MULTIMEDIA COMMUNICATIONS TECHNICAL COMMITTEE
http://www.comsoc.org/~mmc

MMTC Communications - Frontiers

Vol. 11, No. 2, March 2016

CONTENTS

Message from MMTC Chair

Dear MMTC colleagues:

It is time for a change! Our signature E-Letter and R-Letter have been well recognized by our research community over the year. Given the popularity they have gained, we have been advised by IEEE ComSoc to brand them for better exposure. After a long period of consultation and discussion, we have been approved by IEEE ComSoc to brand both letters into one title. From March 2016, we will name the *MMTC E-Letter* as *MMTC Communications – Frontiers*, the *MMTC R-Letter* as *MMTC Communications – Reviews*.

Again, I would like to thank the directors and editors for both letters, for their passionate services that have made our MMTC signature publications one of the greatest successes.

Yonggang Wen
Chair, Multimedia Communications TC of IEEE ComSoc

SPECIAL ISSUE ON INTERACTIVE MULTI-VIEW VIDEO SERVICES: FROM ACQUISITION TO RENDERING

Guest Editors: Erhan Ekmekcioglu, Loughborough University London,
Thomas Maugey, INRIA Rennes Bretagne Atlantique
Laura Toni, EPFL

E.Ekmekcioglu@lboro.ac.uk, thomas.maugey@inria.fr, laura.toni@epfl.ch

Emerging video technologies, such as 360-degree videos[1], virtual reality devices[2][3], and free viewpoint interactive TV[4], have pushed the advent of new immersive and interactive media services that have revolutionized multimedia communications. The users are no longer seen as passive consumers of multimedia content, but have become active players in the communication with multiple dimensions of interactivity: from personalization to social experience. To make online interactive services a reality, media streaming systems need to constantly adapt to users' requests and interaction. To reach this goal, interactive coding and streaming methods have been in the focus of several research communities, including multimedia, communication, and computer vision.

With this Special Issue, we bring together seven papers from these communities, which provide the main challenges and solutions for handling users' interactivity in novel *Interactive Multi-view Video Services* (IMVS). IMVS is predominantly seen as the future in the entertainment multimedia, following the advances in multi-view content generation. Besides entertainment, it will also play a role in increasing the sense of reality and effectiveness in a wide range of areas, such as educational content delivery, telepresence (e.g., remote surgery), manufacturing (e.g., industrial design), and advanced training simulators. In this Special Issue, authors highlight their research findings and perspectives on the different aspects of IMVS: from acquisition to rendering. Authors also cover two multi-view applications describing the technical details of involved steps.

The first two contributions review the open challenges and novel solutions on the coding structure for IMVS, studying the tradeoff between coding efficiency and flexibility in extracting information from coded streams. In "*Merge Frame for Interactive Multiview Video Navigation*", Gene Cheung and Ngai-Man Cheung raise the problem of coding drift caused by view switching in predictive coding. As a solution, they propose the novel concept of merge frame that aims at merging different side information frames to an identical reconstruction.

In the paper titled "*An Information theoretical problem in interactive Multi-View Video services*", Aline Roumy describes novel coding challenges from an information theory perspective. The novelty lies in studying the source data compression with new IMVS-constraints, namely massive numbers of user requests and random access of specific views due to the heterogeneity of the interactive population.

In "*Free Viewpoint Video Streaming: Concepts, Techniques and Challenges*", Árpád Huszák provides an overview of the novel solutions in IMVS from a networking perspective. The main challenge is to provide to the clients the viewpoints of interest with a minimum view-switching delay under limited network resources. The optimal grouping of multicast trees and in-network viewpoint synthesis functionalities are investigated by the author.

Federica Battisti and Patrick Le Callet elaborate on how the Quality of Experience in interactive Free-Viewpoint TV services should be measured in their paper titled "*Quality Assessment in the context of FTV: challenges, first answers and open issues*". The authors summarize the limitations of existing quality assessment methods, mainly developed for 2D sequence, when measuring the quality of free-viewpoint interaction. They also describe a roadmap to develop effective quality measurement protocols.

[1] https://www.google.com/get/cardboard
[2] https://www.oculus.com/en-us/rift
[3] https://www.lytro.com/immerge
[4] http://www.bbc.co.uk/rd/projects/iview

In the paper titled "*3D Visual Attention for Improved Interaction, Quality Evaluation and Enhancement*", Chaminda T.E.R. Hewage describes the use of 3D visual attention in free-viewpoint interactivity, as well as in quality evaluation and quality enhancement. How the visual attention model is extracted, and the differences between the 2D and 3D visual attention models are outlined. Novel use cases are depicted where the visual attention information is exploited for better results.

The last two papers focus on concrete applications of IVMS systems. In "*RE@CT: Immersive Production and Delivery of Interactive 3D Content*", Marco Volino, Dan Casas, John Collomosse and Adrian Hilton build an end-to-end system in which real characters are captured, synthetically reconstructed and animated via computers. The paper presents efficient tools that have been developed in the RE@CT project in order to improve the efficiency and accuracy of the whole processing chain. They finally discuss how such results could be beneficial for the future 3D cinema production.

In the paper entitled "*SceneNet: Crowd Sourcing of Audio Visual Information aiming to create 4D video streams*" written by Dov Eilot et. al., the use case is different, namely a user-generated IMVS. The work relies on the observation that more and more scenes (concerts, sports events, etc.) are captured from different angles by the devices (e.g., smartphones) of people in a crowd. The SceneNet project focuses on gathering and calibrating the acquired data, as well as on reconstructing the 3D point cloud representing the scene.

With this Special Issue we have no intent to present a complete picture on the emerging topic of IMVS systems. However, we hope that the seven invited letters give the audience a flavor of the interesting possibilities offered by IMVS, both in terms of novel and exciting research topics and future applications.

Our special thanks go to all authors for their precious contributions to this Special Issue. We would also like to acknowledge the gracious support from the MMTC E-Letter Board.

Erhan Ekmekcioglu received his Ph.D. degree from the Faculty of Engineering and Physical Sciences of University of Surrey, UK, in 2010. He continued to work in the Multimedia Communications Group as a post-doctoral researcher until 2014. Since 2014 he is with Institute for Digital Technologies at Loughborough University London as a research associate. His research interests include 2D/3D (multi-view) video processing and coding, video transport over networks, quality of experience, emerging immersive and interactive multimedia systems, and video analysis for intelligent applications. He has worked in a number of large scale collaborative research projects on 3DTV. He is a co-author of several peer-reviewed research articles, book chapters, and a book on 3DTV systems.

Thomas Maugey received the M.Sc. degree from the École Supérieure d'Electricité, Supélec, Gif-sur-Yvette, France, and from Université Paul Verlaine, Metz, France, in 2007 in fundamental and applied mathematics. He received the Ph.D. degree in image and signal processing from TELECOM ParisTech, Paris, France, in 2010. From 2010 to 2014, he was a Post-Doctoral Researcher with the Signal Processing Laboratory, Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland. Since 2014, he has been a Research Scientist with INRIA in the team-project SIROCCO, Rennes, France. His research interests include monoview and multiview video coding, 3D video communication, data representation, network coding, and view synthesis.

Laura Toni received the Ph.D. degree in electrical engineering in 2009 from the University of Bologna, Italy. During 2007, she was a visiting scholar at the University of California at San Diego (UCSD), CA. After working at the Tele-Robotics and Application department at the Italian Institute of Technology (2009-2011), she was a Post-doctoral fellow in the EE Department at the University of California, San Diego (UCSD) from November 2011 to November 2012. Since December 2012, she has been a Post-doctoral fellow in the Signal Processing Laboratory (LTS4) at the Swiss Federal Institute of Technology (EPFL), Switzerland.

Her research interests are in the areas of interactive multiview video streaming, wireless communications, and learning strategies for optimization problems.

Merge Frame for Interactive Multiview Video Navigation

Gene Cheung and Ngai-Man Cheung
National Institute of Informatics, Tokyo, Japan, Singapore University of Technology and Design
cheung@nii.ac.jp , ngaiman_cheung@sutd.edu.sg

1. Introduction

Advances in image sensing technologies mean that a dynamic 3D scene can now be captured by an array of closely spaced cameras synchronized in time, so users can individually choose from which viewpoints to observe the scene. In an *interactive multiview video streaming* (IMVS) system [1,2], such view interaction can take place between a server and a client connected via high-speed networks: a server pre-encodes and stores multiview video contents *a priori*, and at stream time a client periodically requests switches to neighboring views as the video is played back in time. See in Fig. 1 a *picture interactive graph* (PIG) for a video with three views that illustrates possible navigation paths chosen by users as the streaming video is played back in time uninterrupted.

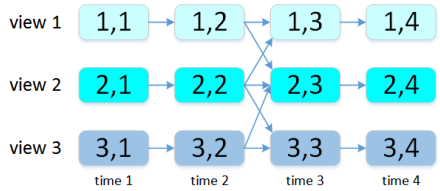


Fig. 1: A *picture interactive graph* (PIG) showing possible view navigation paths for an IMVS system with 3 views

Because the flexibility afforded by IMVS means that a client can take any one of many possible view navigation paths, at encoding time the server does not know which frames will be available at the decoder buffer. This makes differential coding difficult to employ to reduce bitrate. The technical challenge is thus how to facilitate view-switching at stream time while still performing differential coding at encoding time for good compression efficiency.

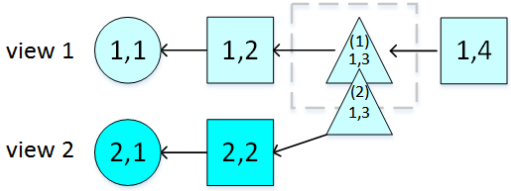


Fig. 2: An example coding structure using SP-frames to enable view-switch at view 1 of instant 3. I-, P- and SP-frames are circles, squares and triangles, respectively.

One possible solution to the view-switching problem is SP-frames in H.264 video coding standard [3]. There are two kinds of SP-frames: *primary* and *secondary* SP-frames. A primary SP-frame is coded like a P-frame, with an extra quantization step after motion compensation so that transform coefficients of each fixed-size code block are quantized to integers. A secondary SP-frame is *losslessly* coded after motion compensation to reconstruct *exactly* the quantized coefficients of the primary SP-frame. Fig. 2 illustrates an example where, at time instant 3, a primary SP-frame $SP_{1,3}(1)$ is encoded to enable switch from view *1* to *1*, and a secondary SP-frame $SP_{1,3}(2)$ is encoded to enable switch from view *2* to *1*. The problem with SP-frames is that the lossless coding employed in secondary SP-frames means that the sizes of secondary SP-frames can be very large—often larger than I-frames.

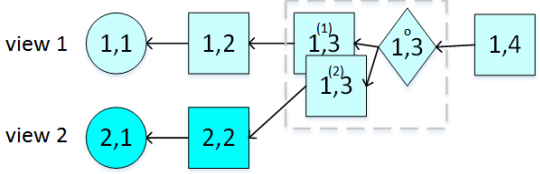


Fig. 3: An example coding structure using optimized target M-frame to enable view-switch at view 1 of instant 3. I-, P- and M-frames are circles, squares and diamonds, respectively.

In a recent work [4,5], using *piecewise constant* (PWC) functions as operators, the authors proposed a new frame type called *merge frame* (M-frame) to efficiently merge slightly different *side information* (SI) frames $\boldsymbol{S}^n$ from

different decoding paths into a unique reconstruction ***M***, so that subsequent frames in time can use the identically reconstructed ***M*** as a common predictor for differential coding. As an example, in Fig. 3, two P-frames $P_{1,3}(1)$ and $P_{1,3}(2)$ of view 1 and time instant 3—these are the SI frames—are first predicted from $P_{1,2}$ and $P_{2,2}$ respectively. An M-frame $M_{1,3}$ of the same time instant is then encoded so that any one of SI frames $P_{1,3}(1)$ and $P_{1,3}(2)$ plus $M_{1,3}$ can result in an identical reconstruction. Subsequent P-frame $P_{1,4}$ can then use $M_{1,3}$ as predictor for differential coding. M-frame thus provides a solution to facilitate view-switches (server sends combo of $(P_{1,3}(1), M_{1,3})$ or combo of $(P_{1,3}(2), M_{1,3})$ depending on user's chosen path), while permitting differential coding to lower bitrate. Experiments in [4,5] show that M-frame can outperform existing switching mechanisms in the literature such as SP-frames [3] in expected and worst-case transmission rate when the probabilities of switching to any views are uniform.

In this paper, we overview the design methodologies of M-frames described in [5] and provide examples of how M-frames can be used practically in IMVS systems.

2. Merge Frame Overview

To reconstruct M-frame, [5] proposed two methodologies. The first is called *optimized target merging*, where the distortion of the reconstructed M-frame ***M*** can be traded off with the encoding rate of ***M***, so long as the identical reconstruction condition from any SI frame $\boldsymbol{S}^n$ is met. The second is called *fixed target merging*, where any SI frame $\boldsymbol{S}^n$ is merged identically to a pre-specified target. We overview the two methodologies here.

In order to merge N different SI frames $\boldsymbol{S}^n$ to a unique reconstruction ***M***, the key idea in [4,5] is to employ a PWC function as a merge operator, whose parameters are explicitly coded in the M-frame, to merge quantized transform coefficients from different SI frames to the same values. Specifically, an SI frame $\boldsymbol{S}^n$ is first divided into fixed-size blocks of K pixels. Each pixel block b is transformed into the DCT domain and quantized into coefficients $X_b^n(k)$. Correct decoding of an M-frame means that the decoder, given only one set of $X_b^n(k)$ from an SI frame $\boldsymbol{S}^n$, can merge $X_b^n(k)$ to identical reconstruction $\overline{X}_b(k)$ via the use of specified PWC functions.

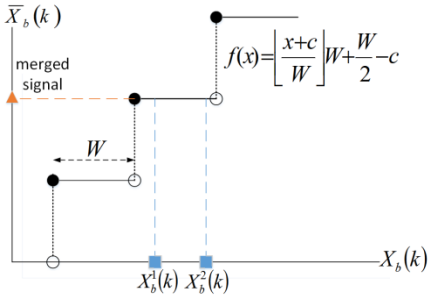


Fig. 4: piecewise constant (PWC) function $f(x)$ to merge quantized coefficients to an identical value

Suppose the floor function $f(x)$ with *shift* c and *step size* W is used to merge coefficients $X_b^n(k)$ of block b from any SI frame $\boldsymbol{S}^n$ to a unique value $\overline{X}_b(k)$:

$$f(x)=\left\lfloor \frac{x+c}{W} \right\rfloor W+\frac{W}{2}-c \qquad (1)$$

That means *any* $X_b^n(k)$ of SI frame $\boldsymbol{S}^n$ must be floored to the same value:

$$\left\lfloor \frac{X_b^1(k)+c}{W} \right\rfloor=\left\lfloor \frac{X_b^n(k)+c}{W} \right\rfloor,\ \forall n\in\{2,\ldots,N\} \qquad (2)$$

(2) is called the *identical merging condition*. Graphically, this also means that they fall on the same step of the floor function, as illustrated in Fig. 4. The optimization is thus to select shift c and step size W for each coefficient k of block b so that (2) is satisfied.

In *fixed target merging*, a desired target value $X_b^0(k)$ is first selected *a priori*, and floor function parameters shift c and step W are then selected to ensure that $X_b^0(k)=f\left(X_b^n(k)\right),\forall n\in\{1,\ldots,N\}$; *i.e.*, coefficients $X_b^n(k)$ of SI frames

$\boldsymbol{S}^n$ merge *identically* to pre-selected $X_b^0(k)$. It is proven [5] that to achieve merging to the desired target, step W_b must satisfy $W_b > 2\max_n \left|X_b^0(k) - X_b^n(k)\right|$ and $c_b = X_b^0(k) \bmod W_b$ for block b. Step size is typically chosen per-frequency for all blocks; W^* is thus chosen as the largest W_b of all blocks b:

$$W^* = 2 + 2\max_b \max_n \left|X_b^0(k) - X_b^n(k)\right| \qquad (2)$$

Thus the coding cost of W^* for all K frequencies is inexpensive. However, shift c_b is chosen per-block per-frequency, and thus the overhead in coding c_b (via arithmetic coding) dominates the coding cost of an M-frame. Since c_b is the remainder of target $X_b^0(k)$, its probability distribution is roughly uniform in $[0, W^*)$, and thus the coding cost of a fixed target M-frame is relatively high.

In *optimized target merging*, there is no pre-selected target value for coefficient $X_b^n(k)$ to converge to: the converged value is selected based on an RD criterion. In this case, it is shown [5] that step W must now satisfy $W > \max_{n,m\in\{1,\ldots N\}} \left|X_b^n(k) - X_b^m(k)\right|$ instead, while c_b is chosen to optimize an RD objective among all values that ensure identical merging condition in (2). This flexibility means that the probability distribution $Pr(c_b)$ can be designed to be skewed (not uniform), resulting in a low coding rate for c_b using arithmetic coding. Thus, in general, an optimized target M-frame is smaller than a fixed target M-frame.

3. Usage of M-frames in IMVS Systems

We now discuss how optimized target M-frames and fixed target M-frames can be used in IMVS systems. When the view-switching probabilities are comparable for all views, the coding structure shown in Fig. 3—called *uniform probability merge* (UPM) structure—is a good design: an optimized target M-frame is constructed to merge differences from P-frames (SI frames) predicted from different navigation paths, which in general is much smaller than a secondary SP-frame.

In the case when the view-switching probabilities are skewed, *e.g.*, when the probability of staying in the same view is exceedingly high, then the UPM structure may not be optimal, because the transmission of an extra M-frame $M^o_{1,3}$ is required for all navigation paths. In this case, authors in [6] proposed a structure called *high probability merge* (HPM) structure using fixed target M-frame (shown in Fig. 5). The most likely path (staying in view 1) involves transmission of a single P-frame $P_{1,3}(1)$. The unlikely path (switching from view 2 to 1) involves transmission of a $P_{1,3}(1)$ and fixed target M-frame $M^f_{1,3}$, where the target for $M^f_{1,3}$ is $P_{1,3}(1)$. While a fixed target M-frame is in general larger than an optimized M-frame, its larger size is offset by the small probability of this navigation path when computing the expected transmission rate.

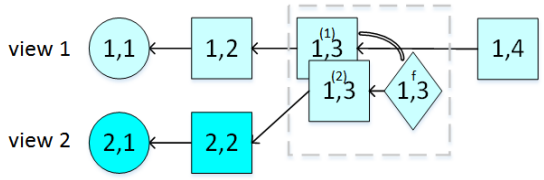


Fig. 5: An example coding structure using fixed target M-frame to enable view-switch at view 1 of instant 3

4. Sample Experimental Results

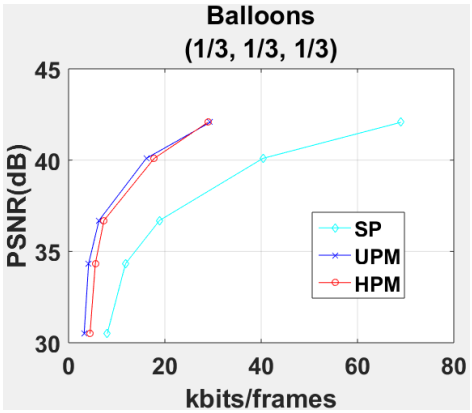


Fig. 6: PSNR versus expected transmission cost for switching to view 2 using SP-frame, UPM and HPM structures. The view switching probabilities are (0.33, 0.33, 0.33).

To demonstrate the superior performance of M-frames and structures that appropriately utilize M-frames, using three views of the multiview video sequence *Balloon*, we computed the expected transmission rates of SP-frames, UPM structure and HPM structure assuming equal view-switching probabilities from view 1 to 2, 2 to 2 and 3 to 2. The resulting PSNR versus expected transmission rate plot is shown in Fig. 6. We observe that for this uniform view-switching probability distribution, as expected UPM performs the best among the three competing schemes at all bitrate regions. This shows the effectiveness of both M-frame and the UPM structure that properly utilizes optimized M-frame.

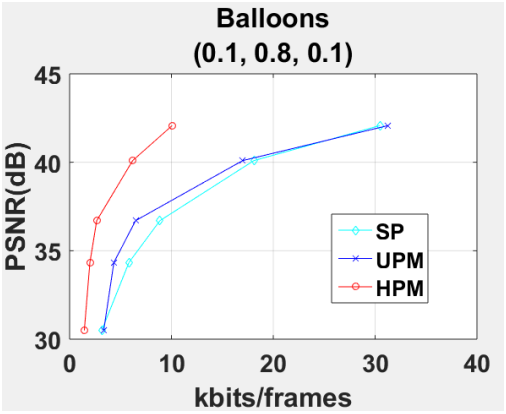


Fig. 7: PSNR versus expected transmission cost for switching to view 2 using SP-frame, UPM and HPM structures. The view switching probabilities are (0.1, 0.8, 0.1).

Conversely, we examine also the RD performance of these three schemes when the view-switching probability distribution is highly skewed (0.1, 0.8, 0.1), where the probability of view 2 to 2 is dominant. The resulting plot of PSNR versus expected transmission rate for the same *Balloon* sequence is shown in Fig. 7. In this case, we observe that HPM is the best structure, outperforming SP-frames and UPM by a wide margin. More extensive results can be found in [6].

5. Conclusion

Designing efficient coding schemes for interactive multiview video streaming (IMVS) systems—where a client can periodically request view-switches from server to navigate to neighboring views as the video is played back in time—is difficult, because at encoding time the server does not know with certainty what frames at the decoder buffer can serve as predictor for differential coding. In this paper, we reviewed recent work on the coding tool and coding structure for IMVS. [4,5] proposed a new design called merge frame (M-frame) to merge different side information (SI) frames to an identical reconstruction, thereby eliminating coding drift caused by view switching. [6] studied different usages of M-frames for different view-switching probabilities. In particular, the authors in [6] proposed new coding structures combining P-frames and fixed and optimized target M-frames that are efficient for skewed view-switching probabilities. For future work, RD optimization methods can be investigated to determine

the optimized coding structures for various view-switching probabilities.

References

[1] G. Cheung, A. Ortega, and N.-M. Cheung, "Interactive streaming of stored multiview video using redundant frame structures," in *IEEE Transactions on Image Processing*, March 2011, vol. 20, no.3, pp. 744–761.

[2] X. Xiu, G. Cheung, and J. Liang, "Delay-cognizant interactive multiview video with free viewpoint synthesis," in *IEEE Transactions on Multimedia*, August 2012, vol. 14, no.4, pp. 1109–1126.

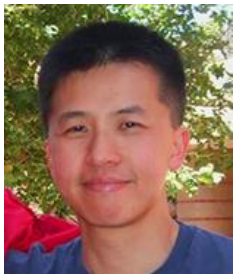
[3] M. Karczewicz and R. Kurceren, "The SP- and SI-frames design for H.264/AVC," in *IEEE Transactions on Circuits and Systems for Video Technology*, July 2003, vol. 13, no.7, pp. 637–644.

