INFLUENCE OF SHOOTING CONDITIONS, RE-ENCODING AND VIEWING CONDITIONS ON THE PERCEIVED QUALITY OF USER-GENERATED VIDEOS

Yohann Pitrey, Patrik Hummelbrunner, Benjamin Kitzinger, Shelley Buchinger, Marcus Barkowsky, Patrick Le Callet, Romuald Pépion

INFLUENCE OF SHOOTING CONDITIONS, RE-ENCODING AND VIEWING CONDITIONS ON THE PERCEIVED QUALITY OF USER-GENERATED VIDEOS

Y. Pitrey[*], *P. Hummelbrunner*[*], *B. Kitzinger*[*], *S. Buchinger*[*], *M. Barkowsky*[+], *P. Le Callet*[+], *R. Pepion*[+]

[*] University of Vienna, Institute for Entertainment Computing, AUSTRIA
[+] Polytech'Nantes, LUNAM Univ. de Nantes, IRCCyN IVC – UMR CNRS 6597, FRANCE

ABSTRACT

User-generated videos are becoming a very popular way for capturing and sharing material. The creation and consumption of these videos is quite different from the traditional professionally designed videos. Consumer-range cameras are often used to generate the videos, in uncontrolled conditions, possibly operated without the same considerations regarding camera stabilization or lighting conditions. If the videos are then shared on a video sharing platform, they often have to go through a re-encoding step that is likely to degrade their original quality. At last, if consumers are watching the videos on a mobile phone, the viewing conditions might have an impact on the perception of quality of the videos. In this paper, we evaluate the whole chain from production to consumption of user-generated videos through a subjective quality assessment experiment conducted on 29 naive viewers. The results of the experiment allow us to identify camera shaking and re-encoding as two main factors of variation of the video quality in this context. However, the influence of the viewing context can not be considered significant, even if the results show that the videos degradations are perceived less severely in a distracting environment.

1. INTRODUCTION

The amount of user-generated video content has been rapidly increasing during the last decade, with platforms such as YouTube[1] allowing users to upload clips captured from their smartphones or other consumer-range camera devices. Regarding many aspects, this type of content is different from professionally produced content [1, 2, 3]. In fact, the whole process from production to consumption is changing from what is usually applied for professional content. First, on the content production side, the quality of the camera used to create video content can vary from high-definition camera recorders to non-specialized devices such as smartphones, which can in turn be derived in various levels of camera quality. Second, the shooting conditions are often not optimal, and the captured clips might be subject to shaking due to handheld cameras, or suffer from deffects in illumination due to bad lighting conditions. Sharing videos on a platform such as YouTube usually implies a re-encoding step on the platform's side in order to match storage and bandwidth requirements. This re-encoding step might lead to a significant difference in quality between the original video and the version broadcasted by the video platform. On the consumer's side, the viewing conditions can change the perception of the video content. Bad viewing conditions and environmental distractions can indeed mask some characteristics of the video and modify its perception. Finally, the viewers themselves might represent a factor of change in the appreciation of the videos. Human observers might rate the same content differently, according to their motivations and expectations [4]. Their level of familiarity with user-generated videos and level of engagement in the content they watch might also modify their perception [5].

In this paper, we present the results of a subjective experiment conducted in order to evaluate the quality of experience on user-generated videos in realistic conditions. We simulate conditions ranging from an almost professional production scenario to several more amateur settings in order to evaluate the influence of five parameters covering the entire process from the creation to the consumption of videos. The five parameters are (1) the quality of the camera device, (2) camera shaking, (3) global illumination, (4) video re-encoding and (5) viewing environment. This experiment is the first step of a series of tests aimed at comparing the responses given to a set of videos by different populations of viewers. It was conducted with high school pupils as observers, whose age group is known to be a significant contributor to user-generated content [6]. However, the focus of this publication is not on the influence of user population, as other experiments need to be conducted in this direction.

The rest of this paper is organized as follows. Section 2 gives more details on the user-generated video production and consumption process and the parameters involved in the quality of experience in such a context. Section 3 describes

[1] www.youtube.com

the experiment we conducted to assess the influence of these parameters on the quality of experience. Section 4 presents the results of the experiment. Finally, Section 5 concludes the paper.

