MULTIMEDIA COMMUNICATIONS TECHNICAL COMMITTEE
IEEE COMMUNICATIONS SOCIETY
http://mmc.committees.comsoc.org/

MMTC Communications – Review

IEEE COMMUNICATIONS SOCIETY

Vol. 11, No. 5, October 2020

TABLE OF CONTENTS

Message from the Review Board Directors

Welcome to the October 2018 issue of the IEEE ComSoc MMTC Communications – Review.

This issue comprises four reviews that cover multiple facets of multimedia communication research including predictive VR rendering, 360 video streaming, 360 video compression, and front-end visual sensing. These reviews are briefly introduced below.

The first paper, published in IEEE INFOCOM 2020 and edited by Dr. Carsten Griwodz, designed a system to reduce the effects of latency on user experience in remotely rendered VR scenes.

The second paper is published in IEEE Transactions on Multimedia and edited by Dr. Mengbai Xiao. It proposes a novel field of view switching mechanism in 360-degree video streaming systems.

The third paper, published in IEEE Transactions on Multimedia and edited by Dr. Roberto G. de A. Azevedo, investigates how to compress the interactive omnidirectional 360-degree content in an efficient way.

The fourth paper, published in IEEE Transactions on Multimedia and edited by Dr. Tiesong Zhao, studies an integrated model reuse and prediction method for digital retina that could potentially benefit the way that computers process and communicate information.

All the authors, nominators, reviewers, editors, and others who contribute to the release of this issue deserve appreciation with thanks.

IEEE ComSoc MMTC Communications – Review Directors

Zhisheng Yan
Georgia State University, USA
Email: zyan@gsu.edu

Yao Liu
Binghamton University, USA
Email: yaoliu@binghamton.edu

Wenming Cao
Shenzhen University, China
Email: wmcao@szu.edu.cn

Phoenix Fang
California Polytechnic State University, USA
Email: dofang@calpoly.edu

Latency Compensation for Remotely Rendering VR Scenes as 360° Video Frames

A short review for "Predictive Scheduling for Virtual Reality"

Edited by Dr. Carsten Griwodz

I-Hong Hou, Narges Zarnaghi Naghsh, Sibendu Paul, Y. Charlie Hu, Atilla Eryilmaz, "Predictive Scheduling for Virtual Reality," IEEE INFOCOM 2020, 10 pages, DOI: 10.1109/INFOCOM41043.2020.9155249

Virtual Reality (VR) is an attractive medium that allows its users highly immersive experiences in real as well as virtual worlds. The market has in recent years provided VR headsets with the same computational power as a mobile phone, enabling them to compute and render views with a very high resolution and with a level of detail comparable to a desktop computer. However, just like mobile phones, these VR headsets cannot compete with the capabilities of a high-end computer, and they are restricted in the amount of content that they can access rapidly for use in the VR scene; due to more limited hardware, there are also details of game physics that are still limited to workstation-class computers that are equipped with better-performing GPUs and are not limited by their energy consumption. Thus, to experience a dynamic VR environment at the highest quality level, these stand-alone VR headsets are forced to rely on remote rendering by more powerful computers and the transmission of the resulting video over a wireless network. The challenge of this approach is that actions by the wearer of the VR headset must first travel to that computer, the VR scene must be rendered, transmitted to the VR headset, and subsequently rendered. The latency budget for doing all of these steps without impairing the quality-of-experience for the user is quite low.

The authors of the article "*Predictive Scheduling for Virtual Reality*" [1] that is featured in this review attack this challenges with a series of innovative steps. They are not attempting to avoid latency, they are making a series of decision to reduce its impact on user experience.

The first step in their approach is to use 360° video instead of rendering a specific view to the VR headset. Although the use of 360° video implies that the VR headset cannot provide a stereoscopic depth experience, it provides the user the freedom of rotating their view freely at a specific position. The approach provides the user with 3 degrees of freedom and removes the need for an accurate signaling of the head orientation, leaving only the head's position as a problem.

