Using Spatio-Temporal Saliency to Predict Subjective Video Quality: A New High-Speed Objective Assessment Metric

Maria Laura Mele[1,2,3(✉)], Damon Millar[3], and Christiaan Erik Rijnders[3]

[1] Department of Philosophy, Social and Human Sciences and Education, University of Perugia, Perugia, Italy
marialaura.mele@gmail.com
[2] ECONA, Interuniversity Centre for Research on Cognitive Processing in Natural and Artificial Systems, Sapienza University of Rome, Rome, Italy
[3] COGISEN Engineering Company, Rome, Italy
{marialaura,damon,chris}@cogisen.com

Abstract. We describe a new Objective Video Quality Assessment (VQA) metric, consisting of a method based on spatio-temporal saliency to model human visual perception of quality. Accurate measurement of video quality is an important step in many video-based applications. Algorithms that are able to significantly predict human perception of video quality are still needed to evaluate video processing models, in order to overcome the high cost and time requirement for large-scale subjective evaluations. Objective quality assessment methods are used for several applications, such as monitoring video quality in quality control systems, benchmarking video compression algorithms, and optimizing video processing and transmission systems. Objective Video Quality Assessment (VQA) methods attempt to predict an average of human perception of video quality. Therefore subjective tests are used as a benchmark for evaluating the performance of objective models. This paper presents a new VQA metric, called Sencogi Spatio-Temporal Saliency Metric (Sencogi-STSM). This metric generates subjective quality scores of video compression in terms of prediction efficacy and accuracy than the most used objective VQA models. The paper describes the spatio-temporal model behind the proposed metric, the evaluation of its performance at predicting subjective scores, and the comparison with the most used objective VQA metrics.

Keywords: Objective video quality assessment · Video compression models · Spatio-temporal saliency · Video quality assessment metrics

1 Introduction

Humans are the end-users of most multimedia applications. Since objective models are unable to perfectly model human vision, the most accurate methodology of video quality assessment is still through subjective perception [1].

M. Kurosu (Ed.): HCI 2017, Part I, LNCS 10271, pp. 353–368, 2017.
DOI: 10.1007/978-3-319-58071-5_27

Predicting subjective quality ratings in a reliable way is one of the main issues facing objective video quality assessment (VQA) models, because subjective tests require high cost and time effort. Moreover, the State of the Art on subjective VQA shows a wide range of evaluation methods. At the moment, the most used methods follow the ITU-R Recommendation BT.500 [2], which proposes standardized presentation formats to measure human participants' mean opinion scores of video quality. The main issue of subjective VQA measurement is that it is often time-consuming and requires the recruitment of a high number of participants to be statistically reliable, thus incurring high costs.

To avoid the cost and delay of subjective VQA, objective VQA is often used. The current objective VQA methods can be classified in three categories: full-reference VQA, reduced-reference VQA, and no-reference VQA. In full-reference VQA methods, an undistorted quality reference video is fully available for comparisons with distorted videos. In reduced-reference VQA methods, only some features of the undistorted quality reference video is used to evaluate the quality of distorted videos. In no-reference VQA methods, the reference video is not available at all [3]. This paper focuses on full-reference methods.

The first section of this paper describes the most commonly used full-reference objective VQA methods i.e., Peak Signal to-Noise Ratio, which is a simple and easy to calculate algorithm but it does not highly correlate with perceived quality subjective evaluations, and the more accurate Structural Similarity Index (for a review on all the existing objective VQA methods see [3]). Neither of these objective VQA metrics are able to calculate whether the relationships among pixels is perceptually salient, so they cannot be applied to evaluate saliency-based compression algorithms. The second section of the paper describes a new metrics, called Sencogi Spatio-Temporal Saliency Metric (Sencogi-STSM), designed by an engineering company called Cogisen (www.cogisen.com). The metric is based on a model using spatio-temporal saliency to account for human visual perception. Sencogi-STSM is compared to the performance of both PSNR and SSIM, taking as a benchmark the subjective evaluation of compressed videos.

2 Quality Assessment Methods

This section describes two quantitative VQA methodologies: the most used objective quality assessment methods and metrics, and the VQA methods and metrics based on saliency models.

2.1 Objective Quality Assessment Methods

Objective VQA methods provide video quality scores without the involvement of participants. Since there is no delay for human testing, objective VQA scores allow practitioners to quickly develop video codecs. Many types of objective VQA methods (e.g. Video Quality Metric (VQM); Visual Information Fidelity (VIF), see [3]), have been proposed in the literature but there is no objective measurement which is able to

predict subjective quality scores in all experimental testing conditions, as the research results of the Video Quality Experts Group (VQEG) show [4].

Two existing objective methods will be described: Peak Signal to Noise Ratio (PSNR) [5] and Structural Similarity index (SSIM) [6]. The selected methods –which are widely cited in the literature and provide the most used measures by practitioners– belong to Image Quality Metrics (IQMs). IQMs attempt to measure the quality of a single static image, and can also be used to measure video quality by treating the video stream as a collection of images, and calculating an aggregate score.