2. BACKGROUND

Camera quality can vary among equiped devices for natural reasons such as price or manufacturer choices in terms of hardware and software. A lot of work exists on evaluating and comparing the quality of camera devices, which is out of the scope of our paper. We address the problem of camera quality from the end user's point of view, by simply using a camera to produce videos. The camera sensor is not the only parameter that can influence the quality of the captured videos. The video codec that is used to process the raw data and make it easier to manipulate by the end user is also quite important. Nowadays, MPEG-4 AVC/H.264 is being more and more implemented on consumer-range devices, for its well-known bitrate/quality performance [7]. Nevertheless, older video cameras are still equiped with previous codecs such as MPEG-4 Part 2 or even MPEG-2, which do not allow for the same level of performance. Additionally, because of the complexity of the H.264 encoding process, some devices do not use the full capabilities of the standard and therefore are only able to reach intermediate qualities.

The impact of camera shaking on the perceived video quality is twofold. First, the uncontrolled motion can give the impression of a poorly edited video, and influence the viewer in getting a negative opinion [8]. The second impact of camera shaking involves both visual attention and video coding, as the fast and erratic motion creates a substantial activity in the video, which might not only make it difficult for a human viewer to focus his attention on the elements of the scene, but also decrease the efficiency of the encoding. Existing work considering camera skaking mostly focuses on algorithms to automatically remove the undesired motion [9, 10]. The assumption is made that shaking has an impact on the perceived quality, but this impact is not clearly evaluated from the viewers point of view. However, a camera stabilization algorithm is of interest only if the viewers can notice a difference between the video before and after treatment. The sensitivity of viewers to shakiness might also depend on the context in which a video is shot. For instance, if the person carrying the camera is walking while shooting, the camera shaking might be understood as part of the message of the video and therefore could be interpreted as a less severe degradation [11].

Illumination corresponds to the overall impression of light that is rendered in a video clip. It is conditioned by the available light and the objects in the scene, as well as by the characteristics of the capturing camera. If the available light is too low, the contents in the scene might be difficult to identify and therefore lead to a quality perveived as bad by observers. On the opposite, if the available light is too strong, the quality might be affected in a similar way. Most current camera devices are equiped with automated exposure adjustment mechanisms in order to adapt to the lighting conditions in the scene. These tools are often software-based and need time to adjust. As a result, rapid illumination changes can not be compensated and may lead to under- or over-exposed frames. Additionally, it is well known that the contrast in a video might influence the impression of quality [12]. As user-generated videos are likely to be shot without a fine control of the lighting conditions (*ie:* using only the available light in the scene), the influence of global illumination and contrast appear as important parameters in the acquisition.

Video sharing platforms usually re-encode the uploaded videos in order to maintain homogeneity of the proposed contents and meet their storage constraints. Sometimes they also provide different versions with various quality levels to ensure better adaptability to user device requirements. Encoding all the videos at the same bitrate has a great advantage for transmission under normal bandwidth conditions. However, one can expect a significant difference in quality between the captured video and the re-encoded video. Particularly, re-encoding videos captured from High-Definition camera recorders may introduce a significant loss in quality. Additionally, it is known that transcoding a video from one standard to another can create visual artifacts [13]. Issues such as frame rate reduction, error drifting and jerkiness do indeed have a severe impact on the quality. As a result, the combined effect of bitrate reduction and transcoding needs to be considered as an important factor of quality degradation in the user-generated videos production chain.

Watching videos on a mobile allows users to use their devices in varying contexts. Depending on the level of distractions, such as background noise or external elements, the attention that a human viewer is able to dedicate to the video experience can vary. Previous work has shown that an experiment conducted in a standardized laboratory and in a realistic mobile consumption context produces significantly different results [11]. Although, The three different mobile contexts that were evaluated gave relatively similar results. The study involved essentially transmission distortions, which known to be considered as quite severe degradations. As the degradations which are likely to be present in user-generated videos are not as severe, it is interesting to confirm the previously observed results about the impact of the viewing context on the perceived video quality.

