Examining the effect of task on viewing behavior in videos using saliency maps

Hani Alers[a], Judith A. Redi[a], Ingrid Heynderickx[a,b]
[a] Delft University of Technology, Mekelweg 4, Delft, The Netherlands 2628 CD;
[b] Philips Research Laboratories, Prof. Holstlaan 4, Eindhoven, The Netherlands 5656 AA

ABSTRACT

Research has shown that when viewing still images, people will look at these images in a different manner if instructed to evaluate their quality. They will tend to focus less on the main features of the image and, instead, scan the entire image area looking for clues for its level of quality. It is questionable, however, whether this finding can be extended to engulf videos considering their dynamic nature. One can argue that when watching a video the viewer will always focus on the dynamically changing features of the video regardless of the given task. To test whether this is true, an experiment was conducted where half the participants viewed videos with the task of quality evaluation while the other half were simply told to watch the videos as if they were watching a movie on TV or a video downloaded from the internet. The videos contained content which was degraded with compression artifacts across a wide range of quality. An eye tracking device was used to record the viewing behavior in both conditions. By comparing the behavior during each task, it was possible to observe a systematic difference in the viewing behavior which seemed to correlated to the video quality.

Keywords: Video quality, visual attention, saliency, eye tracking, viewing behavior

1. INTRODUCTION

Researchers have been studying visual attention deployment for many decades now [1,2]. This knowledge has been shown to be useful in a number of applications (e.g. [3,4]), especially when extracting image saliency information [5] through attention prediction models (e.g. [6,7,8]). One example is its implication in visual quality perception, which has been largely studied in images [9,10,11]. However, when it comes to videos there have been few efforts in trying to understand the relation between task, quality, and viewing behavior [12].

This research expands on earlier work performed on still images, focused on the effect that a task given to the observers can have on their viewing behavior [13]. In that work it was shown that the task does impact several spatial and temporal characteristics of the viewing behavior. On the other hand, similar work conducted on video material [12] has suggested that the viewing behavior is not affected by the given viewing task. Therefore, here we extend that study [13] to video material and focus on the task of scoring the quality of videos.

By following a similar methodology as the one used on still images, it is interesting to see whether the results are duplicated or whether the viewing task indeed has no effect on the viewing behavior. Furthermore, this study also looks at the interaction effects between the video encoding quality and the viewing behavior. The goal is to better understand how humans watch videos for entertainment. This information can then be used to optimize the video encoding algorithms to produce the best viewing experience while consuming less resources. For analyzing the viewing behavior, the methodology used here again extends on previous work done on still images [13] and, therefore, employ the use of saliency maps. The paper explores different algorithms for studying saliency maps [14] in an effort to determine which are more suitable for use in videos.

We designed a psychometric experiment to investigate the effects of task and visual quality on humans viewing behavior. We recorded the eye movements of a panel of observers while they were watching a set of distorted videos. These videos were degraded with compression artifacts across a wide range of video qualities. Half of the viewers was instructed to evaluate the visual quality of each video. The other half was asked to watch the videos (as much as possible) as if they were freely watching them in a home setting. In the following we analyze the resulting eye-tracking data across different

tasks and visual quality levels. To do so, we first convert them into saliency information, producing saliency maps averaged across all participants for each video and under each viewing condition. From (dis)similarities in saliency, we are able to detect analogies and differences in viewing behavior depending on task and visual quality level. The experimental methodology and the protocol are discussed in Section 2. The methodology followed to analyze the data is described in Section 3. Section 4 reports on the analysis of the eye-tracking data, which is then discussed in Section 5. Conclusions and future research are prospected in Section 6.

