Split & Dual Screen Comparison of Classic vs Object-based Video

Maarten Wijnants
Hasselt University – tUL
Expertise Centre for Digital Media
Diepenbeek, Belgium
maarten.wijnants@uhasselt.be

Sven Coppers
Hasselt University – tUL
Expertise Centre for Digital Media
Diepenbeek, Belgium
sven.coppers@uhasselt.be

Gustavo Rovelo Ruiz
Hasselt University – tUL
Expertise Centre for Digital Media
Diepenbeek, Belgium
gustavo.roveloruiz@uhasselt.be

Peter Quax
Hasselt University
– tUL – Flanders Make – EDM
Diepenbeek, Belgium
peter.quax@uhasselt.be

Wim Lamotte
Hasselt University – tUL
Expertise Centre for Digital Media
Diepenbeek, Belgium
wim.lamotte@uhasselt.be

ABSTRACT

Over-the-top (OTT) streaming services like YouTube and Netflix induce massive amounts of video data, hereby putting substantial pressure on network infrastructure. This paper describes a demonstration of the object-based video (OBV) methodology that allows for the quality-variant MPEG-DASH streaming of respectively the background and foreground object(s) of a video scene. The OBV methodology is inspired by research into human visual attention and foveated compression, in that it allows to adaptively and dynamically assign bitrate to those portions of the visual scene that have the highest utility in terms of perceptual quality. Using a content corpus of interview-like video footage, the described demonstration proves the OBV methodology's potential to downsize video bitrate requirements while incurring at most marginal perceptual impact (i.e., in terms of subjective video quality). Thanks to its standards-compliant Web implementation, the OBV methodology is directly and broadly deployable without requiring capital expenditure.

CCS CONCEPTS

• **Information systems → Multimedia streaming**; Web applications; • **Networks → Public Internet**; *Application layer protocols*; • **Human-centered computing → User studies**;

KEYWORDS

video coding, H.264, HTTP Adaptive Streaming, MPEG-DASH, subjective evaluation, Web

ACM Reference Format:
Maarten Wijnants, Sven Coppers, Gustavo Rovelo Ruiz, Peter Quax, and Wim Lamotte. 2019. Split & Dual Screen Comparison of Classic vs Object-based Video. In *Proceedings of the 27th ACM International Conference on Multimedia (MM '19), October 21–25, 2019, Nice, France.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3343031.3350582

MM '19, October 21–25, 2019, Nice, France
© 2019 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-6889-6/19/10.
https://doi.org/10.1145/3343031.3350582

1 INTRODUCTION AND MOTIVATION

Massive amounts of video traffic are being delivered over the contemporary Internet [14] and these traffic volumes are poised to increase even further in the nearby future [5]. However, when watching videos, not all spatial regions of these videos are equally important from a perceptual quality perspective. Evidence abounds in the literature that human visual focus during video consumption is subconsciously drawn to a limited quantity of salient objects or regions that stand out from the video background [2, 7, 15]. For instance, human faces are known to be natural attractors of visual attention [3, 8]. At the same time, complementary research shows that the sampling density and sensitivity of the Human Visual System (HVS) gradually decays as the distance to our eyes' fixation area increases [3, 9]. Stated differently, only those spatial portions of the video where viewers fixate on are processed by the HVS in maximal fidelity, while more distant regions are being assigned ever lower processing bandwidth.

Driven by the just described observations, we have previously introduced the object-based video (OBV) methodology [18] that allows for the adaptive delivery of a visual scene by decomposing it into a background and one or more foreground objects which each can be independently streamed in a quality-variant manner. As such, the OBV methodology supports the introduction of intra-scene quality differences during network delivery. In contrast, a classic *frame-based* video encoder does not grant such versatility, since it will enact a rather uniform bitrate and hence quality distribution over the integral visual scene. The OBV methodology can be regarded as a video-only specialization of the more general object-based media (OBM) paradigm [16] that is currently under active investigation by both academia [13] and industry [10]. This demonstration illustrates the viability of leveraging video-only OBM mechanisms to achieve significant cost reductions (i.e., in terms of network load) that entail no or at most limited repercussions with respect to perceived video quality.