We designed a subjective experiment aiming at evaluating the influence of the parameters enumerated in this section. The experimental material and the test conditions are descirbed in the following section.

3. EXPERIMENTAL DESIGN

The focus of our experiment can be divided in three parts: shooting conditions (which covers the *camera device*, camera shaking or *shakiness* and *illumination*), video re-encoding and viewing conditions. The varied parameters are reviewed in detail in this section.

The video material was captured in various places in Vienna, and includes 8 different scenes with a wide variety of contents, such as indoor and outoor scenes, low and high motion patterns, landscape and human characters. Figure 1 shows an overview of these contents.

3.1. Camera devices

Each video was captured using three different devices: a Samsung HD Camcorder (considered as upper-level camera), a Samsung Galaxy S smartphone (considered as state-of-the-art quality) and a HTC Hero smartphone (considered as lower quality). We were knowingly using devices that are significantly different in terms of recording quality to get a set of video data, which reflects the wide range of quality found on video platforms nowadays. The Samsung HD Camcorder has a native resolution 1280x720 pixels at 50 FPS and uses an H264 AVC encoder with an average bitrate of 6 Mbit/s. The Samsung Galaxy S has a resolution of 720x480 at 30 FPS with an H264 AVC encoder and an average bitrate around 3.50 Mbit/s. The HTC Hero has a camera with relatively lower quality than the two other devices. Its resolution is 352x288 with a MP4V encoder and an average bitrate of 720 Kbit/s.

3.2. Shakiness

Shakiness is produced by unwanted camera motion, often due to the lack of stabilization. Especially in casual filming smaller devices are used, which are harder to stabilize without special equipment. Four levels of shakiness were included in our experiment: *stand* (camera attached to a tripod stand), handheld camera with *low-shaking* and *high-shaking*, and *walking* (camera held by a person walking while filming). These levels were chosen, because they are the most common in user-generated videos. *Low-shaking* videos are videos recorded by a person who concentrates on stabilizing the device without extra equipment to get a good looking clip, whereas *high-shaking* videos are recorded without paying any attention on stabilization and therefore are more shaky. We expected that with increasing shakiness the MOS for this specific clip will decrease.

3.3. Illumination

We classified the video contents into four levels of illumination: *very low* (insufficient overall illumination in the scene), *low* (overall impression of low illumination, but objects of interest can be identified easily), *high* (overall high illumination, objects of interest identified easily) and *very high* (impression of over-exposed scene). Depending on the camera quality of the recording device the lighting in the recording environment had a big impact on the overall quality of the resulting video. For lower quality devices differences in contrast and brightness may cause strong artefacts like increased blurring, blocking and noise. Therefore it is even more important to consider different levels of illumination in respect to the recording device.

3.4. Re-encoding

In order to evaluate the impact of the re-encoding process performed by sharing platforms on the perceived quality, we re-scaled each video to 720x480 pixels and we performed re-encoding at 800 Kbit/s using ffmpeg's implementation of the x264 encoder[2]. This bitrate was chosen because it introduces coding distortions noticeable by a standard human viewer. However, the level of quality of the re-encoded videos remains acceptable so that the coding artifacts are not the dominant form of degradation in the videos. In our experiment, both the original videos and the re-encoded versions were presented to the observers. It is worth noticing that the re-encoded bitrate is higher than the native bitrate for the HTC Hero camera. However, the difference is quite small when compared to the actual bitrate, so we consider the original and re-encoded bitrates as equivalent for this device.