2. METHODOLOGY

2.1 Stimuli

A video database was created which consisted of 25 video segments with a duration of 20 seconds each. Since the main purpose of the experiment is to detect if there was a difference in viewing behavior, we wanted to use stimuli that have clearly identifiable natural saliency regions in order to make it easier to detect any differences that may result from changing the viewing task. It was assumed that highly dynamic scenes can seize the focus of the viewers under natural viewing conditions. Therefore, the video segments were extracted from action based movies (some sample frames are shown in Figure 1). From each video, two distorted versions were produced using an H.264 video encoder. These two versions were degraded to two different levels of quality. The x264 encoder was used as provided by the ffmpegX software [16]. The resulting videos had a resolution of 1280x720 pixels, and a frame rate of 25 frames per second. The coding parameters were not uniform across the videos, to allow a variety of quality levels to be judged by the observers. Eventually, the database included 50 distorted videos spanning a wide range of quality.
Two collections of stimuli were generated (collections I and II), each including only one of the two distorted versions generated for each video. The assignment of videos to collections was made randomly, therefore the videos in each collection spanned roughly the same range of quality. To give an idea of the wealth of this quality range, Figure 1 shows a sample frame on the left of one of the videos encoded with the highest bitrate (1237 bit/s), and another on the right which is among those encoded with the lowest bitrate (209 bit/s).

Figure 1. Video segments used for the experiment were taken from action movies and were chosen to have highly dynamic sequences. Varying the bitrate used to compress the videos from higher bitrates (left 1237 bit/s) to lower ones (right 209 bit/s) gave a wide range of quality for the generated videos.

2.2 The experiment setup

Given that the focus of the experiment was on the viewing task, a between subjects design was chosen, where half the participants looked freely at the videos and the other half was asked to evaluate their visual quality. This meant that every viewer saw each video segment only once, and ensured that there was no memory effect influencing the viewing behavior. With 2 viewing conditions and 2 viewing tasks, the experiment took the form of a 2x2 design requiring 4 groups of observers, as shown in Table 1. Each group counted 12 participants, for a total of 48. Participants included master, graduate, and postgraduate students of the Electrical Engineering, Mathematics and Computer Science (EEMCS) faculty building at the Delft University of Technology (TU Delft).

		Task	
		Scoring	Free looking
Collection	I	Group 1	Group 3
	II	Group 2	Group 4

Table 1. Between-subjects design of experiment to determine the impact of task on viewing behavior

The videos were displayed using a late 2008 MacBook on a 17-inch CRT monitor with a resolution of 1280x960 pixels. The experiment was controlled from a remote computer with its monitor positioned so that it would not interfere with the participant's task. In order to avoid outside elements interfering with the results, the experiment was carried out in a controlled environment inside the Delft Experience Lab located in the EEMCS faculty building at the TU Delft. Only the experimenter and the viewer were present while performing the experiment. The illumination level was kept at a constant level of 70 lux. Eye movements were recorded binocularly at 250 Hz with a video-based infrared eye tracker (SR-Research, EyeLink-II). The eye-tracker data was saved to disk for off-line analysis. The experiment setup is shown in Figure 2.

Figure 2. Experimental setup. The viewer watched the videos on a CRT monitor while wearing a head mounted eye tracking device. The experimenter ran the experiment from a non-intrusive position.

2.3 The experiment protocol

Of the four groups of participants (Table 1), the first two groups (1 and 2) as well as the last two (3 and 4) went through identical protocols but watching a different collection of videos. In all cases, participants were given a printed description of the experiment and a list of the instructions they needed to follow. They were then seated so that their viewing distance measured 60 [cm] from the display plain. After being fitted with the head mounted eye tracker, the experimenter ran a 9-point calibration for the gaze location.

Groups 1 and 2 watched the first and second collection of videos respectively. The performed a scoring task which used a single stimulus numerical scaling setup [15]. After the calibration, observers were shown 4 training videos (Which were not a part of the collections I and II), which were representative of the entire range of quality used in the video collections. The training videos helped the participants in getting acquainted with the user interface of the experiment and gave them an idea of the range of quality they could use for scoring the videos. For every video to be evaluated, during both the training and actual experiment, the following steps were performed. The participants first saw a drift correction screen, which helped the head mounted eye tracker compensate for any shifts in its position. To do that, they simply had to fixate their gaze at a red dot in the center of the screen and press the space bar. They were then shown a 20 second video segment. This was followed by a scoring window with a continuous slider going from 0 to 10 with the labels 'poor' at the lower end and 'excellent' at the other. Once a score was chosen, the process was repeated with the drift-correction screen followed by another video. After the first four videos, a dialog window was shown indicating that the training session was over and the process then continued with the videos from one of the collections. Participants from each group always saw the same 25 video segments from the assigned collection, but the order in which the videos were shown was randomized in order to avoid any bias (i.e., learning or fatigue effects) in the results.

