# ANALYSIS OF VIDEO QUALITY INDUCED SPATIO-TEMPORAL SALIENCY SHIFTS

*Xinbo Wu[1], Zhengyan Dong[1], Fan Zhang[2], Paul L. Rosin[1], Hantao Liu[1]*

[1]School of Computer Science and Informatics, Cardiff University, CF24 4AX
[2]Department of Electrical and Electronic Engineering, University of Bristol, BS8 1UB

## ABSTRACT

Human viewers' eye movements reflect their perceptual responses to visual signals. Previous research has shown that distortions in videos cause spatio-temporal gaze shifts, which means gaze behaviour is related to video quality perception. It would be highly beneficial to understand gaze behaviour of viewing videos of varying perceived quality. However, little is known about the interactions between gaze, video content and distortions. In this paper, based on our eye-tracking database for video quality (SVQ160), we perform systematic analyses to reveal the impact of video content (VC) and time order (TO) on gaze shifts. Findings and quantitative methods for gaze behaviour can be used to develop advanced video quality metrics and video processing algorithms.

***Index Terms—*** Video quality, gaze, saliency, spatio-temporal, eye-tracking

## 1. INTRODUCTION

Videos have become one of the primary medium forms of information communication in our daily lives. However, video quality could be inconsistent due to varied compression and transmission conditions; and visible distortions could significantly affect viewers' experience. It is critical to develop reliable video quality assessment (VQA) methods, which form the backbone of advanced video technologies.

The assessment of image quality is a well-studied area that allows for accurate objective measure of the overall quality of a still image [1, 2]. To objectively assess video quality, existing image quality metrics can be directly applied on individual frames in a video sequence, and a sequence level quality index can be obtained based on simple temporal pooling algorithms [3, 4]. To improve the performance of VQA metrics, more advanced methods have also been developed [5–8], which exploit temporal information within videos to enhance the perceptual relevance of VQA metrics. These approaches tend to offer more robust performance when predicting video quality. A significant current trend in VQA research is to incorporate visual attention, which is an essential aspect of the human visual system (HVS) [9–11].

Human viewers' gaze behaviour when watching videos reflects their perception processes and interpretation of the visual content [12–14]. Attempts have been made in the VQA literature to weight local distortions by existing saliency models [9, 10]. The limitation of these studies is that the understanding of how visual attention plays a role in video quality assessment, especially the interactions between saliency, original content and distortions, is far from complete. To develop perceptually meaningful saliency methods for VQA, it is critical to better understand and characterise gaze behaviour when observers view videos of varying quality and diverse content.

In a previous eye-tracking study [11], a large-scale and reliable eye-tracking database for video quality, namely SVQ160 was created. The study revealed that there is a significant difference in saliency between natural scene (i.e., original and pristine video content) and distorted scene (i.e., video content with visible distortions); and that the distortion/quality induced saliency shifts (QSS) significantly contribute towards the video quality assessment behaviour. In this paper, we conduct systematic analyses on the QSS in terms of the impact of video content (VC) and time order (TO). Building on the characteristics of gaze behaviour and quantitative methods, we make recommendations for future development of saliency methods in VQA metrics and video compression algorithms.

## 2. METHODOLOGY AND DEFINITION

### 2.1. SVQ160 Database

The SVQ160 database [11] represents a large-scale and reliable eye-tracking study that involved 160 human observers and 160 video stimuli degraded with different distortion types at various quality levels. The video stimuli were taken from the LIVE video quality database [15], consisting of 10 uncompressed high quality reference videos and 150 distorted videos (i.e., 15 distorted videos per reference, and distortion types include Wireless, IP, H.264 and MPEG-2). The stimuli cover a diverse range of video content, namely 'bs-Blue Sky', 'mc-Mobile Calendar', 'pa-Pedestrian Area', 'pr-Park Run', 'rb-Riverbed', 'rh-Rush Hour', 'sf-Sunflower', 'sh-Shields', 'st-Station' and 'tr-Tractor', as shown in Fig.1. The videos are about 10 seconds long and have a resolutions of $768 \times 432$ pixels. The eye-tracking study included rigorously designed control mechanisms to eliminate experimental biases and en-