Like in others works, they are using the predictability of human movement in the virtual world to differentiate between a *proactive phase*, in which data based on prediction of the user's behavior is sent to the VR headset, and a reactive phase, called the *deadline scheduling phase*, in which data that is generated based on user action is sent to improve or perhaps even replace the data sent due to predictions. Although the update cycle between frames for the VR headset is defined by the framerate (60 fps are used for discussions in the paper), the latency that is experienced by the user is considerably reduced because the major share of the newly rendered information is sent ahead of time.

Finally, they make use of a hierarchically layered video codec. This allows them to save a considerable amount of resources when in the proactive phase. A low-quality base layer of the 360° degree representation for every reachable head position can be sent first, before increasing the quality for the most likely movement direction. Finally, in the deadline scheduling phase, the actually taken direction can be refined as much as the permitted by the deadline.

Hou et al. [1] demonstrate that it is possible to find one single optimal solution for the assignment of packets to available wireless network capacity both in the proactive phase and the deadline scheduling phase. The decision yields a value for every single packet under the assumption that two facts are known: (1) the packet's value within the hierarchically coded

multi-layer 360° video, and (2) the probability with which this packet is desired. Of course, this probability depends on the knowledge of user behavior for the proactive phase, whereas it is perfectly known in the deadline scheduling phase.

It must be noted that this approach, in the form that is presented by the authors, depends on a VR experience that partitions the movement into a small number of discrete movements. However, this is not in any way a fatal flaw. It is possible to partition the space into a limited number of computed positions and interpolate the remaining positions. Truly free movement without these discrete steps would require view interpolation between different 360° videos, for example Cuda2Video proposed by Zhao et al. [2]. Dedicated hardware support for feature detection as available in recent mobile phones would allow to add it with limited overhead. The proposal by Hou et al. would of course require modifications to ensure that pairs of view with a high likelihood of being used for interpolation should receive similar quality layers. This provides potential for future studies.

It is furthermore noteworthy that the authors of "*Predictive Scheduling for Virtual Reality"* evaluate the value of packets only based on the future probability of the VR user's movement. This is somewhat problematic because it means that a mis-prediction, which will lead to a drop of visual quality due to the limited duration of the deadline scheduling phase, is most likely followed immediately by a high-quality frame. We know from quality studies on video that quickly changing video quality is perceived as flicker by users [3], and that the resulting QoE of flickering video is lower than that of a video that retains the lower quality for a small number of seconds (1-3 seconds). Although considerations for this can be easily added to the proposed technique, it will be a matter of future work to investigate how frequent mis-predictions and the resulting brief quality oscillation can be without being perceived as flicker.

After having demonstrated the possibility of finding an optimal solution, Hou et al. transform it into an online policy, which they subsequently put to the test. First, they conduct a simulation study, finally, even a prototype study.

They retain the somewhat dangerous assumption that the prediction of future movement is independent of a user's movement history. It is unlikely that this is true, because the attention of users is directed by elements in the virtual world. Since this can be used to predict changes in users' viewing directions with high confidence [4], as well as the motion of VR users who navigate with mobile phones [5], it seems reasonable to assume that movement frequently follows visually interesting features also in the case of head-mounted VR usage. However, the omission to mention this in the paper [1] does not affect its validity, since the cited works [4] [5] imply that the probabilities can be derived server-sided directly from the rendered content. In [1], this is known from the previous frame's deadline scheduling phase, and the paper permits an interpretation of forward, backward and sideways motion after applying the head rotations observed in the previous frame.

The simulation study includes a simplification that is as such not present in the optimal algorithm proposed by the authors. In the rather aggressive layer compression that is proposed for the simulation, individual packets are assigned values, and these values are chosen as equal for all packets of a layer. While such codecs are possible, it is not an ideal choice for a hierarchical codec that is inspired by the MPEG family of codecs, which is implied by citing Seeling and Reislein [6]. These codecs tend to require complete slices, which potentially comprise several packets, to be decoded correctly and contribute to a frame's quality. A difficulty arises from the varying number of packets comprising such a slices. A variation of the optimization algorithm that deals with slices or varying size instead of fixed-sized packets would be a valuable future contribution.