PSNR is a full reference QA method able to measure the ratio between the maximum power of a signal and the power of corrupting noise, by performing a pixel-by-pixel comparison of a video-frame before and after it is processed [5]. As a first step, PSNR calculates the Mean Square Error (MSE) of each bit, so that the maximum possible pixel value is squared and divided by MSE, and a logarithm taken of it to give the related PSNR index. PSNR is widely used because it provides a simple measure of the distortion and noise in a processed video-frame, even though it is not able to model human perception in a significant way—all pixels are treated as being of equal importance. Due to its inability to model human vision, PSNR is becoming less useful as modern video codecs increasingly apply human perception rules to eliminate the information that falls beyond the visual perception threshold.

Another QA method used is SSIM, which models human perception by calculating an index of "structural similarity" that aims to emulate how the human visual system perceives quality. In SSIM, video-frame degradation is considered as a change in structural information. The model behind SSIM considers pixels as having strong interdependencies, especially when they are spatially close. Pixel interdependencies are therefore able to convey important information about the structure of visual scene. SSIM calculates three visual components of a frame –luminance, contrast and structure– according to the following weighted combination:

- *Luminance.* High values of luminance are weighed more. The luminance of each pixel is twice the product of average x and y over the sum of the square of average.
- *Contrast.* Locally unique pixel values of contrast are weighed more. The contrast of each point is twice the product of variance x and y over the sum of the square of average.
- *Structure.* The more pixel values change together with their neighbours, the more they are weighed. The structure of each point is the covariance of x and y over the product of the variance x and y.

Variants of SSIM have been proposed [7], such as the Multi-Scale SSIM index (MSSIM), which is a measure based on the multi-scale processing of the early vision system. Both SSIM and MSSIM have shown to be highly predictive of human quality scores, but they are more complex to calculate than PSNR, and they have been both been originally designed for static images, thus they do not properly measure visual distortion among the frames in a video. Moreover, although SSIM/MSSIM is able to measure structural relationships among pixels, they are still unable to measure whether those relationships are perceptually salient. This issue affects the evaluation score especially when salient information is selectively compressed following saliency based

compression algorithms (Saliency-based video compression models use saliency to provide better quality in salient areas by keeping the average distortion levels unvaried). Subjective scores report higher values of video quality compared to saliency-based bit distribution, even though MSSIM does not report any improvement. New objective QA models are still needed that are able to calculate the salient parts of video information.

2.2 Saliency-Based Quality Assessment Methods

Saliency is an attention process that helps humans to focus their cognitive resources on the most pertinent subgroup of data, since our visual system can only process partial amounts of information from the wide stream of information that surrounds us [8]. This selection process functions as a filter regulating the access of salient visual information to high level processing systems in the brain, allowing only salient information to reach our awareness.

Since the human visual system (HVS) is the ultimate assessor of image quality, the effectiveness of an Image Quality Metrics (IQM) is generally quantified by to what extent its quality prediction is in agreement with human judgments [9]. The relationship between salience and quality perception has led to a number of approaches that try to integrate salience into IQA metrics to improve their prediction performance [10]. Saliency-weighted IQAs have successfully improved SSIM and PSNR performance [11].

Most saliency algorithms use spatial properties of an image to predict visual salience. There are more than ten spatial saliency algorithms [12]. One reason that there are so many salience algorithms is that the quality of the salience algorithm is important - Zhang et al. found that the difference in predicting human fixations between saliency models is sufficient to yield a significant difference in performance gain when adding these saliency models to IQMs [12].

Some saliency algorithms use frequency domain properties of the image to determine salient areas [13–18]. Frequency domain saliency algorithms respond to patterns in the image, and are typically modelled on the biological properties of the visual cortex of the human eye. Salience maps generated by frequency domain algorithms can solve many problems typically seen in spatial salience calculation methods [19]. Spatial domain algorithms typically produce low-resolution salience maps, have ill-defined object boundaries from severe downsizing of the input image, and fail to uniformly map the entire salient region.

Most saliency-based perception models described in the literature follow two theoretical approaches to obtain saliency: the bottom-up and the top-down approach. The bottom-up approach follows the visual saliency hypothesis [16], which explains the selection of a fixation site as a feature-guided process, and considers visual attention as a data-driven reaction to visual features. The top-down approach is based on the cognitive control hypothesis [16, 20], according to which visual attention is guided in a top-down way according to the task-related needs of the cognitive system. Visual stimuli are relevant (as they are for the bottom-up theory), but this relevance is determined by cognitive information rather than inherent visual saliency [8]. Bottom-up video compression models predict visual saliency from visual patterns, for example using pixel-level contrast or colour differences from the average video-frame colour [16, 21].

However, perceptual sensitivity may not be able to completely explain human visual attention, because it does not consider other variables related to context or cognition. In order to solve this problem, top-down video compression models aim to predict visual saliency starting from representations of viewers' goals and tasks [22]. A problem with top-down saliency models is that they are meant to calculate the saliency on a visual scene, ignoring what is salient or may become salient due to compression artefacts, e.g. ringing, contouring or aliasing. Zhang at el. found that image quality degradation could give rise to changes in images' salience maps [10].