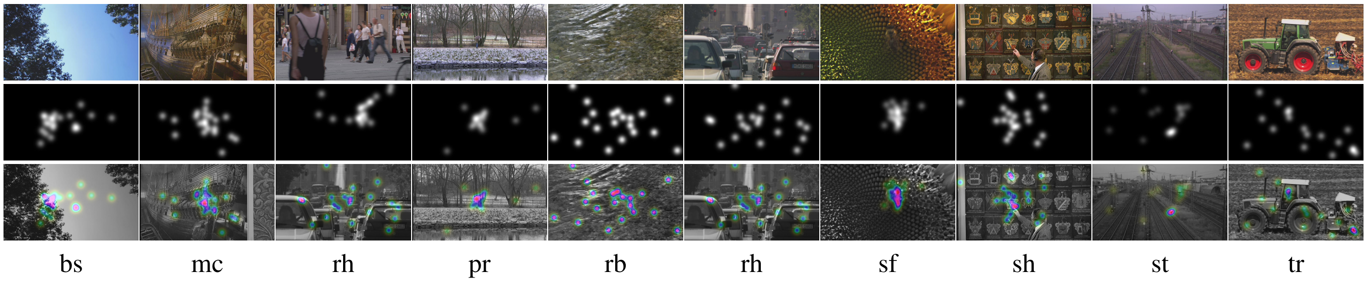


**Fig. 1**. SVQ160 database: first row illustrates content (representative frames) of the original videos, second row shows saliency maps; and third row shows the corresponding heatmaps (saliency maps superimposed on representative frames).

sure reliability of eye-tracking data. Each of the 160 videos received eye-tracking data of 20 viewers.

### 2.2. Definition of quality induced saliency shifts (QSS)

Now, we define a key variable in our study - quality induced saliency shifts (QSS) - as a measure to quantify the difference in saliency between the original and distorted videos. First, a frame-level saliency map is created from fixations obtained from all subjects in the eye-tracking study. The saliency map stimulated the foveal vision of the HVS [11, 16] with the approximate size of the fovea being 2° visual angle. The frame-level saliency map (FSM) is then calculated as:

$$\mathrm{FSM}_{(x,y)} = \sum_{i=1}^{N} \exp\left[-\frac{(x_i - x)^2 + (y_i - y)^2}{\sigma^2}\right] \qquad (1)$$

where $(x_i,y_i)$ stands for the position of the $i$-th fixation point, $N$ is the total number of fixations. Examples of the frame-level saliency maps are shown in Fig.1. Then, the saliency similarity between the original frame and distorted frame is calculated using the Pearson linear correlation coefficient (CC). Note, CC has been proven to be the most appropriate perception-based saliency evaluation metric [17], and is defined as:

$$\mathrm{CC}_{(\mathrm{FSM}\_ref,\mathrm{FSM}\_dis)} = \frac{\mathrm{cov}(\mathrm{FSM}\_ref, \mathrm{FSM}\_dis)}{\sigma_{\mathrm{FSM}\_ref} \times \sigma_{\mathrm{FSM}\_dis}} \qquad (2)$$

where $\sigma_{\mathrm{SM}\_ref}$ and $\sigma_{\mathrm{SM}\_dis}$ denote the standard deviation of $\mathrm{SM}\_ref$ and $\mathrm{SM}\_dis$ respectively, $cov(\mathrm{FSM}\_ref, \mathrm{FSM}\_dis)$ represents the covariance. The value of CC ranges between -1 and 1. The closer of CC to -1 or 1, the higher the similarity between saliency maps; and the closer of CC to 0, the less similarity exists between the two saliency maps. Finally, the quality induced saliency shifts (QSS) can be defined based on equation (2) for each video, i.e., the statistics of frame-based CC over time characterises the spatio-temporal QSS.