3.5. Viewing environment

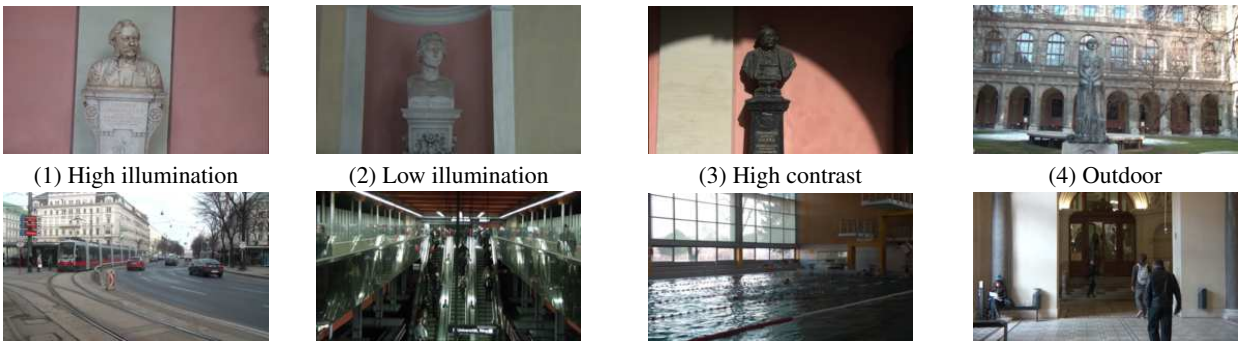
The video sequences were presented on a state-of-the-art mobile phone in three different environments: a *classroom*, a *cafeteria* and a computer *laboratory*. The classroom was empty apart from the test subject, but noises could be heard from people passing in the corridor next to the room. Natural light was used but no direct sunlight entered the room. The cafeteria was busy with people producing a relatively noisy atmosphere. Artificial and natural lights were mixed. The laboratory was empty, and no signficant distracting sound could be heard (the computers were not running). Artificial light was used, with shades closed so that no natural light would enter the room. A group of 10 viewers took the test in each environment.

3.6. Test conditions

The smartphone used for the test had a native resolution of 800x480 pixels. We used the subjective player for android[3] to present the videos in random order to the viewers. The

[2] www.ffmpeg.org, www.videolan.org/x264

[3] http://code.google.com/p/subjectiveplayer



(1) High illumination (2) Low illumination (3) High contrast (4) Outdoor

(5) Motion/High Illumination (6) Motion/Low Illumination (7) Motion/Very low Illumination (8) Motion/High Contrast

Fig. 1. Overview of the 8 different scenes used in the experiment. (1) No motion and overall good lighting. (2) No motion in a darker environment. (3) No Motion and a shadow produces high contrast. (4) Still outdoor scene with many details and graded light condition. (5) Cars driving by in a well lighted outdoor scene rich in detail. (6) People on an escalator and metro passing by in a poorly illuminated scene. (7) People swimming in a pool in a scene with frontlighting and poor illumination. (8) People passing by in university building, which has different lighting conditions in the front and the back of the scene.

test was conducted according to standard observer training and rating procedures such as described in [14]. The observers rated the videos directly on the device using a standard ACR scale with 5 levels. A total of 30 observers participated in the test (13 females, 17 males, aged 15 to 17). They received a short introduction on the purpose of the test, during which the focus was brought on the overall video quality, asking the viewers to focus less on the content of the videos. A training session of 5 clips was performed before the test to help the viewers get used to the voting interface. The main test contained 108 clips of 10 seconds length, for a total duration of the test around 25 minutes, including voting time. The experimental results are described in the next section.

4. EXPERIMENTAL RESULTS

The Mean Opinion Score (MOS) was calculated for each presented configuration. To allow reliable comparison of the MOS values, we computed the 95% intervals of confidence for each MOS. Two MOS values are considered equivalent if their intervals of confidence overlap.

4.1. Camera device

Figure 2 presents the influence of the camera device, camera shakiness and bitrate on the quality of videos. The observed quality for each device at original bitrate confirms our previous ordering, the HD Camcorder being superior quality than the Samsung Galaxy, in turn better than the HTC Hero.