Objective Video Quality Metrics (VQMs) differ from IQMs because human perception of static images is different than moving images. VQMs also differ from IQMs in

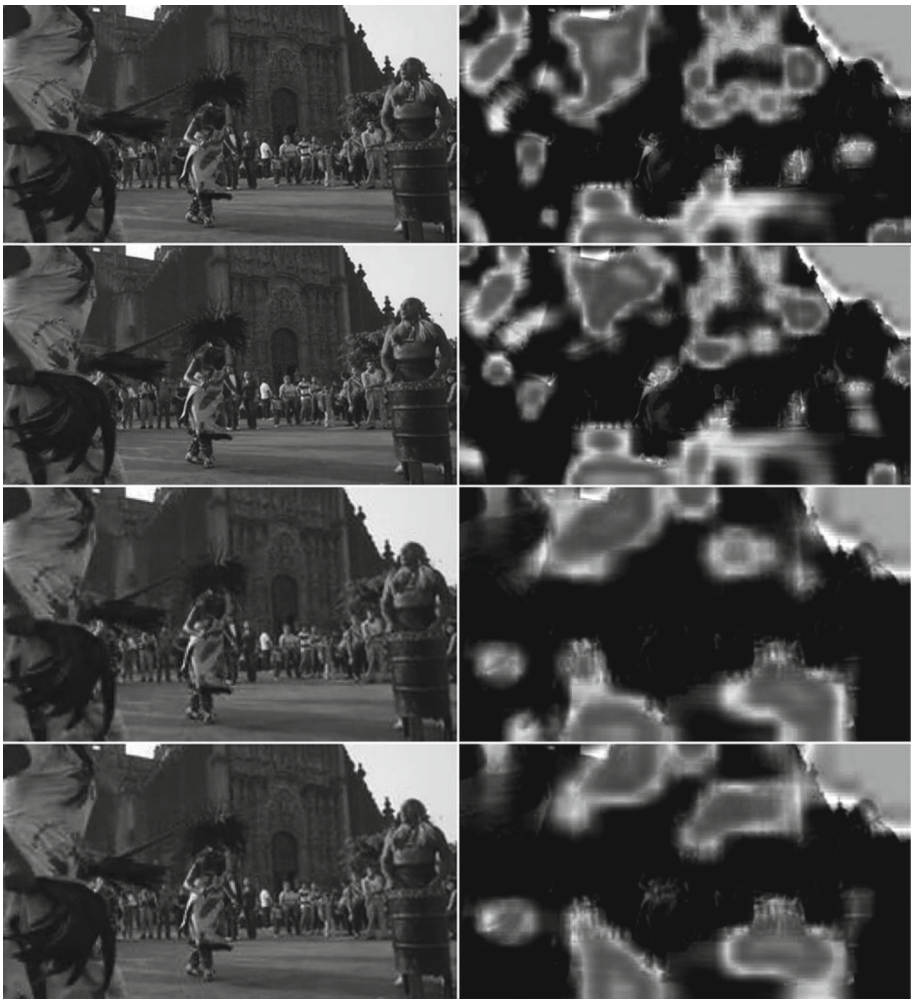

Fig. 1. Video frame at progressively lower resolutions and quality, and the spatial salience map of the frame. The salience map has readily visible changes in response to quality.

that there is a timeliness requirement - processing video can be resource-intensive. The saliency analysis of videos is more complicated than that of still images because there is a spatio-temporal correlation between regions of consecutive frames. The motion of objects changes their importance in a scene and leads to a different saliency map [23]. To address the changing salience of video, some VQMs attempt to incorporate spatio-temporal measures of salience. VQMs that incorporate a measure of salience, perform significantly better than traditional IQAs at predicting subjective image quality [10].

Video compression often produces distortions turning non-salient parts of a visual scene into salient areas. Both bottom-up and top-down video compression models only consider within-frame visual saliency (called "spatial saliency"), thus not properly calculating between-frames spatio-temporal saliency, also called "spatio-temporal saliency".

In the literature, less attention is given on spatio-temporal saliency compared to spatial saliency. Spatio-temporal saliency is mainly studied in cognitive science research, which aims to model perceptual and attentional processes [24–26], and spectral analysis research, which aims to extend frequency domain use of phase data [27, 28]. Applying spatio-temporal saliency to compression may be complicated because of noise produced by camera sensor or compression codec, which can be difficult to discriminate from salient motion. Most of the compression models based on spatio-temporal saliency use global search methods based on a single phenomenon such as motion, optical flow, flicker, or interest points. They impose heavy computational costs because they need to combine many such search algorithms at many scales. Measures of the salience map deformation are a good basis for VQA, because: (1) changes in quality are more visible in salience maps than in video images (Fig. 1); (2) changes in salience can cause a scene to be regarded differently by a viewer (e.g. regarding a different part of a scene) affecting subjective quality; (3) if video has been encoded using a salience-aware codec, that more heavily compresses parts of the frame it predicts as non-salient, then deformations in the salience map may cause the viewer to attend to heavily compressed areas.

