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

R-LETTER

Vol. 6, No. 1, February 2015

IEEE COMMUNICATIONS SOCIETY

CONTENTS

Message from the Review Board Directors

Welcome to the February 2015 issue of the Review Letter (R-Letter) of the IEEE Communications Society Multimedia Communications Technical Committee (MMTC). This issue is brought to you by review board members who independently nominated research papers published within IEEE MMTC sponsored publications and conferences.

We hope that this issue **stimulates your research in the area of multimedia communication** and an overview of all reviews are provided in the following:

The **first paper**, published in the *IEEE Transactions on Multimedia* and *edited by Carl James Debono*, adopts a semantic approach in order to perform classification of video events.

The **second paper**, published in the *Proceedings of the European Conference on Computer Vision* and *edited by Jun Zhou*, describes a way towards better understanding of deep learning architectures.

The **third paper** is *edited by Ramon Aparicio-Pardo and Gwendal Simon* and has been published within the *Proceedings of the 4th ACM Multimedia Systems Conference*. It provides an answer to the question whether multipath can boost network performance of real-time media.

The **forth paper**, published in the *Proceedings of Network and Operating System Support on Digital Audio and Video Workshop (NOSSDAV '14)* and *edited by Benjamin Rainer and Christian Timmerer*, highlights the understanding of interdependencies of MPEG-DASH clients and caches.

The **fifth paper**, published in the *Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* and *edited by Bruno Macchiavello*, provides an approach for improved view synthesis in a 3D camera space.

The **sixth paper** is *edited by Pradeep K. Atrey* and published in *IEEE Transactions on Multimedia*. It describes a video-in-video advertising method called CAVVA which stands for Computational Affective Video-in-Video Advertising.

Finally, the **seventh paper**, published in *Proceedings of Network and Operating System Support on Digital Audio and Video Workshop (NOSSDAV '14)* and *edited by Pavel Korshunov*, provides a perceptual quality assessment of watching tiled video at mixed resolutions.

We would like to thank all the authors, nominators, reviewers, editors, and others who contribute to the release of this issue.

IEEE ComSoc MMTC R-Letter

Director: Christian Timmerer
Alpen-Adria-Universität Klagenfurt, Austria
Email: christian.timmerer@itec.aau.at

Co-Director: Weiyi Zhang
AT&T Research, USA
Email: wzhang@ieee.org

Co-Director: Yan Zhang
Simula Research Laboratory, Norway
Email: yanzhang@simula.no

Classification of Video Events using Semantics

A short review for "Conceptlets: Selective Semantics for Classifying Video Events"
(Edited by Carl James Debono)

M. Mazloom, E. Gavves, and C.G.M. Snoek, "Conceptlets: Selective Semantics for Classifying Video Events," IEEE Transactions on Multimedia, vol. 16, no. 8, pp. 2214-2228, December 2014.

Video classification provides an important tool for the search and retrieval of multimedia content. With the increasing availability of multimedia content, more accurate classifiers are needed. Classification using semantic descriptors is one possible solution that allows users to utilize more natural queries and, if accurate enough, can lead to better search results. The generation of events, that provide the semantics used for classification, can be acquired through a learning process. To do this, the descriptors are compared to a known data set and, depending on some similarity score, a threshold is used to classify the content. Such comparisons can be very computationally intensive, given that a large descriptor set would be needed to cover the possible semantics that describe the content in random video samples. Thus, solutions that reduce computation complexity while providing good classification are necessary.

Determining and classifying events in videos using intelligent algorithms has been the study of various researchers, such as [1] and [2]. Most classifiers in literature are developed for a particular application and demand models that accurately represent the events of interest. This implies that they cannot be used as a general tool for a broader class of events. Recent work shows that bag-of-words representations can be used for recognition [3]. Furthermore, a number of low-level features were used in [4] and [5] to study their performance in event classification. These indicate that more general descriptors that can recognize events from any content can be developed.

Concept detectors have been applied to video but most of the time these require a database against which the video is compared. Examples of these methods are found in [6] and [7]. Since each concept needs to be checked, this leads to high computational complexity and depend on the number of concept detectors. As the number of concept detectors increases, it becomes very difficult to identify what concepts contribute most information for the classification of the video.

The authors of the original paper develop a solution that uses examples for a given event to learn what concepts provide the most information to include in the bank. They refer to this as conceptlet, as it results in a subset of the whole list of possible concepts. Importance sampling simulation is applied to model the selection of the conceptlet out of the bank. The selection of the conceptlet is done through cross-entropy optimization. This results in a sub-optimal solution.

The algorithm developed to implement the sample and search strategy for the conceptlet is an iterative process. Three steps are performed during each iteration of the process. In the first instance, the concept subsets are sampled based on the parameters obtained in the preceding iteration. The second step uses a scoring function which is applied to the results of the first step. The values are then sorted according to their performance and the best one selected. The third step involves the updating of the parameter vector according to the cross-entropy distance.

The authors test the algorithm on three different datasets, namely; (a) the *TRECVID 2010 multimedia event detection* dataset [8], (b) the *TRECVID 2012 multimedia event detection* dataset [8], and (c) the *Columbia Consumer Video* dataset [9]. In their experiments, the authors report the results showing the influence of individual concepts on classification accuracy, assessment of the performance against the scaling up of concepts within the bank, a comparison of the results of the conceptlets against using all concept detectors, and a comparison of the results obtained using the conceptlets with the cross-entropy algorithm with the conceptlets using the minimum redundancy maximum relevancy algorithm [10] or the L1-regularized logistic regression algorithm [11]. The results reported on the first test verify that some individual concepts are better in representing events than others. The second tests indicate that the accuracy of classification of events improves as more concepts are introduced in the bank. The third experiment showed that the conceptlets give better performance than the bank with all the concepts using much less semantic concepts. Finally, the last tests reported indicate that the authors' proposed solution provides better classification precision when compared to the minimum redundancy maximum relevancy algorithm and the L1-regularized logistic regression algorithm at the expense of more time complexity.