2 OBV IMPLEMENTATION

The OBV methodology is exclusively implemented using standardized Web technologies (i.e., HTML5, CSS, JavaScript, WebGL) to yield a portable, multi-platform solution that is accessible via an off-the-shelf Web browser [18]. The methodology takes as input a video that has been disassembled into multiple visual signals, one per background and individual foreground object. In each such signal, pixels that have been segmented away are replaced by a fixed chroma value (typically pristine green). All these signals are then frame-based encoded in multiple qualities using a vanilla video codec (i.c., x264 version 20180118-7d0ff22). From this point on, each back- and foreground object can be independently and adaptively streamed (i.c., via MPEG-DASH). At reception side, the incoming signal bitstreams are subjected to *chroma keying* to transform pixels carrying the predefined chroma value from the content preparation stage into transparent ones. Finally, the original video scene is recomposed by alpha blending the processed signal bitstreams in the appropriate order (i.e., from back- to foreground), on a frame-by-frame basis. For this approach to work, the playback of the scene’s constituent video signals needs to be synchronized, which is achieved by leveraging the W3C Timing Object specification [1, 4]. Both the chroma keying and alpha blending operations are implemented as WebGL shaders to profit from 3D hardware acceleration.

By applying lossy video compression to the segmented bitstreams, color bleeding at the object boundaries might occur. This results in gradient-like instead of discrete color transitions (i.e., from object color to chroma keying color or vice versa), which in turn hampers the performance of purely color-based chroma keying algorithms [18]. This demonstration therefore instead relies on a chroma keying implementation that is grounded on the *erosion* morphological operation [12]. In particular, when decomposing the video scene, a fixed number of pixels around the edge of each object is included in its representation (instead of being replaced by the chroma keying color), with the same amount of pixels being trimmed by the erosion operation. The additional pixels hence serve as spatial buffer to absorb the color contamination caused by lossy compression and are made fully transparent during the chroma keying processing, this way resulting in an artifact-free video recomposition.

3 DEMONSTRATION

The demonstration[1] allows users to perceptually compare, in a playback synchronized fashion, a traditionally encoded version of a video with its corresponding OBV rendition whose background quality can be interactively controlled. Two visualization methods are provided to conduct this perceptual comparison. The first shows the classic and object-based video coding alternatives on dedicated physical displays to enable side-by-side comparison (see top image in Figure 1). The second visualization method merges both video representations in a single HTML5 `<canvas>` on a single screen and allows the user to control a vertical divider (via a range slider widget) that determines how much of respectively the traditional and object-based alternatives are being rendered (see bottom image in Figure 1). In both visualization approaches, the quality setting of the classic video version remains fixed at a CRF[2]value of 23, as does the foreground object in the object-based version. On the other hand, the OBV background quality is interactively adjustable by users (CRF 23 up to CRF 38). Background quality settings are enforced via quality-variant MPEG-DASH streaming (using Media Segments with a one second duration). We use `dash.js` [6] version 2.6.0 as MPEG-DASH client and instruct it to buffer at most a single Media Segment to enable prompt responding to requested background quality switches.