[4] W. Dai, G. Cheung, N.-M. Cheung, A. Ortega, and O. Au, "Rate-distortion optimized merge frame using piecewise constant functions," in *IEEE International Conference on Image Processing*, Melbourne, Australia, September 2013.

[5] W. Dai, G. Cheung, N.-M. Cheung, A. Ortega, and O. Au, "Merge frame design for video stream switching using piecewise constant functions," September 2015, http://arxiv.org/abs/1509.02995.

[6] B. Motz, G. Cheung, N.-M. Cheung, "Designing Coding Structures with Merge Frames for Interactive Multiview Video Streaming," submitted to *22nd International Packet Video Workshop*, Seattle, USA, July 2016.

Gene Cheung received the B.S. degree in electrical engineering from Cornell University in 1995, and the M.S. and Ph.D. degrees in electrical engineering and computer science from the University of California, Berkeley, in 1998 and 2000, respectively. He is now an associate professor in National Institute of Informatics in Tokyo, Japan. He has been an adjunct associate professor in the Hong Kong University of Science & Technology (HKUST) since 2015.



Ngai-Man Cheung received his Ph.D. degree in Electrical Engineering from University of Southern California (USC), Los Angeles, CA, in 2008. He has been an Assistant Professor with the Singapore University of Technology and Design (SUTD) since 2012. His research interests are signal, image and video processing.

An Information theoretical problem in interactive Multi-View Video services

Aline Roumy
Inria, Rennes, France
aline.roumy@inria.fr

1. Introduction

Free-viewpoint television (FTV) [1] is a novel system for watching videos, which allows interaction between the server and the user. More precisely, different views of the same 3D scene are proposed and the user can choose its viewpoint freely. Moreover, the user can change its viewpoint at any time, leading to a free navigation within the 3D scene. (Figure 1 shows one example of navigation path within the views). The goal of FTV is thus to propose an immersive sensation, although the visualization remains 2D.

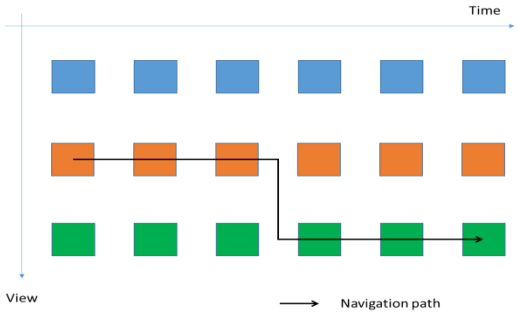


Figure 1. FTV: the user can navigate from one view to another view during playback.

FTV presents nice features that are interesting in sport events like soccer. Indeed, in its application for the 2022 FIFA World Cup[5], Japan announced that the stadium would be equipped with FTV technology.
This soccer example immediately suggests the technical constraints on FTV. First, FTV implies the use of a *large database*. Indeed, the MPEG working group in charge of FTV [2] considers 100 views in order to allow smooth and nice navigation within the scene, which leads to video needing about 120 Gbps (uncompressed HD videos with 50 frames per second and 100 views). In addition, *the number of users is potentially huge* as seen in the soccer example. Finally, in such a scenario, only one view per user is needed at a time, and the choice of this view depends on the user's requests. Therefore, each user *requests a subset of the data*, and this request can be seen as *random* from the sender perspective, as it only depends on the user choice.

2. A novel problem: massive random access to subsets of compressed correlated data

In fact, FTV raises a novel question that has not been yet studied in depth in information theory. This problem can be called massive random access to subsets of compressed correlated data and is illustrated in Figure 2.
This problem can be defined as follows. Consider a database so large that, to be stored on a single server, the data have to be compressed efficiently, meaning that the redundancy/correlation inside the data have to be exploited. The compressed dataset is then stored on a server and made available to users. We consider a scenario in which users want to access only a subset of the data. This is typically the case in FTV. Since the choice of the subset (i.e. the view in FTV) is user-dependent, the request (of a data-subset) can be modeled as a random access. Finally, massive requests are made, meaning that a lot of users may want to access some data at the same time. Consequently, upon request, the server can only perform low complexity operations (such as bit extraction but no decompression - compression).
The novelty of this problem lies in the study of data compression, while adding a new constraint: namely massive and random access. Indeed, classical unconstrained compression does not allow to access part of the compressed data. This is a consequence of the way the encoding map is constructed but also of the optimality that occurs only in the asymptotic regime. More precisely, the source coding map consists in associating to each input sequence an

[5] See the advertisement video at:
https://www.youtube.com/watch?v=KrmbMHJQ_u4

index that is by definition non-separable. Second, optimality occurs when sequences of infinite length are processed. Therefore, accessing part of the data in the classical compression framework can only be achieved in the signal domain but not in the compressed domain.

3. Solutions to the massive random access problem based on classical compression algorithms

Even if incompatible, existing compression algorithms can be adapted to meet the random access constraint. This can be done in one of these three ways:

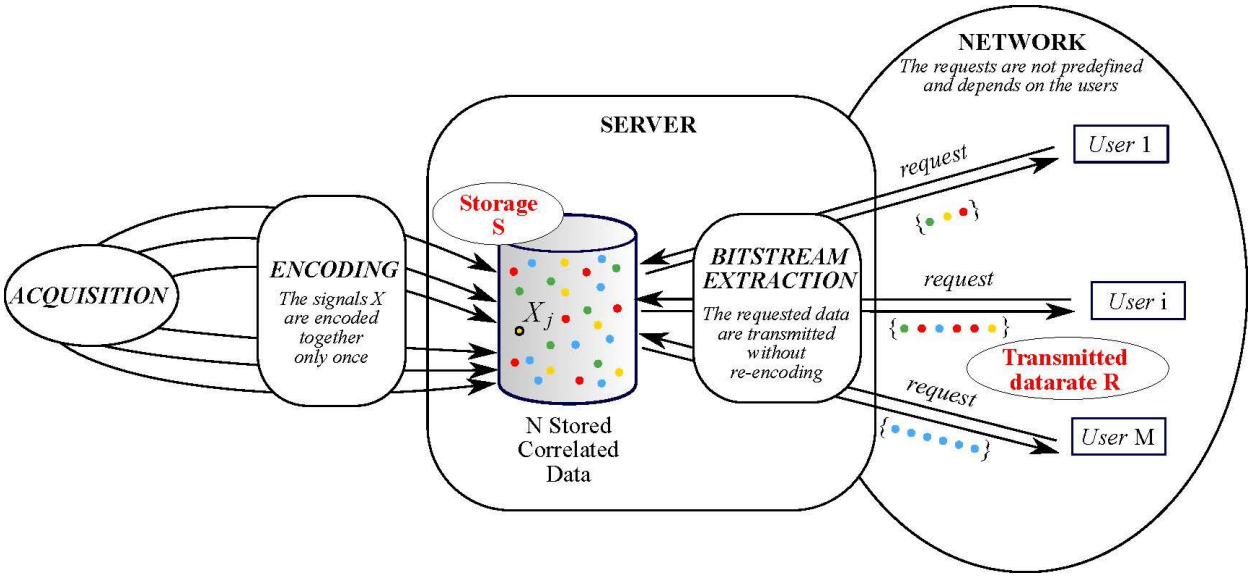


Figure 2. Random access to a database: the user can choose any subset of the compressed correlated data

1. *Send the whole database* as shown in Figure 3. This is the most efficient solution from the compression perspective, but the least efficient from the communication one. Moreover, it might also be infeasible as the required transmission data rate might be larger than the capacity link in many scenarios. As an example, sending 80 views compressed with the best-known compression algorithm for Multiview images (3D-HEVC) requires about 100 Mbit/s [3].
2. *Split the database into chunks and compress each chunk separately* as shown in Figure 4. To allow a smooth navigation, the chunk may only contain a single frame.

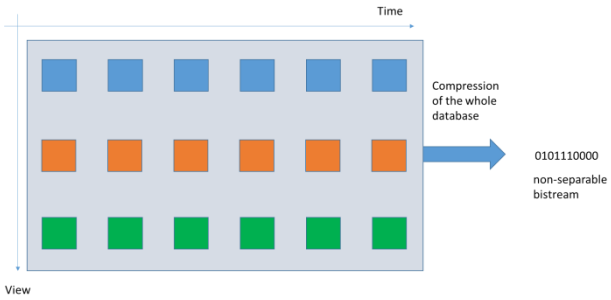


Figure 3. Compression of the whole database.

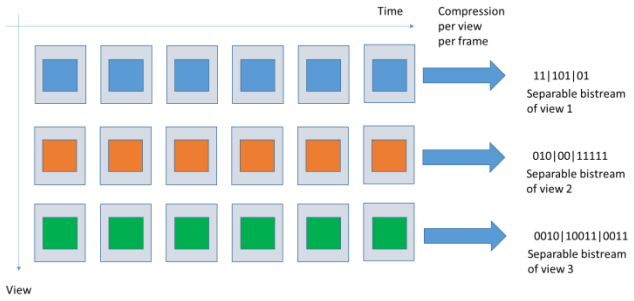


Figure 4. Compression of separate chunks (a chunk is a frame in FTV).

This scheme is inefficient not only from the storage but also the communication perspective. It is inefficient from the communication perspective because the redundancy between the successive frames is not exploited. This requires about *80% more transmission rate.* This loss corresponds to the discrepancy between the inter and the intra frame rates averaged over the Class A and B sequences of the HEVC corpus. Inter coding was performed with GOPsize 8 and IntraPeriod 64.

From the storage perspective, neither the inter-frame correlation nor the inter-view correlation is exploited. The former incurs a loss of 80%, while the latter brings an additional cost of about 40%. Indeed, the latter loss corresponds to the discrepancy observed between simulcast (separate encoding of each view) and multiview encoding [3]. Therefore, about *152% more storage* is required.

3. *Decode the whole database and re-encode the request only*. This is optimal from both compression and communication perspectives. But, in case of massive access, this approach is too complex for the server.

Classical (unconstrained) solutions are either not feasible (case 1 and 3) or suboptimal (case 1 and 2) in a context of massive random access. Therefore, there is a need to design new compression algorithms that take into account the massive random access constraint.

4. Information theoretical bounds for random massive access: the lossless i.i.d. case.

The observation concerning the incompatibility between classical compression algorithms and massive random access (see Section 3) raises the interesting question of the existence of a theoretical tradeoff between massive random access and compression such that a compression algorithm allowing flexibility in the access to the data will always suffer some sub-optimality in terms of compression.

A partial answer to this question is given in our recent work [4], where we showed that there is theoretically no coding performance drop to expect with massive random access, from the communication perspective. The setup considered in [4] is the following. Let $\{X_i(t)\}_{i,t}$ represents the set of frames to be compressed, where $i \in [1,N]$ and $t \in [1,T]$ are the view index and time index, respectively.

For the sake of analysis, a frame (for fixed i and t) is modeled as a random process $X_i(t)$ of infinite length. This infinite length model is indeed a requirement in order to derive information theoretical bounds as compression rates can only be achieved in the asymptotic regime. Note however that the novel video formats (UHD, HDR) tend to produce frames containing a large number of symbols, such that the infinite length assumption is not restrictive.

Now let us assume that at time instants *t-1* and *t*, requested views are *j* and *i*, respectively. Let S_i and $R_{i,j}$ be the storage and the transmission rate (in bits per source symbol) for view *i*, when the previous request is *j*. Let us further assume that, given *i* and *t*, $X_i(t)$ is an independent and identically distributed (i.i.d.) random process and that lossless compression is performed. We now compare three different schemes.

1. Encoding with **perfect** knowledge of the previous request, as shown in Figure 5. This case corresponds to either a non-interactive scheme, where the user cannot choose the view, or to a non-massive random access, where the data are re-encoded at the server upon request. The latter scenario is Scheme 3 described in Section 3. In this case, the necessary and sufficient storage and data rate for lossless compression are:

$$S_i = R_{i,j} = H(X_i(t) \mid X_j(t-1)), \qquad (1)$$

where $H(.|.)$ stands for the conditional entropy. This case is presented for comparison purpose, as the encoding scheme does not satisfy the random massive access constraints.

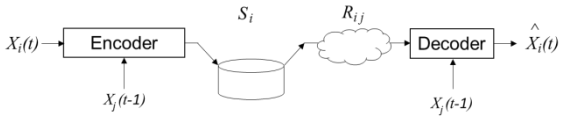


Figure 5. Encoding with **perfect** knowledge of the previous request at the encoder.

2. Encoding with **partial** knowledge of the previous request, as shown in Figure 6. This case corresponds to an interactive scenario. Indeed, at time instant *t*, view *i* is compressed without any explicit reference to a previous request. A trivial upper bound for compression rate is then the unconditional entropy $\mathrm{H}(\mathrm{X}_i(t))$, since the previous request is not known upon encoding. However, compression can be performed under the assumption that one view among the set $\{\mathrm{X}_k(t-1)\}_k$ will be available at the decoder. This allows to reduce both storage and transmission rate from the unconditional entropy $\mathrm{H}(\mathrm{X}_i(t))$ to the conditional entropy in equation (2):

$$\mathrm{S}_i = \mathrm{R}_{i,j} = \max_k H(\mathrm{X}_i(t) \mid \mathrm{X}_k(t-1)), \qquad (2)$$

These rates are necessary and sufficient for lossless source encoding for the scheme depicted in Figure 6, and have been derived in [5]. A practical scheme for video coding has been constructed in [6] based on the insights provided by the achievability part of the theorem in [5].

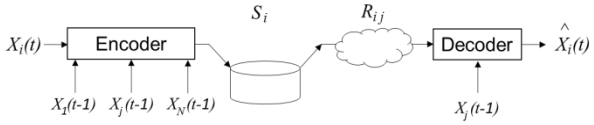


Figure 6. Encoding with **partial** knowledge of the previous request at the encoder.

3. Encoding with **partial** knowledge of the previous request but sending with **perfect** knowledge, as shown in Figure 7. As in Figure 6, compression occurs without any explicit reference to the previous request. This leads to an interactive scheme and requires the same storage as in (2). However, upon request of a particular view at time instant *t*, the server knows the previous request of the same user. [4] shows that this information may be used to lower the transmission data rate from the worst case conditional entropy to the true conditional entropy, see (3), in the case of lossless compression.

$$\mathrm{S}_i = \max_k H(\mathrm{X}_i(t) \mid \mathrm{X}_k(t-1)), \qquad (3)$$

$$\mathrm{R}_{i,j} = H(\mathrm{X}_i(t) \mid \mathrm{X}_j(t-1)). \qquad (4)$$

This result shows that one can efficiently compress the data, while allowing random access to the data. More precisely, it is possible to compress the data in a flexible way such that the server extracts the requested data subset from the compressed bitstream, without the need to decode the whole database and encode the requested subset of data. Surprisingly, the transmission data rate does not suffer any increase even if a flexibility constraint is added. The constraint incurs an additional cost in the storage only.

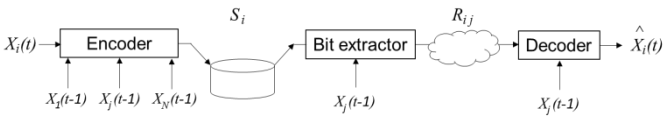


Figure 7. Mixed: **partial** (**perfect**) knowledge of the previous request at the encoder (sender, respectively).

The necessary and sufficient data rates for the three schemes described in this section are shown in Figure 8.

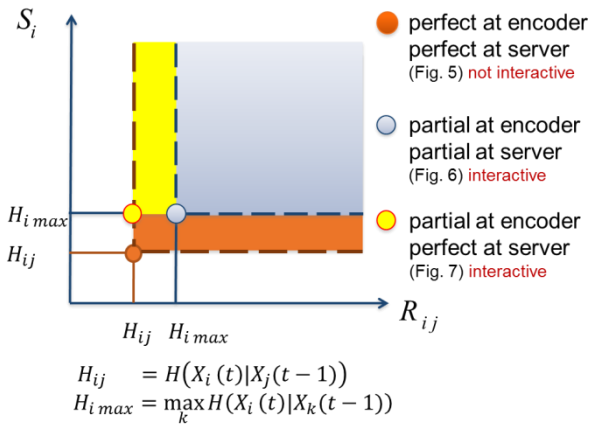


Figure 8. Comparison of the storage and the transmission data rates for the schemes with either perfect (Fig. 5), partial (Fig. 6), or mixed (partial at encoder and perfect at transmitter, Fig. 7) knowledge of the previous user request.

5. Conclusion

In this paper, we reviewed some results on data compression for interactive video communication services. A novel problem has been defined: massive and random access to subsets of compressed correlated data. It was shown that FTV can be seen as an instance of this problem. A very surprising result was stated: from the communication prospective, flexible compression with random access to a subset of the data achieves the same performance as the very complex scheme that performs whole database decompression and data subset compression upon request. This optimality result is, however, only partial, as it concerns lossless (and not lossy) compression for a simplified model with N correlated sources, where each source is i.i.d. There is now a need to extend this preliminary but very promising result.

Acknowledgement

This work has been partially supported by a French government grant given to the CominLabs excellence laboratory (Project InterCom: interactive communication) and managed by the French National Agency for Research (ANR) in the "Investing for the Future" program under reference Nb. ANR-10-LABX-07-01.

The author would like to thank Elsa Dupraz, Michel Kieffer, and Thomas Maugey for stimulating discussions that helped in recognizing that the FTV scenario in [4] is an instance of the more general problem of massive random access discussed in the present paper.

References

[1] M. Tanimoto, M.P. Tehrani, T. Fujii, T. Yendo, "FTV: Free-viewpoint television," *IEEE Signal Processing Magazine*, vol. 27, no. 6, pp. 555–570, Jul. 2012.

[2] Call for Evidence on Free-Viewpoint Television: Super-Multiview and Free Navigation, ISO/IEC JTC1/SC29/WG11 MPEG2015/N15348, June 2015. http://mpeg.chiariglione.org/sites/default/files/files/standards/parts/docs/w15348_rev.docx

[3] G.J. Sullivan, J.M. Boyce, and Y. Chen and J.-R. Ohm, C.A. Segall, and A. Vetro, "Standardized Extensions of High Efficiency Video Coding (HEVC*)," IEEE Journal of Selected Topics in Signal Processing*, vol. 7, no. 6, pp. 1001–1016, 2013.

[4] A. Roumy and T. Maugey, "Universal lossless coding with random user access: the cost of interactivity," in *IEEE International Conference on Image Processing* (ICIP), 2015.

[5] S. C. Draper and E. Martinian, "Compound conditional source coding, Slepian-Wolf list decoding, and applications to media coding," in *IEEE Int. Symp. Inform. Theory*, 2007.

[6] G. Cheung, A. Ortega, and NM. Cheung, "Interactive streaming of stored multiview video using redundant frame structures," *IEEE Trans. on Image Proc*., vol. 3, no. 3, pp. 744–761, Mar. 2011.

Aline Roumy received the Engineering degree from Ecole Nationale Supérieure de l'Electronique et de ses Applications (ENSEA), Cergy, France in 1996, the Master degree in 1997 and the Ph.D. degree in 2000 from the University of Cergy-Pontoise, France. During 2000-2001, she was a research associate at Princeton University, Princeton, NJ. In 2001, she joined Inria, Rennes, France as a research scientist. She has held visiting positions at Eurecom and Berkeley University.
Her current research interests include the area of signal and image processing, coding theory and information theory.

Free Viewpoint Video Streaming: Concepts, Techniques and Challenges

Árpád Huszák
Budapest University of Technology and Economics, Budapest, Hungary
Multimedia Networks and Services Laboratory
huszak@hit.bme.hu

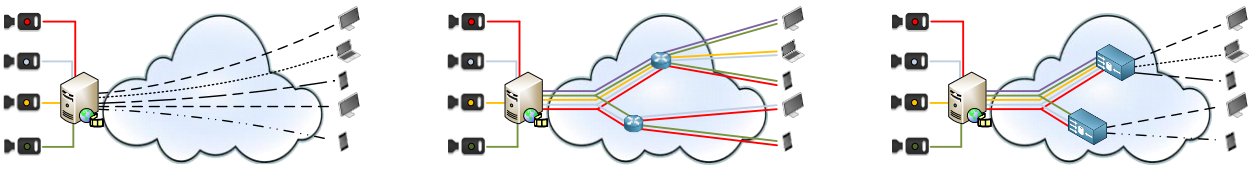
1. Introduction

In contrast to traditional 3D videos, which offer the users only a single viewpoint, Free-Viewpoint Video (FVV) is a promising approach to allow free perspective selection while watching multi-view video streams. The user-specific views dynamically change based on the user's position [1], and they must be synthetized accordingly.
The unique views are synthesized from two or more high bitrate camera streams and the corresponding depth maps [2] that must be delivered over the network and displayed with low latency. By increasing the number of deployed cameras and the density of the camera setup, the free-viewpoint video experience becomes more realistic. But on the other hand, more camera streams require higher network capacity. Therefore, viewpoint synthesis is a very resource hungry process and there is the need to find the best tradeoff between the quality of the synthetized view, which is related to the number of the delivered camera streams, and the processing time of the algorithm.
The required camera streams may change continuously due to the free navigation of viewpoint, hence effective delivery schemes are required to avoid starvation of the viewpoint synthesizer algorithm and keep the network traffic as low as possible. Moreover, packet losses due to congestion and the increased latency can disturb the user experience. In order to support more multi-view videos in IP networks, a simple approach is to minimize the bandwidth consumption by transmitting only the minimal number of camera views, as it was investigated in [3][4][5].

2. FVV architecture models

From architectural point of view, the FVV streaming models can be categorized based on the location of the virtual viewpoint synthesis in the network. The first category depicted in Fig. 1(a) is the server-based model, where all the camera views and corresponding depth map sequences are handled by a media server that receives the desired viewpoint coordinates from the customers and synthesizes a unique virtual viewpoint stream for each user. In this case, only unique free viewpoint video streams must be delivered through the network. The drawback of the server-based solution is that the computational capacity of the media server may limit the scalability of this approach and the service latency is also higher. In case of interactive real-time services, latency is one of the most critical parameters. If remote rendering is used, the control messages must be delivered to the rendering server and the generated stream must be forwarded back to the user, causing significant time gap between triggering the viewpoint change and the synthetized view playout. Moreover, in case of large number of customers, the centralized approach can suffer from scalability issues.

The approach in the second architectural solution, shown in Fig. 1(b), is to deliver reference camera streams and depth sequences directly to the clients to let them generate their own virtual views independently. In this approach the limited resource capacity problem of the centralized media server can be avoided, but huge network traffic must be delivered through the network caused by multiple camera streams. Multicast delivery can reduce the overall network traffic, however the requested camera streams by a user is changing continuously that must be also handled using advanced multicast group management methods. The benefit of client-based architectural model is that it has the lowest latency values, because the viewpoint synthesis is performed locally and the user control can be processed immediately by the rendering algorithm. Unfortunately, rendering FVV video streams at an interactive frame rate is still beyond the computation capacity of most devices, especially in mobile terminals.