### 2.3. Saliency dispersion measure

The saliency dispersion measure [18] provides an algorithm to quantify the degree of saliency dispersion in the spatial domain. The multilevel entropy (ME) of a saliency map (S) is calculated based on Shannon entropy applied to $p \times p$ non-overlapping blocks of the saliency map:

$$\mathrm{ME} = H_{\Sigma}(S) = \frac{1}{P_{\max}} \sum_{P=1}^{P_{\max}} \sum_{B=1}^{N_{\max}} H(B) \qquad (3)$$

where $H$ represents the entropy of a 2-D image block, $P_{max}$ refers to the segmentation level (i.e., $P_{max} = 4$ is empirically determined in [18] and also used here), $N_{max} = P_{max}^2$, and $B$ runs over each block. The lower the multilevel entropy, the more the saliency is concentrated in fewer areas in the spatial domain; otherwise, the higher the entropy, the more the saliency is dispersed throughout the spatial domain.

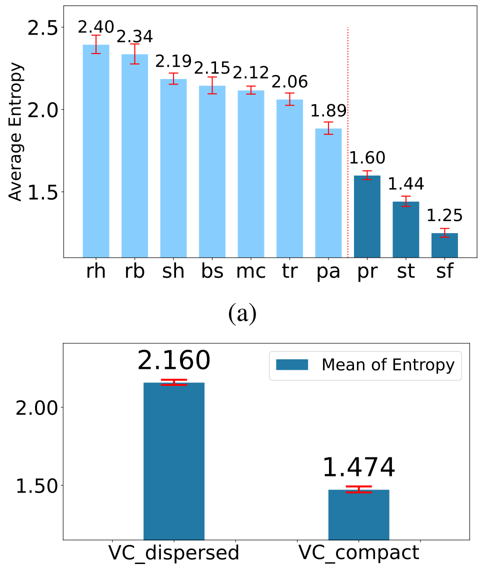
## 3. STATISTICAL ANALYSIS AND RESULTS

### 3.1. Impact of visual content on QSS

*Hypothesis: We hypothesize that the impact of video content (VC) on the quality induced saliency shifts (QSS) is statistically significant.*

We first define the video content (VC) variable as a classification of the saliency dispersion degree of the original content. For each original video, we calculate the sequence-level ME by taking the average of frame-level ME values. Fig.2(a) shows the saliency dispersion degrees for all original videos. Based on observations, we could classify the videos into two groups, i.e., VC_dispersed (including 'rh' to 'pa') represents the dispersed saliency and VC_compact (including 'pr' to 'sf') represents the concentrated saliency. To verify the VC grouping is statistical meaningful, we perform hypothesis testing selecting ME as the dependent variable and the categorical VC group as the independent variable. The Mann-Whitney U test [19] is performed (due to evidence of non-normality as per the Shapiro-Wilk test [19]), and the results ($P < 0.05$) show that the ME of VC_dispersed is statistically significantly higher than that of VC_compact, as shown in Fig.2(b).

Now, for the two distinctive VC classes (i.e., VC_dispersed and VC_compact), we analyse the impact of VC on quality induced saliency shifts (QSS) in terms of the spatio-temporal



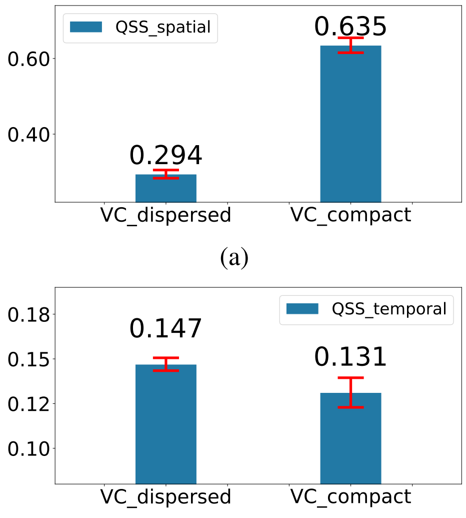
(a)

(b)

**Fig. 2**. (a) Saliency dispersion degrees (measured by multi-level entropy) for all original videos contained in the SVQ160 datatabase. (b) Difference (in saliency dispersion) between two instinctive visual content (VC) classes VC_dispersed and VC_compact. Error bars indicate the 95% confidence interval.

statistics. Note, each original video is associated with 15 distorted videos in the SVQ160 database. We formulate two target variables here: for each distorted video, based on the frame-level CC calculated by equation2, we calculate the mean and standard deviation of CC values over time, which are referred to as QSS_spatial and QSS_temporal, respectively.