The ability to automatically classify video events using semantics is beneficial for the development of better video search and retrieval tools. Further research is needed in this area to improve the accuracy of the descriptor sets and develop better classifiers. Moreover, complexity needs to be reduced for implementations on resource-limited devices and speed up the search time. Exploiting more parallel architectures and use of cloud computing can potentially help in mitigating time complexities when dealing with all the possible combinations of descriptors and the ever increasing amount of video content on the Internet.

References:

[1] L. Xie, H. Sundaram, and M. Campbell, "Event mining in multimedia stream," in *Proceedings of IEEE*, vol. 96, no. 4, pp. 623-647, April 2008.

[2] Y.-G. Jiang, S. Bhattacharya, S.-F. Chang, and M. Shah, "High-level event recognition in unconstrained videos," *International Journal of Multimedia Information Retrieval*, vol. 2, no. 2, pp.73-101, June 2013.

[3] Y.-G. Jiang, J. Yang, C.-W. Ngo, and A. Hauptmann, "Representations of keypoint-based semantic concept detection: A comprehensive study," *IEEE Transactions on Multimedia*, vol. 12, no. 1, pp. 42–53, January 2010.

[4] Y.-G. Jiang, "Super: Towards real-time event recognition in internet videos," in *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval,* June 2012.

[5] G.K. Myers, R. Nallapati, J. van Hout, S. Pancoast, R. Nevatia, C. Sun, A. Habibian, D.C. Koelma, K.E.A. van de Sande, A.W.M. Smeulders, and C.G.M. Snoek, "Evaluating multimedia features and fusion for example-based event detection," *Machine Vision and Applications*, vol. 25, no. 1, 2014.

[6] S. Ebadollahi, L. Xie, S.-F. Chang, and J.R. Smith, "Visual event detection using multi-dimensional concept dynamics," in *Proceedings of the IEEE International Conference on Multimedia and Expo*, July 2006.

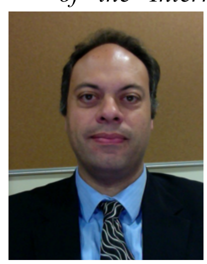
[7] M. Merler, B. Huang, L. Xie, G. Hua, and A. Natsev, "Semantic model vectors for complex video event recognition," *IEEE Transactions on Multimedia*, vol. 14, no. 1, pp. 88–101, January 2012.

[8] NIST TRECVID Multimedia Event Detection (MED) Evaluation Track [Online]. Available: http://www.nist.gov/itl/iad/mig/med.cfm

[9] Y.-G. Jiang, G. Ye, S.-F. Chang, D. Ellis, and A.C. Loui, "Consumer video understanding: A benchmark database and an evaluation of human and machine performance," in *Proceedings of ACM International Conference on Multimedia Retrieval*, April 2011.

[10] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp.1226–1238, December 2005.

[11] A.Y. Ng, "Feature selection, l1 vs. l2 regularization, and rotational invariance," in *Proceedings of the International Conference on Machine Learning*, 2004.



Carl James Debono (S'97, M'01, SM'07) received his B.Eng. (Hons.) degree in Electrical Engineering from the University of Malta, Malta, in 1997 and the Ph.D. degree in Electronics and Computer Engineering from the University of Pavia, Italy, in 2000.

Between 1997 and 2001 he was employed as a Research Engineer in the area of Integrated Circuit Design with the Department of Microelectronics at the University of Malta. In 2000 he was also engaged as a Research Associate with Texas A&M University, Texas, USA. In 2001 he was appointed Lecturer with the Department of Communications and Computer Engineering at the University of Malta and is now an Associate Professor. He is currently the Deputy Dean of the Faculty of ICT at the University of Malta.

Prof. Debono is a senior member of the IEEE and served as chair of the IEEE Malta Section between 2007 and 2010. He was the IEEE Region 8 Vice-Chair of Technical Activities between 2013 and 2014. He has served on various technical program committees of international conferences and as a reviewer in journals and conferences. His research interests are in wireless systems design and applications, multi-view video coding, resilient multimedia transmission, and modeling of communication systems.

Towards Better Understanding of Deep Learning Architecture

A short review for "Visualizing and Understanding Convolutional Networks"
(Edited by Jun Zhou)

*Matthew D Zeiler and Rob Fergus, "*Visualizing and Understanding of Convolutional Networks*", Proceedings of the European Conference on Computer Vision, pp. 818-833, 2014.*

Deep learning has become a phenomenon. Since the milestone paper of Geoff Hinton published in 2006 [1], many researchers have devoted to the advances of theory on this topic [2]. Up to date, deep learning techniques have demonstrated their success in many applications which proves their structural expressive power to a significant extent. For example, on the very challenging ImageNet classification task, deep learning based approaches have outperformed all known conventional image representation methods in terms of the recognition accuracy when large amount of training data are available [3].

Derived from artificial neural networks, the flourish of deep learning research is boosted by three reasons. First, the advances of novel training algorithms enable the efficient handling of complex multiple layers of non-linear processing units for the learning of representation of data. Second, thanks to the development of Internet, large amount of labeled and unlabeled training data become available. Third, parallel computation hardware and software, especially the GPU systems, have greatly speeded up the training of very large models.

Rather than using the manually designed features, deep learning methods automatically learn features from raw pixel values directly. In this way, both low-level and high-level features can be extracted in different layers of the architecture. However, in many cases, the development of deep learning methods is often done empirically, without theoretical analysis on how they achieve good performance. One reason is that analysis tools to foster the understanding of internal behavior of learning models are still missing.

The paper from Zeiler and Fergus aims to address this bottleneck by introducing a visualization technique that projects the feature activations back to the input pixel level. This enables the analysis of feature evolution during training process, identification of potential problems with the model, and further improvement of model architectures.

The reported visualization technique is built on top of a standard deep convolutional neural networks model with eight layers [4]. The first five layers are convolutional networks that contain rectified linear function, max pooling, and contrast normalization steps. The last three layers are fully connected, with the final decision layer being a softmax classifier.