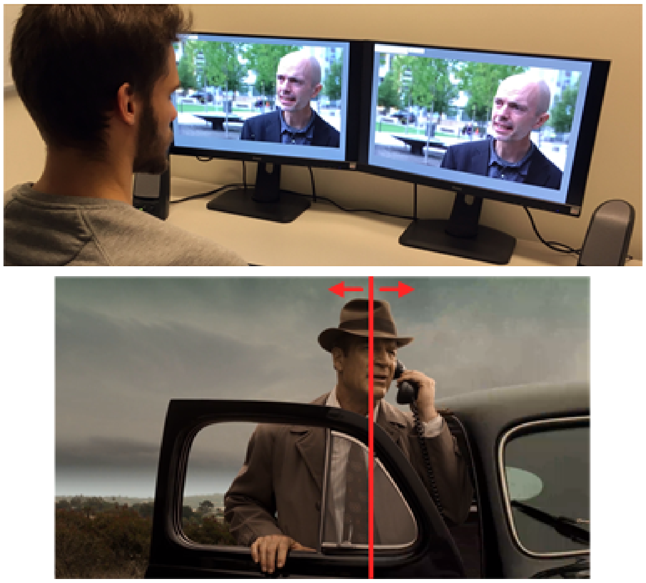

Figure 1: Perceptually comparing classic versus object-based video coding: (top) dual screen, (bottom) split screen.

The demonstration is populated with nine Full HD live action videos with an average duration of 12.3±2.3 seconds [17]. Content-wise, each video involves one talking person (English speech) as foreground object, with those persons standing in front of backgrounds with varying complexity and heterogeneous characteristics (e.g., in-focus versus out-of-focus). All videos represent realistic *talking heads* footage as it commonly appears in video genres like, for example, newscasts, interviews, talk shows and documentaries. For each video in the content corpus, two object-based instantiations exist that differ in terms of scene segmentation accuracy (i.e., pixel-precise segmentation along the foreground object’s *contour* versus a *bounding box*-based[3]representation of the foreground object).

Users can leverage the demonstrator, for example, to identify the background CRF value up to which quality differences between the traditional and object-based versions are not or barely visible, or to search for the highest CRF value that yields a visually still acceptable OBV rendition (even if a background quality difference at this point is noticeable compared to the classic encode). Similarly, the demonstrator allows users to compare the perceptual impact of relying on contour-precise versus bounding boxed foreground object representations. Finally, both the dual and split screen visualization methods include the option to illustrate potential cost savings afforded by the OBV methodology. These cost savings are expressed as both a bitrate percentage and a monetary value (based on an estimate of mobile data cost) relative to the traditional video version.

ACKNOWLEDGMENTS

Maarten Wijnants is funded by a VLAIO Innovation Mandate (project number HBC.2016.0625), co-sponsored by Androme. Sven Coppers is funded by the Special Research Fund (BOF) of Hasselt University (R-8150). We thank Davy Vanacken for his methodological advice.

[1] Running on an Intel Core i5-8259U 2.3 GHz CPU with 32 GB RAM and an Intel Iris Plus Graphics 655 integrated GPU, using Google Chrome version 75 (64-bit) on Windows 10.

[2] Quality degrades as the Constant Rate Factor (CRF) value rises (i.e., quality and CRF value are inversely proportional). The H.264 video codec’s default CRF value is 23 [11].

[3] The bounding boxes change neither position nor size throughout video playback.

REFERENCES

[1] Ingar M. Arntzen and Njål T. Borch. 2016. Data-independent Sequencing with the Timing Object: A JavaScript Sequencer for Single-device and Multi-device Web Media. In *Proceedings of the 7th International Conference on Multimedia Systems (MMSys '16)*. ACM, Article 24, 10 pages. https://doi.org/10.1145/2910017.2910614

[2] Robert B Goldstein, Russell Woods, and Eli Peli. 2007. Where people look when watching movies: Do all viewers look at the same place? *Computers in Biology and Medicine* 37, 7 (08 2007), 957–964. https://doi.org/10.1016/j.compbiomed.2006.08.018

[3] Giuseppe Boccignone, Angelo Marcelli, Paolo Napoletano, Gianluca Di Fiore, Giovanni Iacovoni, and Salvatore Morsa. 2008. Bayesian Integration of Face and Low-Level Cues for Foveated Video Coding. *IEEE Transactions on Circuits and Systems for Video Technology* 18, 12 (December 2008), 1727–1740. https://doi.org/10.1109/TCSVT.2008.2005798

[4] Njål T. Borch, Ingar M. Arntzen, and François Daoust. 2018. Timing Object – Draft Community Group Report. Online, http://webtiming.github.io/timingobject/.