The first hypothesis testing is conducted selecting QSS_spatial as the dependent variable, and the categorical VC group as the independent variable. The Mann-Whitney U test is performed (due to evidence of non-normality as per the Shapiro-Wilk test), and the results ($P < 0.05$) show that the QSS_spatial of VC_dispersed (dispersed saliency) is statistically significantly lower than that of VC_compact (concentrated saliency), as shown in Fig.3(a). This suggests that when watching the distorted videos of VC_dispersed there are significant gaze shifts relative to the original video content (i.e., $CC = 0.294$ as shown in Fig.3(a)). The evidence here has implications for the VQA metrics that contain saliency prediction component, note the saliency predicted from the original videos cannot reflect the saliency of the distorted videos. However, for the videos of VC_compact, viewers' gaze is less effected by the distortions (i.e., $CC = 0.635$ as shown in Fig.3(a)). For the VQA metrics or video compression algorithms, the distortions occurring in the non-salient areas could be less penalized for overall quality.



(a)

(b)

**Fig. 3**. Difference (in spatial saliency shifts (a) and temporal saliency shifts (b)) between two instinctive visual content (VC) classes VC_dispersed and VC_compact. Error bars indicate the 95% confidence interval.

The second hypothesis testing is conducted selecting QSS_temporal as the dependent variable, and the categorical VC group as the independent variable. The Mann-Whitney U test is performed (due to evidence of non-normality as per the Shapiro-Wilk test), and the results ($P < 0.05$) show that the QSS_temporal of VC_dispersed (dispersed saliency) is statistically significantly higher than that of VC_compact (concentrated saliency), as shown in Fig.3(b). This indicates that the quality induced saliency shifts are more consistent over time for the videos of VC_compact than VC_dispersed, meaning how viewers' gaze is affected by distortions is strongly time (temporal domain) dependent. Also, such temporal variations pose challenges for accurately predicting saliency of distorted videos of VC_dispersed.

### 3.2. Impact of time order on QSS

*Hypothesis: We hypothesize that the impact of time order (TO) on the quality induced saliency shifts (QSS) is statistically significant.*

In order to illustrate the variations of gaze behaviour in time order (TO), we divide a video into 10 successive blocks of time (i.e., each time bock represents one second of video playback). For each time block, the mean of frame-level CC values (see equation (2)) over all distorted videos (i.e., 150) contained in the SVQ160 database is calculated to characterise the quality induced saliency shifts (QSS). Fig.4(a) shows the QSS in time order. We could use a polynomial fit to approximate the gaze behaviour as shown in Fig.4(a).

Based on observations of Fig.4(a), we formulate the time order into three semantic categories, including TO_beginning

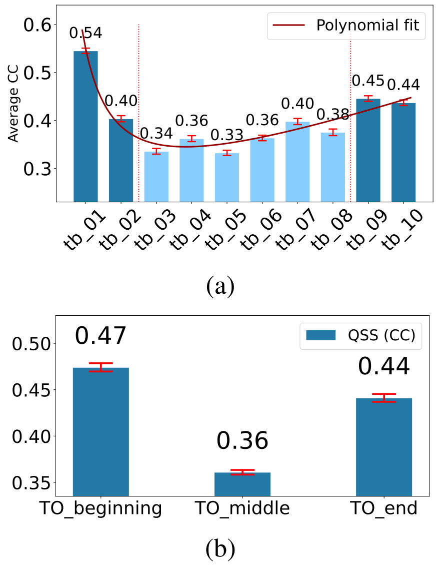


**Fig. 4**. (a) QSS (measured by frame-level CC over all distorted videos contained in the SVQ160 database) in time order. (b) Difference (in QSS measured by CC) between three semantic time order (TO) categories: TO_beginning, To_middle and To_end. Error bars indicate the 95% confidence interval.