To help the understanding the feature activities in the intermediate layers, a deconvolutional network [5] is integrated into the forward network to map features back to pixels. Given an input image and features computed throughout the layers, deconvolutional process involves a series of operations, e.g., unpooling, rectification, and filtering, to reconstruct the activity in the neighboring layers backwards, after setting all other activations in the layer to zero and using the feature maps as input to the corresponding deconvolutional layer.

Visualization is presented in two aspects, feature visualization and feature evolution during training. The former shows how different structures excite feature maps, and the hierarchical nature of the features in the network. The latter discovers that strong activations are generated by change in images, and that the training of lower and upper layers converges in different speed. This influences the decision on when training shall stop.

The utility of the method is demonstrated by several examples. These include architecture selection such as determination of layer size, and occlusion sensitivity, i.e., whether the model can identify objects. By applying these findings, authors show that they can significantly improve the classification accuracy on several benchmark datasets including the ImageNet.

In summary, this paper is interesting as it provides a practical analysis tool to help the understanding of internal behavior of deep learning networks. This will greatly benefit the research community and industries in developing more powerful image classification systems.

References:

[1] G. Hinton, S. Osindero, and Y. The, "A fast learning algorithm for deep belief nets", *Neural Computa-*

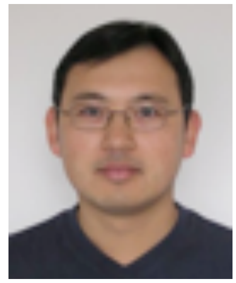
tion, Vol. 18, pp. 1527 – 1554, 2006.

[2] L. Deng and D. Yu D, "Deep Learning: Methods and Applications", *Foundations and Trends in Signal Processing*, Vol. 7, pp. 197-387, 2014.

[3] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. Berg, L. Fei-Fei, "ImageNet Large Scale Visual Recognition lenge". *arXiv*:1409.0575, 2014.

[4] A Krizhevsky, I. Sutskever, G. Hinton, "Imagenet classification with deep convolutional neural networks". *Advances in Neural Information Processing Systems,* pp. 1106-1114, 2012.

[5] M. Zeiler, G. Taylor, R. Fergus, "Adaptive deconvolutional networks for mid and high level feature learning" *International Conference on Computer Vision*, pp. 2018-2025, 2011.



Jun Zhou received the B.S. degree in computer science and the B.E. degree in international business from Nanjing University of Science and Technology, China, in 1996 and 1998, respectively. He received the M.S. degree in computer science from Concordia University, Canada, in 2002, and the Ph.D. degree in computing science from University of Alberta, Canada, in 2006.

He joined the School of Information and Communication Technology in Griffith University as a lecturer in June 2012. Prior to this appointment, he had been a research fellow in the Australian National University, and a researcher at NICTA. His research interests are in statistical pattern recognition, interactive computer vision, and their applications to hyperspectral imaging and environmental informatics.

Can Multipath Boost the Network Performances of Real-time Media?

A short review for "MPRTP: Multipath Considerations for Real-time Media"
(Edited by Ramon Aparicio-Pardo and Gwendal Simon)

Varun Singh, Saba Ahsan, and Jörg Ott, "MPRTP: Multipath Considerations for Real-time Media", in Proc. 4th ACM Multimedia Systems Conference (MMSys '13), Oslo, Norway, Feb. 2013

There are multiple routes between two hosts in the current Internet. This statement tends to be even truer when considering the flattening Internet topology, where Internet Service Providers (ISPs) have multiple options to reach a distant host. It is also truer with the multiple network interfaces available in the modern mobile devices and the multiple wireless network accesses that co-exist in the urban environment. The question now is about the exploitation of these multiple routes. The network protocols that are in used today stick to the traditional monopath paradigm. Yet, scientists have shown that leveraging multipath can bring many advantages, including better traffic load balancing, higher throughput and more robustness.

This paper, which is already two years old, studies multipath opportunities for the specific case of conversational and interactive communication systems between mobile devices (e.g. Skype). These applications are especially challenging because the traffic between communicating hosts should meet tight real-time bounds. The idea of this paper is to study whether the most widely used network protocol for the applications, namely Real Time Transport Protocol (RTP), can be turned into a multipath protocol. They thus propose a backwards-compatible extension to RTP called Multipath RTP (MPRTP). This paper describes the protocol, which has then been proposed as an IETF drafts [1,2].

In short, this paper presents the MPRTP extension and evaluates its performance in several scenarios. First, the authors comment the main challenges that an extension of RTP protocol must face in order to split a single RTP stream into multiples subflows. Second, the authors present the protocol details as well as the algorithms that are considered to solve these challenges. Third, simulations are conducted to evaluate the performance of the proposal.

Authors point out that a MPRTP protocol should be able to adapt to bandwidth changes on the paths by redistributing the traffic load among them in a smooth way to avoid oscillations. This is especially important in the case of mobile communications where quick capacity changes are common. To guarantee fast adaptation, the authors propose packet-scheduling mechanisms that do not abruptly reallocate traffic among congested and non-congested paths if a path becomes suddenly congested.

Other important issue is the variation on packet inter-arrival time (packet-skew) among the different paths. The fact of having multiple diverse paths make harder to estimate the right buffer size to prevent this issue. To overcome this problem the authors propose an adaptive playout buffer, which individually considers the path skew in each path. They also privilege the selection of paths with similar latencies.

The choice of suitable transmission paths should consider the path characteristics in terms of QoS metrics as losses, latency or capacity. The authors propose several extensions to the RTP protocol, including a new RTP reporting message (where the receiver provides QoS data per sub-flow) and a scheduling algorithm (where the sender uses these reports to decide a traffic distribution among the available paths).

All the aforementioned extensions are always designed to be backwards compatibility, i.e. traditional RTP hosts can interoperate with hosts equipped with MPRTP extensions in single-path scenarios.