"*Predictive Scheduling for Virtual Reality"* is rounded off with a prototype study that pits the authors' prediction algorithm and its heuristic against existing implementations. The improved quality and scalability compared to the existing solution for splitting the coding effort between a server and the VR client can be clearly demonstrated by the reported data and visualized in an example screenshot.

In summary, Huo et al. have presented a new approach for assigning wireless network resources to reduce the latency between user action and rendering its results in a video presentation of a virtual world, while also increasing the latency. Their work inspires to take

this further in a multitude of directions, including a study of motion interpolation, the prediction of rotations, and further investigations of QoE effects that could be suppressed by considering past as well as future movement decisions. Considerations of stereoscopic effects, which are highly desirable for head-mounted VR displays, could also be investigated with the article as a starting point. With the proliferation of head mounted displays, we can hope that more researchers consider these challenges and add more in-depth studies that are as complete and well-rounded as this paper.

References:

[1] I.-H. Hou, N. Z. Naghsh, S. Paul, Y. C. Hu, and A. Eryilmaz, "Predictive Scheduling for Virtual Reality," in *IEEE INFOCOM 2020 - IEEE Conference on Computer Communications*, 2020, pp. 1349–1358.

[2] Q. Zhao, L. Wan, W. Feng, J. Zhang, and T.-T. Wong, "Cube2Video: Navigate Between Cubic Panoramas in Real-Time," *IEEE Trans. Multimed.*, vol. 15, no. 8, pp. 1745–1754, Dec. 2013.

[3] P. Ni, R. Eg, A. Eichhorn, C. Griwodz, and P. Halvorsen, "Flicker effects in adaptive video streaming to handheld devices," in *Proceedings of the 19th ACM international conference on Multimedia - MM '11*, 2011, p. 463.

[4] C.-L. Fan, S.-C. Yen, C.-Y. Huang, and C.-H. Hsu, "Optimizing Fixation Prediction Using Recurrent Neural Networks for 360° Video Streaming in Head-Mounted Virtual Reality," *IEEE Trans. Multimed.*, vol. 22, no. 3, pp. 744–759, Mar. 2020.

[5] X. Hou, J. Zhang, M. Budagavi, and S. Dey, "Head and Body Motion Prediction to Enable Mobile VR Experiences with Low Latency," in *2019 IEEE Global Communications Conference (GLOBECOM)*, 2019, pp. 1–7.

[6] P. Seeling and M. Reisslein, "Video Transport Evaluation With H.264 Video Traces," *IEEE Commun. Surv. Tutorials*, vol. 14, no. 4, pp. 1142–1165, 2012.

Carsten Griwodz is professor at the University of Oslo. His research interest is the performance of multimedia systems. He is concerned with streaming media, which includes all kinds of media that are transported over the Internet with a temporal demands, including stored and live video as well as games and immersive systems. To achieve this, he wants to advance operating system and protocol support, parallel processing and the understanding of the human experience. He was area chair of ACM MM 2019 and 2014, and general chair of ACM MMSys and NOSSDAV (2013), co-chair of ACM/IEEE NetGames (2011), NOSSDAV (2008), SPIE/ACM MMCN (2007) and SPIE MMCN (2006), TPC chair ACM MMSys (2012), and systems track chair ACM MM (2008). More information can be found at http://mlab.no.

Agile FoV Switch in DASH-based 360-degree Video Streaming Systems

A short review for "A Fast FoV-Switching DASH System Based on Tiling Mechanism for Practical Omnidirectional Video Services"

Edited by Mengbai Xiao

J. Song, F. Yang, W. Zhang, W. Zou, Y. Fan and P. Di, " A Fast FoV-Switching DASH System Based on Tiling Mechanism for Practical Omnidirectional Video Services," IEEE Transactions on Multimedia, vol. 22, no. 9, Sep. 2020.