3 Sencogi Spatio-Temporal Saliency Metric (Sencogi-STSM)

A new saliency-aware VQA metric called Sencogi-STSM has been developed by Cogisen. The metric is able to predict subjective evaluation of quality for compression models without using cohorts of human viewers. Unlike most objective VQA models, Sencogi-STSM is able to evaluate the quality of videos compressed by saliency-based codecs.

3.1 Cogisen's Video Compression Algorithm

The VQA metric is based on saliency algorithms that Cogisen developed for video compression. Cogisen's video compression algorithms were designed for low bandwidth video applications, such as mobile, that have low video resolution/quality. At low resolutions, it can be challenging for video encoders to calculate saliency, because there are not enough pixels to calculate edges and contrasts. Although low-resolution is difficult to compress, low bandwidth is particularly important for devices that have limited

processing capacity and data bandwidth, such as smartphones. Smartphones are becoming the dominant device used for video recording and playing [29]. Smartphone video is also frequently used for live video streaming, such as video chat communication, where latency and delays are easily apparent, so suitable video compression algorithms should meet tight speed and low bandwidth targets.

Cogisen's salience-enabled video compression algorithms were developed for real-time live video, where each frame is compressed in the time between subsequent frames, which requires very fast saliency calculations. Four different types of saliency algorithms are simultaneously run on a real-time video stream and combined to drive the codec's variable macro-block compression. Cogisen's saliency is used in a different way than other salience-based video compression algorithms: many algorithms use saliency to variably drive compression level, to find an acceptable quality trade-off, where videos can have a lower subjective quality in non-salient parts in order to obtain extra compression gains. In Cogisen's implementation, the saliency algorithms are tuned for threshold rather than trade-off. Using a saliency threshold ensures there is no visible loss anywhere in a video. The use of four salience drivers ensures that information removal in one domain does not introduce salient artefacts in another domain.

3.2 A New Saliency-Aware Video Quality Assessment

The salience algorithms from Cogisen's video compression were used to create Sencogi-STSM, a new saliency-aware VQA. The four types of salience computed are:

- *Pixel Noise Detection*, which discerns between pixel noise and motion. Camera sensors produce noise, which appears as random bit changes in the frame. In some situations where a small part of the sensor's dynamic range is used (e.g. low light conditions) pixel noise can be the majority of the change between frames. Pixel noise is the first type of saliency to be calculated, because spatio-temporal algorithms cannot discern genuine scene motion from sensor pixel noise.
- *Static Saliency,* which is saliency within a video frame.
- *Spatio-Temporal Saliency*, which is saliency of the motion between frames. Some types of motion are more salient than others. Once the pixel noise has been identified, the spatio-temporal saliency gives an indication of how strong the video compression can be in different parts of the frame. Spatio-temporal saliency, in particular the prediction of spatio-temporal saliency artifacts, was found to be the most influential factor in subjective image quality, especially in low bandwidth implementations. In low quality videos, any reduction in quality or resolution may result in distortions such as blurry edges due to ringing artifacts or shadowing effects behind the motion. At even lower resolutions and quality levels, a moving object may not even be recognizable but it will be a blob. Cogisen's spatio-temporal saliency algorithm is able to detect those parts that might become salient due to new pixel noise artifacts, by calculating the correlation between the original high quality saliency map and the saliency map of the compressed video.
- *Delta-Quality Saliency*, which calculates whether a quality change is noticeable subjectively by a user, affecting the natural scene salience [30]. If part of a video

becomes better or worse quality, it can attract attention, depending on the amount of quality change. We term this induced saliency "Delta-Quality Saliency". It is a separate saliency calculation for each macro-block that is correlated to the amount of compression change that would lead to the video quality being perceived as changed.

The salience maps are weighted by tunable thresholds, then added to form an overall salience map (Fig. 2). A video quality score is obtained by comparing the overall salience maps of the compressed and reference videos (see Fig. 3) using SSIM.

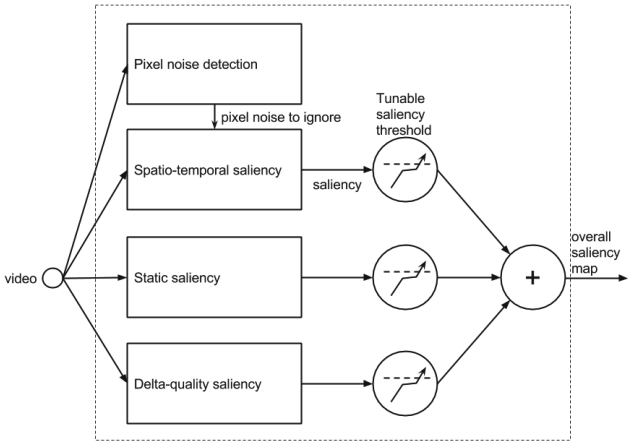

Fig. 2. Figure shows how the different saliency types are combined.