Participants in groups 3 and 4 were again shown collections I and II respectively. They followed the same protocol described above except that they were not given the scoring window after each video. Of course, the experiment instructions they had were also adjusted so that they were told to only watch the videos as if they were watching TV or a video downloaded from the Internet.

3. ANALYZING THE DATA

The eye tracker collected fixation and saccade information. Smooth pursuit eye behavior exhibited when the viewer followed the movements of objects on the screen was registered by the eye tracker as fixations. In order to perform a precise spatial analysis of attention deployment, we decided to transform fixation data into saliency information, adapting the procedure proposed in [14]. First, per each video sampling point we grouped in a single fixation map all the fixation locations of all observers. This resulted into as many fixation maps as sampling points per each video, giving a too fine granularity for the purposes of our analysis. Thus, videos were divided in coarser slots of 1 second each, and fixation maps were averaged over these time slots, resulting in 20 fixation maps per video. These fixation maps were finally converted to saliency maps to better reflect the characteristics of human vision. In particular, a Gaussian patch with a width σ approximating the size of the fovea (about 2° visual angle) was applied to each fixation. A mean saliency map that takes into account all fixations of all subjects was calculated as follows:

$$S_{t,i}(k,l) = \sum_{j=1}^{T} \exp\left[-\frac{(x_j - k)^2 + (y_j - l)^2}{\sigma^2} \right]$$

where $S_{t,i}(k, l)$ indicates the saliency map for stimulus I_i for the time slot $t \in [1,20]$ of size MxN pixels (i.e. $k \in [1, M]$ and $l \in [1, N]$), (x_j, y_j) indicates the spatial coordinates of the jth fixation ($j=1 \ldots T$), T is the total number of all fixations over all subjects in that time period, and σ indicates the standard deviation of the Gaussian. The intensity of the resulting saliency map is linearly normalized to the range $[0, 1]$. Each of these maps specifies the saliency distribution over a specific time slot of a specific video. In other words, it is a representation of the probability, pixel by pixel, that at a given time slot, for a given video and task, the average observer will fixate on a specific pixel. This process was repeated for each 1 second time slot for each video sequence. Hence, we obtained in total 20 (slots) x 25 (videos) x 2 (collections) x 2 (tasks) = 2000 saliency maps.

4. RESULTS

The scores collected from groups 1 and 2 for each encoded video sequence were processed to calculate one mean-opinion-score (MOS) [17] representing the subjective quality level of that video segment. Graphs shown in Figure 3 illustrate the range of subjective quality used in the videos. The left graph shows the MOS values for the two degraded versions of each video segment included in collections I and II. The graph clearly shows that the videos in both groups were well distributed across the used range of quality. It also conveys that the difference in quality between the two encoded versions of each video segment varies across the generated collections. The graph on the right side of Figure 3