(time blocks 1-2), TO_middle (time blocks 3-8), and TO_end (time blocks 9-10). Hypothesis testing is performed selecting CC as the dependent variable, and the categorical TO group as the independent variable. The Mann-Whitney U test is performed (due to evidence of non-normality as per the Shapiro-Wilk test) for each comparison, including TO_beginning versus TO_middle, TO_middle versus TO_end, and TO_beginning versus TO_end. The results ($P < 0.05$) show that the difference of each comparison is statistically significant, as shown in Fig.4(b). Overall, this indicates that in the beginning of video playback, viewers' gaze is less affected by distortions than the rest of viewing time. A plausible reason is center-bias, which is the tendency of observers to preferentially look towards the centre of images (image center-bias [20]), or their first fixations tend to be near the centre of an object (object center-bias [21]). A viewing strategy was observed in previous studies that viewers tend to look at locations closer to the centre immediately after the beginning of the scene [22, 23]. In the middle of viewing, there are significant saliency shifts due to the occurrence of distortions, meaning viewers might be most sensitive to distortions during these times. The impact of distortions on gaze behaviour significantly decreases towards the end of viewing (i.e., a significant increase of CC as shown in Fig.4(b) from TO_middle to TO_end). This could be attributed to the fact that viewers would have learned to tolerate the distortions to some extent, which leads to the changes in gaze behaviour from middle to the end of viewing.

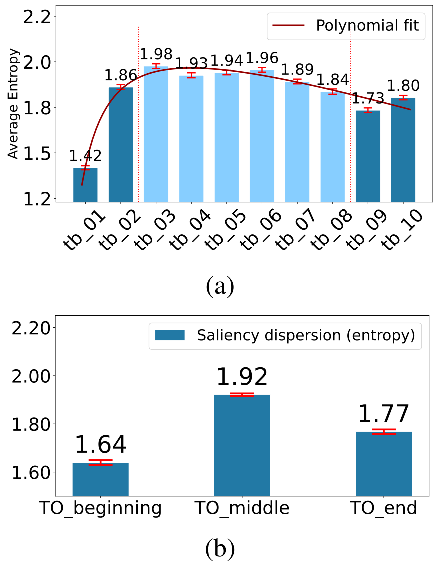


**Fig. 5**. (a) Saliency dispersion (measured by entropy over all videos contained in the SVQ160 database) in time order. (b) Difference (in saliency dispersion) between three semantic time order (TO) categories: TO_beginning, To_middle and To_end. Error bars indicate the 95% confidence interval.

Fig.5(a) illustrates the average entropy (see equation (3)) in time order (i.e., 10 time blocks) over all original and distorted videos (i.e., 160) contained in the SVQ160 database. The Mann-Whitney U test is performed (due to evidence of non-normality as per the Shapiro-Wilk test) to analyse the impact of different TO groups (i.e., TO_beginning, TO_middle, and TO_end) on the saliency dispersion measure (i.e., entropy). The results ($P < 0.05$) show that the difference of each comparison between TO groups is statistically significant, as shown in Fig.5(b). This means viewers' gaze tends to be more dispersed in the middle of viewing than the beginning/end of viewing. This reflects the fluctuation of distortion induced gaze shifts in time oder, which is consistent with the findings in Fig.4. The time order effects on distortion perception could be potentially used (e.g., as a time order-aware weighting function) to improve VQA metrics or video compression algorithms.

## 4. CONCLUSION

In this paper, we have investigated quality induced saliency shifts (QSS) - a highly relevant attribute of video quality. Our statistical analyses on the large-scale eye-tracking database (SVQ160) reveal that video content classification and time order have significant impacts on the spatio-temporal characteristics of QSS. Findings can be used to facilitate the development of advanced algorithms for video quality assessment and video coding.

# References

[1] Z. Wang and A. C. Bovik, “Modern image quality assessment,” in *Modern Image Quality Assessment*, 2006.

[2] X. Yang, F. Li, and H. Liu, “A survey of dnn methods for blind image quality assessment,” *IEEE Access*, vol. 7, pp. 123 788–123 806, 2019.

[3] S. Chikkerur, V. Sundaram, M. Reisslein, and L. J. Karam, “Objective video quality assessment methods: A classification, review, and performance comparison,” *IEEE Transactions on Broadcasting*, vol. 57, no. 2, pp. 165–182, 2011.

[4] Z. Wang, H. Sheikh, and A. Bovik, “Objective video quality assessment,” in *The Handbook of Video Databases: Design and Applications*. CRC Press, 2003, pp. 1041–1078.