In recent years, the 360-degree videos are rapidly moving towards the mainstream due to the immersive experience provided to users. However, this experience comes at a cost of prohibitively high bandwidth demands. The resolution of 360-degree videos is suggested to be 8K or even higher for a good experience [1] and thus makes streaming this type of video over Internet more challenging. Since the audience of 360-degree video only watches a small portion of the omnidirectional scene, the approach to effectively reducing the bandwidth requirement is the viewport adaptive streaming, i.e., the content in the user's field of view (FoV) is streamed with high quality while the invisible part is delivered with low quality. As a result, one of the most effective solutions is the tile-based framework. In this framework, the frames are spatially divided into tiles that are encoded independently. So, the tiles with different quality levels could be selected on the fly for constituting a complete scene, leading to high flexibility. However, implementing an efficient and satisfying tile-based streaming system is hard. First, multiple decoders need to be setup for decoding the tiles, consuming too many resources. Second, the visual quality in the FoV only adapts at the end of a segment if the user's viewport changes, causing unsatisfying quality of experience (QoE).

In this paper, the authors design and implement a Fast FoV-Switching DASH system for 360-degree videos (FFS-360DASH), which is a full-functional tile-based streaming system. In FFS-360DASH, motion-constrained tile set (MCTS) [2] is integrated into the encoder of high efficiency video coding (HEVC) [3], so that tiles can be encoded, transmitted, and decoded independently with one pair of encoder and decoder. Moreover, a tile-merge method is proposed to reuse the visually overlapped tiles in an FoV switch. For the challenge that the visual quality only adapts at the end of a segment, FFS-360DASH prepares multiple copies that start at different time for a segment. Once the client detects an FoV switch, the corresponding copy could be then embedded into the stream, realizing agile adaptation.

More specifically, a 360-degree video is encoded into two layers, namely, the basic layer (BL) and the enhanced layer (EL). In the BL, the full video is encoded at a low bitrate and then is separated into chunks. For the EL, the 360-degree video is spatially divided into tiles and then encoded into segments at multiple quality levels. In FFS-360DASH, the video is partitioned into 88 tiles, in which the tiles close to the polar are down-sampled to eliminate the redundant pixels generated in equirectangular projection (ERP). When encoding the tiled video, the MCTS scheme is incorporated into the HEVC encoder so that inter prediction between tiles is not allowed. This sacrifices some extent of coding efficiency but allows tiles to be independently transmitted and decoded. More importantly, by aligning the tile boundary with the slice boundary, modifying the slice headers, and regenerating the sequence parameter set as well as the picture parameter set, a set of selected tiles could be merged into a Merged EL chunk (MEL) that is decodable on a single hardware decoder. However, the order of tiles in an MEL cannot be changed for correctly decoding. So, in order to reuse the overlapped tiles after the user's viewport change, FFS-360DASH manipulates the merged stream by replacing the invisible tiles

with the new ones, keeping the overlapped tiles intact. For the case that the tile number varies in the new viewport, FFS-360DASH fills the empty slot with its neighbor and discards the content after decoding. It worth noting that the MEL always has 30 slots for tiles, which is the largest possible number of a viewport.

To realize the agile quality adaptation after viewport change, FFS-360DASH prepares 1 original chunk and N−1 derivatives for a video segment with N frames. The derivatives start at different frames and have decreasing frame numbers. With this method, FFS-360DASH could select the most proper derivative and start playing high-quality content anytime. However, it cannot be implemented if the derivatives are encoded in a closed-GOP structure. The Instantaneous Decoder Refresh (IDR) frame of the derivative would clean all states in the decoder, making other P-tiles within the viewport undecodable. The derivatives in FFS-360DASH are encoded in an open GOP structure, where the first frame is a Clear Random Access (CRA) frame that allows other P-tiles to use reference frames before the current one.

FFS-360DASH is a full-functional streaming system built upon HTTP/2. In the system, three connections are established. One connection is used to download the BL chunks, and the EL chunks in the predicted viewport (PEL) are downloaded via another connection. The last connection is activated to deliver instant EL chunks (IEL) only if the user's viewport deviates from the prediction. When streaming the video, individual buffers are installed for the connections. The BL chunks are hold in a long buffer while the PEL and IEL chunks are hold in two short buffers. The bandwidth is fully allocated to download BL chunks if the long buffer is below a threshold, guaranteeing the video is always playable. Only if the long buffer has been sufficiently filled, the bandwidth is used to download EL chunks. In FFS-360DASH, the truncated linear prediction method [4] is adopted for viewport estimation with history used to predict the bandwidth in near future [5].