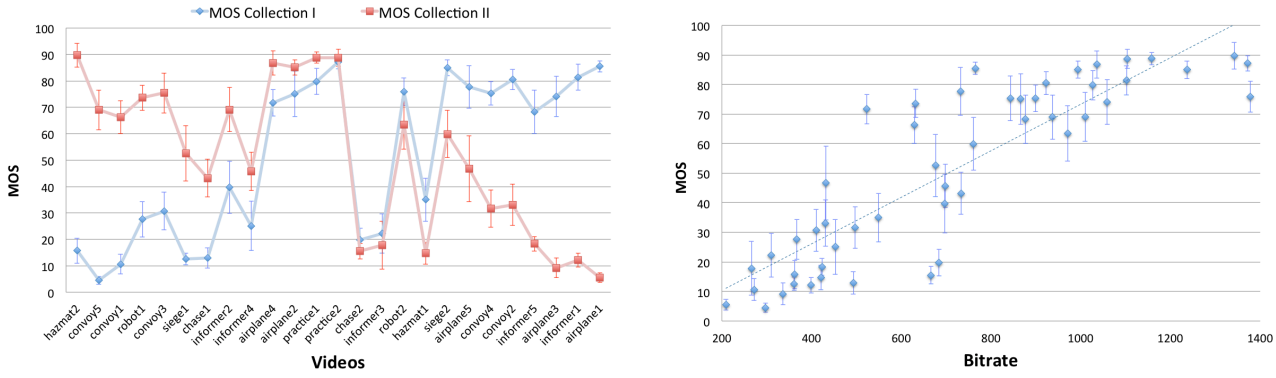

Figure 3. Two graphs illustrating the range of quality in the used videos. On the left are the MOS values for the two versions of videos in each collection sorted by the difference in MOS. On the right the MOS for all 50 videos is plotted against the bitrate used to encode the videos.

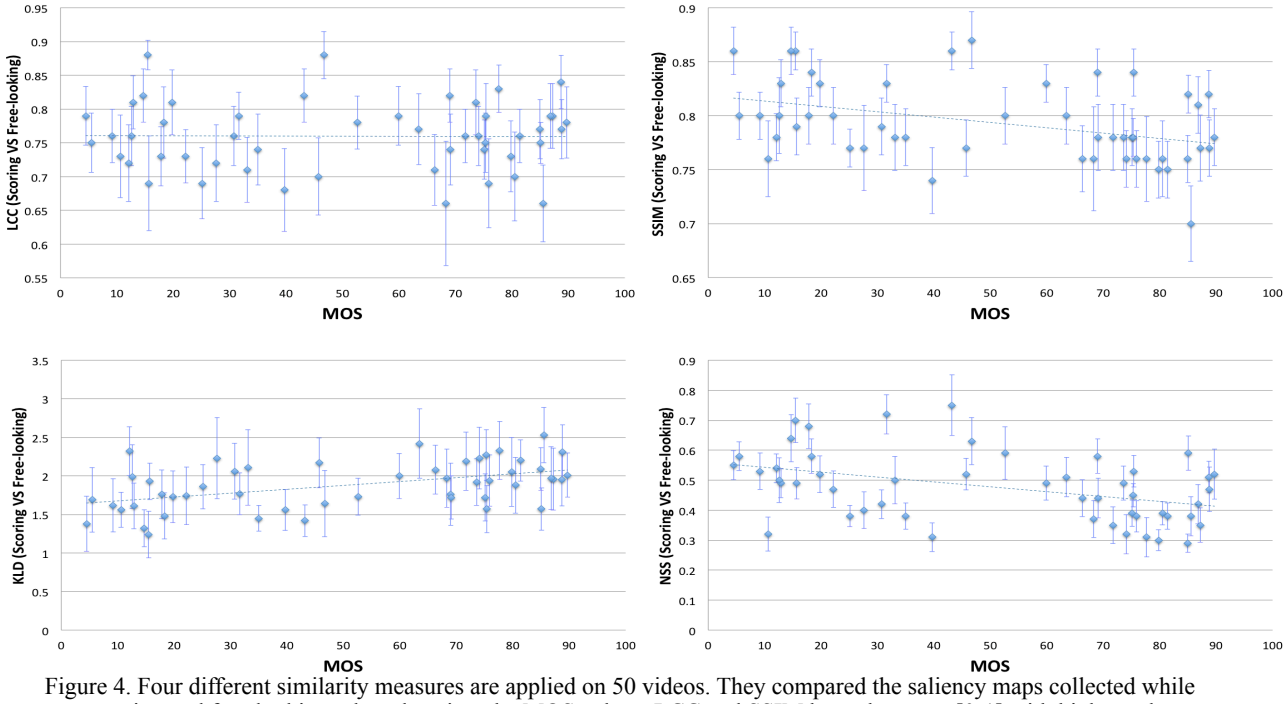

Figure 4. Four different similarity measures are applied on 50 videos. They compared the saliency maps collected while scoring and free-looking, plotted against the MOS values. LCC and SSIM have the range [0-1] with higher values indicating more similarity. With NSS, a value of 0 represents no similarity with higher values representing more similarity. KLD is the opposite starting at 0 for perfect similarity, with higher values meaning less similarity.

plots the MOS as a function of the video bitrate. Despite the wide spread that can be observed in the middle of the scale, a linear relation can be observed between the two values approximated by the stapled line plotted in the graph.