[5] M. Barkowsky, J. Bialkowski, B. Eskofier, R. Bitto, and A. Kaup, “Temporal trajectory aware video quality measure,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 3, no. 2, pp. 266–279, 2009.

[6] M. H. Pinson and S. Wolf, “A new standardized method for objectively measuring video quality,” *IEEE Transactions on broadcasting*, vol. 50, no. 3, pp. 312–322, 2004.

[7] K. Seshadrinathan and A. C. Bovik, “Motion tuned spatio-temporal quality assessment of natural videos,” *IEEE transactions on image processing*, vol. 19, no. 2, pp. 335–350, 2009.

[8] P. V. Vu, C. T. Vu, and D. M. Chandler, “A spatiotemporal most-apparent-distortion model for video quality assessment,” in *2011 18th IEEE International Conference on Image Processing*. IEEE, 2011, pp. 2505–2508.

[9] X. Feng, T. Liu, D. Yang, and Y. Wang, “Saliency inspired full-reference quality metrics for packet-loss-impaired video,” *IEEE Transactions on Broadcasting*, vol. 57, no. 1, pp. 81–88, 2011.

[10] D. Ćulibrk, M. Mirković, V. Zlokolica, M. Pokrić, V. Crnojević, and D. Kukolj, “Salient motion features for video quality assessment,” *IEEE Transactions on Image Processing*, vol. 20, no. 4, pp. 948–958, 2011.

[11] W. Zhang and H. Liu, “Study of saliency in objective video quality assessment,” *IEEE Transactions on Image Processing*, vol. 26, no. 3, pp. 1275–1288, 2017.

[12] F. Dobrian, V. Sekar, A. Awan *et al.*, “Understanding the impact of video quality on user engagement,” *ACM SIGCOMM computer communication review*, vol. 41, no. 4, pp. 362–373, 2011.

[13] J. Wang, P. Antonenko, and K. Dawson, “Does visual attention to the instructor in online video affect learning and learner perceptions? an eye-tracking analysis,” *Computers & Education*, vol. 146, p. 103779, 2020.

[14] X.-b. Zhang, C.-T. Fan, S.-M. Yuan, and Z.-Y. Peng, “An advertisement video analysis system based on eye-tracking,” in *2015 IEEE International Conference on Smart City/SocialCom/SustainCom (SmartCity)*, 2015, pp. 494–499.

[15] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, “Study of subjective and objective quality assessment of video,” *IEEE transactions on Image Processing*, vol. 19, no. 6, pp. 1427–1441, 2010.

[16] O. Le Meur, A. Ninassi, P. Le Callet, and D. Barba, “Overt visual attention for free-viewing and quality assessment tasks: Impact of the regions of interest on a video quality metric,” *Signal Processing: Image Communication*, vol. 25, no. 7, pp. 547–558, 2010.

[17] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand, “What do different evaluation metrics tell us about saliency models?” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 3, pp. 740–757, 2019.

[18] W. Zhang, R. R. Martin, and H. Liu, “A saliency dispersion measure for improving saliency-based image quality metrics,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 6, pp. 1462–1466, 2018.

[19] A. Field, *Discovering statistics using IBM SPSS statistics*. SAGE Publications, 2013.

[20] B. W. Tatler, “The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions,” *Journal of Vision*, vol. 7, no. 14, pp. 4–4, 11 2007.

[21] J. M. Henderson, “Eye movement control during visual object processing: effects of initial fixation position and semantic constraint.” *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, vol. 47, no. 1, p. 79, 1993.

[22] P.-H. Tseng, R. Carmi, I. G. M. Cameron, D. P. Munoz, and L. Itti, “Quantifying center bias of observers in free viewing of dynamic natural scenes,” *Journal of Vision*, vol. 9, no. 7, pp. 4–4, 07 2009.

[23] L. O. M. Rothkegel, H. A. Trukenbrod, H. H. Schütt, F. A. Wichmann, and R. Engbert, “Temporal evolution of the central fixation bias in scene viewing,” *Journal of Vision*, vol. 17, no. 13, pp. 3–3, 11 2017.