Extensive experiments are carried out when the bandwidth is set constant, variable, and based on real-world traces. The experiments demonstrate the effectiveness of FFS-360DASH. When compared to a peer system that runs under the traditional DASH framework, FFS-360DASH can adapt the bandwidth fluctuations and viewport changes in a more agile way, reducing the recovery duration of high-quality video by approximately 90%.

In summary, FFS-360DASH is proved to be a practical streaming system for 360-degree videos. The ingenious implementation in the encoder makes the tile-based 360-degree video streaming framework far more promising in the real world. In addition, users' QoE is effectively improved as in FFS-360DASH, high video quality could be randomly accessed at any time instant.

References:

[1] D. You, B. Seo, E. Jeong, and D. H. Kim, "Internet of Things (IoT) for Seamless Virtual Reality Space: Challenges and Perspectives," *IEEE Access*, vol. 6, no. 40, pp. 439–449, 2018.

[2] A. Zare, A. Aminlou, M. M. Hannuksela, and M. Gabbouj, "HEVC-compliant Tile-based Streaming of Panoramic Video for Virtual Reality Applications," *ACM Multimedia*, 2016, pp. 601–605.

[3] G. J. Sullivan, J. Ohm, W. Han, and T. Wiegand, "Overview of the High Efficiency Video Coding (HEVC) Standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1649–1668, Nov. 2012.

[4] L. Sun, F. Duanmu, Y. Liu, Y. Wang, Y. Ye, H. Shi, and D. Dai, "Multi-path Multi-tier 360-degree Video Streaming in 5G Networks," *ACM MMSys*, 2018, pp. 162–173.

[5] T. C. Thang, H. T. Le, A. T. Pham, and Y. M. Ro, "An Evaluation of Bitrate Adaptation Methods for HTTP Live Streaming," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 4, pp. 693–705, 2014.

Mengbai Xiao, Ph.D, is a Professor in the School of Computer Science and Technology at Shandong University, China. He received the Ph.D degree in Computer Science from George Mason University in 2018, and the M.S. degree in Software Engineering from University of Science and Technology of China in 2011. He was a postdoctoral researcher at the HPCS Lab, the Ohio State University. His research interests include multimedia systems, parallel and distributed systems. He has published papers in prestigious conferences such as ACM Multimedia, ACM ICS, IEEE ICDE, IEEE ICDCS, IEEE INFOCOM.

Towards Efficient Compression of Interactive Omnidirectional Visual Content

A short review for "Fine Granularity Access in Interactive Compression of 360-degree Images based on Rate Adaptive Channel Codes"

Edited by Roberto G. de A. Azevedo

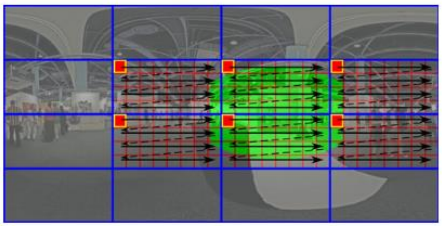
N. Mahmoudian-Bidgoli, T. Maugey, A. Roumy, "Fine granularity access in interactive compression of 360-degree images based on rate adaptive channel codes," IEEE Transactions on Multimedia, Early access, 2020. doi: 10.1109/TMM.2020.3017890.

Recent advances in camera and display hardware coupled with new coding and streaming algorithms have been allowing the widespread adoption of streamed virtual reality (VR) services by end-users. Particularly, the consumption of VR media through head-mounted-displays (HMD) allows an increased sense of presence. As one of the media formats that can support captured VR content, omnidirectional (or 360-degree) videos have been attracting a lot of attention in the multimedia communication research.