As previously stated, we are interested not only in the impact of task on visual attention but also on that of the quality level. Thus, It is useful for the analysis to sort the collected data into groups depending on the MOS quality levels. We redistributed the videos from the two collections into a High Quality (HQ) and Low Quality (LQ) groups. For each video, the version that received the lower MOS is collected in the LQ group and the one with the higher quality is assigned to the HQ group. By taking the 2 tasks into consideration, we end up with 4 sets of data as shown in Table 2.

		Task	
		Scoring	Free looking
Quality	Higher	S-HQ	F-HQ
	Lower	S-LQ	F-LQ

Table 2. The redistributed data sets used in the analysis of the results

In order to see whether the quality scoring task (the independent variable) affected the viewing behavior, the saliency maps collected under each task are compared to measure the level of similarity among them. Many approaches for analyzing similarities between saliency maps have been proposed in similar research studying attention data in still images [14]. In the sake of being thorough, 4 of the measures proposed in the literature [14] are computed here for the saliency maps in order to find the most appropriate measures capable of highlighting differences in the viewing behavior in videos. These approaches are: Linear Correlation Coefficient (LCC), Kullback-Leibler divergence (KLD), Normalized Scanpath Saliency (NSS), and the Structure Similarity Index (SSIM) [18]. A value of LCC = 1 indicates identical maps, while LCC = 0 indicates uncorrelated maps. This is also the case for SSIM with a range of [0-1] and higher scores indicating more similarity. The NSS will return a value greater than zero if there is a greater correspondence between the two saliency maps than expected by chance. The NSS value of zero would mean there is no such correspondence and a value of less than zero would mean there is anti-correspondence between the saliency maps. Finally, the KLD is a

positive quantity that increases with the dissimilarity of the maps, and KLD = 0 only in the case of identical maps. More details on what each of these method measures and how they are calculated can be found in the literature [14].

First the LCC was computed for each pair of saliency maps corresponding to the same video and time slots for each of the two tasks. In other words, the LCC was calculated for each video for each saliency map in S-HQ and the corresponding saliency map in the F-HQ dataset. The same process is then repeated for the S-LQ vs F-LQ, S-HQ vs S-LQ, and F-HQ vs F-LQ data sets. In this way, we had, per each second and per each encoded video, an indication of the similarity of viewing behavior under different tasks and under different quality levels. This process was then repeated using the other three similarity measures (SSIM, KLD, and NSS). Figure 4 shows the similarity values for the task effect for all 4 similarity measures. Each data point represents the average value over 20 time slots for each of the 50 encoded video. The error bars represent the 95% confidence interval for the 20 time slots for each video segment. The similarity scores are plotted against the subjective MOS quality and a trend line is plotted representing the best linear function fitting the data.

Since the KLD scale is reversed (*higher* values indicate *less* similarity), the trend lines of the SSIM, KLD, and NSS seen in Figure 4 indicate that the viewing behavior becomes less similar as the quality level increases. The trend line for the LCC values is virtually flat and neither support nor oppose this observation. When it comes to explaining the effect of the task however, it may be helpful to have a reference measure to compare the data against.