(a) server-based (b) client-based (c) distributed

Figure 1. FVV streaming model categories based on the location of the virtual viewpoint synthesis

The third model is a distributed approach, Fig. 1(c), where the viewpoint rendering is done in locations distributed over the network. The user is not connected directly to the media server, but asks for the most appropriate proxy server for a synthetized stream from the desired viewpoint. Remote rendering provides a simple but effective solution, because both bandwidth and computation problems can be solved by synthetizing virtual views remotely on a powerful server at the price of increased latency [6]. Even if the distributed rendering solution can handle some of the bandwidth and computational limitations, new questions arises, *e.g.*, how to optimally design the FVV network architecture.
Possible answers to these questions have been proposed in [7]. Our aim was to find the optimal deployment locations of the distributed viewpoint synthesis processes in the network topology by allowing network nodes to act as proxy servers with caching and viewpoint synthesis functionalities. The other goal was to propose viewpoint prediction based multicast group management method in order to prevent the viewpoint synthesizer algorithm from remaining without any camera streams.

3. Optimized FVV network topology

The distributed approach provides a tradeoff between the server-based architecture and the client-based one, because it can avoid bandwidth and computational resource overloads and handles the user requests in a scalable way. Our goal was to optimize the FVV service topology by minimizing the traffic load without overloading the computational and other resources of the network components. In order to find the optimal arrangement of the distributed viewpoint synthesis model, the network architecture must be overviewed first.
The path between the media server and each client can be divided into two parts: *i)* from the media server to the proxy server, where the real camera streams are delivered and *ii)* from proxy server to the client, where the user specific views are transferred [7]. By locating the viewpoint synthesis functionality closer to the camera sources, the high bitrate camera streams will use less network links, therefore occupying less total bandwidth in the network. On the other hand, the proxy servers will have to serve more clients, so the total network traffic of the unique user specific streams will be higher.
In order to analytically investigate the optimal hierarchical level of the proxy servers, *k*-ary tree is considered. The depth of the tree is *D*, with the source at the root of the tree, while all the receivers are placed at the leaves and the viewpoint synthesis are performed in the proxy servers located δ hops from the root as illustrated in Fig. 2.

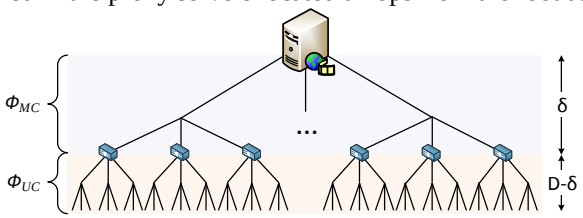


Figure 2. Distributed *k*-ary tree network topology

The goal is to determine the proxy locations to minimize the overall number of link usage:

$$\min\left\{\Phi_{UC}+\Phi_{MC}\right\} \qquad (1)$$

where Φ_{MC} stands for the overall number of multicast links from the media server to the proxy server and Φ_{UC} is the number of unicast links used to deliver user specific streams from proxy server to the client, respectively.
To calculate the number of multicast links (Φ_{MC}) we adapt the results of Phillips et al. [8] to the multi-view video scenario. The unicast part (Φ_{UC}) is easier to calculate. There are D-δ unicast hops from proxy to client as shown in Fig. 2, hence the total number of hops is $\Phi_{UC}=M\cdot(D-\delta)$, where M is the number of users. Assuming n proxy servers placed at level δ in the hierarchical tree, the summarized network resources can be calculated as follows, where c stands for the number of deployed cameras:

$$\Phi_{MC}+\Phi_{UC}=c\cdot\sum_{l=1}^{\delta}k^{l}\left(1-\left(1-k^{-l}\right)^{n}\right)+M\left(D-\delta\right) \quad (2)$$

The number of FVV cameras and the number of users influence the optimal proxy server location. In order to show

how the number of cameras modifies the traffic load in a *k*-ary tree network, we set *k*=3, number of users *M*=1000 and network depth *D*=8. The numbers of occupied links in the delivery paths are shown in Fig. 3.
By increasing the number of cameras, the number of hops and the traffic load increase in the multicast part of the network (Φ_{MC}). Thus, it is worth locating the proxy servers closer to the camera sources. In an opposite case, where there are only three cameras, the lowest number of link usage can be achieved if the view synthesis is performed at level δ=6 that is further from the media server. The *k*-ary tree topology is a simplified layout for analytical investigation. In fact, finding the optimal proxy server locations in dynamically changing network is extremely difficult.

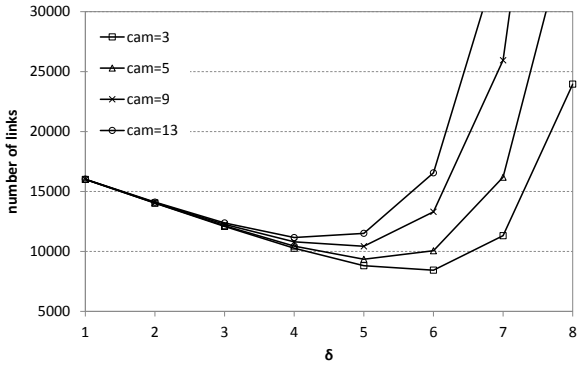


Figure 3. Overall link usage in *k*-ary tree network

The distributed architecture combined with multicast routing can solve the network overload problems and keep the traffic load as low as possible. However the increased latency of control messages can decrease the experienced user quality.

4. FVV multicast

In order to support more multi-view videos in IP networks, a simple approach is to minimize the bandwidth consumption by transmitting only the minimal number of views required. Multicast transmission is effective to reduce the network load, but continuous and frequent viewpoint changes may lead to interrupted FVV service due to the multicast group join latencies. To prevent the user's viewpoint synthesizer algorithm from starving, effective multicast group management methods must be introduced that can rely on viewpoint prediction. Therefore, our aim was to propose a viewpoint prediction based group management solution to minimize the probability of the synthesis process starvation.
Current IP multicast routing protocols (*e.g.*, PIM-SM) exploit shortest path tree logical layout for point-to-multipoint group communication that significantly reduces the network bandwidth.
In case of multicast free viewpoint video streaming each camera view is encoded and forwarded on a separate channel to the users. The separate channels (camera views) can be accessed by joining the multicast group that contains the needed camera source. Users can switch views by subscribing to another multicast channel, while leaving their present one.
If the multicast group change (leaving the old multicast group and joining the new one) happens only when the screen playout of the new virtual view is due, there will be an interruption in the FVV experience, since the lately requested camera view stream will not be received on time to synthetize the new view. Therefore, our aim was to propose a viewpoint prediction based solution for camera stream handoffs to minimize the probability of the synthesis process starvation.
To prevent the user's viewpoint rendering algorithm from starvation, the multicast group join message must be sent in time in order to provide all camera streams that may be requested in the near future. The join message must be sent when the viewpoint coordinates reach a predefined threshold coordinate value. While the viewpoint of the client is within the threshold zone, it will become a member of three multicast groups (*e.g.*, blue, green and yellow), as illustrated in Fig. 4. When the viewpoint coordinates leave the threshold zone, the client should receive only the two required camera streams.

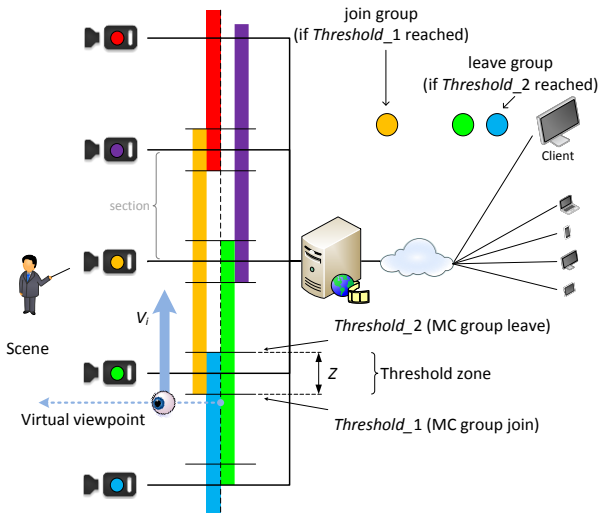


Figure 4. Multicast group join thresholds

An optimization goal can be to keep the threshold area as narrow as possible to reduce the number of multicast group memberships, so that the overall network traffic is reduced, but keep it wide enough to avoid playout interruption during viewpoint changes. In order to find the optimal threshold values, the multicast groups join latency and viewpoint movement features must be considered. Different algorithms can be used for viewpoint estimation such as linear regression or Kalman-filter [7]. To determinate the threshold values and the zone width (Z) of the viewpoint coordinates that trigger the multicast join and leave processes, the required time duration (T_D) from sending a multicast join message to receiving the first I-frame of the camera stream and the estimated viewpoint velocity (v) are used.

$$Z \geq v \cdot T_D \qquad (3)$$

Controlled threshold zone setup can minimize the starvation effect. The comparison of viewpoint velocity values and the caused starvation ratios are presented in Fig. 5.

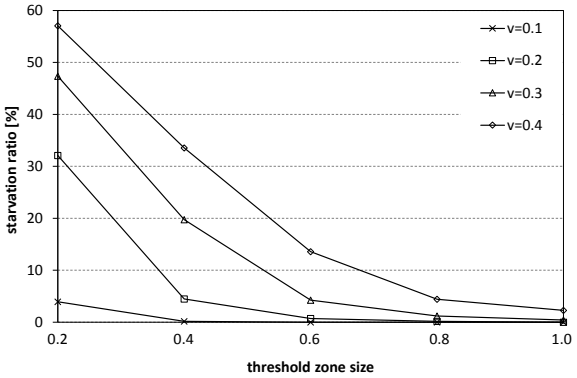


Figure 5. Starvation ratio in case of different velocity values and threshold zone sizes

According to the obtained results, setting the threshold zone too narrow can make the starvation ratio reach to as much as 57%, which renders the FVV service unacceptable. However, using adaptive threshold size can make the synthesizer algorithm get the camera views in time in more than 95% of the cases.

5. Conclusions

Both stream delivery and viewpoint generation are resource hungry processes leading to scalability issues in a complex network with a large number of users. The delivery of high bitrate camera views and depth images required

for viewpoint synthesis can overload the network without multicast streaming, while at the same time, late multicast group join messages may lead to the starvation of the FVV synthesis process. Distributed viewpoint synthesis approach and prediction based multicast group management schemes can offer scalable solutions for new FVV services and hopefully it can become a popular interactive multimedia service of the near future.

Acknowledgement
The author is grateful for the support of the Hungarian Academy of Sciences through the Bolyai János Research Fellowship.

References

[1] Gurler, C.G.; Gorkemli, B.; Saygili, G.; Tekalp, A.M., "Flexible Transport of 3-D Video Over Networks," Proceedings of the IEEE, vol.99, no.4, pp.694,707, April 2011.
[2] Christoph Fehn, "Depth-image-based rendering (DIBR), compression, and transmission for a new approach on 3D-TV", Proc. of SPIE, Vol. 5291, Stereoscopic Displays and Virtual Reality Systems, pp. 93-104, May 2004.
[3] E. Kurutepe, A. Aksay, C. Bilen, C. G. Gurler, T. Sikora, G. B. Akar, and A. M. Tekalp, "A standards-based, flexible, end-to-end multi-view video streaming architecture", in Proc. Int. Packet Video Workshop, Lausanne, Switzerland, Nov. 2007, pp. 302–307.
[4] Li Zuo; Jian Guang Lou; Hua Cai; Jiang Li, "Multicast of Real-Time Multi-View Video," Multimedia and Expo, 2006 IEEE International Conference on , vol., no., pp.1225,1228, 9-12 July 2006.
[5] T.-Y. Ho, Y.-N. Yeh, and D.-N. Yang, "Multi-View 3D Video Delivery for Broadband IP Networks," IEEE International Conference on Communications (IEEE ICC), June 2015.
[6] L. Toni, G. Cheung and P. Frossard, "In-Network View Re-Sampling for Interactive Free Viewpoint Video Streaming, Proceedings of IEEE ICIP, Quebec City, Canada, September 2015.
[7] Árpád Huszák, "Advanced Free Viewpoint Video Streaming Techniques", International Journal on Multimedia Tools and Applications, Springer, ISSN 1573-7721, pp 1-24, November 2015
[8] Graham Phillips, Scott Shenker, Hongsuda Tangmunarunkit, "Scaling of multicast trees: comments on the Chuang-Sirbu scaling law", SIGCOMM '99, New York, USA, 1999.

Árpád Huszák received his M.Sc. degree in 2003 as Electrical Engineer from the Budapest University of Technology and Economics (BUTE) at the Department of Telecommunications (Dept. of Networked Systems and Services since 2013) and completed his Ph.D. in 2010. Currently he is with the Department of Networked Systems and Services as assistant professor, but previously he also worked for the Mobile Innovation Center Hungary (MIK) and Ericsson Hungary. He has been involved in many European projects (FP6-IST, FP7-ICT, and Celtic). His research interests focus on network protocols, mobile computing and adaptive multimedia communications.

Quality Assessment in the context of FTV: challenges, first answers and open issues

Federica Battisti [a] and Patrick Le Callet [b]
[a]Roma Tre University, Rome, Italy; [b]IRCCyN UMR CNRS, Polytech Nantes, France
federica.battisti@uniroma3.it; patrick.lecallet@univ-nantes.fr

1. Introduction

Traditional acquisition systems record the scene from only one point of view while free-viewpoint television (FTV) or free-viewpoint video (FVV) allows the rendering of a complete representation of the scene. To this aim, it is necessary to use several input cameras, ideally as many as the possible viewing positions of the user. Due to the complexity and cost of such camera set up, a tractable approach consists in acquiring or transmitting only few viewpoints, while intermediate or missing points can be obtained by view interpolation. This is possible thanks to view synthesis techniques that exploit the geometric information of a scene, available for example in MVD (Multi-view Video plus Depth) sequences.

FTV is an important step towards interactivity since observers can freely change their point of view while exploring the scene. Such capability comes with new technological constraints that must be assessed not only in terms of perceived quality, but especially from the Quality of Experience (QoE) point of view [1]. In fact, FTV brings new issues in QoE assessment, due to the variety of use cases in which this technique can be employed. Furthermore, the free navigation capability is a completely new task in which user experience and QoE assessment methodology have both to be investigated, revisiting traditional approaches.

In the following, an overview of the artefacts caused by virtual views generation, together with some options for content exploration are presented. Then, the state of the art on testing protocols and objective quality metrics to assess the quality of rendered FTV views is presented. Finally, we describe available datasets that could support the activities of further objective quality measures development.

2. Navigation and FTV: generating new perceptual artefacts

As previously mentioned, one of the greatest challenges in FTV is to provide to the viewers the capability of exploring a scene by freely changing point of view (available or synthesized), as they would normally do in real life. Unfortunately, the process of new viewpoints creation, introduces several artefacts. In particular, it introduces new types of distortion compared to the usual ones considered by the quality assessment community. These are caused mainly by the view synthesis process that can be even emphasized when combined with some artefacts due for example to coding texture and\or depth information of the reference views [2][3][4].

Among them, we can mention flickering, geometry and depth distortions, presence of blurry regions, object shifting, and warping distortions [5].

Beside the need of properly addressing the artefacts generated by view synthesis, the interactivity allowed by the FTV framework poses new challenges in understanding the impact of these impairments on the experience of navigation.

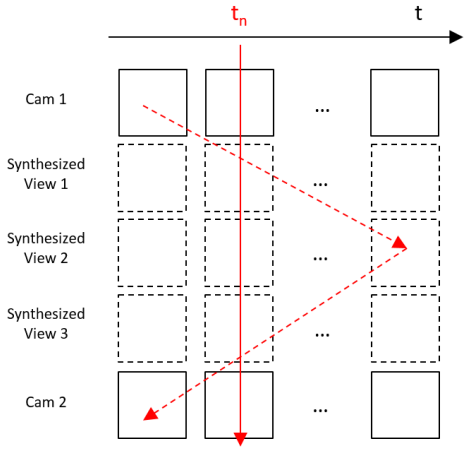


Figure 1. Example of different ways of navigating the FTV content.

Let us consider the simple case depicted in Figure 1. Two original views (Cam 1 and Cam 2) are considered and

from these, three new viewpoints are synthesized.
In this scenario, the user can navigate through the scene by changing viewpoint in several ways: for example, it is possible to fix a time t_n and explore the scene at that specific time, or to navigate through time and views. Changing viewpoint provides an experience equivalent to the exploration of the real 3D scene from several viewpoints. It corresponds at looking at different video streams (recorded and synthesized) inducing not only DIBR related artefacts (*i.e.*, occlusions, incorrect matting) but also non smoothness of the transition among viewpoints in time (*i.e.*, flickering). This changing viewpoint possibility leads to different ways of content fruition and different user experiences thus posing new hard challenges on subjective and objective quality assessment.

3. Towards new protocols to test the effect of coding/transmission and rendering technology on quality of Navigation

The characteristic of good protocols for assessing the quality of media with observers is to obtain reliable and reproducible results. This implies that, to derive MOS (Mean Opinion Score), similar experiments are conducted in well controlled conditions with several observers. This approach is by nature not fully compatible with free interactivity. In fact, in FTV each user can change point of view and everyone has a personalized fruition of the content. Therefore, each user evaluates a different test set and the collected opinion scores are not comparable.
Nevertheless, an interesting intermediate step relies on assessing the quality of single rendered view (image or video content) without allowing switching conditions between views. Focusing only on the presence of DIBR-specific artefacts, the design of methodologies must consider a specific use case [6] to define the suitable requirements such as viewing conditions, characteristics of the source signals, test methods, number of subjects, test duration, and definition of techniques to be used to analyze the test results. A first analysis has investigated the reliability of 2D subjective protocols for assessing the quality of synthesized views [7]. The results of the experimental tests show that ACR-HR (Absolute Category Rating with Hidden Reference) and PC (Pair Comparison) [8] are valid methodologies for comparing the effect of various synthesis algorithms on quality when considering a sufficient number of observers.
A protocol for evaluating the impact of different algorithms for depth map coding for new viewpoints generation has been proposed in [9]. This represents a first step towards being able to deal with specific characteristics of the FTV content such as the presence of jitter while switching views or the consistency of synthesis artefacts along views.
To meet consolidated requirements such as reproducibility, representativeness, and consistency between observers, the idea is to constrain the scenario of interactivity while stressing the most plausible artefacts that interactivity may bring. To cope with these requirements, a predefined trajectory through time and views is defined so that the users can explore the scene from multiple viewpoints but on a fixed path (as in Figure 1). Videos reflecting a possible path in an interactive context can then be generated. In this way, all subjects/observers evaluate the same test set and usual MOS can be easily obtained.
Consequently, this protocol offers to test the effects of distortions caused by view switching that are typically observed in the free interactivity scenario. Nevertheless, this approach does not allow addressing other aspects of QoE such as responsiveness of a system to user expectation. This could be done through extensive user studies but for other aspects of the QoE of the system under test. Note also that in [9], trajectories are generated at a freezing time, *i.e.*, switching between different views acquired at the same time. An interesting study could consider navigation in both time and view directions.

4. Objective measures of quality of navigation: not yet there

Derive reliable objective measures of the image quality produced by FTV systems are highly desirable for efficient technological optimization or benchmarking. Those measures, in order to be effective, should be able to account for the peculiarities of the FTV content. This means that, for understanding the features that impact on the perceived quality, all the steps of the FTV chain, from acquisition\creation to rendering, should be taken into account. In fact, the first problems can occur in the acquisition step where the synchronization and the calibration of the cameras need to be taken into account. After acquisition, the 3D scene representation can cause artefacts due to limited information on the 3D scene geometry but also due to problems related to the reconstruction of shapes and textures (*i.e.*, inconsistent silhouettes due to wrong camera calibrations, color inconsistency across views). Beside this, as mentioned earlier, the synthesis process can introduce distortions due to mismatching textures and the presence of occlusions. Preliminary studies, without considering view switching, analyzed the possibility of using 2D full reference Image Quality Metrics/measures, IQM, such as PSNR and SSIM, included in MeTriX MuX [10]. As expected, these metrics are not able to predict the MOS (correlation coefficient lower than 0.18) due to their limits in addressing the specific artefacts created by the view synthesis process [11]. Those results suggest the definition of specific guidelines [11] to be used in the design of new quality metrics for 3D synthesized view assessment. Based

Database	Characteristics	Subjective score	Testing methodology	On line
DIBR Images	96 images, 3 MVD video sequences, 7 DIBR algorithms, 4 new viewpoints	yes	ACR-HR, PC	yes [18]
MCL 3D Database	693 stereoscopic image pairs, 9 MVD sequences, one DIBR algorithm, distortions on texture and depth	yes	PC	yes [19]
DIBR Videos	102 video sequences, 3 MVD video sequences, 7 DIBR algorithms, 4 new viewpoints	yes	ACR-HR	yes [20]
SIAT Synthesized Video Quality Database	140 video sequences, 10 MVD sequences, 14 texture/depth quantization combinations	yes	ACR-HR	yes [21]
Free-Viewpoint synthesized videos	264 video sequences, 6 MVD sequences, 7 depth map codecs	yes	ACR-HR	yes [22]
High-Quality Streamable Free-Viewpoint Video	5 videos recorded through a dense set of RGB and IR video cameras	no	-	yes [23]
Tanimoto FTV test sequences	7 video sequences (5 still camera, 2 moving camera)	no	-	yes [24]

Table 1. Main characteristics of available datasets.

on the analysis of the performances of twelve IQMs, the authors propose some improvements for the design of new quality metrics. In particular, they suggest that a key factor to be considered is the location of the distortions created by the synthesis process along contours. A first attempt to apply these findings to SSIM reveals an increase in the correlation with the MOS from 0.18 to 0.84 if the consistency of the contours shift is considered, and from 0.18 to 0.78 when the amount of distorted pixels is included in quality assessment.

Some work towards the definition of metrics showing higher correlation with MOS has recently been done in [4][13][14][15][16]. These metrics are characterized by a common approach and they deal with synthesis-related artefacts even if in different domains. They rely on the fact that the synthesized views present a non-natural feeling due to the presence of non-regular geometrical distortions, especially perceivable along the edges. All these metrics show an increased PCC with MOS that is lower than 0.8 for all metrics except for [14] for which PCC can reach 0.9. The quality of single rendered views can then be reliably assessed with objective measures. Nevertheless, switching among views implied by interactivity is still to be considered. In [17] the correlation between the Differential Mean Opinion Score (DMOS) and the values predicted by the usual quality measures is analyzed on the data obtained in [9]. The achieved results show that none of the considered 2D IQMs is able to reliably predict the MOS (PCC lower than 0.26) even if the correlation increases of 3% when content characteristics are taken into account. Up to now, no better objective measures have been proposed. In this case, inconsistencies among views, generating unnatural geometric flicker along time and views, must be considered leaving large space for improvements.