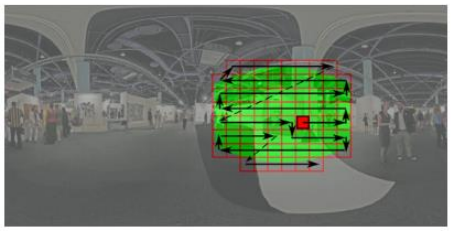
In their paper [1], Mahmoudian-Bidgoli, Maugey, and Roumy advance the state-of-the-art in 360-degree imaging coding by proposing a new interactive compression scheme that aims at replacing the commonly used tile-based methods. In tile-based omnidirectional video streaming, the projected 360-degree content (e.g., using equirectangular or cube map projection [2]) is spatially divided into non-overlapping tiles, which are then independently compressed. During runtime, based on the user's head motion information and viewport prediction algorithms, the client is able to independently request only the tiles that are part of the viewport (or request the ones that are not part of the viewport with less quality). The aim of the authors is to propose a coder that is able to achieve the oracle transmission rate, i.e., obtained as if the user head motion was known at the encoder.

Fig. 1 compares both the tile-based approach (a) and the proposed coding scheme (b). The proposed scheme is based on predictive coding [3]. As can be seen in Fig. 1(a) each tile (rectangles with blue borders) is divided into blocks (rectangles with red borders). Predictive coding is commonly used so that the reference blocks (red rectangles with yellow borders) are independently coded. The remaining blocks are encoded/decoded in a fixed order (black arrows).

In the proposed scheme, a block (rectangles with red borders in Fig.1(b)) can be decoded using any combination of the neighboring blocks that are available at the client side. At the encoder, for each block (rectangles with red borders in Fig.1(b)), a set of predictions is computed, one per set of neighboring blocks, that might be available at the decoder. Instead of reference blocks from the tile-based approach, the authors propose the use of access blocks (red shaded in Fig.1(b)) that can be decoded independently. At the decoder, given a user request (green area in Fig.4(b)) an access block (AB) should be decoded, and then a prediction for neighboring blocks is calculated. As clear in Fig.4, while in the tile-based approaches many blocks that are not in the viewport must be requested/decoded, in the proposed scheme only the blocks in the viewport are requested/decoded.



(a) Tile-based approach



(b) Proposed coding scheme

Fig.1 Comparison between the tile-based approach and [1] proposal.

The authors discuss in details: i) how to place the ABs in the omnidirectional image so that any requested viewport has at least one SB; ii) propose a set of prediction functions based on the neighbors; and iii) discuss how to optimize the decoding order to reach a better storage and transmission efficiency. From a set of ablation studies, they: i) show that the transmission rate is not dependent on the location of ABs; ii) analyze the different decoding orders, concluding that when the camera orientation is horizontal, horizontal predictions should be favored. Moreover, the paper compares the proposed approach to different tile structures (e.g., regular non-overlapping tiles and irregular tiles), to the exhaustive storage approach (i.e., that considers considering all the predictions for each block), and to the theoretical/oracle encoder. Both they analyze the accumulated transmission rate for a user during successive requests and the usefulness of the transmitted data. The results of the experiment clearly show that the proposed scheme performs better than the tile-based approaches.

In conclusion, Mahmoudian-Bidgoli, Maugey, and Roumy propose an interesting and promising approach that can improve the storage and transmission rates of omnidirectional visual content. Although the paper focuses only on images, they have opened new research possibilities that can be explored on the video context as well and in view of the recent video standards and adaptive streaming approaches. Optimizing the block partitioning scheme, intra/inter- prediction schemes, etc. can improve even more the proposed approach, and are some of the envisioned future works. Finally, it will also be interesting to see new developments on how their proposal affect current viewport-based adaptive streaming algorithms.

References:

[1] N. Mahmoudian-Bidgoli, T. Maugey, A. Roumy, “Fine granularity access in interactive compression of 360-degree images based on rate adaptive channel codes,"' IEEE Transactions on Multimedia, Early access, 2020. doi: 10.1109/TMM.2020.3017890.

[2] R. G. A. Azevedo, N. Birkbeck, F. De Simone, I. Janatra, B. Adsumilli and P. Frossard, “Visual Distortions in 360° Videos,” in IEEE Transactions on Circuits and Systems for Video Technology, vol. 30, no. 8, pp. 2524-2537, Aug. 2020, doi: 10.1109/TCSVT.2019.2927344.

[3] M. Wien, High Efficiency Video Coding: Coding Tools and Specifica- tion, ser. Signals and Communication Technology. Springer Berlin Heidelberg, 2014.