We next take the free-looking data collected while viewing the high quality segments (F-HQ) as a reference. With no assigned task and the relatively higher level of quality, it represents the closest saliency data to the natural saliency of the original 25 video segments. We use this reference data to examine the similarity of the viewing behavior for the lower

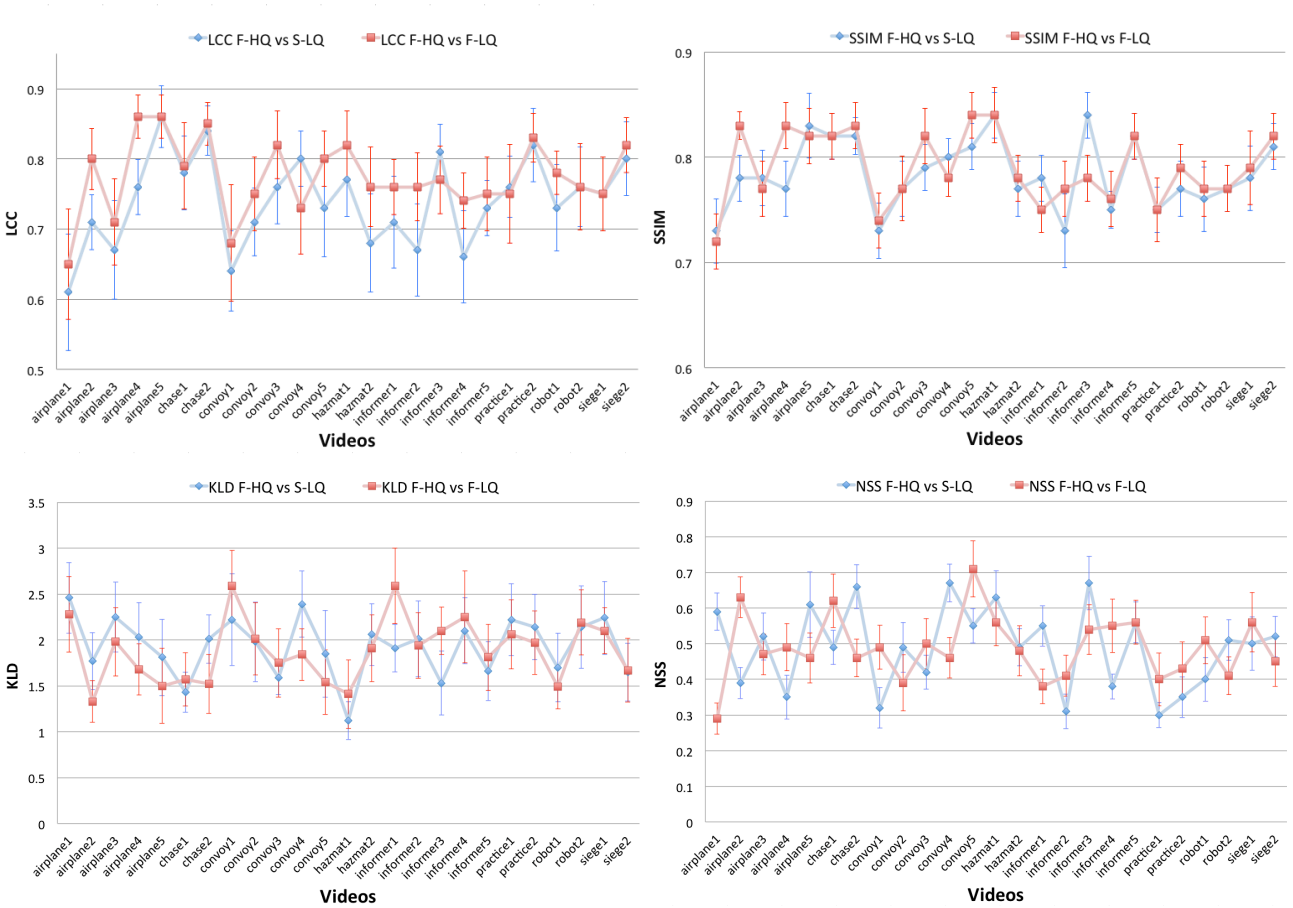

Figure 5. By taking the free looking high quality as a reference. The similarity with the low quality videos is measured for both the free looking and the scoring conditions

quality versions of the same video segments under free-looking (F-LQ) and scoring (S-LQ) conditions. Figure 5 shows this comparison using the 4 similarity measures.