[5] Cisco. 2019. Visual Networking Index: Forecast and Trends, 2017 - 2022. Online, https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white-paper-c11-741490.html.

[6] dash.js. 2019. A reference client implementation for the playback of MPEG DASH via Javascript and compliant browsers. Online, https://dashif.org/dash.js/.

[7] Michael Dorr, Thomas Martinetz, Karl R. Gegenfurtner, and Erhardt Barth. 2010. Variability of eye movements when viewing dynamic natural scenes. *Journal of Vision* 10, 10 (August 2010), 1–17. https://doi.org/10.1167/10.10.28

[8] Ulrich Engelke, Hagen Kaprykowsky, Hans-Jürgen Zepernick, and Patrick Ndjiki-Nya. 2011. Visual Attention in Quality Assessment. *IEEE Signal Processing Magazine* 28, 6 (November 2011), 50–59. https://doi.org/10.1109/MSP.2011.942473

[9] Ulrich Engelke, Romuald Pépion, Patrick Le Callet, and Hans-Jürgen Zepernick. 2010. Linking Distortion Perception and Visual Saliency in H.264/AVC Coded Video Containing Packet Loss. In *Proceedings of Visual Communications and Image Processing (VCIP 2010)*. https://doi.org/10.1117/12.863508

[10] Michael Evans, Tristan Ferne, Zillah Watson, Frank Melchior, Matthew Brooks, Phil Stenton, Ian Forrester, and Chris Baume. 2017. Creating Object-Based Experiences in the Real World. *SMPTE Motion Imaging Journal* 126, 6 (August 2017), 1–7. https://doi.org/10.5594/JMI.2017.2709859

[11] FFmpeg. 2019. Encode/H.264. Online, https://trac.ffmpeg.org/wiki/Encode/H.264.

[12] Rafael C. Gonzalez and Richard E. Woods. 2018. *Digital Image Processing, 4th Edition*. Pearson.

[13] Jie Li, Thomas Röggla, Maxine Glancy, Jack Jansen, and Pablo Cesar. 2018. A New Production Platform for Authoring Object-based Multiscreen TV Viewing Experiences. In *Proceedings of the 2018 ACM International Conference on Interactive Experiences for TV and Online Video (TVX '18)*. ACM, 115–126. https://doi.org/10.1145/3210825.3210834

[14] Sandvine. 2018. Global Internet Phenomena Report. Online, https://www.sandvine.com/2018-internet-phenomena-report.

[15] Meijun Sun, Ziqi Zhou, Dong Zhang, and Zheng Wang. 2018. Hybrid convolutional neural networks and optical flow for video visual attention prediction. *Multimedia Tools and Applications* 77, 22 (November 2018), 29231–29244. https://doi.org/10.1007/s11042-018-5793-z

[16] Marian F. Ursu, Ian C. Kegel, Doug Williams, Maureen Thomas, Harald Mayer, Vilmos Zsombori, and Mika L. Tuomola. 2008. ShapeShifting TV: interactive screen media narratives. *Multimedia Systems* 14, 2 (July 2008), 115–132. https://doi.org/10.1007/s00530-008-0119-z

[17] Maarten Wijnants, Sven Coppers, Gustavo Rovelo Ruiz, Peter Quax, and Wim Lamotte. 2019. Talking Video Heads: Saving Streaming Bitrate by Adaptively Applying Object-based Video Principles to Interview-like Footage. In *Proceedings of the 27th ACM International Conference on Multimedia (MM '19, to appear)*. ACM.

[18] Maarten Wijnants, Gustavo Rovelo, Peter Quax, and Wim Lamotte. 2016. A Pragmatically Designed Adaptive and Web-compliant Object-based Video Streaming Methodology: Implementation and Subjective Evaluation. In *Proceedings of the 24th ACM International Conference on Multimedia (MM '16)*. ACM, 1267–1276. https://doi.org/10.1145/2964284.2964300