5. Datasets to follow up

The subjective and objective quality assessment processes require the availability of suitable stimuli.
Currently, the community made available few datasets to push this effort forward and summarizes their characteristics. Among these, two include still images while the other five contain video sequences. As detailed in Table 1, the main aspects that have been addressed are the artefacts caused by the application of different DIBR algorithms and the ones that are produced by texture and\or depth map coding. These characteristics are suited to address a real scenario in which there is the need of creating new viewpoints from available data. On the other hand, in [23], the authors present a video dataset recorded through a dense set of RGB and IR video cameras. In this case, no view synthesis is performed and no coding artefacts are considered. The videos are available for download and can be used for experiencing free navigation at high quality. From the analysis performed we can conclude that datasets are yet far from covering all the recent progresses and the possible artefacts that can be encountered in full interactive systems and new efforts need to be devoted to content creation. An immediate requirement relies in the assessment of view switching condition along time. Among the available datasets, the one most approaching view switching fluidity is provided in [22], but it relies on a time freeze condition which is letting aside the combination of both temporal and synthesis artefacts. The availability of a dataset able to address navigability from lag point of view is also extremely desirable.

6. Conclusions

In this paper, the new challenges of quality assessment in the FTV framework have been presented. The creation of FTV content and its navigation open new issues that need to be addressed for properly defining testing

methodologies and objective quality metrics. While promising approaches have been proposed on the front of subjective assessment, there is still some room for further investigation of the interactive scenarios, exploring noticeably more trajectories in the multi view/time space. Concerning objective measures, good tools are available for consistent evaluation of a single rendered view (still image or video). It is of course one little part of the whole as there is not yet satisfying objective measure for tackling the in-between views navigation case. Hopefully, existing datasets can be used to address this challenge which is noticeably under consideration in groups like VQEG Immersive Media Group (IMG) and IEEE P3333.1.

References

[1] W. Chen, J. Fournier, M. Barkowsky, P. Le Callet, "Quality of experience model for 3DTV", in Proc. SPIE Stereoscopic Displays and Applications XXIII, 8288 (59), pp.1-6, 2012.

[2] J. Y. Lee, J. Lee, and D.-S. Park, "Effect of a synthesized depth view on multi-view rendering quality", in Proc. of ICIP, 2011.

[3] Y. Zhang, "Quality assessment of synthesized view caused by depth maps compression", in Proc. of CISP, 2014.

[4] F. Battisti, E. Bosc, M. Carli, P. Le Callet, and S. Perugia. "Objective image quality assessment of 3D synthesized views", in Signal Processing: Image Communication, 30:78–88, jan 2015.

[5] F. Devernay, A. R. Peon, "Novel view synthesis for stereoscopic cinema: detecting and removing artifacts", in Proc. of the First International Workshop on 3D Video Processing Service, 2010, pp.25–30.

[6] ISO/IEC, JTC1/SC29/WG11, Use Cases and Requirements on Free-viewpoint Television (FTV), Editors: M. P. Tehrani, S. Shimizu, G. Lafruit, T. Senoh, T. Fujii, A. Vetro, M. Tanimoto, Geneva, Switzerland, October 2013.

[7] E. Bosc, P. Le Callet, L. Morin, M. Pressigout, "Visual quality assessment of synthesized views in the context of 3D-TV. 3D-TV System with Depth-Image-Based Rendering in Architectures, Techniques and Challenges", Springer, pp.439-474, 2013, 978-1-4419-9964-1.

[8] ITU-T, Subjective video quality assessment methods for multimedia applications, Geneva, Rec. P910, 2008.

[9] E. Bosc, P. Hanhart, P. Le Callet, and T. Ebrahimi. "A quality assessment protocol for free-viewpoint video sequences synthesized from decompressed depth data", in Prof. of QoMEX, 2013.

[10] MeTriX MuX Visual Quality Assessment Package, available online at http://foulard.ece.cornell.edu/gaubatz/metrix_mux, last visited April 2016.

[11] E. Bosc, R. Pepion, P. Le Callet, M. Koppel, P. Ndjiki-Nya, M. Pressigout, and L. Morin. Towards a New Quality Metric for 3-D Synthesized View Assessment. IEEE J. Sel. Top. Signal Process., 5(7):1332–1343, nov 2011.

[12] E. Bosc, R. Pepion, Patrick Le Callet, Muriel Pressigout, and Luce Morin, "Reliability of 2D quality assessment methods for synthesized views evaluation in stereoscopic viewing conditions", in Proc. 3DTV-CON, 2012.

[13] D. Sandic-Stankovic, D. Kukolj, P. Le Callet, "DIBR synthesized image quality assessment based on morphological wavelets", in Proc. QoMEX, 2015.

[14] D. Sandić-Stanković, D. Kukolj, and P. Le Callet, "Multi–Scale Synthesized View Assessment Based on Morphological Pyramids", in Journal of Electrical Engineering, 67(1), jan 2016.

[15] M. Solh, G. Al Regib, and J. M. Bauza, "3VQM: A vision-based quality measure for DIBR-based 3D videos", in Proc. ICME, 2011.

[16] P. Conze, P. Robert, L. Morin, "Objective view synthesis quality assessment", in Proc. of SPIE Conference Series, vol.8288, 2012.

[17] P. Hanhart, E. Bosc, P. Le Callet, and T. Ebrahimi. "Free-viewpoint video sequences: A new challenge for objective quality metrics", in Proc. MMSP'14, 2014.

[18] Subjective scores for DIBR algorithms using ACR-HR & Pair comparison, available online at http://ivc.univ-nantes.fr/en/databases/DIBR_Images/, last visited April 2016.

[19] MCL 3D Database, available online at http://mcl.usc.edu/mcl-3d-database/, last visited April 2016.

[20] DIBR Video quality assessment, available online at http://ivc.univ-nantes.fr/en/databases/DIBR_Videos/, last visited April 2016.

[21] SIAT Synthesized Video Quality Database, available online at http://codec.siat.ac.cn/SIAT_Synthesized_Video_Quality_Database/index.html, last visited April 2016

[22] Free-Viewpoint synthesized videos with subjective quality measures, available online at http://ivc.univ-nantes.fr/en/databases/Free-Viewpoint_synthesized_videos/, last visited April 2016.

[23] High-Quality Streamable Free-Viewpoint Video, available at http://research.microsoft.com/en-us/um/redmond/projects/fvv, last visited April 2016.

[24] Tanimoto FTV test sequences, available at http://www.tanimoto.nuee.nagoya-u.ac.jp/MPEG-FTVProject.html, last visited April 2016.

Federica Battisti received the Laurea Degree (Master) in Electronic Engineering and the PhD degree from Roma Tre University in 2006 and 2010 respectively. Her research interests include signal and image processing with focus on subjective quality analysis of visual contents. She is currently assistant professor at the Department of Engineering at Roma Tre University.

Patrick Le Callet received M.Sc. degree PhD degree in image processing from Ecole Polytechnique de l'Universite de Nantes. Since 2006, he is the head of the Image and Video Communication lab at CNRS IRCCyN. He is mostly engaged in research dealing with the application of human vision modeling in image and video processing. His current interests are 3D image and video quality of experience, watermarking techniques and visual attention modeling and applications. He is currently Full Professor at Ecole polytechnique de l'Université de Nantes.

3D Visual Attention for Improved Interaction, Quality Evaluation and Enhancement

Chaminda T.E.R. Hewage
Department of Computing & Information Systems, Cardiff Metropolitan University, Cardiff, UK
chewage@*cardiffmet.ac.uk*

1. Introduction

Viewers tend to focus into specific Regions of Interest (RoI) in an image, driven by the task or the low level information of the image/video. Therefore, visual attention is one of the major aspects to understand the overall Quality of Experience (QoE), user perception and interaction. For instance, visual attention cues can be used in spatial navigation (as one of the most prominent FTV scenarios). This letter investigates how 3D visual attention can be used for better interaction, quality assessment and processing of 3D image/video. Furthermore, open challenges in integrating visual attention for 3D interaction is also discussed.

Our eye vision represents an important channel for perceiving our environment. In addition, our gaze direction can convey what we currently attend to, for instance, looking at somebody while addressing this person in a conversation. Eye tracking experiments are widely employed to investigate user eye gaze positions during consumption of visual information. The collected eye movement data are then post-processed to obtain Fixation Density Maps (FDM) or saliency maps. Visual attention models have emerged in the recent past to predict user attention in image, video and 3D video [1-3]. These attention models predict user eye movements based on low level image features such as spatial frequency, edge information, etc. Visual attention models can therefore be used in image processing applications (e.g. post processing, image quality evaluation, image retargeting) to improve the interactivity and engagement with the multimedia content and the task being considered. However, the usage of these models in quality assessment/improvement and improved interaction has not been thoroughly investigated up to date.

There are two major approaches to analyze user visual attention, namely: free viewing task (i.e., bottom-up approach) and task oriented (top-bottom approach). The former approach is driven by low level image features such as spatial and temporal frequencies. The top-bottom approach is driven by the task (e.g., target identification in shooting games). Several other factors influence visual attention such as sociocultural background, context, duration, etc.

The attention of users during 3D viewing can be influenced by several factors including spatial/temporal frequencies (as in the 2D visual attention described above), depth cues, conflicting depth cues, etc. A comprehensive analysis of visual attention in 3D, and of the weaknesses of existing models and their usage is discussed in [4]. The studies on visual attention in 2D/3D images found out that the behaviors of viewers during 2D viewing and 3D viewing are not always identical. For instance, the study in [5] for 2D/3D images has shown that added depth information increases the number of fixations, eye movement throughout the image and shorter and faster saccades. This observation is also complemented by the investigation carried out by Hakkinen et al. [6], which showed that eye movement during 3D viewing is more distributed. In contrast to these observations, in [7] Ramasamy et al. found out that the spread of fixation points are more confined in 3D viewing than in 2D viewing. These observations have direct influences in how we perceive 3D video. Therefore, effective 3D video interaction, quality evaluation and QoE enhancement schemes could be designed based on these observations. The proposed image processing methods in the literature exploit these visual attention models for better interaction, 3D QoE evaluation and processing.

Modeling visual attention in 2D viewing is mainly driven by spatial and temporal frequencies of the image as suggested by various studies [8][9]. However, for 3D images/video, depth cues need to be added to the existing image features in order to generate a robust 3D saliency map. Most of the reported 3D visual attention models in the literature [10][11] are therefore based on scene depth information in addition to motion and spatial characteristics. 3D visual attention models can be divided into two main categories, as shown below:

- Depth weighted 3D saliency model (see Figure l);
- Depth saliency based 3D saliency model (see Figure 2).

The depth weighted model weighs the generated 2D saliency model based on the depth information in order to obtain the 3D visual saliency map. The second method generates two visual saliency maps: the first one for 2D image information and the second one for the corresponding depth map of the scene (see Figure 2). Then both saliency maps are combined into one 3D saliency map based on the selected weights as described in (1). For instance, the 3D saliency based model described in [1], considers both current image information and prior

knowledge. However, this 3D saliency model does not take into account the temporal activity of the scene.

$$SM_{3D} = W_1 \text{ x } SM_{Depth} + W_2 \text{ x } SM_{2D} \quad (1)$$

where W_1 and W_2 are weights assigned for the depth saliency model (i.e., SM_{Depth}) and 2D saliency model (i.e., SM_{2D}) respectively.

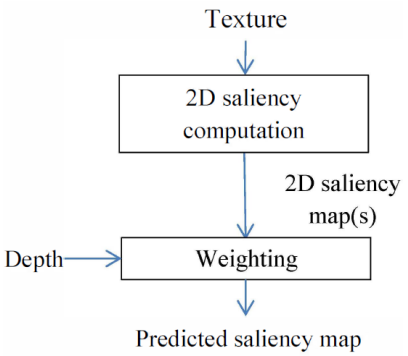


Figure 1. Depth weighted 3D saliency model

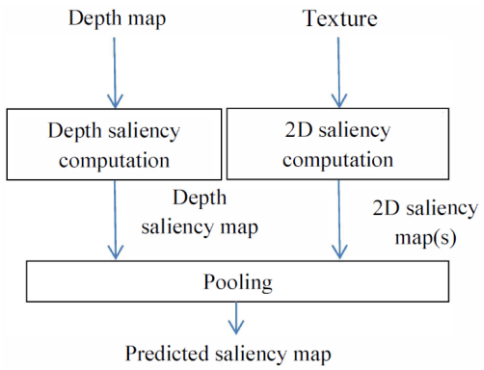


Figure 2. Depth saliency based 3D saliency model.

In this paper, we discuss how we could exploit 3D visual attention to interact, measure and improve 3D QoE. The following sections elaborate how we could exploit 3D visual attention models for better interaction (Section 2), to measure QoE (Section 3) and to improve 3D video perception in general (Section 4). Section 5 concludes the paper.

2. Visual attention driven 3D interaction

Visual attention has been widely studied in psychology and cognitive science research. The understanding of visual attention immensely helps application developers to process and render 3D content in order to maximize the utility or experience. For instance, non-gamers may not perform well in 3D game environments, or they do not pick up an important item because they don't notice it. Visual attention research results can be used to inform designers on how to compose colors and placements of objects to stimulate attention and eliminate these problems. In addition to gaming there are other applications where visual attention can improve the engagement and interactivity. An example is sub-title and logo placement in 3D video services [21]. In order to be noticed, those should be in the right depth location. Otherwise, users won't be able to experience the content to the maximum satisfaction. Moreover, some users may attend to the objects with positive parallax whereas other users may prefer objects with negative parallax. Therefore, when rendering or capturing 3D content, attention should be paid to what users may focus on during the viewing. Furthermore, if user attention can be predicted, system resources can be optimized by anticipating the delivery of novel views for rendering. For instance, in the framework of free viewpoint video, if the visual attention can predict the next possible view by analyzing user gaze, the system can pre-download or render the next possible view in advance. Therefore this allows resource optimization while providing the best user QoE. The eye tracking is considered as one of the main supporting technologies to realize spatial navigation in FVV. Data visualization in Big Data applications can also benefit from visual attention details [22], *e.g.*, finding out at what depth level each data point should be placed in order to attract more attention from the users.

2.1 3D User Interfaces (3D-UIs)

Another important application of visual attention is 3D virtual interfaces. The effectiveness of interaction with 3D interfaces can be improved by integrating gaze input [23-25]. For instance, gaze information can be used to provide an immersive experience with 3D virtual environments (e.g. 3D gaming, virtual interactive training, scientific 3D visualizations, as well as social networking environments (e.g., Second Life)).

Eye tracking can be employed to find user gaze information for such applications. This information together with traditional inputs such as *mouse* input can be used to improve the user interaction in 3D virtual environments. There are two advantages of using eye tracking in 3D-UIs. First, eye tracking can be used to study how users interact with 3D virtual environments. On the other hand, eye tracking can be used to provide an additional input to directly interact with the 3D environment. One of the constraints for the development of gaze-based interaction techniques is the lack of low-cost, reliable, and convenient eye trackers. However, in the recent past, eye tracking devices have become more affordable. In addition, the systems develop into more lightweight and more comfortable setups, which include long range eye tracking systems (e.g., [26-28]). This provides a great potential for the integration of gaze input in a professional or even everyday 3D application context.

The availability of affordable and user friendly eye tracking technologies has enabled gaze incorporated interaction in various user contexts. Especially the gaze input together with the other input modalities has demonstrated the highest potential [29-31]. For instance, Figure 3 illustrates a combination of gaze input with a handheld to interact with virtual 3D scenes shown on a distant display may provide a natural and yet efficient interaction.

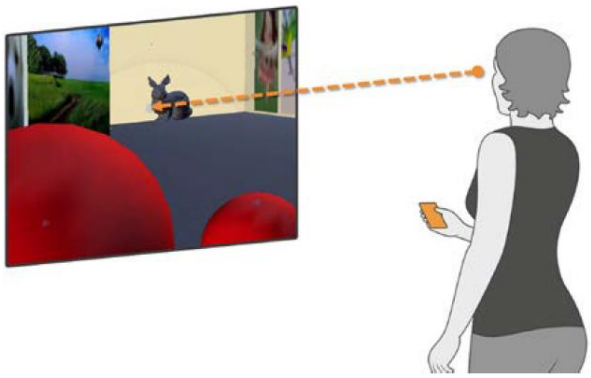


Figure 3. Gaze-supported steering and selection in virtual 3D environments (e.g., [31]).

3. Visual attention driven 3D quality assessment

There are still unanswered questions such as whether quality assessment is analogous to attentional quality assessment and also how we could integrate attention mechanisms into the design of QoE assessment methodologies. 2D image/video quality assessment presented in [12], investigated the impact of different RoIs in image quality evaluation. However, a thorough study has not yet been conducted to identify the relationship between 3D image/video attention models and 3D image/video quality evaluation. The COST action presentation in [13] identifies three main approaches to integrate visual attention into image/video quality evaluation (see Figure 4, 5 and 6). Similar to the integrated model described in Figure 6, attentive areas identified by visual attention studies can be utilized to extract image features which can be used to design No-Reference (NR) and Reduced-Reference (RR) quality metrics for real-time 3D video application. The use of extracted features to design RR 3D image/video quality metrics has been undertaken in previous research [14-17]. Furthermore, the use of 3D visual saliency information could be used to further reduce the amount of side-information for real-time quality evaluation.

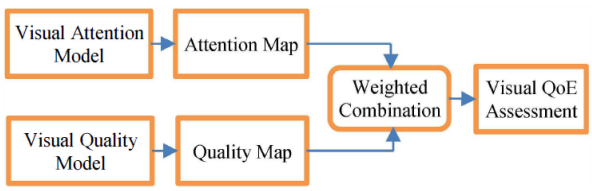


Figure 4. Direct combination [13].

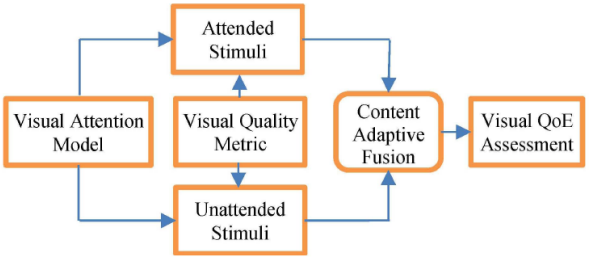


Figure 5. Divided integration [13].

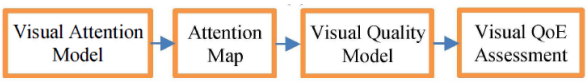
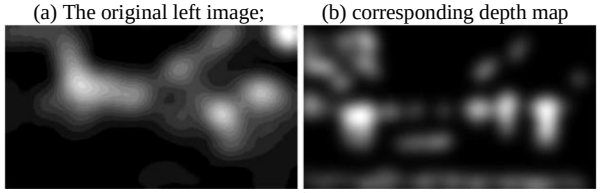


Figure 6. Integrated combination [13].

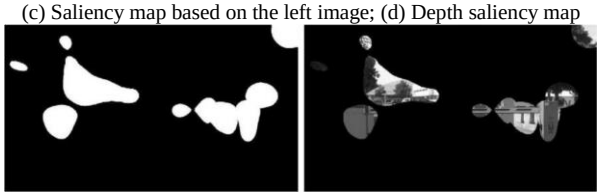
4. Visual attention driven 3D video processing

Since visual attention models can predict the highly attentive areas of an image or video, these can be integrated into video coding at the source-end. The proposed RoI based encoding methods for 2D/3D video has shown improved quality at a given bitrate compared to conventional encoding methods [18][19]. For instance the RoI based encoding method proposed and evaluated in [19] shows that by protecting combined edges of color plus depth based 3D video, the overall quality of the rendered views can be improved. This study also incorporates an Unequal Error Protection (UEP) mechanism to protect different image regions. However, only a few visual attention based RoI encoding methods have been reported for 3D video applications to date. For instance, the work described in [32] utilizes 3D visual attention information to encode and provide unequal error protection for 3D video transmission over unreliable channels. Figure 7 graphically illustrates the identification of attentive areas in [32].

(a) The original left image; (b) corresponding depth map



(c) Saliency map based on the left image; (d) Depth saliency map



(e), Combined 3D saliency map (binary); (f) Identified saliency area

Figure 7. Sample image from the *Barrier* HD 3D sequence [20][32]

5. Conclusion

In this paper we discuss the usage of 3D visual attention for better interaction, quality evaluation and processing. Moreover, the use of visual attention or gaze as a supported input to interact with 3D virtual environments is also elaborated. The integration of visual attention models in 3D quality evaluation and processing applications is also conversed with insights into the future research directions.

References

[1] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in Proc. of IEEE Inter. Conference on Computer Vision and Pattern Recognition, June 2007, pp. 1-8.

[2] Y. Zhang, Gangyi Jiang, Mei Yu, Ken Chen, "Stereoscopic Visual Attention Model for 3D Video", Adv. in Multimedia Modelling: Lecture Notes in C. S. vol. 5916, pp. 314-324, 2010.