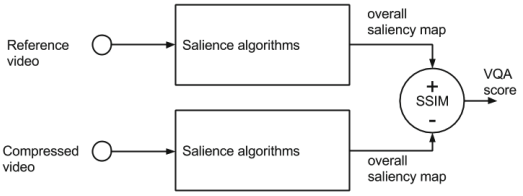

Fig. 3. Figure shows how video quality is measured as a change in saliency map.

4 Performance Evaluation of the Sencogi Spatio-Temporal Saliency Metric

4.1 Methodology

The evaluation of the performance of the Sencogi-STSM followed three phases. In Phase 1, a subjective model was followed to create a benchmark database. In Phase 2, objective VQA scores were calculated by applying both the most used objective

VQA metrics (PSNR, and SSIM), and Sencogi-STSM. In Phase 3, we compared the subjective quality scores obtained in Phase 1, to the objective score obtained in Phase 2.

4.2 Phase 1. Subjective Video Quality Assessment Database

In order to create a subjective video quality database for benchmarking the evaluation of the Sencogi-STSM, the subjective opinion scores were calculated of videos compressed at different Constant Rate Factor (CRF) values, and by two different compression methods. Constant Rate Factor is a setting that instructs the encoder to attempt to achieve a certain output quality by reducing the bitrate. The range of the quantizer scale is 0–51: where 0 is lossless, 23 is default, and 51 is worst possible. A lower value is a higher quality and a normal range is 18–28. CRF 18 is considered to be visually lossless [31]. Reference videos were compressed by two video compression models: ×264 (which does not include a salience model) and the ×264 codec with compression weighted by a salience model that was previously proven to increase compression without affecting subjective scores. The saliency-based video compression model has been recently validated and evaluated [32, 33]. Video compression was performed at two compression levels: Constant Rate Factors (CRF) 21 and CRF 27. The experimental design was 5 (reference videos) × 2 (compression methods) × 2 (compression levels). Subjective opinion scores assigned to each compression level were collected to create a VQA database.

4.2.1 Materials

Five benchmark videos (called "Big Bucks Bunny", "Bouncing balls", "Netflix ritual dance", "Crowd run" and "Tears of steel") with high technical complexity related to the current compression methods were selected. The selected videos lasted less than 10 s and were in the uncompressed YUV4MPEG 4:2:0 format. Only one video was in the MP4 format ("Bouncing Balls") because it was unavailable in an uncompressed format. All videos were 426 × 224 landscape resolution. The raw source of each file was encoded into the MPEG4 format. Reference videos were compressed with a visually lossless CRF value of 10. CRF 10 reference videos were then compressed using both the standard H264 compression and the saliency based model, each video was compressed to two levels: CRF 21 and CRF 27.

4.2.2 Procedure

The Single Stimulus Continuous Quality Scale (SSCQS) method with hidden reference removal was used [2]. The SSCQS method presents one video at a time to the viewer. An example of a high quality video is presented only once at the beginning of the test. Reference high quality videos are randomly shown during the test as a control condition, and participants are not aware of that. The sequence presentations are randomized to ensure that the same video is not presented twice in succession (the randomization is performed when the survey is developed – every user receives the same randomized sequence). As the presentation of each trial ends, observers evaluate the quality of each

video using a grading scale of integers in the range 1–100. The scale was marked numerically and divided into five equal portions, which were labelled with adjectives: "Bad", "Poor", "Fair", "Good", and "Excellent". The position of the slider is automatically reset after each evaluation. The survey was created and administered using a web-based survey software tool called SurveyGizmo (www.surveygizmo.com), following an online-based methodology, whose validity was previously assessed by the authors [32].

4.2.3 Subjects

Thirty-nine participants (mean age 31.6 years old, 70.9% male, 17.9% expert viewers, 58.9% indoor with artificial lights, 41.1% indoor with natural lights) completed the subjective test in a single session on November 4, 2016. The pre-screening of the subjective test scores consisted of determining if the participants met the preliminary requirements (no vision impairments, only personal computers, no smartphone or tablets, maximum brightness on, bandwidth speed higher than 40 megabits/seconds). Six outliers were removed.

4.2.4 Results of Subjective Video Quality Assessment

The Mean Opinion Scores assigned to the reference videos were used to calculate the Difference Mean Opinion Scores (DMOS) between each compressed video and the relating reference using the following formula:

$$d_{ij} = r_{i\mathrm{ref}(j)} - r_{ij}$$

where r_{ij} is the raw score for the i-th subject and j-th image, and $r_{i\mathrm{ref}(j)}$ denotes the raw quality score assigned by the i-th subject to the reference image corresponding to the j-th distorted video [35].

Scale assessment. Internal consistency was supported by Cronbach's alpha (alpha = 0.969), Spearman Brown split-half value (rho = 0.932) (Cronbach's Alpha = 0.951 for the first half and alpha = 0.931 for the second half), meaning that all the items of the scale measured the same dimension.