An exhaustive battery of simulations is conducted to evaluate the MPRTP performance in a broad range of scenarios: (i) path properties (losses, delays, and capacities) vary along time; (ii) paths share a common bottleneck, and (iii) MPRTP is deployed over mobile terminals using WLAN and/or 3G paths. These evaluations show that (1) the dynamic MPRTP performance is not far from the static performance for single and multipath cases, (2) MPRTP successfully offloads traffic from congested paths to the other ones keeping some proportional fairness among them, and (3) on lossy links multipath is more robust and produces fewer losses with respect to single path.

Overall, this paper addresses a significant problem (how to make a real-time UDP-based protocol multipath) with a comprehensive study. It is one of the first attempts to exploit multipath functionalities in the framework of multimedia communications, and especially with tight real time limitations. This paper thus perfectly completes the works that have been done by the network community on multipath TCP protocols. That being said, many problems related to multipath multimedia protocols are still open. Among others, let us cite rate-adaptive streaming and multi-view video in the context of multipath.

References

[1] V. Singh, T. Karkkainen, J. Ott J., A. Ahsan, L. Eggert, Multipath RTP (MPRTP), AVTCore, IETF, Dec, 2014, https://tools.ietf.org/html/draft-ietf-avtcore-mprtp-00.

[2] V. Singh, J. Ott, T. Karkkainen, R. Globisch, and T. Schierl. Multipath RTP (MPRTP) attribute in Session Description Protocol, 2012. IETF Draft, draft-singh-mmusic-mprtp-sdp-extension.

Gwendal Simon is Associate Professor at Telecom Bretagne. He received his Master Degree in Computer Science in 2000 and his PhD degree in Computer Science in December 2004 from University of Rennes 1 (France). From 2001 to 2006 he was a researcher at Orange Labs, where he worked on peer-to-peer networks and social media innovations. Since 2006, he has been Associate Professor at Telecom Bretagne, a graduate engineering school within the Institut Mines-Telecom. He has been a visiting researcher at University of Waterloo from September 2011 to September 2012. His research interests include large-scale networks, distributed systems, optimization problems and video delivery systems.

Ramon Aparicio-Pardo is a postdoctoral research fellow at Telecom Bretagne. He received a Master Degree in Telecommunications Engineering and a Ph.D. Degree in Information and Communication Technologies from Universidad Politécnica de Cartagena (UPCT), Spain, in 2006 and 2011, respectively. After that, he completed a postdoctoral fellowship with Orange Labs in the academic year 2012-2013 in Lannion, France. Since 2013, he works as a postdoc researcher in the Dept. RSM at Telecom Bretagne. He has visited the Networks Lab, University of California, Davis, lead by Prof. Biswanath Mukherjee, from August to December 2011. His research interests include planning, design and evaluation of communication networks, particularly, optical, multimedia and wireless networks by means of mathematical optimization and combinatorics.

MPEG-DASH and Caches: Understanding the interdependency of MPEG-DASH clients and Caches

A short review for "Caching in HTTP Adaptive Streaming: Friend or Foe?"
(Edited by Benjamin Rainer and Christian Timmerer)

Danny H. Lee, Constantine Dovrolis, and Ali C. Begen, „Caching in HTTP Adaptive Streaming: Friend or Foe?", In Proceedings of Network and Operating System Support on Digital Audio and Video Workshop (NOSSDAV '14), ACM, New York, NY, USA, pp. 31-36.

MPEG Dynamic Adaptive Streaming over HTTP (DASH) is an emerging standard for over-the-top streaming using HTTP [1]. In the last few years this standard has gained more and more popularity because it standardizes only the manifest – referred to as Media Presentation Description (MPD) – that describes the different available representations of multimedia content. All the other components are left open and are subject to research or industry competition. For example, the adaptation logic is not standardized. This component is responsible for selecting an appropriate representation of the multimedia content according to the MPD. MPEG-DASH further foresees that the multimedia content is present in time-bounded chunks (e.g., 2 seconds, 4 seconds, 6 seconds, etc.) called segments [2].

The authors investigate the oscillation of the multimedia representation that can occur if a cache (e.g., proxy with cache) is between the client and the content provider when using a rate-based adaptation logic. Suppose that a cache is between a client and the content provider. The problem occurs if the available bandwidth between the client and the cache is greater than between the cache and the content provider, and if the current representation is available in the cache but the other representations are not cached. Then, it may occur that the adaptation logic switches to another representation, due to the high bandwidth available. The newly selected representation requires more bandwidth than the previous once. Since, the cache is not holding any segments of this representation, the next segment has to be fetched from the content provider. The client measures the bandwidth again during downloading the next segment and the adaptation logic detects the lower available bandwidth and switches back to its initial representation. Nevertheless, this representation can be fed from the cache and the game starts over. The selected representation begins to oscillate from segment to segment. This problem has been earlier investigated by [3].

The authors propose a new video-aware cache called **Vi**deo **S**haping **I**ntelligent **C**ache (ViSIC). The idea of the proposed approach is to shape the traffic such that it does not exceed the path bandwidth. This shall ensure that oscillations, as described before, are minimized. The authors introduce an algorithm that measures the bandwidth of the client and the bandwidth between the cache and the corresponding content provider. The cache shapes the request for the segments by selecting another representation.

The authors further evaluate their approach against a simple cache and a scenario without a cache. The results are analyzed with respect to the requested representation in terms of bitrate, switching stability (i.e., the ratio between consecutive representation switches and the number of segments), and the buffer fullness. The results clearly show that the ViSIC provides the best performance for the selected metrics. In particular, ViSIC is able to mitigate the oscillation effect.

Nevertheless, the evaluation presented in this paper uses only a single client and only one cache between the client and the content provider. In the open Internet there are several clients that may hit the same caches and there are several caches within the path from a client to the content provider. In [3] at least two clients were used but this is still not enough to fully understand the occurring interdependencies. Therefore, in [4] a completely different approach is presented that tries to mitigate the oscillation effect by deriving the representation from the variation of the playback buffer.

In general, however, there is the need for a test-bed that reflects the reality to a certain extent. Although there exists a dataset for MPEG-DASH content [5], there is currently a lack of a dataset that comprises statistics about DASH traffic on caches, clients, and origin servers. Such a dataset would allow simulations with realistic data and may provide new insights on the interdependencies of many DASH clients and caches.