4.2. Shakiness

Globally, the standing camera is best accepted by the viewers for each device. Then the quality globally decreases

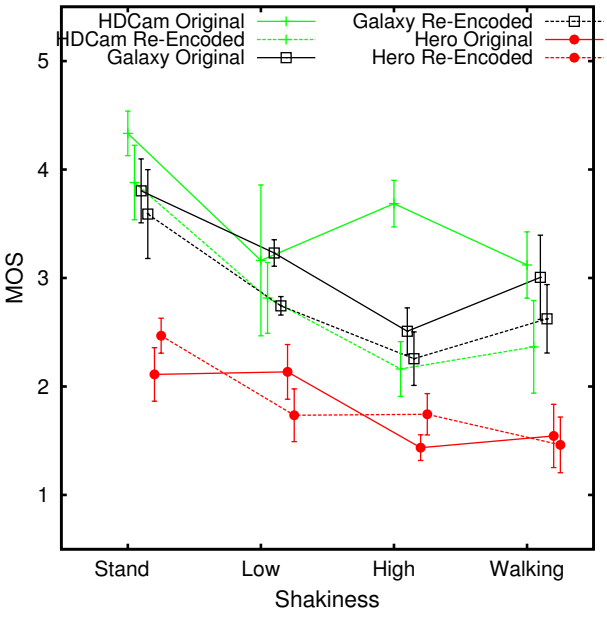


Fig. 2. Influence of camera, bitrate and shakiness on the MOS.

when the level of shakiness increases. Nevertheless, some interesting behaviours can be identified that are not consistent with regard to the camera devices. For the Samsung Galaxy, the walking person gives significantly higher quality than the high motion pattern. However, analyzing the video visually allows us to identify a higher shakiness in the walking version. One possible explanation is that the waking person moves and therefore introduces meaningful movement in the video. The viewers might understand this as part of the design of the video and penalize the quality

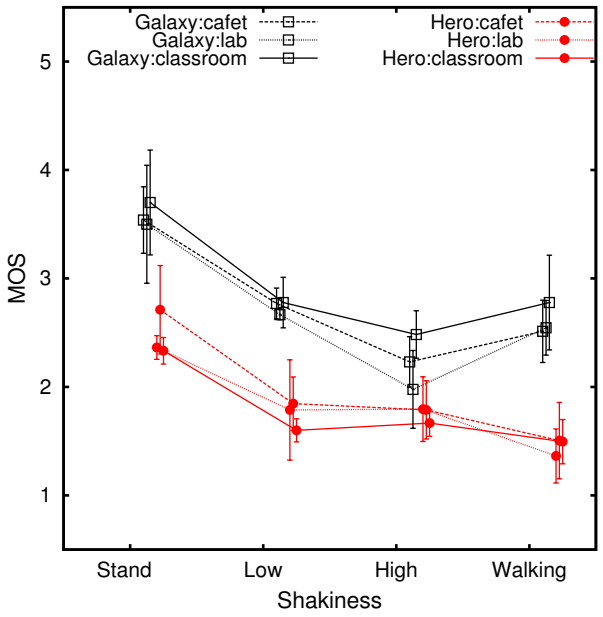


Fig. 3. Influence of the viewing environment on the MOS.

less severely. Another explanation could be a saturation effect in the perception of shakiness by the viewers.

4.3. Illumination

We could not identify any significant influence of illumination in the subjective scores. We observed a slight increase of the scores when the global illumination increases, possibly due to the fact that the objects in the scene become more distinct. The classification of videos into four illumination levels was performed manually. Some scenes were difficult to classify, as the objects of interest were hard to determine or were only partly illuminated. A possible explanation might also be the influence of illumination being masked by the influence of more dominant parameters such as shakiness or bitrate, as will be described in the next subsection.

4.4. Re-encoding

The original bitrates naturally give a better quality than the re-encoded bitrates for the HD Camcorder and for the Samsung Galaxy. Although, the original superiority of the HD Camcorder seems to be voided by the re-encoding process, showing the limited interest of high definition devices if the created content is uploaded on a video sharing platform. Even after re-encoding, the HTC Hero does not reach equivalent quality to the two other devices, showing a poor camera quality.

4.5. Viewing Enviroment

Figure 2 presents the scores obtained by the Samsung Galaxy and HTC Hero in each viewing environment for the re-encoded videos. Globally, no significant impact of the viewing environment can be identified. A more detailed analysis however allows us to distinguish between different behaviours for the two devices. For the HTC Hero, the cafeteria gives higher quality than the two other environments, only for the lowest level of shakiness. The results for the Samsung Galaxy are different: classroom gets higher scores than cafeteria, laboratory is in between. Although the difference is not always significant, we suggest that the cafeteria might draw the attention away from the distractions of the videos, which do not appear as severely distorted as in the more quiet classroom.