[3] J. Wang, M.P. Da Silva, P. Le Callet, V. Ricordel, "Computational Model of Stereoscopic 3D Visual Saliency", IEEE Tran. on Image Proce.,

vol. 22, no. 6, pp. 2151-2165, 2013.
[4] Q. Huynh-Thu, M. Barkowsky, P. Le Callet, "The Importance of Visual Attention in Improving the 3DTV Viewing Experience: Overview and New Perspectives", IEEE Transactions on Broadcasting, vol. 57, no. 2, pp. 421-431, 2011.
[5] L. Jansen, S. Onat, and P. Konig" "Influence of disparity on fixation and saccades in free viewing of natural scenes", Journal of Vision, vol. 9, no. 1, pp. 1-19, Jan. 2009.
[6] J. Hiikkinen, T. Kawai, J. Takatalo, R. Mitsuya, and G. Nyman, "What do people look at when they watch stereoscopic movies?", in Proc. SPIE Con! Stereoscopic Displays and Applications XXI, vol. 7524, San Jose, January 2010.
[7] C. Ramasamy, D. House, A. Duchowski, and B. Daugherty, "Using eye tracking to analyze stereoscopic, filmmaking", in Froc. SIGGRAPH 2009: Posters, 20l0.
[8] M. Treisman and G. Gelade, "Feature integration theory of attention", Cognitive Psychology, vol. 12, pp. 97-136, 1980.
[9] O. Le Meur and P. Le Callet, "What we see is most likely to be what matters: visual attention and applications", in Proc. IEEE International Conference on Image Processing, Cairo, Nov 2009, pp. 3085-3088.
[10] A. Maki, P. Nordlund, and J.O. Eklundh, "Attentional scene segmentation: Integrating depth and motion from phase", Computer Visi. and Image Understa., vol. 78, pp. 351-373, 2000.
[11] N. Ouerhani and H. HUgli, "Computing visual attention from scene depth", in Proc. International Conference on Pattern Recognition, Barcelona, 2000, pp. 375-378.
[12] O. Le Meur, A. Ninassi, P. Le Callet and D. Barba, "Overt visual attention for free-viewing and quality assessment tasks. Impact of the regions of interest on a video quality metric", Elsevier, Signal Processing: Image Comm., vol. 25, no. 7, 2010.
[13] J. You, Visual Attention Driven QoE: Towards Integrated Research, Norwegian University of Science and Technology, Qualinet COST Action: Doc-Qi0179., WG2, Prague, Feb 2012
[14] C.T.E.R. Hewage, and M.G. Martini, "Edge based reduced-reference quality metric for 3D video compression and transmission", IEEE Journal of Selected Topics in Signal Processing vol. 6, no. 5, pp. 471-482, Sept 2012.
[15] C.T.E.R. Hewage, & M.G. Martini, "Reduced-reference quality assessment for 3D video compression and transmission", IEEE Transactions on Consumer Electronics, vol. 57, no. 3, pp. 1185-1193, Aug. 2011.
[16] C.T.E.R. Hewage, S.T. Worrall, et al, "Quality evaluation of color plus depth map-based stereoscopic video", IEEE Journal of Selected Topics in Signal Proce., vol. 3, no. 2, pp. 304-318, 2009.
[17] C.T.E.R. Hewage, and M.G. Martini, "Perceptual quality driven 3D video communication", *Scholar's Press*, 2013.
[18] M.G. Martini, and C.T.E.R. Hewage, "Flexible Macroblock Ordering for context-aware ultrasound video transmission over mobile WiMAX", International Journal of Telemedicine and Applications, 20lO, ISSN (print) 1687-6415.
[19] C.T.E.R. Hewage, and M.G. Martini, "ROI-based transmission method for stereoscopic video to maximize rendered 3D video quality", In: Stereoscopic Displays and Applications XXIII; Jan 2012, California, U.S.A. (Proceedings of SPIE, no. 8288).
[20] M. Urvoy, M. Barkowsky, et al, "NAMA3DS1- COSPADl: Subjective video quality assessment database on coding conditions introducing freely available high quality 3D stereoscopic sequences," 4th International Workshop on Quality of Multimedia Experience (QoMEX), July 2012, pp. l09-114.
[21] M., Lambooij, et al, "The impact of video characteristics and subtitles on visual comfort of 3D TV." Displays, vol. 34, no. 1, pp. 8-16, 2013.
[22] Z. Liu, and J.T. Stasko. "Mental models, visual reasoning and interaction in information visualization: A top-down perspective." Visualization and Computer Graphics, IEEE Transactions on, vol. 16, no. 6, pp. 999-1008, 2010.
[23] E. Castellina, and F. Corno, "Multimodal Gaze Interaction in 3D Virtual Environm.". In *Proc. COGAIN '08*, pp. 33–37, 2008.
[24] N. Cournia, J. D. Smith, and A.T. Duchowski, "Gaze- vs. hand-based pointing in virtual environments" In *Proc. CHI EA '03*, pp. 772-773, 2003.
[25] S. Stellmach, L. Nacke, R. Dachselt, "3D Attentional Maps - Aggregated Gaze Visualizations in Three-Dimensional Virtual Environments". In Proc. AVI'10, ACM, pp. 345-348, 2010.
[26] J. S. Babcock, and J. B. Pelz, "Building a light-weight eye tracking headgear" *In Proc. ETRA '04, ACM*, pp. 109-114, 2004.
[27] C.H. Morimoto, and M. R. M. Mimica, "Eye gaze tracking techniques for interactive applications". *Comput. Vis. Image Underst*. vol. 98, no. 1, pp. 4-24, 2005.
[28] C. Hennessey, and F. Jacob, "Long range eye tracking: bringing eye tracking into the living room." *In Pro. of the Symposium on Eye Tracking Research and Applicati*., pp. 249-252. ACM, 2012.
[29] S. Bardins, T. Poitschke, S. Kohlbecher, "Gaze-based interaction in various environments". In Proc. VNBA '08, pp. 47-54, 2008.
[30] S. Stellmach, and R. Dachselt, "Investigating Gaze-supported Multimodal Pan and Zoom". In *Proc. ETRA'12*, ACM, 2012.
[31] S. Stellmach, and R. Dachselt, "Look & Touch: Gaze-supported Target Acquisition". In *Proc. CHI'12*, ACM, 2012.
[32] C.T.E.R Hewage, J. Wang, M.G. Martini, and P. Le Callet. "Visual saliency driven error protection for 3D video." In IEEE Multimedia and Expo Workshops (ICMEW).

Chaminda T.E.R. Hewage received the B.Sc. in Engineering (Hons) from University of Ruhuna, Sri Lanka. Then he worked as a Telecommunication Engineer for Sri Lanka Telecom Ltd, Sri Lanka. He was awarded the Ph.D. from University of Surrey, UK, in 2009 for his work on perceptual quality driven 3D video over networks. He joined I-Lab University of Surrey as a Research Fellow in 2009 and then joined WMN Research group in Kingston University – London as a Senior Researcher. Since 2015, he has been with Cardiff Metropolitan University as a Senior Lecturer. His current research interests include 2D/3D video processing, communication and QoE evaluation. He is a Fellow of HEA (UK) and Member of IEEE.

RE@CT: Immersive Production and Delivery of Interactive 3D Content

Marco Volino, Dan Casas, John Collomosse and Adrian Hilton
Centre for Vision, Speech and Signal Processing, University of Surrey, UK
{*m.volino, j.collomosse, a.hilton*}*@surrey.ac.uk, dan.casas@gmail.com*

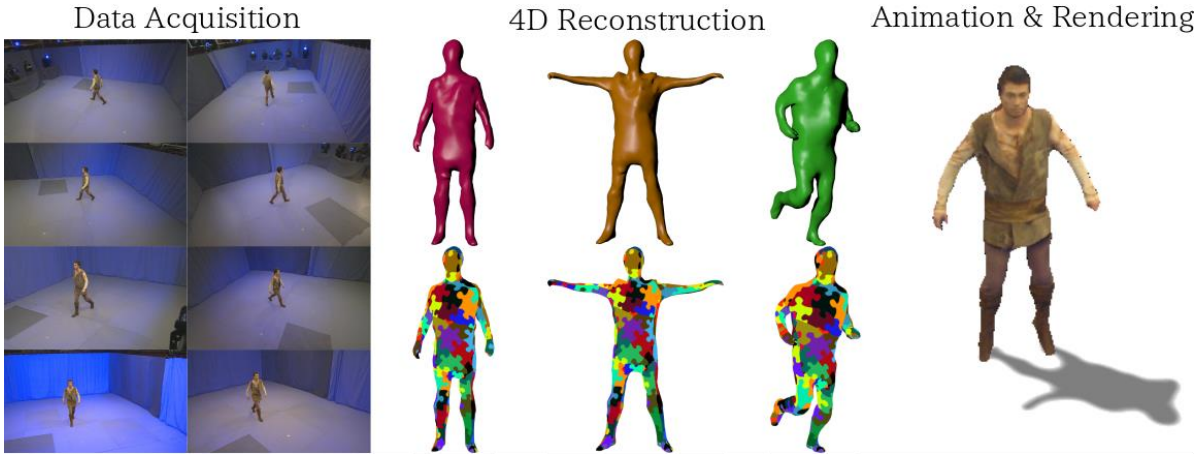


Figure 1. An overview of the RE@CT production pipeline. Data is acquired in a multiple camera studio, 4D reconstruction is performed enforcing a common mesh structure over all reconstructed frames which allows efficient real-time character animation and rendering.

1. The RE@CT Project

The RE@CT project introduced a new production methodology to create film quality interactive characters from markerless video-based capture of actor performance. Advances in graphics hardware have produced interactive video games with photo-realistic scenes. However, interactive characters still lack the visual appeal and subtle details of the real actor performance as captured in video. In addition, existing production pipelines for authoring animated characters are highly labor intensive.

The RE@CT project revolutionized the production of realistic characters and significantly reduced costs by developing an automated process to extract and represent animated characters from actor performance captured in a multiple camera studio. The key innovations in RE@CT are the analysis and representation of 4D video to allow reuse for real-time interactive animation. This enables efficient authoring of interactive characters with video quality appearance and motion.

RE@CT brought together world-leading researchers from academic and industrial institutions through the European Union's seventh framework programme. Institutions based in France, Germany and UK included Artefacto, the British Broadcasting Corporation (BBC) R&D, Fraunhofer Heinrich Hertz Institute (HHI), Inria Rhône-Alpes Grenoble, Oxford Metrics Group (OMG) Vicon, and the Centre for Vision, Speech and Signal Processing (CVSSP) at the University of Surrey.
Some example use cases of the RE@CT system are: (i) Production of video-realistic characters based on real actor performance and appearance for use in games, broadcast and film without the requirement of extensive artist time; (ii) cost effective production of interactive 3D content for teaching, training and social media available on a variety of platforms [2]; (iii) future use in sports to enhance broadcast visualization and analysis.

This paper gives an overview of technologies that were developed as part of the RE@CT project with a particular focus on efficient real-time video-based character animation.

2. The RE@CT Production Pipeline

This section presents an overview of the RE@CT production pipeline from data acquisition and 4D reconstruction to real-time character animation and rendering, as outlined in Figure 1.

2.1. Data Acquisition

The RE@CT capture studio is fitted with up to 15 static fixed-focus cameras evenly positioned around the studio to give full 360 degree coverage of a subject. Synchronized, full HD resolution video streams are captured at conventional frame rates. Cameras are located approximately 2.5 meters above the ground to give an effective capture volume of 3x3x2.5 meters.

A novel LED wand based calibration system was designed and developed to provide real-time feedback to a studio operator and ensure even sampling of the capture volume. Camera parameters are jointly optimized against the detected LED markers, which results in robust camera calibration.

To aid background/foreground segmentation, the studio was fitted with retro-reflective Chroma cloth [13]. Blue LED light rings are placed around each camera lens illuminating the Chroma cloth background from the viewpoint of the capture cameras. This allows automatic extraction of clean foreground silhouettes.

2.2. 4D Reconstruction

An approximation of the underlying 3D geometry is first obtained via the visual hull, given by the intersection of the extracted silhouettes. The visual hull is refined using photometric features through multiple view stereo reconstruction techniques, see [11] for further details. This is performed on a frame-by-frame basis and results in an *unstructured* collection of 3D meshes with varying numbers of vertices and inconsistent mesh topology. This makes reuse of reconstructed data a challenging problem.

To overcome this, RE@CT introduced several techniques to enforce a consistent mesh topology across all reconstructed frames [1, 3]. Figure 1 (center, bottom) demonstrates this by applying a patterned texture to the time varying meshes, notice that the pattern remains fixed to the 3D surface even in the presence of large non-rigid deformations.

Allain *et al* developed a volumetric framework in which meshes are decomposed into discrete 3D volumes based on centroidal Voronoi tessellation. By conserving volume during deformations, the method is robust when tracking large deformations resulting from fast motions, see [1] for further details. Budd *et al* solve the problem in a non-sequential non-rigid Laplacian deformation framework. Unstructured meshes were organized into a tree structure ordered in terms of shape similarity. This allows alignment across multiple sequences, which is important for data-driven animation. Alignment is performed in a pairwise fashion between parent and child tree nodes using the mesh at the tree root node as a template. This reduces the total accumulated error when compared to sequential frameworks, see [3] for further details.

This process results in a database of character motions that share a common mesh topology. The combination of a temporally consistent mesh sequence and the captured camera frames is referred to as *4D video*. This *structured* data is the primary input for character animation and rendering techniques. During a capture session, typical game character motions are captured including walk/jog/run cycles and vertical and horizontal jumps of various lengths and heights.

2.3. Character Animation and Rendering

The RE@CT project introduced several innovations in the area of character animation and rendering which are described in the following sections.

2.3.1. Multiple Layer Texture Representation

The multiple camera capture process inherently produces huge amounts of data, which is prohibitive to the practical application of 4D video. A significant bottleneck with view-dependent rendering approaches, *e.g.* [12], is that they require access to the captured frames at render-time. This is bandwidth and memory intensive and restricts rendering techniques to high performance computers.

To address this problem, the Multiple Layer Texture Representation (MLTR) was developed to reduce the storage requirements of the captured video while improving the efficiency of view-dependent rendering. The MLTR resamples the multiple camera images at each time instance into a hierarchal set of texture maps, Figure 2 (top row), which gives a storage reduction of >90% compared to the original captured data. Also encoded into the MLTR is the

camera assignment, Figure 2 (bottom row), which identifies the camera from which the texture was sampled. View-dependent rendering is made computationally efficient by pre-computing computationally expensive operations in the render cycle, *e.g.* depth testing, leaving only a minimal amount of texture blending to be performed at runtime based on the virtual viewpoint. This allows view-dependent rendering to be performed efficiently in a WebGL environment and on mobile devices [9].

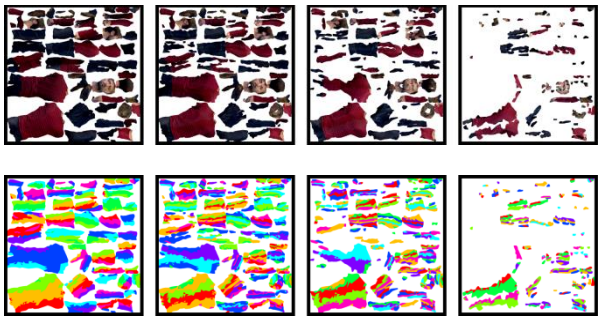


Figure 2. MLTR for a single multiple camera frame. Appearance data (top row) is ordered by visibility therefore the most visible polygons are resampled into the first layer in descending order (L-R). Camera assignment of each polygon in every layer is also encoded into the MLTR (bottom row).

The spatial/temporal sampling from the multiple view image sequence is also optimized to encourage further redundancy in the MLTR data [14]. This process results in a further reduction of storage requirements by >95% compared to the captured data, see [14] for further details.

2.3.2. 4D Parametric Motion Graphs

Given a 4D motion database, a graph-like structure referred to as 4D Parametric Motion Graph [6] was introduced to synthesize novel animations, see Figure 3. Inter- and intra-transition points are identified to allow seamless blending and interpolation between captured motions.

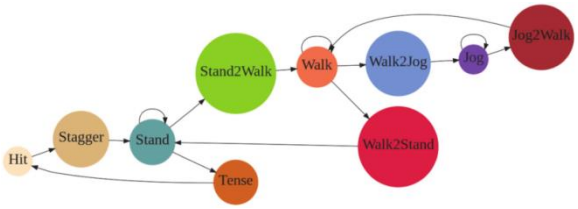


Figure 3. Parametric motion graph [8].

Semantically related motions, *e.g.* walk/jog, low/high jump, short/long jump, were group together creating the graph nodes. Intermediate motion can be obtained by interpolating corresponding vertex positions between two related frames, shown in Figure 4. Often this parameterization can take semantic meaning, *e.g.* walking speed or length of a jump. Transitions between nodes, which form the edges of the graph, were found by finding similar pose and model appearance across nodes, which enables seamless transitions between parametric motions. Figure 5 shows an example animation of the RE@CT infantry character in a game environment, smoothly transitioning between the captured motions.

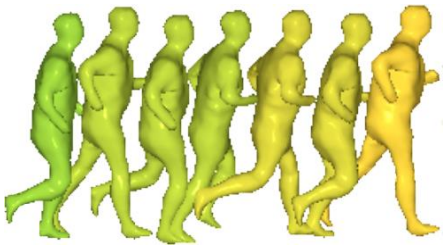


Figure 4. Geometry blending from walk (green) to jog motion (yellow) (L-R) with intermediate colors representing the blending weight [6].

Methods were also developed to extend the range of captured motions by bridging the gap between 4D video data and skeletal motion capture databases [8]. This work allows 4D video based animation to be driven by skeletal motion capture queries and allows 4D video data to leverage the diverse range of skeletal motion capture data, see [8] for more details.

The presented character animation techniques were adapted to run in a WebGL environment [15], which allows online delivery of interactive characters. This paves the way for realistic web-based content derived from 4D video. See [15] for further details and [2, 16] for interactive demonstrations of this work.

Figure 5. RE@CT infantry character animation [4].

2.3.3. 4D Video Textures

Whilst geometry can be interpolated to produce an intermediate motion, the same approach cannot simply be applied to texture the model, as this results in visual artefacts. 4D Video Textures (4DVT) [4] used optical flow correspondence to make local corrections at run-time based on the virtual viewpoint to prevent visual artefacts occurring. The key idea behind the approach is to projectively texture the intermediate geometry using the selected frames from two related motions; optical flow is then computed between these rendered images and the resulting optical flow vectors are interpolated based on desired blending weight, Figure 6, see [4] for further details.

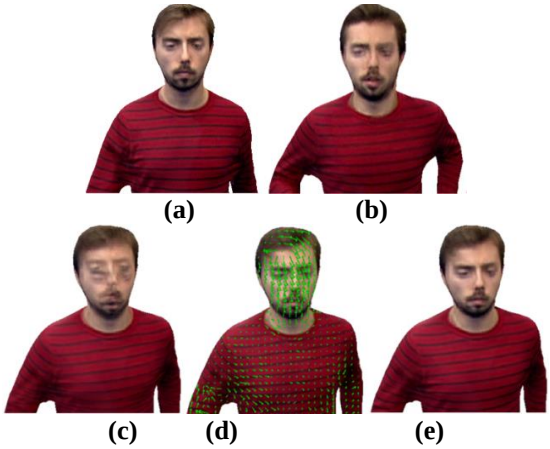


Figure 6. 4D Video Textures. (a,b) Textured walk and jog motions; (c) linear interpolation of geometry and appearance resulting in ghosting artefacts; (d) optical flow correspondence between walk and jog when texturing interpolated geometry; (e) interpolated geometry and texture with optical flow correction [4].

To overcome the computationally expensive rendering pipeline of 4DVT, 4D Model Flow [5] pre-computes the optical flow correspondence offline between related frame pairs. These are loaded at render-time and utilized in the same way as in 4DVT. This requires additional memory and storage overheads, but is able to achieve comparable visual results at significantly higher frame rates.

3. Conclusions

This paper has presented an overview of selected innovations developed over the course of the RE@CT project. Full details about the RE@CT project can be found on the project webpage [10]. Datasets captured during RE@CT are made available to the research community for further study. These can be obtained from the CVSSP dataset repository [7].

Acknowledgements

RE@CT was funded by the European Union's Seventh Framework Programme (FP7/2007-2013) under grant number 288369.

References

[1] Allain, B., Franco, J.-S., & Boyer, E. (2015). An efficient volumetric framework for shape tracking. In 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 268–276). IEEE.
[2] BBC Taster Dancer Demo: www.bbc.co.uk/taster/projects/dancer
[3] Budd, C., Huang, P., Klaudiny, M., & Hilton, A. (2012). Global Non-rigid Alignment of Surface Sequences. International Journal of Computer Vision, 102(1-3), 256–270.
[4] Casas, D., Volino, M., Collomosse, J., & Hilton, A. (2014). 4D video textures for interactive character appearance. Computer Graphics Forum, 33(2), 371–380.
[5] Casas, D., Richardt, C., Collomosse, J., Theobalt, C., & Hilton, A. (2015). 4D Model Flow: Precomputed Appearance Alignment for Real-time 4D Video Interpolation. Computer Graphics Forum, 34(7), 173–182.
[6] Casas, D., Tejera, M., Guillemaut, J.-Y., & Hilton, A. (2012). 4D parametric motion graphs for interactive animation. Proceedings of the ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games, 1(212), 103–110.
[7] CVSSP Data repository: http://cvssp.org/data/cvssp3d
[8] Huang, P., Tejera, M., Collomosse, J., & Hilton, A. (2015). Hybrid Skeletal-Surface Motion Graphs for Character Animation from 4D Performance Capture. ACM Transactions on Graphics, 34(2), 1–14.
[9] Imber, J., Volino, M., Guillemaut, J.-Y., Fenney, S., & Hilton, A. (2013). Free-viewpoint video rendering for mobile devices. In Proceedings of the 6th International Conference on Computer Vision / Computer Graphics Collaboration Techniques and Applications - MIRAGE '13 (p. 1). New York, USA: ACM Press.
[10] RE@CT Project Website: http://react-project.eu
[11] Starck, J., & Hilton, A. (2007). Surface Capture for Performance-Based Animation. Computer Graphics and Applications, IEEE, 27(3), 21–31.
[12] Starck, J., Kilner, J., & Hilton, A. (2009). A Free-Viewpoint Video Renderer. Journal of Graphics, GPU, and Game Tools, 14(3), 57–72.
[13] Thomas, G. (2006). Mixed reality techniques for TV and their application for on-set and pre-visualization in film production. International Workshop on Mixed Reality, (January), 31–36.
[14] Volino, M., Casas, D., Collomosse, J., & Hilton. A. Optimal Representation of Multiple View Video. Proceedings of the British Machine Vision Conference. BMVA Press, September 2014.
[15] Volino, M., Huang, P., & Hilton, A. (2015). Online interactive 4D character animation. In Proceedings of the 20th International Conference on 3D Web Technology - Web3D '15 (pp. 289–295). New York, New York, USA.
[16] WebGL Demo: http://cvssp.org/projects/4d/webGL/

Marco Volino is a research fellow within CVSSP. He obtained a PhD in Computer Graphics in 2016 under the supervision of Prof. Adrian Hilton and a MEng in Electronic Engineering at the University of Surrey. He has previously worked at Sony BPRL, BBC R&D, and the USC's Institute for Creative Technologies solving problems related to computer vision, graphics and animation.

Dan Casas is a postdoc in the Graphics, Vision and Video group at the Max Planck Institute in Saarbrucken, Germany. Dan received his PhD in Computer Graphics in 2014 from the University of Surrey (UK), supervised by Prof. Adrian Hilton. His dissertation introduced novel methods for character animation from multi-camera capture that allow the synthesis of video-realistic interactive 3D characters.

John Collomosse is a senior lecturer (asst. prof.) within CVSSP at the University of Surrey. John's research fuses Computer Vision, Graphics and Machine Learning to tackle Big Data problems in visual media, with particular emphasis on post-production in the creative industries and intelligent algorithms for searching and rendering large visual media repositories.

Adrian Hilton is Professor of computer vision and Director of CVSSP at the University of Surrey. He currently leads research investigating the use of computer vision for applications in entertainment content production, visual interaction, and clinical analysis. His research interests include robust computer vision to model and understand real-world scenes.

SceneNet: Crowd Sourcing of Audio Visual Information aiming to create 4D video streams

D. Eilot[1], Y. Schoenenberger[2], A. Egozi[3], E. Hirsch[1], T. Ben Nun[1], Y. Appelbaum-Elad[1], J. Gildenblat[1], E. Rubin[1], P. Maass[3], P. Vandergheynst[2], C. Sagiv[1]
[1]SagivTech Ltd., [2]EPFL, [3]University of Bremen
chen@sagivtech.com

1. Introduction

If you visited a rock concert recently, or any other event that attracts crowds, you probably recognized how many people are taking videos of the scenario, using their mobile phone cameras. The aim of SceneNet is to use the power of crowd sourcing and to combine individual videos from multiple viewpoints of the same scene to create a 3D video experience that can be shared via social networks. A typical SceneNet scenario can be a rock concert, a sports event, a wedding ceremony, breaking news events and any other multiple mobile users' crowded event.