The research on the interdependencies of DASH clients and caches is not only important in IP based networks. Information Centric Networking (ICN) has been proposed as one possible architecture of the future Internet [6]. The architecture of ICN foresees that each node inherently comprises a cache. Therefore, MPEG-DASH faces the same problems in ICN as in IP based networks which is subject to current and future research.

References:

[1] ISO/IEC 32009-1:2014, "Information technology -- Dynamic adaptive streaming over HTTP (DASH) -- Part 1: Media presentation description and segment formats," *2014*.

[2] I. Sodagar, "The MPEG-DASH Standard for Multimedia Streaming Over the Internet," *IEEE MultiMedia*, vol. 18, no. 4, pp. 62–67, 2011.

[3] C. Mueller, S. Lederer, Christian Timmerer, "A Proxy Effect Analysis and Fair Adaptation Algorithm for Multiple Competing Dynamic Adaptive Streaming over HTTP Clients," In Proceedings of the IEEE Conference on Visual Communications and Image Processing Conference (VCIP 2012) (Kiyoharu Aizawa, Jay Kuo, Zicheng Liu, eds.), IEEE, San Diego, CA, USA, pp. 6, 2012.

[4] X. Zhu, Z. Li, R. Pan, G., H. Hu, "Fixing multi-client oscillations in HTTP-based adaptive streaming: A control theoretic approach," IEEE 15th International Workshop on Multimedia Signal Processing (MMSP), pp. 230-235, 2013

[5] S. Lederer, C. Mueller, C. Timmerer, "Dynamic Adaptive Streaming over HTTP Dataset," In Proceedings of the Third Annual ACM SIGMM Conference on Multimedia Systems, ACM, New York, NY, USA, pp. 89-94, 2012.

[6] V. Jacobson, D. K. Smetters, J. D. Thornton, M. F. Plass, N. H. Briggs, R. L. Braynard, "Networking named content, " In Proceedings of the 5th international conference on Emerging networking experiments and technologies, ACM, New York, NY, USA, pp. 1-12, 2009.

Benjamin Rainer is a PhD candidate at the Institute of Information Technology (ITEC), Alpen-Adria-Universität Klagenfurt, Austria. His research interests include Inter-Destination Multimedia Synchronization, Security and Risk Management, Content Centric Networking, and Quality of Experience. He received his MSc in 2012 from the Alpen-Adria-Universität Klagenfurt.

Christian Timmerer is an Associate Professor at Alpen-Adria-Universität Klagenfurt, Austria. His research interests include the immersive multimedia communication, streaming, adaptation, and Quality of Experience. He was the general chair of WIAMIS'08, QoMEX'13 and ACM MMSys'16. He participated in several EC-funded projects, notably DANAE, ENTHRONE, P2P-Next, ALICANTE, SocialSensor, and COST IC1003 QUALINET. He also participated in ISO/MPEG work for several years, notably in the area of MPEG-21, MPEG-M, MPEG-V, and MPEG-DASH. Follow him on http://www.twitter.com/timse7 and subscribe to his blog http://blog.timmerer.com. In 2012 he co-founded bitmovin.net to provide professional services around MPEG-DASH.

Improved View Synthesis in a 3-D Camera Space

A short review for "Expansion hole filling in depth-image-based rendering using graph-based interpolation"
(Edited by Bruno Macchiavello)

Yu Mao; Cheung, G.; Ortega, A.; Yusheng Ji, "Expansion hole filling in depth-image-based rendering using graph-based interpolation," Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on , vol., no., pp.1859,1863, 26-31 May 2013.

The goal of free viewpoint navigation is to allow the receiver, of a video communication, to freely choose any view from which to observe a three-dimensional (3-D) scene [1]. Therefore, in a free viewpoint system the ability to create different virtual views is necessary. With the advent of depth sensors, it is common to represent visual data of a 3-D scene in a texture-plus-depth format [2]. The depth maps are the per-pixel distance between the captured objects and the capturing sensor. The texture-plus-depth format allows depth-image-based rendering (DIBR) techniques, such as 3-D warping [3]. DIBR is a pixel-to-pixel mapping from a reference view to the desired virtual view. In DIBR, the synthesized location of a pixel is derived from the corresponding depth pixel in the reference view. However, certain locations are not visible at the reference view, which leads to disocclusions. Therefore, DIBR is used jointly with image inpainting algorithms in order to fill the holes created by disoccusions.

In immersive applications, such as video conference, a viewer in a certain position observes a real-time synthesized image on a 2-D display, whose rendering perspective is adaptively changed in response of the up-to-date tracked head position of the viewer. The resulting motion parallax effect can enhance the viewer's depth perception in the 3-D scene [4]. Besides left-right head movement (x-dimensional motion) there is also front-back motion (head movement along the z-dimension). Motion along the z-dimension can significantly change how the objects in the scene are viewed. When the virtual viewpoint is located much closer to the 3-D scene than the reference view, objects will increase in size. A large increase in object size means that a patch of pixels sampled from an object surface in the reference view will be scattered to a larger spatial area, resulting in expansion holes.

Several previous works focus on inpainting algorithms for disocclusions holes during DIBR rendering [5, 6]. This work is different from previous approaches, since the authors addressed the problem of expansion holes.

First, a method to determine if the hole is an expansion or dissoclusion hole is presented. To achieve such classification, the virtual view is divided into non-overlapping square blocks. Each block is decomposed into depth layers, and the pixels in each layer are processed separately. In order to identify to which depth layer each pixel belongs, a histogram of depth values is constructed for the current block. Then, the local minima in the histogram are used as layer-dividing boundaries. When processing a certain layer, all synthesized pixels of higher layers are treated as empty (holes); this allows erasing a synthesized background pixel from an expansion hole of the foreground object. For each empty pixel, the four closest synthesized neighbors are selected, given that each neighbor should be in a different quadrant. These neighbors are mapped back to the reference view, and the pair-wise distance between them is computed. If two or more neighboring pairs are considered to be very close (the distance between them is below certain threshold), then the empty pixel is considered to belong to an expansion hole.