By simply looking at Figure 5, it is difficult to see a systematic effect difference in the similarity values under each task. Therefore the data is examined statistically using a one sample T-test. First with LCC, the delta between the LCC values (calculated for each of the F-LQ and S-LQ against the corresponding reference F-HQ saliency maps) is calculated for each time slot for each video. These delta values are then averaged across all time slots giving 25 values representing the average delta in viewing behavior in the 25 videos across all time-slots. Using a one sample T-test shows that there is a significant difference in LCC values for the viewing behavior from a test value of 0 with ($M=0.03$, $SD=0.04$) $t(24)=4.09$, $p<0.001$. By following the same steps with the rest of the similarity measures we find that SSIM also shows a significant difference with ($M=0.05$, $SD=0.03$) $t(24)=8.37$, $p<0.001$. However, no significant difference is found using the rest of the similarity measures with KLD giving ($M=-0.05$, $SD=0.31$) $t(24)=-0.77$, $p=0.45$ and NSS giving ($M=-0.002$, $SD=0.145$) $t(24)=-0.008$, $p=0.993$.

5. DISCUSSION

Looking at the graphs shown in Figure 4 one can safely say that all similarity measures have a similar level of performance considering the spread of the data points and the size of the resulting confidence intervals. It is therefore difficult to single out any of them as the best or worst. Hence, the analysis of the results looks at all the four similarity measures.

Figure 4 also shows that there is a trend of lower similarity in behavior as the quality of the videos increases. In other words, when looking at a better quality video, the viewing task has a greater influence on behavior. One possible explanation is that higher quality videos contain less artifacts, and therefore requires the viewers to actively search for clues of quality by ignoring the natural salient regions and scanning the entire video area. This trend, however, is not visible when analyzing the data using LCC (represented by the flat trend line in Figure 4 top left), which may be an indication that it is not a very strong trend.

As mentioned before, freely viewing the higher quality versions of the video segments (F-HQ) gives the closest viewing characteristics to natural saliency. For that reason it is used as the reference value for the analysis shown in Figure 5. To see whether the task has an effect on the viewing behavior, we measure the similarity of the reference data (F-HQ) to that collected with F-LQ. This is then repeated for the reference data (F-HQ) and S-LQ to see if changing the task gives a different result. Since the difference of quality is equal in both cases, we assume that it does not play a significant role in this comparison.

Looking at the LCC values (top left of Figure 5), one can see that the average LCC value for S-LQ often falls below that of F-LQ. A statistical t-test proves that this trend is systematic for both the LCC and the SSIM values. The positive mean value in both cases indicate that the the viewing behavior is less similar when the task is changed and that the difference is measurable. This effect can intuitively be explained as the viewers watching the videos in a similar manner when they are looking freely at the videos and that giving them a different task (scoring the video quality) changes their viewing behavior to a measurable extent. Unfortunately, it is not possible to observe the same effect using the KLD and NSS measures.

This result falls in line with earlier research performed on still images [13]. However, it differs from findings reached in earlier research performed on videos [12]. This may be due to the highly dynamic character of the video segments used in this experiment which makes it easier to detect deviations from the natural scene saliency. It may also be the result of the different method used for analyzing the data, since even within this data set not every used similarity measure was able to detect the difference in viewing behavior.

Finally, the above analysis show that different similarity measures are better suited for detecting different trends in the data. Only the SSIM was sensitive enough to detect both effects measured here. However it is difficult to definitively recommend one similarity measure as the most suitable for analyzing saliency of video viewing behavior. Clearly more work needs to be done using the data collected from this experiment [19] and similar efforts in this field.

6. CONCLUSIONS

In this paper we examined the effect of the given task on the viewing behavior when watching videos. By tracking the eye movements of the observers under both conditions it was possible to generate saliency maps representing the viewing characteristics under each condition. A set of different possible measures for saliency similarity [14] was used to analyze the data. From these measures, the SSIM was the only one sensitive enough to detect all observed effects.