Figure 1. The SceneNet pipeline

The SceneNet pipeline (Figure 1) starts at the mobile devices where the video streams are acquired, pre-processed and transmitted along with a tagging identity to the server. At the server, the various video streams are registered and synchronized and then submitted to 3D reconstruction to generate a multi-view video scene that can be edited and shared by the users. The main achievement of SceneNet is the ability to demonstrate the entire pipeline for dynamic scenes. In the rest of this paper we will describe the main components of the SceneNet pipeline and some examples of the entire SceneNet flow.

2. Creating the Mobile Infrastructure

We aimed to develop the infrastructure on the cellular device for the acquisition, processing and transmission of the video streams to the server. An Android application was developed for capturing video streams from the device camera, pre-processing, compressing and uploading them to the server. The device tagging software collects all the device sensors data: location, altitude, rotation, device unique ID, IP address, device name, battery status information and camera parameters such as focal length. This data is encoded as an XML document and transmitted to the server.

One of the expected bottlenecks within the context of the SceneNet pipeline is the transfer of data between the mobile platforms and the server. In order to generate a good 3D reconstruction not all available data is required. A high quality data with reasonable redundancy is sufficient, as long as we have enough information on the scene from various angles.

The Compact Coding Scheme (CCS) we developed defines the flow of the data in a way that minimizes the required bandwidth of the cellular network, allowing the required data to reach the SceneNet servers and eliminating data which is not needed or wanted. The CCS is composed of video quality estimation and input stream selection algorithms.

Nevertheless, the 3D reconstruction is a computationally intensive task. The computation can be further reduced by reconstructing only objects of interest in the scene or using as input only those videos that contain specific people or objects. In this kind of a content aware 3D reconstruction, object detection technology can be applied on the input videos, and only objects of interest detected in the 2D images, will be processed and reconstructed into 3D scenes. Doing this directly on the mobile devices has the advantages of possibly sparing the need to upload the videos. The state of the art today in object detection is Deep Learning (DL). Applying DL techniques in SceneNet involves a compute challenge: DL networks usually run on high end GPUs that reside on servers. Thus, deploying a DL

network on a mobile device equipped with a much smaller compute power presents a major challenge. We have used state of the art DL networks applied YOLO[1], a recent "thin" network, on Nvidia's Tegra X1.

3. Spatial Calibration

An algorithm was developed for estimating the camera location in an efficient way that is robust to noise and outliers. The main novelty of the proposed method is using the device's rotation matrix obtained through the operating system API. The relative motion, i.e., the rigid rotation and translation, between any two cameras or two images taken at different times, is encapsulated in a 3x3 real matrix known as the fundamental matrix. The fundamental matrix can be estimated from a set of corresponding points between the two images. Corresponding points are obtained by image feature matching which was implemented on the mobile device. The algorithm developed has been extensively tested. In all the experiments the developed algorithm outperformed state-of-the-art techniques with large margin, in both accuracy and efficiency. A comparative experiment was conducted, where the calibration algorithm developed was compared to the state-of-the algorithm of Dalalyan and Keriven[2]. In this experiment, both implementations (Dalalyan *et al* Matlab code is available on-line) were run on the same dataset, feature points and rotation matrices[3]. The results on two datasets, Fountain-P11[4] and HerzJesu-P25[5] are reported in Table 1. The accuracy of the calibration is measured as the normalized distance between the estimated camera locations and the ground-truth values. As can be seen, while the accuracy of the results of both algorithms is comparable on the Fountain-P11, the efficiency of our algorithm is much better. On the MathcesHerzJesuP25 dataset, the results of the algorithm developed in this work outperform the other algorithm both in accuracy and in efficiency.

Table 1. Spatial calibration experimental results.

Fountain-P11 dataset:		
Algorithm implemented	**Accuracy**	**time(sec)**
SceneNet 3-pt alg.	0.000128	15.3
Dalalyan & Keriven	0.000342	265.0
MathcesHerzJesuP25		
	Accuracy	**time (sec)**
SceneNet 3-pt. alg.	0.000968	11.7
Dalalyan & Keriven	0.040161	269.23

The ability to interactively control the viewpoint, while watching a video, is an exciting application of SceneNet. In order to obtain a free viewpoint video we have developed an image interpolation algorithm to generate the non-existing images of a virtual camera moving in a specific trajectory while viewing first a static scene, and then a dynamic scene. An example is illustrated in Figure 2, where two cameras are depicted with their given images and the missing image in the virtual camera (empty pyramid) needs to be estimated. We developed a geometry-based approach that uses, except from the available images, the 3D structure computed.

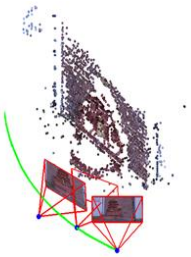


Figure 2. Algorithm for synthesis images of a virtual camera located on an artificial trajectory (green line).

We demonstrate the results of the novel-view synthesis with the Fountain-P11 dataset[6]. The virtual camera path that was defined is shown (in green) in Figure 2. Figure 3 shows two existing images and the novel synthesized image (the right image), and in addition, shows a close-up view of the fountain. More detailed description on the spatial calibration can be found in Egozi *et al*[7].

Figure 3. The results of applying our novel-view synthesis algorithm on the Fountain-P11 dataset. The left and the middle images are part of the dataset, and the right image is a synthesis image on the green path that is illustrated in Figure 2.

4. 3D Reconstruction and Display

We developed the 3D reconstruction based on the patch-based Multi-View Stereo (MVS) algorithm. As demonstrated in[8,9,10], patch-based MVS provides very accurate results without calibration and without relying on strong constraints. The pipeline for the 3D reconstruction is the following: first a dense set of multi-scale binary descriptors is computed for each frame and a pairwise matching is established (using epipolar based sampling and cross-validation). Then from these pairwise matches multiple independent point clouds are reconstructed, and finally the points are filtered using a global reprojection error minimization. The resulting point cloud is dense with respect to the image resolution, *i.e*. a pair of images of N pixels each will generate a point cloud of O(N) points. By design, the algorithm is highly parallelizable: description is parallelizable at the image level, pairwise matching at the pair level and at the pixel level and filtering at the point level. This results in a complexity governed by a set of a huge number of small tasks, which is ideal for GPU-accelerated computation. In Figure 4 we show the reconstruction result from two frames and in Figure 5 the reconstruction results from multiple frames.

Figure 4. Reconstruction results from two frames.

Figure 5. Reconstruction results using all 11 frames.

Point cloud denoising algorithms were developed for static and dynamic use cases. For the spatial denoising of a static point cloud, we construct a graph out of the point cloud by connecting each point to its closest neighbours. Then the graph structure is used to remove outliers coming from wrong matches at the image level. This degree filtering method removes points from the point cloud. Another algorithm that has been developed smooths out the noise in the position. Here points are moved rather than removed by solving a convex optimization problem. Dynamic point cloud denoising has also been tackled. The same outlier removal is used but for the smoothing step at a given frame, the graph is created not only from the point cloud at that frame, but we also connect points to the point clouds from adjacent times. This improves the noise removal because the additive white Gaussian noise averages out to zero over long time series. For the visualization task, work has been done on generating video frames from a point cloud. The approach taken is to exploit the nearest neighbors graph structure of the point cloud. A fully parallelizable algorithm has been developed to learn the underlying manifold given the graph (created from the point cloud). Then, for each pixel that needs to be generated ray tracing is used and thus the frame pixels are a sampling of the reconstructed manifold.

A dynamic scene point cloud visualizer was developed. The point cloud video visualizer is a graphical user interface allowing the user to explore point cloud videos by controlling the movement of a virtual camera (Figure 6).

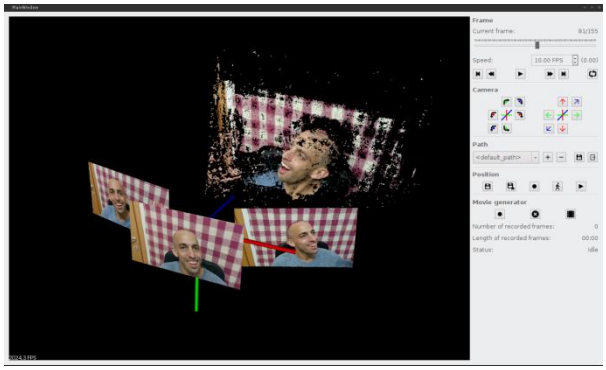


Figure 6. A dynamic 3D scene visualizer developed along with the capability to select the view point of interest.

A second visualizer developed in this project is a web-browser based dynamic point cloud visualization tool, which can be played on any remote computer or mobile device with internet connection. It is based on an open-source code for stereo 3D static point cloud visualization (threejs.org) and can play either a regular video or a stereo video, which can be used by Virtual Reality glasses to provide a real stereoscopic dynamic point cloud.

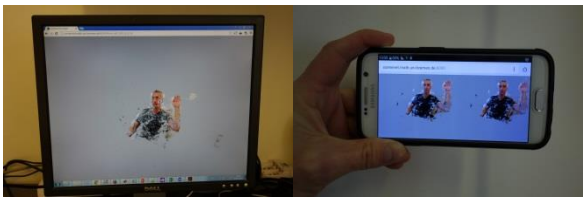


Figure 7. A dynamic point cloud shown on a desktop (single image on the left) and on a mobile device on the right (stereo rendering for virtual reality glasses).

Figure 7 demonstrates a dynamic point cloud shown on a desktop (single image) and on a mobile device (stereo rendering for virtual reality glasses). More details on the 3D reconstruction can be found in our 3DTV paper[11].

5. Computational Acceleration

The SceneNet pipeline involves an extremely compute intensive flow on both the mobile devices used for the acquisition and the server side for the registration, 3D reconstruction and visualization.
The main activity on the server side was to migrate the most time consuming parts of the 3D reconstruction to GPU based platform and speedups of the building blocks of the algorithm are of 1-2 orders of magnitude. The whole 3D reconstruction process, running on 11

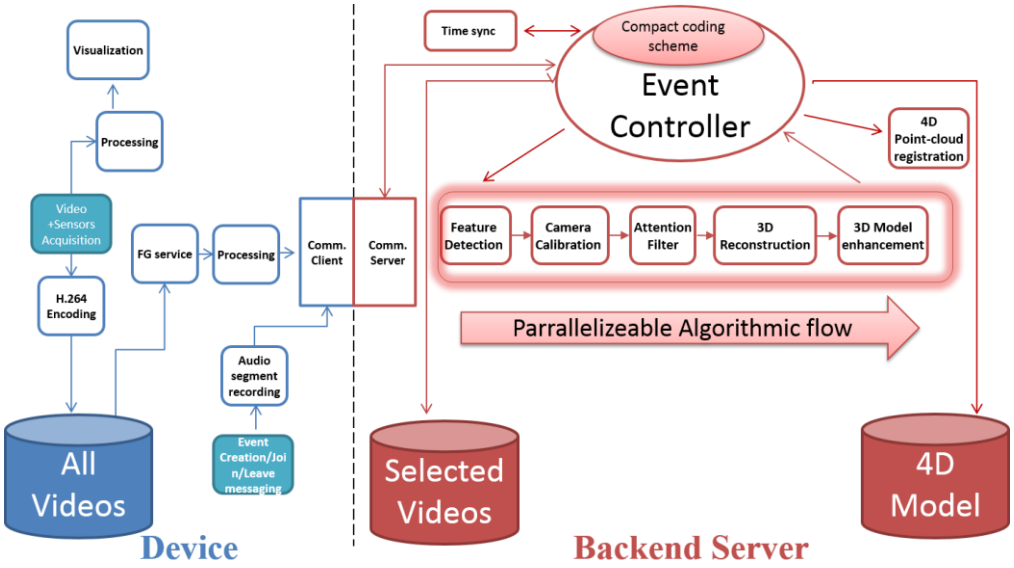


Figure 8: The SceneNet system for acquisition and processing.

input 3072 x 2048 images, utilizing the CPU, took above 2 minutes. The 3D reconstruction process, running on 11 input 3072 x 2048 images, utilizing the K20Xm GPU, took about 8 seconds. In order to gain further acceleration we designed a scalable system: going from a CPU only system, to a CPU-GPU system, to a multi GPU system in a single computer and finally a distributed system with multiple multi GPU computers on the cloud.

6. The SceneNet system

The design of the SceneNet architecture is divided into two parts: (1) acquisition and processing, (2) distribution and visualization. The acquisition and processing design includes the acquisition on the mobile device through transmission to the backend server, to generation of the 4D model on the backend server. The distribution and visualization design includes the flow from existing 4D model to the visualization on many customers' devices (mobile devices or PC's). The system developed allows a fully automatic flow of the SceneNet pipeline - acquisition, transmission, processing and visualization of a dynamic 3D scene generated from inputs from mobile devices. In Figure 8 we present the design of the SceneNet system for acquisition and processing.

The first version of the SceneNet system was a free hand acquisition system where several videos are taken around the object of interest for the 3D reconstruction of static scenes. The next stage involved generating a series of 3D models in time for static models, while the cameras position and orientation changes. The final challenge was to add time synchronization, to allow the reconstruction of dynamic scenes.

Following are video clips generated to showcase the evolution of the SceneNet system:

1. Moving camera with still object simulating still cameras and still objects can be seen in http://youtu.be/AOTYA1JiLu4.
2. A demonstration of a dynamic scene with hand held cameras and moving actors can be seen in https://youtu.be/qjPYr4NTvtI.

7. Conclusion

In this paper we present the main results of the FP7 project SceneNet that merges several 2D video inputs into a unified 3D experience for static and dynamic scenes. In the scope of the research project we provided a proof of concept of the SceneNet system. The open challenges of the SceneNet project revolve around improving the calibration process, especially for dynamic scenes, further improving the visualization utility and acceleration of both the on device and server parts of the SceneNet pipeline.

Acknowledgements

This project is funded by the European Union under the 7th Framework Program (FET-Open SME), Grant agreement no. 309169. The authors would like to thank Matan Milo and Nizan Sagiv for their contributions.

References

[1] YOLO: Real Time Object Detection. http://pjreddie.com/darknet/yolo/
[2] Arnak S. Dalalyan, Renaud Keriven: Robust estimation for an inverse problem arising in multiview geometry. *J. Math. Imaging Vision, 43(1), pp. 10–23.*
[3] Dalalyan A., Keriven R., Robust Estimation for an Inverse Problem Arising in Multiview Geometry http://imagine.enpc.fr/~dalalyan/3D.html.
[4] http://cvlabwww.epfl.ch/data/multiview/fountain_dense.html.
[5] http://cvlabwww.epfl.ch/data/multiview/herzjesu_dense_large.html.
[6] http://cvlabwww.epfl.ch/data/multiview/denseMVS.html.
[7] Egozi, A., Eilot, D., Maass, P., Sagiv,C.: *A robust estimation method for camera calibration with known rotation.* Applied Mathematics, 6(9): 1538-1552, 2015.
[8] Furukawa *et al* , Accurate, Dense and Robust Multi-view Stereopsis, IEEE Transactions on Pattern Analysis and Machine Intelligence, 32(8), 1362-1376, 2007.
[9] Furukawa et. al , Towards Internet-scale Multi-view Stereo, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1434-1441, 2010.
[10] Fua *et al*., Efficient large-scale multi-view stereo for ultra high-resolution image sets, http://cvlab.epfl.ch/research/surface/emvs
[11] Y. Schoenenberger, J. Paratte and P. Vandergheynst, "Graph-based denoising for time-varying point clouds," *3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON), 2015*, Lisbon, 2015, pp. 1-4.

Dov Eilot received his BSc. and MSc. in Biomedical engineering from the Technion, Israel. Dov has 15 years of experience in algorithms and software development and computer vision. Dov served as SceneNet's technical leader.

Yann Schoenenberger received a BSc. and a MSc. in Communication Systems from EPFL and is currently pursuing a PhD at EPFL under the supervision of Prof. Pierre Vandergheynst. Yann's involvement in SceneNet was primarily in the signal processing and point cloud algorithms aspects.

Amir Egozi received the B.Sc., M.Sc. and Ph.D. degrees in computer and electrical engineering from the Ben-Gurion University. He held a Post-Doc position at the department of applied mathematics in the University of Bremen. Currently, he works as a senior algorithm research in MedyMatch Technology Ltd.

Eyal Hirsch is the Mobile GPU leader and a GPU Computing expert at SagivTech. Prior to that, Eyal was a member of AMD's OpenCL team in Israel and a team leader at Geomage, a leading software company in the Oil &

Gas field. Eyal also served as a team leader in Cyota, who was later sold to RSA.

Tal Ben-Nun received his B.Sc. and MSc. in computer science at the Hebrew University of Jerusalem where he is a PhD candidate. His research interests are parallel and distributed algorithms, nonlinear optimization, and massively parallel architectures.

Yossi Appelbaum-Elad holds a MSc. in computer science from the Hebrew University of Jerusalem. Yossi is a computer vision algorithm developer and an entrepreneur. Yossi also teaches Karate and self-defense for women.

Jacob Gildenblat received his BSc. and MSc. in Electrical Engineering focusing on Computer Vision from the Tel Aviv University. In SagivTech Jacob focuses in Computer Vision, Machine learning and GPU Computing.

Eri Rubin serves as SagivTech's VP R&D. Before that, Eri was a Team Leader of CUDA Development at OptiTex and worked as a Senior Graphics Developer for IDT-E Toronto. Eri completed his MSc. in Parallel Computing at the Hebrew University in Jerusalem.

Peter Maass received a Diploma in Mathematics from Heidelberg University and his PhD on 3D X-ray tomography at Technical University, Berlin. Since 1999 he is full professor and director of the center for industrial mathematics at Bremen University. He has been vice president of the German Mathematical Society and he is adjunct professor at Clemson University, South Carolina.

Pierre Vandergheynst received M.S. in physics and Ph.D. in mathematical physics from the Université catholique de Louvain. He was a Postdoctoral Researcher and Assistant Professor at EPFL where he is now a Full Professor. Pierre co-founded two start-ups and holds numerous patents.

Chen Sagiv received a BSc. in Physics and Mathematics, a MSc. in Physics and PhD. in Applied Mathematics from the Tel Aviv University and did her post doc at the University of Bremen. Chen is co-CEO of SagivTech. Chen Served as the coordinator of SceneNet.

SPECIAL ISSUE ON MULTIMEDIA COMMUNICATIONS IN 5G NETWORKS

Guest Editors: [1]Honggang Wang, [2]Guosen Yue
[1]Dept. of Electrical and Computer Engineering, University of Massachusetts Dartmouth, USA
[2]Futurewei Technologies, USA
[1]hwang1@umassd.edu, [2]yueguosen@gmail.com

It has been seen that a new mobile generation has appeared approximately every ten years since the first generation (1G) system. Since the 4G LTE systems were widely deployed recently, the attentions of both industrial and academic researchers have been drawn to the next generation, i.e., the 5th Generation (5G) network with much higher efficiency and lower latency, as well as new use cases such as internet of things (IoT). It is envisioned that the 5G network will be rolled out by 2020. However, reality bites and deploying the new network is costly. It is then a question whether the operators will have motivations to deploy 5G network in the next several years particularly when they have found that the currently deployed 4G network is underutilization. Some killer applications and services would be needed to generate new and/or additional revenues for the mobile operators so that it could make the operators hunger for the new network. One of the important services is the multimedia content over the mobile network that usually consumes large bandwidth. The purpose of this special issue is to bring some new and key enabling technologies in the areas of multimedia communication.

This special issue features five invited articles to identify some interesting research issues and advances in multimedia communications over the future wireless and cellular network. These papers, although not many, cover a wide range of topics on Multimedia Communications in 5G Networks, which we hope to draw broad interests to the readers who research in different areas.

The first article titled "Security Enhancement for Wireless Multimedia Communications by Fountain Code" by Qinghe Du, Li Sun, Houbing Song, and Pinyi Ren presents an overview on the fountain codes with its diverse applications to the security enhancement for the multimedia delivery over wireless channels. In particular, the article introduced recent research outputs on the fountain-code based wireless security enhancing approaches, with some discussions on dynamic code construction, power control, cross-layer cooperative jamming, and content-aware data delivery. In the end, the authors also summarize some unsolved issues and open problems on the fountain codes for secure transmission.

In the second paper "SDN based QoS Adaptive Multimedia Mechanisms and IPTV in LayBack" by Akhilesh Thyagaturu, Longhao Zou, Gabriel-Miro Muntean, and Martin Reisslein, an innovative Software Defined Networking (SDN)-based QoS adaptive mechanism is presented for unicast transmissions and a flow duplication method is proposed for IPTV multimedia multicast transmissions within the context of the LayBack (a layered backhaul) architecture. The proposed two-level QoS adaptive mechanisms for unicast enables an SDN-based network reconfiguration of the backhaul to accommodate the changes in the heterogeneous wireless links and the multitude of device capabilities. The proposed flow duplication method for IPTV multicast transmission eliminates the need for complex protocols in conventional multicast networking thus simplifies the distribution of multimedia content to a large number of users.

The third paper titled "Multimedia Streaming in Named Data Networks and 5G Networks" by Syed Hassan Ahmed, Safdar Hussain Bouk and Houbing Song provides a survey on the recent advancements in the emerging field of Named Data Networks (NDN) and its feasibility check with the promising 5G network architectures. The multimedia streaming support for NDN and 5G is discussed. Moreover, in the paper, the authors identify the gray areas and provide a road map for the research community working in this area.

The fourth paper "RtpExtSteg: A Practical VoIP Network Steganography Exploiting RTP Extension Header" by Sunil Koirala, Andrew H. Sung, Honggang Wang, Bernardete Ribeiro and Qingzhong Liu introduces a pristine concept of steganography for embedding the secret information inside RTP Extension field. The paper also discusses the embedding secret data inside other RTP fields and provided information about the terms used for the

proposed RtpExtSteg implementation. The results from system implementation have shown the success and effectiveness of the proposed method.

In the last paper, "Video Transmission in 5G Networks: A Cross-Layer Perspective" by Jie Tian, Haixia Zhang, Dalei Wu and Dongfeng Yuan, first the impacts of the network parameters at each protocol layer on the video transmission quality are analyzed. Then a comprehensive cross-layer video transmission framework is developed. Based on the developed cross-layer framework, the authors also provide an interference-aware cross-layer video transmission scheme for 5G distributed interference-limited network scenarios.

We would like to thank all the authors for their contributions and great efforts. We hope you enjoy reading this special issue that is devoted to multimedia communications in future mobile networks.

Honggang Wang is an associate professor at UMass Dartmouth. His research interests include Wireless Health, Body Area Networks (BAN), Cyber and Multimedia Security, Mobile Multimedia and Cloud, Wireless Networks and Cyber-physical System, and BIG DATA in mHealth. He has published more than 100 papers in his research areas, including more than 30 publications in prestigious IEEE journals.