The intuition behind this method is that if the empty pixel's closest neighbors in the virtual view are nearby pixel in the reference view, then the empty pixel is very likely to be inside the convex set spanned by those neighbors in the reference view.

Once, the expansion holes are identified, two interpolation methods for image inpainting are presented. The first one is a simple linear interpolation. The three nearest non-empty pixels are used to construct a plane, then the empty pixel is interpolated using the constructed plane and its own pixel coordinates. The second, and more complex, method uses Graph-based interpolation. A graph is created for each block, where pixels in the block are nodes in the graph. The weight of edges in the graph can be computed in two different manners: *(i)* it is set to be inverse proportional to the difference in texture value of the connected pixels, if the pixels are both synthesized pixel, or *(ii)* it is set to be inverse proportional to the coordinate distance, if at least one of the pixels is empty. From this graph a set of basis vectors is obtained. In order to do so, a graph Laplacian is defined as the difference between the Adjacency and Degree matrix of the constructed graph. Then, eigen-decomposition is performed on the graph Laplacian in order to obtain the eigen-vectors, which are used as basis vectors. If a signal is projected onto these basis vectors, it becomes the spectral decomposition of the signal given the constructed graph.

Finally, the authors propose a Linear Programming formulation [7] in order to interpolated the empty pixels, such that the difference with the original signal at the synthesized pixel locations is minimized, given the basis vectors. Such formulation can be solved using any one of a set of known Linear Programming algorithms [8]

Experimental results show a significant improvement of up to 3.76dB for Linear Interpolation and 4.25dB for Graph-based interpolation over the inpainting algorithm employed in the reference software VSRS v. 3.5. It is important to mention that in the experiments the authors used only a single pair of texture and depth maps, which is not the situation for which the VSRS software was created. Moreover, the authors modified two sets in the Middlebury database to simulate a significant motion along the *z*-dimension. No real motion was tested. Nevertheless, these results are promising and can help to improve view synthesis in a 3-D camera space.

References:

[1] Tanimoto, M.; Tehrani, M.P.; Fujii, T.; Yendo, T., "Free-Viewpoint TV," Signal Processing Magazine, IEEE , vol.28, no.1, pp.67,76, Jan. 2011

[2] P. Merkle, A. Smolic, K. Mueller, and T. Wiegand, “Multiview video plus depth representation and coding,” in IEEE International Conference on Image Processing, San Antonio, TX, October 2007.

[3] W. Mark, L. McMillan, and G. Bishop, “Post-rendering 3D warping,” in Symposium on Interactive 3D Graphics, New York, NY, April 1997.

[4] S. Reichelt, R. Hausselr, G. Futterer, and N. Leister, “Depth cues in human visual perception and their realization in 3D displays,” in SPIE Three-Dimensional Imaging, Visualization, and Display 2010, Orlando, FL, April 2010.
7

[5] K.-J. Oh, S. Yea, and Y.-S. Ho, “Hole-filling method using depth based in-painting for view synthesis in free viewpoint television (FTV) and 3D video,” in Picture Coding Symposium, Chicago, IL, May 2009.

[6] I. Daribo and B. Pesquet-Popescu, “Depth-aided image inpainting for novel view synthesis,” in IEEE International Workshop on Multimedia and Signal Processing, Saint-Malo, France, October 2010.

[7] G. Cheung, A. Kubota, and A. Ortega, “Sparse representation of depth maps for efficient transform coding,” in IEEE Picture Coding Symposium, Nagoya, Japan, December 2010.

[8] S. Boyd and L. Vandenberghe, Convex Optimization, Cambridge,
2004.

Bruno Macchiavello is an assistant professor at the Department of Computer Science of the University of Brasilia (UnB), Brazil. He received his B. Eng. degree in the Pontifical Catholic University of Peru in 2001, and the M. Sc. and D.Sc. degrees in electrical engineering from the University of Brasilia in 2004 and 2009, respectively. Prior to his current position he helped develop a database system for the Ministry of Transport and Communications in Peru. He also was co-organizer of a special session on Streaming of 3D content in the 19th International Packet Video Workshop (PV2012). His main research interests include video and image coding, image segmentation, distributed video and source coding, multiview and 3D video processing.

CAVVA: A Video-in-video Advertising Method

A short review for "CAVVA: Computational Affective Video-in-Video Advertising"
(Edited by Pradeep K. Atrey)

Karthik Yadati, Harish Katti, Mohan Kankanhalli, " CAVVA: Computational Affective Video-in-Video Advertising," IEEE Transactions on Multimedia, Vol. 16, pp. 15-23, Feb 2014.

Online video collections such as YouTube have been growing rapidly in the public domain. The ever-increasing user base for such platforms has turned them into lucrative markets for advertising various kinds of products/services and brought tremendous interest in online computational advertising. Computational advertising has brought forth problems for which solutions draw from diverse domains of varied fields like economics, machine learning, optimization and statistics. Computational advertising in online videos also draws from general online advertising in search engines, where context sensitive text is presented to users as they search for queries and surf the web. These methods evolved from content-agnostic random advertisement placement to placing relevant advertisements based on keywords in the web page (contextual advertising). Such changes can be seen in advertisements placed in online videos as well. YouTube initially had advertisements which were placed only at the beginning or at the end of the video called the pre-roll and post-roll advertisements respectively. Now, we see a lot of advertisements which are inserted during the play time of the video. Though context has been usually discussed in relation to visual content and semantics, successful computational advertising requires that we also understand and model emotion driven (affective) aspects of consumer behavior.

In this paper, the authors have proposed a in-stream video advertising strategy, CAVVA (Computational affective video-in-video advertising). The authors have looked at computational advertising in a more holistic perspective that the video-in-video advertisements should be harmonious with not only the content of the video, but also the subjective experience of the viewer. In particular, their approach is to actively put humans in the loop along content analysis by measuring viewers' emotional state through pupillary dilation responses.

