std_{space} \left[Sobel\ (F_n) \right] \right\} \tag{1}$$

The TI (Temporal perceptual Information) is also based on a luminance plane and calculates the motion. The motion is considered to be the difference between the pixel values at the same location in space but at successive frames: $M_n(i, j)$. $M_n(i, j)$ is therefore a function of time (n) and it is defined as:

$$M_n(i, j) = F_n(i, j) - F_{n-1}(i, j) \tag{2}$$

where $F_n(i, j)$ is the pixel value at the i^{th} row and j^{th} column of n^{th} frame in time.

The measure of TI is calculated as the maximum over time (max_{time}) of the standard deviation over space (std_{space}) of $M_n(i, j)$ over all i and j.

$$TI = \max_{time}\left\{std_{space}\left[M_n(i, j)\right]\right\} \tag{3}$$

SI and TI are usually computed for the whole sequence, so each scene is described by two parameters. Higher values of SI and TI represent sequences which are more difficult to decode and are more likely to suffer from impairments. In order to choose scenes which will span as wide range of information to decode as possible, usually SI and TI are put in the TI(SI) plot and the scenes with the uttermost values are selected.

3. SI and TI new approach

Authors conducted Single Stimulus Continuous Quality Evaluation method [2, 3, 4], using 4 sequences (each 15 seconds long) coded with 13 GOP, all three possible GOP structures (with 1, 2 or without B frames) and with 5 levels of bitrate in a range of 2 Mbps to 5 Mbps. Hence, the test material was 30 minutes long and contained 15 variants of coding each of 4 test sequences. 45 subjects participated in the research. The voting signal was sampled at 2 Hz frequency.

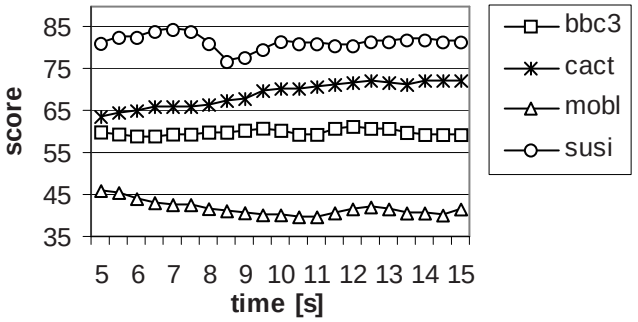

Fig.2. The average score given in time
to 4 sequences across the whole bitrate range

The interesting observation was that the lowest grade was always given to the sequence ‘mobl’ or ‘bbc3’, while ‘cact’ scene used to get scores close to the easiest to decode – ‘susi’ (fig. 2). According to SI and TI parameters, ‘cact’ was the sequence with the highest TI value and should contain clearly visible impairments (fig. 1a), which were supposed to affect the mean score given by observers.
This phenomenon impelled authors to investigate the variability of SI and TI in time. For this purpose both parameters were calculated on fame by frame basis. The intriguing discovery was that the high level of TI for ‘cact’ sequence was caused by one extraordinary peak, which falls on the frames with scene cut (fig. 3). Although it may cause some problems with coding, observers seem not to react to this incident at all (fig. 2).

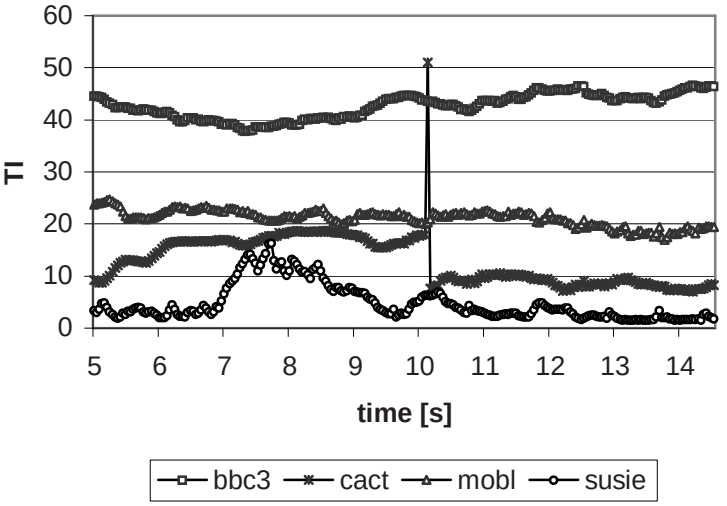

Fig.3. The TI (computed for each frame) distribution in time for 4 test sequences

As the initial role of SI and TI parameters was to reflect perceptual amount of information, authors propose a slight modification in the way those parameters are computed, so that the values were in accordance with the level of scene complication perceived by the observer:

$$Q_3.SI = \text{Upper quartile}_{time}\left\{ std_{space}\left[Sobel(F_n) \right] \right\} \quad (4)$$

$$Q_3.TI = \text{Upper quartile}_{time}\left\{ std_{space}\left[M_n(i,j) \right] \right\} \quad (5)$$

Fig. 1 shows that Q_3.SI and Q_3.TI placed the "cact" sequence in a new position on the plot – now it can be identified as the scene just slightly more difficult to decode than "susi", while "mobl" and "bbc3" kept their position of critical for the coding system. Hence Q_3.SI and Q_3.TI reflect perceptual amount of information in a better way in comparison with traditional SI and TI.

4. Conclusions

As the end-user of the video coding system is the observer himself, on each step of investigation it is important to mind the features of human visual perception, which is disposed to average the stimuli rather than to react to short time values. Hence, the idea of computation the upper quartile values of information in a scene seems to reflect human perception in a more adequate way than the maximum value. Still the conception of evaluating the amount of perceptual information seems to be imperfect and should be under investigation in the future. Before it's done, authors advise to take into consideration the upper quartile value or to study the plots of spatial-temporal information on a frame-by-frame basis. The use of information distributions over a test sequence also permits better assessment of scenes in case of continuous assessment, which is a new mainstream in the area of subjective quality evaluation of compressed video.

References

[1] ITU-Telecommunications Standardization Sector: Two criteria of video test scene selection, Geneva, 2-5 December 1994.

[2] ITU-T Recommendation P.911 (1996), Subjective audiovisual quality assessment methods for multimedia applications.

[3] Ostaszewska A., Żebrowska-Łucyk S., Kłoda R.: Metrology tools in subjective quality evaluation of compressed video, Mechatronics Robotics and Biomechanics Trest, Czechy 2005.

[4] Ostaszewska A., Żebrowska-Łucyk S., Kłoda R.:Metrology properties of human observer in compressed video quality evaluation, XVIII IMEKO WORLD CONGRESS, Metrology for a Sustainable Development Rio de Janeiro, Brazil 2006.