Guosen Yue received the Ph.D. degree in electrical engineering from Texas A&M University, College Station, TX, USA, in 2004. He was a Senior Research Staff with NEC Laboratories America, Princeton, NJ, USA, where he conducted research on broadband wireless systems and mobile networks. His research interests are in the general areas of wireless communications and signal processing. For 2013 to 2015, he was with Broadcom Corporation, Matawan, NJ, USA, as a System Design Scientist. Since May 2015, he has been with Futurewei Technologies. He has served as an Associate Editor of the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS and the Symposium Co-Chair of the 2010 IEEE International Conference on Communications (ICC).

Security Enhancement for Wireless Multimedia Communications by Fountain Code

Qinghe Du[1], *Li Sun*[1], *Houbing Song*[2], *and Pinyi Ren*[1]
[1] *Department of Information and Communications Engineering,*
Shaanxi Smart Networks and Ubiquitous Access Research Center,
Xi'an Jiaotong University, China
[2]*Department of Electrical and Computer Engineering, West Virginia University*
Montgomery, WV 25136-2437, USA
{*duqinghe, lisun*}*@mail.xjtu.edu.cn, h.song@ieee.org, pyren@mail.xjtu.edu.cn*

1. Introduction

Multimedia services with great diversities will occupy the majority of market share in future mobile networks. How to provide delay-bounded and reliability-acceptable multimedia services has been constantly attracting research efforts. With the continuous innovation of wireless transmission technologies, the assurance of delay as well as reliability requirements has been a less challenging problem. Yet when the mobile communications system and network has evolved to the 5th Generation (5G) featured by heterogeneous, dense, and ubiquitous access [1]. The security issues for multimedia content delivery have become more critical than ever. The kept-increasing computation capability integrated in mobile devices makes eavesdropping and deciphering much easier tasks than before. Aside from the traditional encryption techniques at high protocol layer, security enhancement via signal design and resource allocation techniques offers powerful ways in degrading eavesdroppers' received signal quality, and thus strengthen anti-eavesdropping capability together with encryption techniques.

Current approaches for security enhancement via signal design and resource allocation, typically dwelled at physical (PHY) layer, can be summarized into two categories: 1) degrading the eavesdropper's channel quality; 2) increasing the legitimate receivers' signal-to-noise ratio. These goals can be achieved by using artificial noise, MIMO techniques, directional antennas, etc., which typically require accurate channel state information (CSI) from legitimate users and sometime the eavesdroppers. Therefore, in realistic systems where accurate CSI is hard to obtain, the effectiveness of these approaches are severely degraded.

Recent research shows that Fountain code [2][3][4][5][5][6][7], a well-known rateless erasure-correcting code with low encoding and decoding complexity, can be applied to enhance wireless security. The name "Fountain code" comes from its transmission style. The transmitter can persistently generate coded packets on top of a number equal-length information packets until the receiver accumulates sufficient number of packets for successful decoding, which is analogous with drinking from water fountain. Fountain code at the packet level, was initially proposed for multicast transmissions [4], where it has the capability to repair diverse packet loss patterns in different receivers. In the meantime, fountain code enabled asynchronous bulk-data download from multiple distributed servers without feeding back the desired sequence number of packets. The well-known fountain codes include Reed-Solomon (RS) code [2], LT code [3], and Raptor code [5]. 3GPP has adopted the Raptor code in technical specification for multimedia broadcast/multicast services (MBMS) [8]. As each fountain-coded packet alone cannot directly give the original content, the receiver has to obtain sufficiently many coded packets to recover the original bulk data. In favor of this feature, fountain code is motivated with new function: secure wireless transmissions [9][10][11], as long as the legitimate user can accumulate sufficient number of packets for decoding before the eavesdropper does.

One major concern for application of fountain code to multimedia communications lies in the contradiction between delay-sensitivity of multimedia streaming and the possible long-delay caused by bulk data decoding. However, it is worth noting that buffering for keep streaming fluency is widely adopted in multimedia services. In such a case, initial and successive buffering can be converted to the fashion of periodical block-based buffering. The block size, even not large for steaming, would be already sufficiently large to implement fountain codes in many cases. This letter would like to conduct a concise introduction to the application of fountain codes to wireless secure multimedia services, and hope to motivate more better yet simple solutions for security assurance in multimedia communications.

2. Fundamental principle of Fountain codes

The fundamental framework for Fountain codes [3][4][6][7] is described as follows. The data stream to be

transmitted is divided into blocks, and each block is further partitioned into packets with equal length (which are called information packets). Fountain-coded transmission is in a block by block fashion, and we thus can focus on the transmission of one block in the following sections.

The summation of two packets is defined as bit-by-bit XOR between the two packets, and clearly this operation generates a coded packet (called parity check packet, or check packet in short) with the same length. In Fountain codes, a check packet results from the summation over several selected information packets, the number of which, called degree of this check packet, is a random variable following certain distribution. Given the degree of a check packet, the selection of information packet follows the uniform distribution. The transmitter persistently generates coded packet and then sends the coded packets one by one to the receiver. Upon receipt of a new coded packet, the receiver attempt to decode through the low-complexity iterative approach [3][6][7], which can be also explained as a special case of the belief propagation scheme. The iterative decoding gradually recovers more and more information packets. Once the entire block is recovered, the receiver will inform the transmitter to stop generating more redundancy packets.

The merits for fountain codes include several folds. First, existing results showed us that Fountain code with optimized design can yield very little overhead for large data block. That is, the average code rate is close to one, making fountain code an extreme efficient approach for error correction in packet level [5]. Second, it does not require packet reordering at the receiver end, and thus enable data download from distributed providers [4]. Third, the fountain code construction can be conducted based on the packet loss status for further performance improvement [6]. Most recently, the merits of fountain codes have been used for secure wireless transmissions, which will be elaborated on in the following sections.

3. Application of Fountain code to enhance security for wireless multimedia communications

The basic principle to use Fountain codes to enhance security is to expedite the transmission of each fountain-coded block, such that the eavesdropper cannot obtain enough packets for entirely decoding, resulting in failure of the data interception. This goal can be typically implemented by taking advantage of features uniquely related to legitimate users' transmissions, such as the CSI, packet loss pattern, data content properties, etc., while assuring these information, even known to the eavesdroppers, does not offer any benefits for them. We summarize a number of such designs, covering the code construction, resource allocation, content-aware design, to faciliate fountain-code based security.

Adaptive fountain code design for security-enhanced unicast and multicast

The essential part for the basic fountain code is the design of degree distribution of the check packet, for either unicast or multicast transmissions. However, when packet loss information can be fed back, such as how many information packets have not been recovered, the transmitter is able to determine the degree via maximizing the probability of loss repairing by the current check packet to be generated and transmitted [6]. Further, given the knowledge of uncovered packets' sequence numbers for desired user or users, the transmitter can deterministically construct the check packet beneficial to legitimate user but not necessarily to the eavesdropper. For example, in unicast, the check packet is constructed as the summation over one unrecovered information packet and all other recovered information packets [11], which assures the successful repair of one unknown information packet for the legitimate receiver. Since the eavesdropper has different loss pattern caused by independent channel fading/error, such design maximally prevent the eavesdropper from recovering any packets in this transmission. This is because as long as this summation involves two information packets unrecovered at the eavesdropper, it cannot effectively help eavesdropper intercept the data. The design with similar ideas also fit multicast well, which, however, still remains as an open problem.

Resource allocation for fountain-coded secure wireless transmissions

Power control over fading channels is a classic topic in wireless transmissions, where water-filling scheme is to maximize the average transmission rate and channel inversion offers sustainable constant-rate data delivery. For the wireless transmissions subject to eavesdropping, the eavesdropper enjoys the broadcast nature of wireless channels. However, in light of small-scale fading, the channels of eavesdropper and legitimate users might be highly different from time to time. Then, power control based on the legitimate user's CSI very likely degrades the channel quality of the eavesdropper, such that it is unlikely for the eavesdropper to complete decoding first before the legitimate user does. Reference [9] devised this strategy. Specifically, the channel inversion power allocation is adopted to match the equal-length fountain-coded packet delivery. As a result, the interception probability decays very quickly with

the increase of fountain-coded block size. This work motivates the new way for power allocation towards security. While it limits the transmission to constant rate, multi-rate enabled secure transmissions requires further attention.

Cross-layer security assurance by integrating fountain codes with cooperative jamming

Cooperative relaying is an efficient paradigm for end-to-end data delivery in multimedia communications. For anti-eavesdropping cooperative transmission, the cross-layer scheme can be exploited, combining fountain coding at the application layer and cooperative jamming at the PHY layer [10]. Specifically, the fountain-coded packets are transmitted over the wireless channels, and a cooperative jammer broadcasts jamming signals as well. To guarantee the legitimate receiver to successfully decode first, the transmitter can rotate the signal constellation and exploit the intrinsic orthogonality between the in-phase and quadrature components of a complex signal, such that the received signal-to-noise-plus-interference ratio (SINR) at the eavesdropper is greatly deteriorated by the jamming signals, whereas the adversary impact from jamming on the signal quality at the legitimate terminals can be significantly reduced under the thorough designs of jamming as well as rotation. This approach can be thoroughly tuned to gain optimized tradeoff among delay, security, and reliability, making it flexible for multimedia communications with diverse requirements.

Content-aware secure multimedia delivery aided by fountain codes

Content type can be of great assistance in secure multimedia transmissions, which, however, been overlooked for a long time. A typical realization of such design, is aided by fountain codes for the medical image transmission. A medical image usually consists of region of interest (ROI) and background (BG). Compared with BG, ROI contains important diagnostic information and needs higher security and reliability requirements. Towards this end, the source image packets can be divided into ROI source packets and BG source packets accordingly. Then, ROI and BG source packets will be fountain-coded separately to obtain ROI and BG check packets. For both of these two types of coded packets, adaptive resource allocation related to the legitimate channel is optimized to ensure a higher packet reception rate at the legitimate receiver. But in the meantime, more resources would be dedicated to ROI rather than BG packets, such that higher security is assured for ROI packets. To shorten the delay of multimedia services, the transmitter can also apply superposition encoding to embed BG packets into ROI packets when the legitimate channel's quality is good. This, in fact, further degrades the reception rate at the eavesdropper, enhancing the security via content-aware design.

3. Concluding Remarks

We presented a highly brief overview on fountain codes with its diverse applications. Further motivating by recent research results, fountain-code based wireless security enhancing approaches are introduced, with specific discussions on dynamic code construction, power control, cross-layer cooperative jamming, and content-aware data delivery. It is for sure that there will be more innovative approaches in building a more secure wireless transmission systems. Some unsolved issues and open problems for secure transmission are summarized in the following.

1) Sliding-window based secure fountain codes

Although the applicability of fountain code for multimedia communications is justified by buffering, it is hard to be applied for real-time streaming with very small buffer size. To solve this problem, we hope to confine the information packet block within a sliding window, where the recovered and new information packets can be pushed out and in window, respectively, thus significantly degrading the delay.

2) Joint source-fountain coding.

Content-aware fountain-code based transmission does not essentially integrate the importance of the source data into code construction. It is highly desirable to make use of the source properties to facilitating security. For example, we expect the eavesdropper can at most intercept less crucial part of data via joint source-fountain coding. .

3) Tradeoff energy and secrecy efficiencies

The major of current research on wireless security faces a challenging problem: spending considerable resources, such as transmission energy, in trading off for secrecy of only a small amount of data. Fountain codes, in light of its blocked data structure, bring a way to increase the transmission efficiency, but is still insufficient. It can be expected that trading off between secrecy and energy efficiencies would be a very important problem and topic in the near future.

Acknowledgement

The work in this paper was supported by the National Natural Science Foundation of China under the Grant No. 61431011 and the Fundamental Research Funds for the Central Universities

References

[1] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano; A.C.K. Soong, J. C. Zhang, "What Will 5G Be?" IEEE Journal on Selected Areas in Communications, vol. 32, no. 6, pp. 1065 - 1082, Jun. 2014.
[2] J. Nonenmacher, E. Biersack, and D. Towsley, "Parity-based loss recovery for reliable multicast transmission," IEEE/ACM Trans. Netw., vol. 6, no. 4, pp. 349-361, Aug. 1998.
[3] M. Luby, "LT codes," in Proc. IEEE 43rd Annu. Symp. Found. Comput.Sci., Nov. 2002, pp. 271-280.
[4] J. W. Byers, M. Luby, and M. Mitzenmacher, "A digital fountain approach to asynchronous reliable multicast," IEEE J. Sel. Areas Commun., vol. 20, no. 8, pp. 1528-1540, Oct. 2002.
[5] A. Shokrollahi, "Raptor codes," IEEE Trans. Info. Theory, vol. 52, no. 6, pp. 2551-2567, Jun. 2006.
[6] X. Zhang and Q. Du, "Adaptive Low-Complexity Erasure-Correcting Code Based Protocols for QoS-Driven Mobile Multicast Services Over Wireless Networks," IEEE Trans. Veh. Tech., vol. 55, no. 5, pp. 1633-1647, Sep. 2006.
[7] D. J. C. MacKay, Information Theory, Inference, and Learning Algorithms, Cambridge University Press, 2006.
[8] 3GPP TS 25.346: Introduction of the Multimedia Broadcast/Multicast Service (MBMS) in the Radio Access Network .
[9] H. Niu, M. Iwai, K. Sezaki, L. Sun, and Q. Du, "Exploiting fountain codes for secure wireless delivery," IEEE Communications Letters, vol. 18, no. 5, pp. 777-780, May 2014.
[10] L. Sun, P. Ren, Q. Du, and Y. Wang, "Fountain-coding aided strategy for secure cooperative transmission in industrial wireless sensor networks," accepted to appear, IEEE Trans. Industrial Informatics, Dec. 2015, Online Available. DOI: 10.1109/TII.2015.2509442.
[11] W. Li, Q. Du, L. Sun, P. Ren, Y. Wang, "Security Enhanced via Dynamic Fountain Code Design for Wireless Delivery," IEEE WCNC 2016, Apr. 3-6, 2016, Doha, Qatar.

Qinghe Du received his Ph.D. degree from Texas A&M University, USA. He is currently an Associate Professor of Department of Information & Communications Engineering, Xi'an Jiaotong University, China. His research interests covers widely on wireless communications and networks, with emphasis on 5G Networks, security, big data, QoS provisioning, multimedia, etc. He has published over 100 technical papers. He received the Best Paper Award in IEEE GLOBECOM 2007. He is serving as an Associate Editor of IEEE Communications Letters and an Editor of the KSII Transactions on Internet and Information Systems.

Li Sun received his Ph.D. degrees in Information and Communications Engineering from Xi'an Jiaotong University, China in 2011. He is currently an Associate Professor at Department of Information and Communications Engineering, Xi'an Jiaotong University. His research interests include cooperative relaying networks, wireless physical-layer security, and D2D communications. He has published more than 50 technical papers. He is serving as an Editor of the KSII Transactions on Internet and Information Systems and a TPC Co-Chair for IEEE ICC 2016 Workshop on Novel Medium Access and Resource Allocation for 5G Networks. He received 2013 IEEE Communications Letters Exemplary Reviewers Award.

Houbing Song received the Ph.D. degree in electrical engineering from the University of Virginia, Charlottesville, VA, in August 2012. He is currently an Assistant Professor and the founding director of the West Virginia Center of Excellence for Cyber-Physical Systems (WVCECPS) sponsored by the West Virginia Higher Education Policy Commission, and the Security and Optimization for Networked Globe Laboratory (SONG Lab). His research interests lie in the areas of cyber-physical systems, internet of things, cloud computing, big data, connected vehicle, wireless communications and networking, and optical communications and networking. Dr. Song has published more than 80 academic papers in peer-reviewed international journals and conferences. Dr. Song is an associate editor for several international journals, including IEEE Access, KSII Transactions on Internet and

Information Systems, and SpringerPlus, and a guest editor of several special issues. Dr. Song has served as the General Chair for several international workshops and conferences.

Pinyi Ren Pinyi Ren received his Ph.D. degrees from Xi'an Jiaotong University, China. He is a Professor of Information & Communications Engineering Department, Xi'an Jiaotong University, in 2001. His current research interests include information security, 5G networks, cognitive radio networks, modeling of fading channels, etc. He has published more than 100 technical papers, and received the Best Letter Award of IEICE Communications Society in 2010.

SDN based QoS Adaptive Multimedia Mechanisms and IPTV in LayBack

Akhilesh Thyagaturu, Longhao Zou, Gabriel-Miro Muntean, and Martin Reisslein
{*athyagat, reisslein*}*@asu.edu, longhao.zou3@mail.dcu.ie, and gabriel.muntean@dcu.ie*

1. Introduction

Software Defined Networking (SDN) has been the key technology for revolutionizing the present state of computer networking. With the unparalleled growth of cellular technologies towards the evolution of 5G systems [1], new requirements and challenges have originated at the cellular backhaul. In particular, small cell deployments have proliferated to meet the data demands and rapid developments of applications, especially in the areas related to the Internet of Things (IoT) and cloud computing. Uncoordinated deployments of ultra-dense networks suffer from interferences between adjacent cells [2], and the numbers of inter-connections within the backhaul increase as the number of cells deployed within the network grows. In order to address some of the critical challenges in the cellular backhaul, we present a novel cellular backhaul architecture, namely the **Layered Backhaul (LayBack) architecture**. In this short letter, we present 1) a method for SDN based QoS adaptive multimedia streaming, and 2) an Internet protocol television (IPTV) [3] multicast transmission technique based on SDN, within the context of LayBack.

The LayBack Architecture

Small cell base stations can simultaneously support a multitude of radio access technologies (RATs), such as Wi-Fi, LTE, and WiMAX. Present day handheld devices are capable of connecting to multiple RATs at the same time. To satisfy the data demands and to accommodate the growth of the numbers of users and devices, deployments of small cells are becoming inevitable. In such situations, backhauling of small cells becomes highly challenging. Layered architectures provide a distinctive advantage in accommodating rapidly advancing wireless technologies in independent layers. For example, a bare SDN switching layer can be replaced with SDN supported fog/edge computing switches or caching enabled switches for content delivery networks. Figure 1 illustrates the high level overview of our proposed SDN-based LayBack architecture. In LayBack, connected radio nodes are identified with their advertised messages sent to the architecture core, the SDN-controller layer. An advertised message may contain RAT capabilities (e.g., supported RAT, center frequency) and functionalities (e.g., supported bands for carrier aggregation). Based on the messages received from a radio node, the SDN-controller configures a backhaul specific to the radio node. In contrast to other proposed architectures [4-9], LayBack consistently decouples the SDN-based backhaul from the radio access network (RAN) entities (e.g., LTE eNBs and WiFi access points).

Generally, each environment has its own characteristics, for example a theater may require networking computationally intensive interactive 3D applications. Computations can be done on cloud servers which reside at the theater premises. An architecture should provide a platform to easily support the management and service for cloud servers within the same environment. Therefore, we believe that micro-architectures that are capable of working independently and are tailored specifically to specific environments would result in good user experience. LayBack supports such microarchitecture deployments adaptive to requirement changes for specific environments.

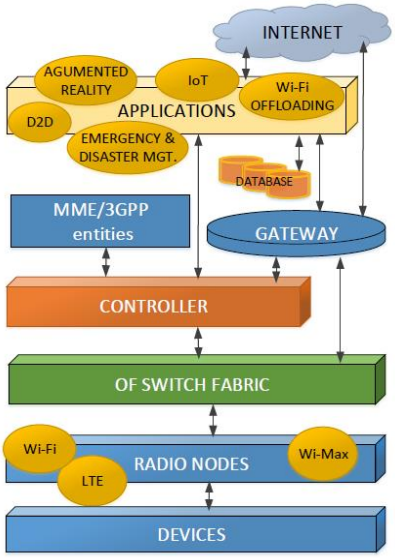


Figure 1. Overview of the SDN-based Layered Backhaul (LayBack) Architecture

SDN Controller and Applications Layer

Applications are programs that are executed by the SDN-controller for delegating instructions to individual SDN-switches, and for communicating with other network entities, such as the 3GPP MME. Network functions, such as Wi-Fi offloading, can be implemented as simple SDN applications.

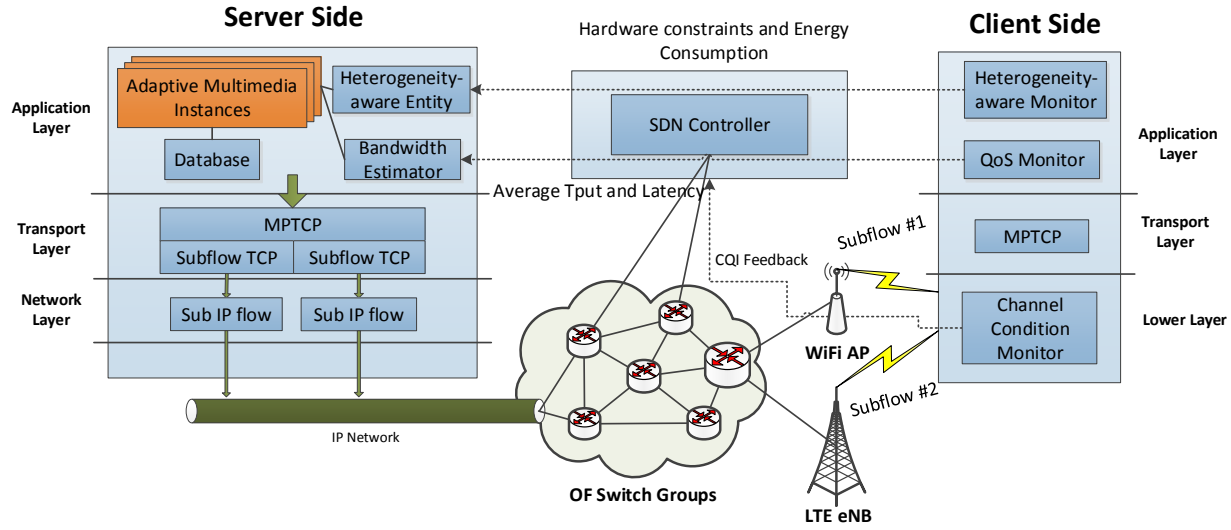


Figure 2: SDN based QoS aware adaptive multimedia transmission mechanism in heterogeneous RAN and devices

SDN-Switch Layer

SDN-Switches are capable of functions, such as forwarding a packet to any port, duplicate a packet on multiple ports, modifying the content inside a packet, or dropping the packet. The switches can be connected in a mesh to allow disjoint flow paths for load balancing.

Gateways Layer

Gateway functions are programmed by the SDN controller to perform multiple network functionalities to

simultaneously establish connectivity to heterogeneous RANs, such as LTE and Wi-Fi.