Using these similarity measures, it was possible to see a trend of the task having a stronger effect on viewing behavior if the video has a higher level of encoding quality. Viewers seem to focus more on searching for clues of the image quality if no clear artifacts are present. It was also possible to detect a systematic difference in the viewing behavior where viewers deviated more from their natural viewing behavior when they were given the task of scoring the quality of the videos.

With regards to future work, we are looking deeper into the video segments using the calculated similarity measures to find specific scenes that exhibit higher sensitivity to the given task and try to identify common characteristics in these scenes. Additionally, it may be interesting to look into applying other image saliency analysis techniques to see how they perform on this data. The saliency data generated in this experiment has also been made available on the Internet for other researchers in the field [19] to offer them the chance to use it in related research.

REFERENCES

[1]. Buswell, G., “How people look at pictures,” Oxford, England: University of Chicago Press, 1935

[2]. Yarbus, A., “Eye movements and vision,” New York: Plenum Press, 1967

[3]. Liu H., and I. Heynderickx, "Visual Attention in Objective Image Quality Assessment: based on Eye Tracking Data," IEEE Transactions on Circuits and Systems for Video Technology

[4]. Adams M.D., “The JPEG-2000 Still Image Compression Standard”, ISO/IEC JTC1/SC29/WG1 (ITU-T SG8), 2001

[5]. Koch C., and S. Ullman, “Shifts in Selection in Visual Attention: Toward the Underlying Neural Circuitry,” Human Neurobiology, vol. 4, no. 4, pp. 219-27, 1985

[6]. Itti L., C. Koch, and E. Niebur, “A model of saliency-based visual attention for rapid scene analysis,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 20, no. 11, 1998.

[7]. Le Meur O., P. Le Callet, and D. Barba, "Predicting visual fixations on video based on low-level visual features," Vision Research, vol. 47, no. 19, 2007.

[8]. Bruce, N.D.B., Tsotsos, J.K., “Saliency, Attention, and Visual Search: An Information Theoretic Approach,” Journal of Vision Vol .9, no.3, 2009

[9]. Vuori, T., Olkkonen, M., Pölönen, M., Siren, A., and Häkkinen, J. “Can eye movements be quantitatively applied to image quality studies?,” In Proceedings NordiCHI 2004. ACM Press, New York, NY, 2004, 335-338

[10].Ninassi, O. Le Meur, P. L. Callet, D. Barba, and A. Tirel, “Task impact on the visual attention in subjective image quality assessment,” Proc. EUSIPCO-06, 2006.

[11].Redi, J.A., Liu, H., Zunino, R. and Heynderickx, I., “Interactions of visual attention and quality perception”, In: IS&T/SPIE Electronic Imaging 2011 and Human Vision and Electronic Imaging XVI. Vol 7865. 2011.

[12].Le Meur O., 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,” Elsevier, Signal Processing: Image Communication, 2010

[13].Alers, H., and I. Heynderickx 2011. How The Task Of Evaluating Image Quality Influences Viewing Behavior. Proceedings of QoMEX, 2011

[14].Redi, J, Heynderickx I. “Image Quality And Visual Attention Interactions: Towards A More Reliable Analysis In The Saliency Space,” Proceedings of QoMEX, 2011

[15].ITU-R Recommendation BT.500-11, “Methodology for the subjective assessment of the quality of television pictures,” Geneva, 2002.

[16].http://www.ffmpegx.com

[17].Sheikh H.R., Sabir M.F. and Bovik A.C., "A statistical evaluation of recent full reference image quality assessment algorithms", IEEE Transactions on Image Processing, vol. 15, no. 11, 3440-3451 (2006).

[18].Wang Z., Bovik A.C., Sheikh H.R. and Simoncelli E.P., "Image quality assessment: from error visibility to structural similarity," IEEE Transactions on Image Processing , vol. 13, no. 4, 600- 612 (2004).

[19].Alers, H., H. Liu, J. Redi and I. Heynderickx, "TUD Video Quality Database: Eye-Tracking Release 2", http://mmi.tudelft.nl/iqlab/video_task_eye_tracking_1.html