Context based advertising has been explored earlier for video-in-video advertising in VideoSense [1], where advertisements were placed at algorithmically chosen points in the video based on the global relevance of the video and the advertisement, as well as local relevance defined as similarity between video content and the advertisement. In CAVVA, the authors presented a holistic angle to video-in-video advertising and extend the notion of context to model the affective states of human viewers. The basis of their method stems from the motivation that humans are emotional creatures and consumer decision making processes involve emotional aspects in addition to rational thought. The ideas presented in the paper are also well supported in consumer psychology [2]. CAVVA addresses the seemingly conflicting objectives of minimal disturbance to the user caused by advertisement insertion and maximal engagement with the content.

Affective content can evoke specific emotions in viewers and emotion has been modeled often using the circumplex model [3] of affect which is an emotion space where affect is measured in two dimensions, first being arousal and refers to the intensity of the emotion, and the second being valence, referring to the type of emotion. The authors model affective states of viewers using the following rules:

- In low arousal, low valence (unpleasant) program context, viewers treat the subsequent advertisements as pleasant, opposite to their evaluation of the program [4].
- In high arousal, high valence (pleasant) program context, viewers treat the subsequent advertisements as pleasant, similar to their evaluation of the program [4].
- A positive commercial viewed in the context of a positive program is treated as pleasant, when compared to the same commercial viewed in a negative program context [5].
- Human beings try to overcome their negative mood and they try to maintain their positive mood.

These rules are realized as a non-linear objective function which can be optimized in order to identify the advertisement insertion points and to simultaneously select appropriate advertisements. The input to CAVVA is a video and the target set of video advertisements, the output being a video which has advertisements placed at points in the video, which are selected by optimizing the objective function. The key steps that bring about this transformation in CAVVA are:

- Performing scene segmentation of the video as each scene change point is considered as a potential advertisement insertion point in our method.

- Employing the circumplex model of affect and using the method proposed in [3] we compute the arousal and valence scores for each scene, similar scores are computed for the advertisements as well.
- The arousal and valence scores computed in the previous step are provided as input to the optimization framework that identifies the appropriate advertisement insertion points and select the corresponding advertisements.

The authors have evaluated their method on videos and advertisements collected from YouTube. They considered two baseline methods viz; pre-roll/post-roll advertising and VideoSense [1] and also performed subjective evaluation to test the ecological validity of CAVVA, where they asked the users to rate their viewing experience in terms of the disturbance caused because of the advertisement insertion. CAVVA performed favorably in comparison to state-of-art in terms of user disturbance and user engagement with the content. In terms of recall, performance of the proposed method is at par with VideoSense [1].

A possible future direction for CAVVA would be to incorporate the combination of affect and contextual relevance to improve the user's experience as well as maximize the reach of advertising. Additionally, social activity surrounding the video such as user comments and tweets can help gauge the user's emotional state and in-turn provide appropriate advertising to the user.

In summary, the proposed method explores the viewer's emotional in computational video advertising and performs on par or better than the existing video-in-video advertising strategies.

Acknowledgement:

The R-Letter Editorial Board thanks the authors of the paper for providing a summary of its contributions.

References:

[1] T. Mei, X.-S. Hua, and S. Li, "Videosense: a contextual in-video advertising system," IEEE Trans. Cir. and Sys. for Video Technol., vol. 19, no. 12, pp. 1866–1879, Dec. 2009.

[2] K. S. Coulter, "The effects of affective responses to media context on advertising evaluations," Journal of Advertising, vol. 27, no. 4, pp. 41–51, 1998.

[3] J. A. Russell, "A circumplex model of affect," Journal of Personality and Social Psychology, vol. 39, no. 6, pp. 1161–1178, 1980.

[4] M. A. Kamins, L. J. Marks, and D. Skinner, "Television commercial evaluation in the context of program induced mood: Congruency versus consistency effects," Journal of Advertising, vol. 20, no. 2, pp. pp. 1–14, 1991.

[5] T. S. Feltham and S. J. Arnold, "Program involvement and ad/program consistency as moderators of program context effects," Journal of Consumer Psychology, vol. 3, no. 1, pp. 51–77, 1994.

Pradeep K. Atrey is an Assistant Professor at the State University of New York, Albany, NY, USA. He is also an (on-leave) Associate Professor at the University of Winnipeg, Canada. He received his Ph.D. in Computer Science from the National University of Singapore. He was a Postdoctoral Researcher at the MCR Lab, University of Ottawa. His current research interests are in the area of Security and Privacy with a focus on multimedia surveillance and privacy, multimedia security, secure-domain cloud-based large-scale multimedia analytics, and social media. He has authored/co-authored over 95 research articles at reputed ACM, IEEE, and Springer journals and conferences. Dr. Atrey is on the editorial board of several journals including ACM Trans. on Multimedia, ETRI Journal and IEEE Communications Society Review Letters. He was also guest editor for Springer Multimedia Systems and Multimedia Tools and Applications journals. He has been associated with over 50 international conferences/workshops in various roles such as General Chair, Program Chair, Publicity Chair, Web Chair, Demo Chair and TPC Member.

Watching tiled video at mixed resolutions

A review for "Mixing Tile Resolutions in Tiled Video: A Perceptual Quality Assessment"
(Edited by Pavel Korshunov)

Hui Wang, Vu-Thanh Nguyen, Wei Tsang Ooi, and Mun Choon Chan. "Mixing Tile Resolutions in Tiled Video: A Perceptual Quality Assessment," in Proceedings of Network and Operating System Support on Digital Audio and Video Workshop (NOSSDAV), pp. 25-30, 2014.

High definition (HD) TV, currently adopted standard of consumer video, is being replaced by ultra high definition (UHD) TV reaching 4K and 8K resolutions, which offer higher immersiveness and higher quality of experience. Typically, such high quality video is consumed on a large TV or a monitor. However, a growing amount of video content is consumed using mobile devices with screens constrained in size and resolution, limited network bandwidth, and low computation power to stream higher definition video. Therefore, it is challenging to provide for mobile users a quality of experience that is available for TV viewers.