Opinion Scores. The mean opinion scores (MOS) were calculated for each subject. The Difference Mean Opinion Scores (DMOS) were obtained by calculating the difference between the MOS of reference videos and the MOS of the related processed videos (H264 DMOS TOT = 15.46; saliency based compression DMOS TOT = 14.52; H264 DMOS CRF 21 = 2.90; saliency based compression DMOS CRF 21 = 0.06; H264 DMOS CRF 27 = 12.5; saliency based compression DMOS CRF 27 = 14.16).

4.3 Phase 2. Objective Video Quality Assessment

The quality of each reference and compressed videos (used in Phase 1 to assess the subjective perception of quality) was measured by the following VQA metrics: (1) PSNR; (2) SSIM; (3) Sencogi-STSM.

4.3.1 Results of Objective Video Quality Assessment

Table 1 shows the total results for each objective metric (Means: PSNR = 35.898, SSIM = 0,951, Sencogi-STSM = 3.19).

Table 1. Objective VQA metrics for compressed video

	"Big bucks bunny" video		"Tears of steel" video		"Neflix ritual dance" video		"Crowd run" video		"Bouncing balls" video	
	H264	Saliency-based compression	H264	Saliency-based compression	H264	Saliency-based compression	H264	Saliency-based compression	H264	Saliency-based compression
PSNR	42.80	42	36.917	36.740	35.484	35.084	32.176	32.102	33.048	32.614
SSIM	0.983	0.981	0.964	0.963	0.933	0.931	0.934	0.935	0.945	0.941
Sencogi-STSM	3.325	3.218	3.167	3.177	2.877	2.888	3.313	3.339	3.305	3.288

4.4 Phase 3. Prediction Performance of Objective Models

Phase 3 consisted of four comparative analyses between the objective metrics calculated in Phase 2 and the subjective scores calculated in Phase 1 (Fig. 4). This phase followed the methodology recommended by the ITU Telecommunication Standardization Sector [33].

4.4.1 Procedure

The performance of all objective models was tested by using the following metrics:

- The *Spearman Rank Order Correlation Coefficient* (SROC) measures the prediction monotonicity of an objective metric, that is to say, the index in which objective scores are able to predict subjective scores.
- The *Pearson Linear Correlation Coefficient* (PLCC) measures prediction accuracy, that is to say the capability to predict the subjective scores with low error. The Pearson linear correlation it is usually calculated after applying a nonlinear regression with a logistic function as recommended by the ITU Telecommunication Standardization Sector.
- The *Outlier Ratio* (OR) is defined as the percentage of the predictions number that falls outside 2 times the standard deviation of subjective DMOS.
- The *Root Mean Square Error* (RMSE) measures prediction accuracy like the Pearson linear correlation [39].

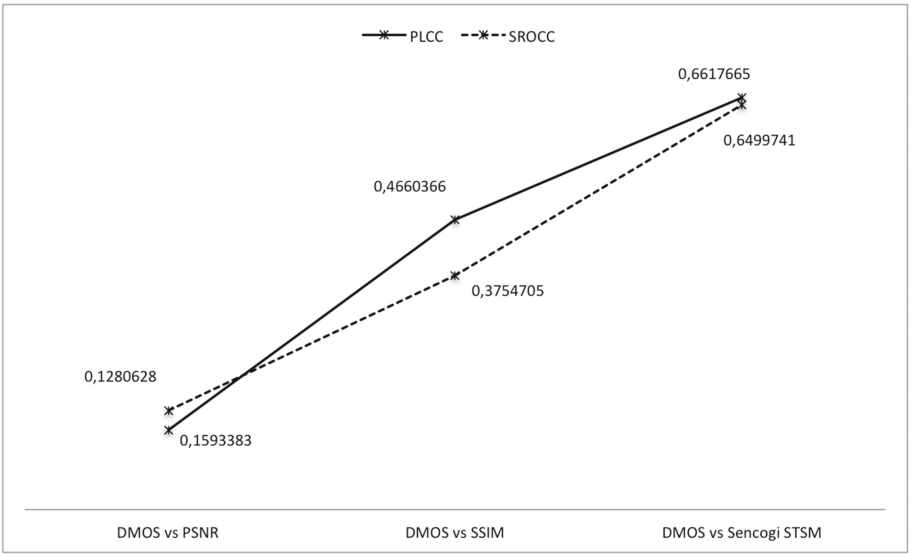

Fig. 4. Comparison among correlations between DMOS and objective metrics

4.4.2 Results of Objective Video Quality Assessment

- *Spearman Rank Order Correlation* (SROC). Results on both CRF 21 and 27 show a significant positive correlation between Sencogi-STSM values and DMOS values ($rho = 0.650$, $p < 0.01$). No significant correlation between both PSNR ($rho = 0.159$, $p > 0.05$) and SSIM ($rho = 0.375$, $p > 0.05$) values and DMOS values was found.