Roberto G. A. Azevedo, Ph.D., is a post-doctoral researcher at EPFL working on visual perception, quality of experience, compression, and streaming of immersive media technologies. He holds a Ph.D. (2015) and M.Sc. (2010) degrees in Informatics from PUC-Rio, and the degree of Computer Scientist from the Federal University of Maranhão (UFMA) (2008). From 2008 to 2018, Roberto actively contributed to the specifications and reference implementation for the standards of the Brazilian Digital TV System and ITU-T Recommendations for IPTV middleware, currently adopted by more than 19 countries in Latin America, Africa, and Asia. Roberto’s main research interests are focused on immersive and interactive media, with roots on the intersection of the broad areas of: multimedia systems, human-computer interaction, and computer graphics.

An Integrated Model Reuse and Prediction Strategy for Digital Retina

A short review for "Towards Efficient Front-end Visual Sensing for Digital Retina: A Model-Centric Paradigm"

Edited by Tiesong Zhao

Y. Lou, L. Duan, Y. Luo, Z. Chen, T. Liu, S. Wang and W. Gao. " Towards Efficient Front-end Visual Sensing for Digital Retina: A Model-Centric Paradigm," IEEE Transactions on Multimedia, early access. 15, Jan. 2020.

The increasing number of surveillance cameras in urban area have brought big visual data that benefits intelligent processing of smart cities. To monitor the current dynamics in real-time, the digital retina technology [1] is developed. However, with massive front-end cameras, the joint processing and communication of visual data is still a challenging task.

In digital retina, the interaction between front-end and back-end can be performed with video and feature streams. The video streams are directly captured by ubiquitous cameras, compressed by video encoders (e.g. H.26x [2] , VP-x and AVSx series) and then transmitted via wired/wireless networks. The feature streams are generated by extracting sematic information from video data, coded by feature encoders (e.g. MPEG CDVS/CDVA [3]) and further delivered for computer processing. In general, a video stream provides more intuitive and exhaustive demonstration while a feature stream can be easily delivered and analyzed by city brain.

The authors propose that the interaction via model streams is also necessary and even essential in digital retina. The models are learned at the central server of city brain and then delivered to the front-end cameras for feature extraction and compression. As such, the models behave as the core components in city brain. The generation, utilization and communication of models are thus critical for a model-centric solution in digital retina. To this aim, the authors propose a novel model generation, utilization and communication paradigm by exploiting the cross-domain and inter-model relationships in the construction of digital retina.

Inspired by the popular transfer learning [4], the authors present a multi-model reuse method that transfers the existing models to new model generation, with a mild assumption that the pretrained and target models can be characterized by Convolutional Neural Networks (CNNs). An adaptive weighting method is newly proposed to improve the performance of model reuse for different knowledge correlations between source and target models.

The models are then compressed to save the storage and transmission bitrates. The authors design an end-to-end codec of deep learning models, with a similar methodology to the video coding. First, the Difference of Models (DoM) is incorporated to reduce the redundancy after model reuse. Second, the DoM is quantized for higher compression efficiency, which also results in a lossy compression scheme. After comparison, the linear quantization and vector quantization [5] outperforms other methods. Third, the authors present different quantization levels to support incremental and adaptive delivery of deep learning models.

For completeness, the authors also provide theoretical analysis of the multi-model reuse scheme, which proves the model can benefit from the source models with a theoretical guarantee on expected risk. In experimental section, the authors evaluate their methods on four different person ReID datasets, Duke, Market1501, MSMT17 and CUHK03. Comparison with the state-of-the-art approaches show the superiority of the proposed model reuse and prediction strategy.

Nowadays, the capture, compression and transmission of data mainly focus on the 'natural' information (e.g. multimedia) that can be perceived by human users. While for computer processing, the communication system for

features is still an emerging application. In this work, the authors propose a model communication framework for smart cities, which would inspire the researches in both artificial intelligence and communication techniques.