To increase the quality of experience of mobile video consumers, a zoomable video streaming was proposed [2–5] that allows viewers to zoom into the video and view selected regions in finer details and at higher resolutions. To achieve that, video frames can be split into several fixed size tiles of different resolutions, which are independently pre-encoded on the server as a set of different smaller video-tiles. By streaming different combinations of these video-tiles the server can allow a user to have a seamless access to different regions of the original video and at different resolutions.

This paper considers the scenario when a live event is broadcasted to multiple users using a tiled based video of multiple resolutions. To allow different users zoom into different regions of interests of the same live video stream, the authors propose to use the same number of tiles for each resolution of the video. It means that an original video is encoded at the server side into versions with different resolutions, and all these versions are split in the same number of tiles. The same amount of tiles is required at the client side to decode each video frame. Within a frame, however, different tiles could come from different resolution streams. If a tile comes from a stream with resolution lower or higher than requested level, it will be scaled up or down accordingly. In a zoomable video, when a user zooms into a region of interest within the video, the server will first determine the tiles that cover this region, and then associate each tile with an appropriate stream version, depending on their popularity and the resource constraints.

The proposed *mixed resolutions tiling* scheme has the following two essential advantages in tiled video streaming. First, adjusting a tile by scaling it up or down to fit the resolution of a user allows reducing the overall size of video streamed from the server to the users. Next, by intelligently allocating resolution version to each tile, video bandwidth can be reduced without loss in perceived video quality. For instance, tiles requested by many users can come from high-resolution streams, while tiles requested by only a few users can come from low-resolution streams to save the overall bandwidth.

The authors then investigated what is the resolution limit to which a tile can be degraded without a significant loss in visual quality. To that end, the authors conducted a subjective assessment with 50 subjects by using an online-based system. They used three video sequences in original HD resolution (1920×1080 MPEG sequences 'crowd run', 'old town cross', and 'rush hour'), which they compared to the versions consisting of combinations of tiles chosen from the set of five different resolutions, with the lowest being three times less than HD (640×360). The degraded versions were constructed by mixing two resolution levels (half tiles of each). To evaluate the impact of tile size, two ways of tiles splitting was considered, in 16×9 tiles and in 80×45 tiles. The subjective study focused on identifying two perceptual thresholds of video quality difference: Just Noticeable Difference (JND) and Just Unacceptable Difference (JUD) [1] were also computed.

The subjective results demonstrated the feasibility of mixing tiles of different resolutions in the same video, with 80% of subjects not noticing the difference even if the resolution of tiles is decreased by 17% (14–20% in bandwidth reduction) and 85% accepting the decrease in tile-resolution by 33% (25–30% in bandwidth reduction), even though, this decrease was noticed by about half of the subjects. The results showed (expectedly) that mixed resolution tiling scheme is content dependent and is more suitable for video with lower motion. Using smaller tiles also showed to be less noticeable but also less efficient in terms of bandwidth savings and encoding quality compared to when larger tiles are used.

In summary, the paper provides an interesting insight into the perception of video that is constructed out of tiles of different resolutions. Besides zoomable video applications for mobile devices, this study can also be considered in a more general streaming context, showing that a tiled scheme can be effectively used in adaptive streaming techniques, such as DASH. The paper shows that a simple division of a video in tiles encoded at different qualities can be very effective in adaptive streaming, since it can significantly reduce the overall bandwidth with little loss in perceived visual quality.

References:

[1] D Varuna SX De Silva, Warnakulasuriya Anil Chandana Fernando, Gokce Nur, Erhan Ekmekcioglu, and Stewart T Worrall. 3d video assessment with just noticeable difference in depth evaluation. *In Image Processing (ICIP), 2010 17th IEEE International Conference on*, pages 4013–4016. IEEE, 2010.

[2] Wu-chi Feng, Thanh Dang, John Kassebaum, and Tim Bauman. Supporting region-of-interest cropping through constrained compression. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP)*, 7(3):17, 2011.

[3] Aditya Mavlankar, Pierpaolo Baccichet, David Varodayan, and Bernd Girod. Optimal slice size for streaming regions of high resolution video with virtual pan/tilt/zoom functionality. *In EUSIPCO*, 2007.

[4] Aditya Mavlankar, David Varodayan, and Bernd Girod. Region-of-interest prediction for interactively streaming regions of high resolution video. *In Packet Video*. IEEE, 2007.

[5] Ngo Quang Minh Khiem, Guntur Ravindra, Axel Carlier, and Wei Tsang Ooi. Supporting zoomable video streams with dynamic region-of-interest cropping. *In ACM Multimedia systems Conference*, 2010.

[6] Ray van Brandenburg, Omar Niamut, Martin Prins, and Hans Stokking. Spatial segmentation for immersive media delivery. *In ICIN*. IEEE, 2011.

Pavel Korshunov is a postdoctoral researcher in Multimedia Signal Processing Group at EPFL. He received his Ph.D. in Computer Science from National University of Singapore (NUS). He is a recipient of ACM TOMCCAP Nicolas D. Georganas Best Paper Award in 2011, two top 10% best paper awards in MMSP 2014, and top 10% best paper award in ICIP 2014, he is a co-editor of the new JPEG XT standard for HDR images, and has over 50 publications. His research interests include computer vision and video analysis, video streaming, video and image quality assessment, crowdsourcing, high dynamic range imaging, ultra-high definition imaging, focus of attention, and privacy issues in video surveillance systems.

Paper Nomination Policy

Following the direction of MMTC, the R-Letter 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 R-Letter Editorial Board Directors:

Christian Timmerer (christian.timmerer@aau.at), Weiyi Zhang (wzhang@ieee.org), and Yan Zhang (yanzhang@simula.no).

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 R-letter quality, a board editor will be assigned to complete the review letter (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 letter.

R-Letter Best Paper Award

Accepted papers in the R-Letter 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://committees.comsoc.org/mmc/rletters.asp

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.