5. CONCLUSION

We presented the results of a subjective experiment aiming at identifying the parameters affecting the perceived quality of user-generated videos. We were interested in identifying more precisely the influence of the quality of the camera device, the level of camera shaking, the global illumination in the scene, the re-encoding of the videos that is usually performed by the video sharing platforms, and the environment in which the videos are watched on state-of-the-art mobile phones. The results we presented lead to interesting conclusions that can be used as guidelines for the creation of video clips in the context of a diffusion on video sharing platforms. First, we demonstrated that avoiding camera shakiness leads to significantly better perceived quality. Second, we showed that when videos are re-encoded at the same bitrate, the benefits of a high quality camera can be lost. Our experiment could not identify any significant influence of the global illumination on the quality. This might be explained either by the presence of more dominant parameters in our data. Finally, we identified that distracting viewing environments might lead to better perceived quality, although in a limited fashion. Future work shall evaluate the impact of the voting population, by reconducting similar experiments on different groups of observers, with specific demographic features or levels of involvement with the video content.

6. REFERENCES

[1] Y. Borghol, S. Mitra, S. Ardon, N. Carlsson, D. Eager, and A. Mahanti, "Characterizing and modelling popularity of user-generated videos," *Performance Evaluation*, vol. 68, no. 11, pp. 1037–1055, Nov. 2011.

[2] C. Meeyoung, K. Haewoon, P. Rodriguez, A. Yong-Yeol, and M. Sue, "Analyzing the video popularity

characteristics of large-scale user generated content systems," *IEEE/ACM Transactions on Networking*, vol. 17, no. 5, pp. 1357 –1370, 2009.

[3] Meeyoung Cha, Haewoon Kwak, Pablo Rodriguez, Y.Y. Ahn, and Sue Moon, "I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system," in *ACM SIGCOMM conf. on Internet measurement*, 2007, pp. 1–14.

[4] T. Hossfeld, R. Statz, and S. Egger, "SOS : The MOS is not enough," *QoMEX*, 2011.

[5] P. Kortum and M. Sullivan, ," *Human Factors: The Journal of the Human Factors and Ergonomics Society*.

[6] S. Buchinger, S. Kriglstein, and H. Hlavacs, "A Comprehensive View on User Studies : Survey and Open Issues for Mobile TV," *EuroITV (ACM)*, 2009.

[7] Richardson, Iain E., ," in *H.264 and MPEG-4 Video Compression*. John Wiley and Sons, 2003.

[8] G. Abdollahian, C.M. Taskiran, Z. Pizlo, and E.J. Delp, "Camera motion-based analysis of user generated video," *IEEE Trans. on Multimedia*, vol. 12, Jan 2010.

[9] G. Spampinato, A. Castorina, A. Bruna, and A. Capra, "Camera shaking effects reduction by means of still sequence stabilization and spatio-temporal filtering," in *Consumer Electronics, ICCE.*, 2009, pp. 1 –2.

[10] M. Ondrej, Z. Frantisek, and D. Martin, "Software video stabilization in a fixed point arithmetic," in *Applications of Digital Information and Web Technologies, ICADIWT*, 2008, pp. 389 –393.

[11] S. Jumisko-Pyykkö and M. Hannuksela, "Does context matter in quality evaluation of mobile television?," in *ACM MobileHCI*, 2008, pp. 63–72.

[12] J. You, T. Ebrahimi, and A. Perkis, "Visual attention tuned spatio-velocity contrast sensitivity for video quality assessment," in *IEEE ICME*, 2011, pp. 1–6.

[13] L. Goldmann, F. De Simone, F. Dufaux, T. Ebrahimi, R. Tanner, and M. Lattuada, "Impact of video transcoding artifacts on the subjective quality," in *QoMEX*, 2010.

[14] ITU-R, "BT.500-10 Methodology for the subjective assessment of the quality of television pictures," 1974.