References:

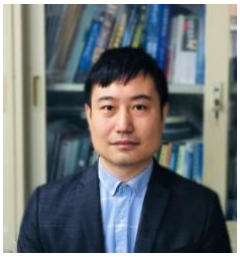
[1] W. Gao and Y. Tian, "Digital Retina: Revolutionizing Camera Systems for the Smart City," Science China: Information Science, vol. 48, no. 8, pp. 1076-1082, 2018.

[2] G. Sullivan, J. Ohm, et al, "Overview of the High Efficiency Video Coding (HEVC) Standard," IEEE Transactions on Circuits and Systems for Video Technology, vol. 22, no. 12, pp. 1649-1668, 2012.

[3] L.-Y. Duan, V. Chandrasekhar, et al, "Compact Descriptors for Video Analysis: The Emerging MPEG Standard," IEEE Multimedia, vol. 26, no. 2, pp. 46-54, 2019.

[4] S. J. Pan, and Q. Yang, "A Survey on Transfer Learning," IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 10, pp. 1345-1359, 2010.

[5] R. M. Gray and D. L. Neuhoff, "Quantization," IEEE Transactions on Information Theory, vol. 44, no. 6, pp. 2325–2383, 1998.



Tiesong Zhao, Ph.D, is a Minjiang Distinguished Professor in Fuzhou University, Fujian, China. He received the B. S. and PhD degree from the University of Science and Technology of China and City University of Hong Kong, in 2006 and 2011, respectively.

His research interests include multimedia signal processing, coding, quality assessment and transmission. Due to his contributions in video coding and transmission, he received the Fujian Science and Technology Award for Young Scholars in 2017. He has also been serving as an Associate Editor of IET Electronics Letters since 2019.

Paper Nomination Policy

Following the direction of MMTC, the Communications – Review platform aims at providing research exchange, which includes examining systems, applications, services and techniques where multiple media are used to deliver results. Multimedia includes, but is not restricted to, voice, video, image, music, data and executable code. The scope covers not only the underlying networking systems, but also visual, gesture, signal and other aspects of communication. Any HIGH QUALITY paper published in Communications Society journals/magazine, MMTC sponsored conferences, IEEE proceedings, or other distinguished journals/conferences within the last two years is eligible for nomination.

Nomination Procedure

Paper nominations have to be emailed to Review Board Directors: Zhisheng Yan (zyan@gsu.edu), Yao Liu (yaoliu@binghamton.edu), Wenming Cao (wmcao@szu.edu.cn), and Phoenix Fang (dofang@calpoly.edu). The nomination should include the complete reference of the paper, author information, a brief supporting statement (maximum one page) highlighting the contribution, the nominator information, and an electronic copy of the paper, when possible.

Review Process

Members of the IEEE MMTC Review Board will review each nominated paper. In order to avoid potential conflict of interest, guest editors external to the Board will review nominated papers co-authored by a Review Board member. The reviewers' names will be kept confidential. If two reviewers agree that the paper is of Review quality, a board editor will be assigned to complete the review (partially based on the nomination supporting document) for publication. The review result will be final (no multiple nomination of the same paper). Nominators external to the board will be acknowledged in the review.

Best Paper Award

Accepted papers in the Communications – Review are eligible for the Best Paper Award competition if they meet the election criteria (set by the MMTC Award Board). For more details, please refer to http://mmc.committees.comsoc.org/.

MMTC Communications – Review Editorial Board

Multimedia Communications Technical Committee Officers

Chair: Jun Wu, Fudan University, China
Steering Committee Chair: Joel J. P. C. Rodrigues, Federal University of Piauí (UFPI), Brazil
Vice Chair – America: Shaoen Wu, Illinois State University, USA
Vice Chair – Asia: Liang Zhou, Nanjing University of Post and Telecommunications, China
Vice Chair – Europe: Abderrahim Benslimane, University of Avignon, France
Letters & Member Communications: Qing Yang, University of North Texas, USA
Secretary: Han Hu, Beijing Institute of Technology, China
Standard Liaison: Guosen Yue, Huawei, USA

MMTC examines systems, applications , services and techniques in which two or more media are used in the same session. These media include, but are not restricted to, voice, video, image, music, data, and executable code. The scope of the committee includes conversational, presentational, and transactional applications and the underlying networking systems to support them.