- *Pearson Linear Correlation Coefficient* (PLCC). Results on both CRF 21 and 27 show a significant positive correlation between objective measures and DMOS subjective scores for both Sencogi-STSM ($r = 0.662$, $p < 0.01$) and SSSIM ($r = 0.466$, $p < 0.05$). No significant correlation between PSNR and DMOS was found ($r = 0.128$, $p > 0.05$). Comparison of both Spearman's (SROC) and Pearson's (PLCC) correlation among PSNR, SSSIM and Sencogi-STSM and DMOS values.
- *Outlier Ratio* (OR). Only 7% of the values predicted by both SSIM (OR = 0.65) and Sencogi-SMST (OR = 0.70) fall outside ±2 of the standard deviation of subjective DMOS, whereas all PSNR values (OR = 1) fall outside ±2 of the SD of subjective DMOS.
- *The Root Mean Square Error* (RMSE). Paired t test showed that SSIM scores ($t(10) = 10.32$, $p = 0.000$) and Sencogi-STSM scores ($t(10) = 12.66$, $p = 0.000$) are more statistically significant than PSNR scores. Moreover, Compared to PSNR and SSIM, Sencogi-STSM prediction scores have significantly lower RMSE than SSIM scores ($t(10 = 2.29$, $p = 0.048$) with Sencogi-STSM RMSE = 9.045; PSNR RMSE = 29.898, SSIM RMSE = 10.201.

5 Discussion

Based on the analyses presented in this work, the new Sencogi-STSM metric is an effective metric for predicting the subjective quality scores of videos. A significant positive Spearman's correlation uniquely between the Cogisen's metric scores and DMOS scores highlights that Sencogi-STSM is the only metric that was able to show an increase of prediction associated with an increase of subjective DMOS in a statistically relevant way, compared to PNSR and SSIM performance. Both Sencogi-STSM and SSIM were able to predict estimates of the subjective scores with a minimum average error, but Sencogi-STSM had a prediction accuracy that was significantly better than both SSIM and PSNR. The improvements in prediction performance found with Sencogi-STSM over the classic SSIM and PSNR metrics, is likely because the method is weighted on perceptual quality, so that the most salient parts of each video-frame affect the VQA metric more than the less salient ones.

6 Conclusion

A new Video Quality Assessment (VQA) metric was developed, called Sencogi Spatio-Temporal Saliency Metric (Sencogi-STSM). Sencogi-STSM is based on a spatio-temporal saliency model that is able to better predict subjective perception scores of video compared to the most used objective VQA metrics, because it uses a saliency model of human visual perception. Sencogi-STSM combines noise detection, the saliency within a video-frame, the saliency of the motion between video-frames, and the delta-quality saliency indicating where a quality change man be noticed by a human viewer. We have assessed the performance of Sencogi-STSM at predicting subjective scores, and compared that performance with the most used VQA metrics, i.e. PSNR and

SSIM. We found that Sencogi-STSM more accurately predicts subjective scores than the most used objective VQA models. The difference between Sencogi-STSM and the most used VQA models (such as PSNR and SSIM) is that Sencogi-STSM uses salience to decide how important each part of a frame is, in terms of quality perception. Future works could be focused on improving the saliency model by combining bottom-up spatio-temporal saliency to top-down saliency, accordingly to task-centred and contextual factors.

References

1. Pedram, M., Abbas, E.-M., Shahram, S.: Subjective and objective quality assessment of image: a survey. CoRR abs/1406.7799 (2014)
2. BT.500: Methodology for the subjective assessment of the quality of television pictures (n.d.). https://www.itu.int/rec/R-REC-BT.500-7-199510-S/en. Accessed 30 Jan 2017
3. Chikkerur, S., Sundaram, V., Reisslein, M., Karam, L.J.: Objective video quality assessment methods: a classification, review, and performance comparison. IEEE Trans. Broadcast. **57**, 165–182 (2011)
4. Brunnstrom, K., Hands, D., Speranza, F., Webster, A.: VQeg validation and ITU standardization of objective perceptual video quality metrics [Standards in a Nutshell]. IEEE Sign. Process. Mag. **26**, 96–101 (2009)
5. Zhang, X., Zhang, X., Silverstein, D.A., Farrell, J.E., Wandell, B.A.: Color image quality metric S-CIELAB and its application on halftone texture visibility. In: Proceedings of IEEE COMPCON 1997. Digest of Papers (n.d.). doi:10.1109/cmpcon.1997.584669
6. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**, 600–612 (2004)
7. Wang, Z., Simoncelli, E.P., Bovik, A.C.: Multiscale structural similarity for image quality assessment. In: The Thrity-Seventh Asilomar Conference on Signals, Systems and Computers (2003). doi:10.1109/acssc.2003.1292216
8. Duchowski, A.: Eye Tracking Methodology, Theory and Practice, vol. 373. Springer Science & Business Media, New York (2007)
9. Wang, Z., Bovik, A.C.: Modern image quality assessment. In: Synthesis Lectures on Image, Video, and Multimedia Processing, vol. 2, pp. 1–156 (2006)
10. Zhang, L., Shen, Y., Li, H.: VSI: a visual saliency-induced index for perceptual image quality assessment. IEEE Trans. Image Process. **23**, 4270–4281 (2014)
11. Larson, E.C., Chandler, D.M.: Unveiling relationships between regions of interest and image fidelity metrics. In: Visual Communications and Image Processing 2008 (2008). doi: 10.1117/12.769248
12. Zhang, W., Borji, A., Wang, Z., Le Callet, P., Liu, H.: The application of visual saliency models in objective image quality assessment: a statistical evaluation. IEEE Trans. Neural Netw. Learn. Syst. **27**, 1266–1278 (2016)
13. Achanta, R., Estrada, F., Wils, P., Süsstrunk, S.: Salient region detection and segmentation. In: Gasteratos, A., Vincze, M., Tsotsos, John K. (eds.) ICVS 2008. LNCS, vol. 5008, pp. 66–75. Springer, Heidelberg (2008). doi:10.1007/978-3-540-79547-6_7

14. Hou, X., Zhang, L.: Saliency detection: a spectral residual approach. In: 2007 IEEE Conference on Computer Vision and Pattern Recognition (2007). doi:10.1109/cvpr.2007.383267
15. Harel, J., Koch, C., Perona, P.: Graph-based visual saliency. In: Advances in Neural Information Processing Systems, pp. 545–552 (2007)
16. Itti, L., Koch, C.: A saliency-based search mechanism for overt and covert shifts of visual attention. Vision Res. **40**, 1489–1506 (2000)
17. Li, J., Levine, M., An, X., He, H.: Saliency detection based on frequency and spatial domain analyses. In: Proceedings of the British Machine Vision Conference 2011 (2011). doi:10.5244/c.25.86
18. Ma, Y.-F., Zhang, H.-J.: Contrast-based image attention analysis by using fuzzy growing. In: Proceedings of the Eleventh ACM International Conference on Multimedia, MULTIMEDIA 2003 (2003). doi:10.1145/957092.957094
19. Achanta, R., Hemami, S., Estrada, F., Susstrunk, S.: Frequency-tuned salient region detection. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition (2009)
20. Wolfe, J.M.: Visual search in continuous, naturalistic stimuli. Vision Res. **34**, 1187–1195 (1994)
21. Mishra, A.K., Aloimonos, Y., Cheong, L.-F., Kassim, A.A.: Active visual segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **34**, 639–653 (2012)
22. Judd, T., Ehinger, K., Durand, F., Torralba, A.: Learning to predict where humans look. In: 2009 IEEE 12th International Conference on Computer Vision (2009). doi:10.1109/iccv.2009.5459462
23. Yubing, T., Cheikh, F.A., Guraya, F.F.E., Konik, H., Trémeau, A.: A spatiotemporal saliency model for video surveillance. Cogn. Comput. **3**, 241–263 (2011)
24. Muddamsetty, S.M., Sidibe, D., Tremeau, A., Meriaudeau, F.: Spatio-temporal saliency detection in dynamic scenes using local binary patterns. In: 2014 22nd International Conference on Pattern Recognition (2014). doi:10.1109/icpr.2014.408
25. Bruce, N.D.B.: Saliency based on information maximization. In: Advances in Neural Information Processing Systems, p. 155 (2006)
26. Mahadevan, V., Vasconcelos, N.: Spatiotemporal saliency in dynamic scenes. IEEE Trans. Pattern Anal. Mach. Intell. **32**, 171–177 (2010)
27. He, X., Gao, X., Zhang, Y., Zhou, Z.-H., Liu, Z.-Y., Fu, B., Hu, F., Zhang, Z.: Intelligence Science and Big Data Engineering. In: 5th International Conference on Image and Video Data Engineering, ISciDE 2015, Suzhou, China, 14–16 June 2015, Revised Selected Papers. Springer, Heidelberg (2015)
28. Bian, P., Zhang, L.: Biological plausibility of spectral domain approach for spatiotemporal visual saliency. In: Köppen, M., Kasabov, N., Coghill, G. (eds.) ICONIP 2008. LNCS, vol. 5506, pp. 251–258. Springer, Heidelberg (2009). doi:10.1007/978-3-642-02490-0_31
29. Cisco: Cisco VNI mobile forecast (2015–2020). Cisco (2016). http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html. Accessed 31 Jan 2017
30. Redi, J., Liu, H., Zunino, R., Heynderickx, I.: Interactions of visual attention and quality perception. In: Human Vision and Electronic Imaging XVI (2011). doi:10.1117/12.876712
31. FFMPEG: https://trac.ffmpeg.org/wiki/Encode/H.264. Accessed 13 Feb 2017
32. Mele, M.L., Millar, D., Rijnders, C.E.: The web-based subjective quality assessment of an adaptive image compression plug-in. In: 1st International Conference on Human Computer Interaction Theory and Applications, HUCAPP, Porto, Portugal (2017)

33. Mele, M.L., Millar, D., Rijnders, C.E.: Validating a quality perception model for image compression: the subjective evaluation of the Cogisen's image compression plug-in. In: Kurosu, M. (ed.) HCI 2016. LNCS, vol. 9731, pp. 350–359. Springer, Cham (2016). doi: 10.1007/978-3-319-39510-4_33
34. ITU-T Study Group 12: Contribution COM 12. Evaluation of new methods for objective testing of video quality: objective test plan 1998