2. QoS-aware Adaptive Multimedia Transmissions

In an ultra-dense small cell networks, heterogeneity can be associated with wireless technologies (e.g., Wi-Fi and LTE), as well as devices (e.g., smart phones and tablets), along with a multitude of capabilities in computing and display. A device can have simultaneous connectivity via multiple wireless technologies. Multipath-TCP (MPTCP) can be used to accommodate the simultaneous connections due to heterogeneities in the wireless technologies. A protocol overview of the server and client for the adaptive mechanism is shown in Figure 2. Generally, adaptive mechanisms of *unicast* multimedia transmissions involve application layer (single level) adaptation only. In a cellular network and wireless environment, quality of service (QoS) variations perceived by the user can be due to various reasons: 1) network congestion, 2) wireless link disturbances due to interferences, poor signal quality, mobility etc., 3) device power saving mechanisms, and 4) overload due to number of devices connected to the access point or the base station causing exhaustion of wireless physical layer resources. Application layer adaptive techniques react immediately to the changes in the QoS parameters at the devices due to the aforementioned reasons which may tend to be temporary or occur only for very short durations (enough to disrupt the buffer playout). Therefore, sometimes, single level application layer adaptive mechanisms underutilize the network resources.

An SDN based two stage adaptive mechanism

We propose a two stage SDN based adaptive technique at both the network and application layers with a global perspective of the requirements due to heterogeneity in the wireless technologies and devices. A high level overview of the SDN-based two level adaptive mechanism is illustrated by the flow chart in Figure 3.

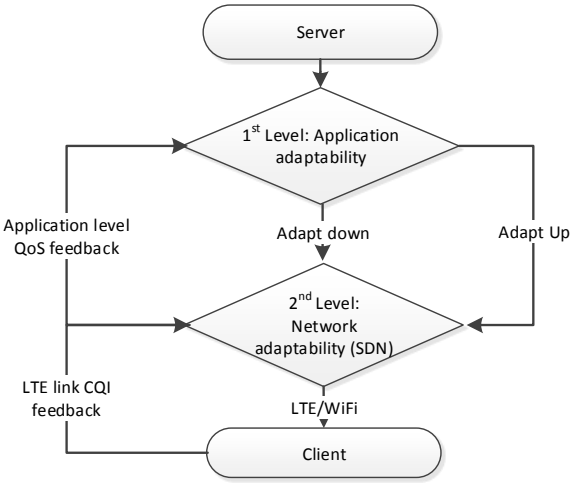


Figure 3: Two stage SDN-based network and application layer adaptive technique

Several global parameters are fed back from the user's device to the SDN controller, including wireless signal quality (CQI in LTE), device capabilities (hardware and power constraints), content type, and service server details. Based on the reported parameters from the device connected network and current availability and reservations, the SDN controller allocates network resources in the cellular backhaul and forwards the information to multimedia servers for service allocation. Any changes in the network, such as congestion, wireless link changes due to interference, or mobility of the device will be analyzed at the SDN controller and decisions to configure the backhaul are made with respect to the global availability and requests in order to achieve the desired QoS with adaptation from both network and applications.

3. SDN-based IPTV Multicast Transmissions

Multicast multimedia transmissions are typically used to distribute the content to large numbers of users over the IP network. Multicast reduces redundant transmissions from the servers and therefore is a popular choice for multimedia distribution applications, such as IPTV. However, conventional IPTV techniques involve numerous

complex distributed network protocols, such as MPLS, IGMP, and RSVP. Zeadally et al. [10] have extensively discussed the challenges of IPTV multimedia distribution.

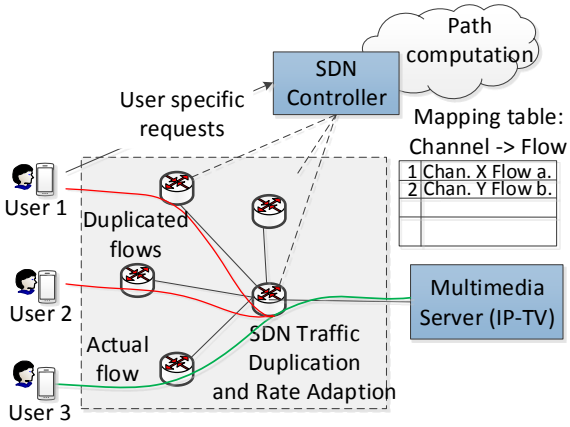


Figure 4: SDN traffic duplication for IP-TV multicast transmissions

As illustrated in Figure 4, we present an SDN-based mechanism for IP-TV multicast transmissions. We exploit the property of SDN traffic duplication to create the multicast flows, eliminating the need for complex protocols in the conventional multicast IPTV transmissions [10]. A user can interact with the SDN controller to switch between the different flows for content reception. Authentication can be provided by the SDN controller. However, the controller may still use protocols such as RADIUS and DIAMETER for communicating with the authentication, authorization, and accounting (AAA) servers. A mapping table which maps the requested service/channel with the flow existing inside the network is maintained at the SDN controller. When a user requests new content, the SDN controller identifies the flow corresponding to the requested content and the traffic is duplicated at the SDN switch nearest to user location, along with the header manipulation for the duplicated packets, such that a flow path is created to the requesting user. If the table lookup for the flow corresponding to the requested content fails, then the SDN controller initiates a new connection with the IPTV multimedia server and the mapping table is updated with new entries. Similarly, when there are no users for a particular service, the SDN controller communicates with the IPTV multimedia server to stop the flow and the corresponding entries in the table are deleted.

4. Conclusion

In this letter we have described an innovative SDN-based QoS adaptive mechanism for unicast transmissions and a flow duplication method for IPTV multimedia multicast transmissions within the context of the LayBack (a layered backhaul) architecture. We have discussed two level QoS adaptive mechanisms for unicast, which enable an SDN-based network reconfiguration of the backhaul to accommodate the changes in the heterogeneous wireless links and the multitude of device capabilities. Our proposed method for IPTV multicast transmission eliminates the need for complex protocols, such as IGMP and MPLS, in conventional multicast networking and simplifies the distribution of multimedia content to a large number of users.

References

[1] Demestichas, P. Georgakopoulos, A. Karvounas, D. Tsagkaris, K., Stavroulaki, V., Jianmin Lu, Chunshan Xiong and Jing Yao, "5G on the Horizon: Key Challenges for the Radio-Access Network," *IEEE Vehicular Technology Magazine* , vol. 8, no. 3, pp. 47-53, Sept. 2013.

[2] Polignano, M., Mogensen, P., Fotiadis, P., Chavarria, L., Viering, I., and Zanier, P., "The Inter-Cell Interference Dilemma in Dense Outdoor Small Cell Deployment," *in Proc. IEEE Vehicular Technology Conference (VTC Spring)*, pp. 1-5, May 2014.

[3] "ITU-T IPTV Global Standards Initiative", 2015. [Online]. Available: http://www.itu.int/en/ITU-T/gsi/iptv/Pages/default.aspx

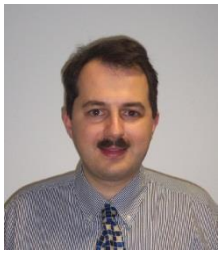
[4] Wang D., Katranaras, E., Quddus, A., Fang-Chun Kuo; Rost, P., Sapountzis, N., Bernardos, C. J., Cominardi, L., and Berberana, I., "SDN-based joint backhaul and access design for efficient network layer operations," *in Proc. IEEE Europ. Conf. on Networks and Communications (EuCNC)*, pp. 214-218, July 2015.

[5] Auroux, S., Draxler, M., Morelli, A., and Mancuso, V., "Dynamic network reconfiguration in wireless DenseNets with the CROWD SDN architecture," *in Proc. IEEE Europ. Conf. on Networks and Communications (EuCNC)*, pp. 144-148, June 2015.

[6] Wang, D., Zhang L., Qi, Y., Quddus, A.u., "Localized Mobility Management for SDN-Integrated LTE Backhaul Networks," *in Proc. IEEE Vehicular Technology Conference (VTC Spring)*, pp. 1-6, May 2015.
[7] Wang, D., Zhang, L., Qi, Y., and Quddus, A.u., "Localized mobility management for SDN-integrated LTE backhaul networks," in *Proc. IEEE Vehicular Technology Conf. (VTC Spring)*, pp. 1-6, May 2015.
[8] Zhang, S., Kai, C., and Song, L., "SDN based uniform network architecture for future wireless networks" *in Proc. Int. Conf. on Computing, Communication and Networking Technologies (ICCCNT)*, pp. 1-5, July 2014.
[9] Oliva, L., De, A., Perez, X. C., Azcorra, A., Giglio, A. D., Cavaliere, F., Tiegelbekkers, D., Lessmann, J., Haustein, T., Mourad, A., and Iovanna, P., "Xhaul: toward an integrated fronthaul/backhaul architecture in 5G networks," *IEEE Wireless Communications*, vol. 22, no. 5, pp. 32-40, 2015.
[10] Zeadally, S., Moustafa, H., Siddiqui, F., "Internet Protocol Television (IPTV): Architecture, Trends, and Challenges," *IEEE Systems Journal*, vol. 5, no. 4, pp. 518-527, Dec. 2011**.**

Akhilesh Thyagaturu is currently pursuing his Ph.D. in Electrical Engineering at Arizona State University, Tempe. He received the M.S. degree in Electrical Engineering from Arizona State University, Tempe, in 2013. He received B.E degree from Visveswaraya Technological University (VTU), India, in 2010. He worked for Qualcomm Technologies Inc. San Diego, CA, as an Engineer between 2013 and 2015.

Longhao Zou (S'12) received the B.Eng. degree in Telecommunication Engineering with Management from Beijing University of Posts and Telecommunications, Beijing, China in 2011. He is currently working towards the Ph.D. degree in Performance Engineering Lab in the School of Electronic Engineering, Dublin City University (DCU). Dublin, Ireland. His research interests include mobile and wireless communications, multimedia adaptive streaming, and user quality of experience and resource allocation. He is a reviewer for international journals and conferences and a member of the IEEE Young Professionals, IEEE Communications Society and IEEE Computer Society.



Gabriel-Miro Muntean (S'02-M'04) received the Ph.D. degree from the School of Electronic Engineering, Dublin City University (DCU), Dublin, Ireland, in 2003 for his research on quality-oriented adaptive multimedia streaming over wired networks. He is currently a Senior Lecturer with the School of Electronic Engineering, DCU, co-Director of the DCU Performance Engineering Laboratory, and Consultant Professor with Beijing University of Posts and Telecommunications, China. He has published over 200 papers in prestigious international journals and conferences, has authored three books and 16 book chapters, and has edited six other books. His current research interests include quality-oriented and performance-related issues of adaptive multimedia delivery, performance of wired and wireless communications, energy-aware networking, and personalized e-learning. Dr. Muntean is an Associate Editor of the IEEE Transactions on Broadcasting, Associate Editor of the IEEE Communication Surveys and Tutorials, and a reviewer for other important international journals, conferences, and funding agencies. He is a member of the ACM, IEEE and the IEEE Broadcast Technology Society.

Martin Reisslein (A'96–S'97–M'98–SM'03–F'14) is a Professor in the School of Electrical, Computer, and Energy Engineering at Arizona State University, Tempe. He received the M.S.E. degree in electrical engineering from the University of Pennsylvania, Philadelphia, in 1996. He received his Ph.D. in systems engineering from the University of Pennsylvania in 1998. Dr. Martin Reisslein has extensive experience in developing and evaluating network architectures and protocols for access networks. He has published over 120 journal articles and over 60 conference papers in the areas of multimedia networking over wired and wireless networks as well as access networks.

Multimedia Streaming in Named Data Networks and 5G Networks

Syed Hassan Ahmed, Safdar Hussain Bouk and Houbing Song
School of Computer Science & Engineering, Kyungpook National University, Korea.
Department of Electrical and Computer Engineering, West Virginia University, WV, USA.
{hassan,bouk}@knu.ac.kr, h.song@ieee.org

1. Introduction

Named Data Network (NDN) has been recently proposed as an alternative to traditional IP-based networking. In NDN, the data is retrieved by a data name, instead of a host identifier (locational identifier). This new type of access methodology rapidly and efficiently disseminates data/content in combination with the in-network caching, depending on the policy. For a practical use of NDN, many network properties studied in IP-based networking are being considered, and new types of ICN architecture components are being designed. NDN is known to be the third revolution in the field of telecommunications. On the other hand, the advancements in networking technologies, mobile services and Internet of Things (IoT) have triggered the need of 5th generation (5G) in the cellular technologies. Some known projects from academia for 5G research are 20BAH [1] and 5GNOW [2]. According to the recent standardization efforts, it is not hard to assume that we will have a real deployment of 5G in next 5 years, where it will support the data rates of 10 to 100 times higher than the current cellular networks. This rapid data rates are foreseen due to the massive content retrieval applications supported by mobile devices during the recent years. Those applications provide HD quality video streaming and other Big Data contents.

No doubt, cellular technologies have evolved quickly from the limited narrowband voice services to the predominant candidate for accessing the varying multimedia services. Today, video streaming has been popular among all the multimedia streaming applications for mobile users. Therefore, few attempts have been made to develop various applications including multimedia streaming supported by NDN and 5G networks, so that in the future users may not witness streaming delays and so on.

In the following sections, we therefore, survey the recent advancements in the emerging field of NDN and its feasibility check with the promising 5G network architectures. Additionally, multimedia streaming support for NDN and 5G will be deliberated. Furthermore, we identify the gray areas and provide a road map for the research community working in the same domain.

2. Named Data Networks and 5G Networks

The Named Data Networking (NDN) is an extension of a well grown-up project Content Centric Networking (CCN) by from PARC [3]. The NDN is providing its part to further enhance CCN project and is funded by US Future Internet Architecture Program. The notion of the NDN is to transform the existing shape of the Internet protocol stack by replacing the narrow thin waist with named data, and under this waist different technologies for connectivity can be used, such as IP. A layer called strategy layer provides the mediation between the underlying technologies and the named data layer.

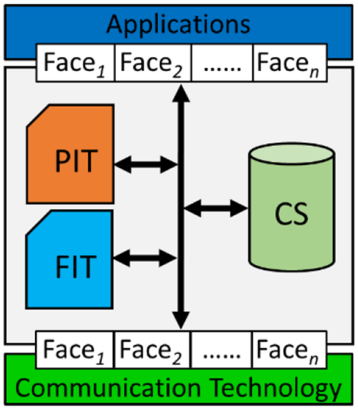


Figure. 1. NDN Overview

In NDN, there are two types of messages used for requesting and routing information, one is Interest message that a subscriber issues for a certain content and in a return publisher provides Data to that subscriber. There are two data-structures (routing tables): Forwarding Information Base (FIB), Pending Interest Table (PIT) and a Content Store (CS) maintained at each content router (CR) in a hop-by-hop fashion for sending interests and replying with content (refer Fig. 1.). The FIB maps the interests received to forward them to interface(s) for publishers or other CRs having content in their caches. The PIT keeps track of the interests for which content is expected to be arriving. Lastly, the CS acts as a cache for the CR to keep the replica of the content that went through that CR. The order of importance among FIB, PIT and CS is that CS reserves the highest priority, then PIT and lastly comes FIB in the priority list. When an interest is received for a content, first a CR checks it into its CS, if found there, then it is returned on the interface at which the interest was received and interest is deleted. If content is not in CS then CR checks its PIT that whether the interest message for the same content has been received earlier, if so the interface over which interest was received is kept as a record in an entry in PIT keeping back track of the interface so that multicast delivery can be performed for multiple subscribers asking for the same content. If there is no entry in PIT then CR makes an entry of it in PIT, and forwards it to other CRs by creating an entry in FIB for this interest. Handling subscriber mobility in NDN is quite similar to that of CCN approach in which the subscriber reissues interest messages from new location, but the content against old interest messages are delivered to old CR. The publisher mobility is difficult to handle because the FIB entries have to be restated when a publisher moves from one CR under a new one. It becomes harder to handle publisher's mobility in very dynamic networks such as MANET. For this purpose, NDN employs Listen First Broadcast Later (LFBL) protocol. In LFBL, a subscriber floods the interest message to all publishers. Any publisher having content against the interest checks the medium that whether any other publisher has already replied with the content or not, if not then sends the content to subscriber.

5G Networks: The 5G mobile telecommunication standard has been conceived for a while now as a successor to the current 4G cellular standards (Long-Term Evolution Advanced or LTE-A) to address the forecasted traffic volume growth, and the inflexibility and difficulties in innovation plaguing the hardware-based architectures at the front end and core of these networks. Aspiring 5G Key Performance Indicators (KPIs) demand for reliable, low latency and high-density networks those capable of supporting Ultra High-Definition (UHD) video. In addition to these, 5G standard proposes better or higher performance and societal KPIs. Detailed descriptions of some of the KPIs in 5G are discussed in [4]. The main KPIs identified include:

1. Traffic volume density
2. Experienced end user throughput
3. Latency
4. Reliability
5. Availability and retain the ability

Energy consumption and cost were included as KPIs with economic relevance. Latency, reliability and availability/retain ability are 3 KPI's important to the current 5G architecture, which will ultimately cater for the end users' throughput with increased traffic volume in applications like video streaming. Since, the 5G networks have Heterogeneous network Architecture, Heterogeneous Terminals and Heterogeneous Spectrum (H3ATS), the integration of new radio concepts such as massive MIMO, ultra-dense networks, moving networks, direct device-to-device communication, ultra-reliable communication, massive machine communication, and others, and the exploitation of new spectrum bands will allow support of the expected dramatic increase in the mobile data volume while broadening the range of application domains that mobile communications can support beyond 2020.

3. Multimedia Streaming in NDN and 5G

This section surveys few recent proposed architectures for video streaming supported by NDN and 5G:

It is also expected that 5G will be enabling NDN routers as a part of mobile networks. For example, in 2013 [5], Adaptive mobile video streaming and sharing in wireless NDN (AMVS-NDN) scheme has been proposed. In this scheme, the dynamic video streaming services is offered based on Dynamic Adaptive Streaming via HTTP (DASH) architecture over different interfaces such as 3G/4G, Wi-Fi and NFC for both real-time and on-demand videos. In reference to DASH stipulations, the goal of using DASH in a highly dynamic condition is to provide a rich quality of experience QoE to users. In DASH, each video is encoded with different resolutions and bit rates, lowest can be 150Kbps for mobile devices and highest can be 6Mbps for devices supporting high definition resolutions. The video is divided into a variable or fixed length chunks or segments customarily between 2-30 seconds in length. A

metadata file called media presentation description (MPD) file contains the information about streams of video and audio such as list of segments, segment size, frame rates, bit rates, and compression scheme etc. To identify multiple copies of same video having different properties, the MPD file serves as basis along with a namespace designed to support this architecture. In AMVS-NDN, the device brings in use heterogeneous network interfaces it has such as 3G/4G, Wi-Fi and NFC. The idea is to reduce the amount of cellular traffic by using interfaces for mobile data offloading through the cellular network and then caching and forwarding through Wi-Fi or other local technologies. To download a video, a consumer issues an interest and, in return get the MPD file. After obtaining the MPD file, the consumer issues the interest and starts downloading the first segment with a conservative approach. In this approach, the very first segment is encoded with the lowest bit rate. While receiving the video segments, the quality of the video is estimated dynamically adapts the video encoding rates according to available bandwidth or signal strength. Following the NDN architecture, a consumer first tries to seek the video segment by sending interest on local network through Wi-Fi. If there is no reply for desired segment, then it forwards the interest in cellular networks. Once the segment against that interest is received, it caches the segment and shares it with other nodes connected through Wi-Fi. For evaluation purposes, a wireless NDN test-bed based on Juni's JPW-8000 WiMAX Base Station is implemented. The application for video streaming uses Android 4.1 platform over HTC EVO 4G+ mobile phone. The experimental results exhibit that AMVS-NDN outperforms both pure streaming via NDN and DASH via NDN in terms of traffic load reduction and average video quality.

Furthermore, in [6], the authors have designed a NDN Video architecture and implemented on top of CCNx that supports both real-time and on-demand video streaming services. The design goals of architecture include efficient QoS, streaming of specific chunks of video based on time code-based namespace, improved synchronization of streams at consumer-side using time code-based namespace, and on-the-go archival of live streams. Gstreamer framework based application is developed for capturing, rendering, and streaming of video. The fundamental facet of NDN Video architecture is its namespace. The namespace designed for NDN Video helps publisher uniquely identify every chunk of the multimedia content. It also helps the consumer to easily solicit a specific chunk based over time in the stream. The names of NDOs are built in order to specify information about the encoding algorithm, video content and sequence number associated to a given chunk. To select the most suitable chunk in a running video stream, a consumer can make use of the time code provided by the publisher in namespace while publishing the multimedia content. The consumer can ask regular consecutive chunks of segments following the randomly accessed part of the video. The real-time streaming is supported by keeping the consumer from sending overwhelmed interests for content that is not yet generated. For this, the estimation procedure at the consumer - side is used. This estimation procedure estimates the time of interests for chunks and rate of content packets generated by the publisher by making use of information stored in a specific field of content packet. The interest not satisfied are retransmitted during a time window whose duration is equal to play out delay. For evaluation, a real-time implementation of NDNVideo is written in Python with open source PyCCN and CCNx. The NDNVideo demonstrates improved QoS of live and pre-recorded videos, also, the packet loss is very less.

Similarly, towards 5G wireless networks, numerous portable terminals (smart phones, tablets, laptops, etc.) and networks with different wireless access technologies (GSM, LTE, WLAN, etc.) will be interconnected. Heterogeneity is a promising future in 5G, which is expected to characterize the mixed usage of access points with different access technologies and spectrum with licensed and unlicensed. Regarding the existing works in 5G radio access networks, most are related to network framework, interference management and cognitive radio. Only a few schemes focus on multimedia traffic management. A promising approach to meet the upcoming 70%-80% mobile traffic generated in indoor is via the deployment of small cells, compared with macro-cellular networks with unacceptable fading of signal in indoor. For the increasingly complex H3ATS scenarios in 5G, it is challenging for 3G/4G small cell networks with traditional centralized network architecture managing multimedia traffic efficiently. Multimedia services exploiting content captured via multiple cameras are increasingly popular, which will be the dominant traffic in the upcoming 5G network. They include visual reality systems, free viewpoint TV, 3D Video etc. On one hand, the transmission of overlapping frames (OFs) in camera streams is redundancy. On the other hand, the bandwidth limited small cell networks cannot cope with the explosive growth in multi-view video (MVV) traffic with high data rate requirements. Thus, it is urgent to solve those two problems jointly for managing MVV traffic efficiently under the demanded QoS. As traditional traffic management schemes for MVV, most are limited to the wired network. In [7], the impact of the placement of in-network processing functions is investigated. In [8], different architectural options for the location of video processing functions are evaluated. Recently, a cloud-clone migration [9] and placement of caching content [10] are evaluated. Although one hop wireless link is considered, few of existing schemes focus on wireless MSCNs. Moreover, in [11], a study of MVV traffic placement in

multicast small cell networks (MSCNs) under H3ATS is done. Firstly, an MVV stream architecture called ORMVVT is proposed by dividing the original streams into overlapping streams (OSs) and non-overlapping streams (NOSs) for separately scheduling. Secondly, a network architecture named MMTS with the consideration of device-to-device (D2D) and Traffic Offloading Point (TOP) is proposed for MSCNs, extending traffic sources from existing small cells to users and network controlled TOPs. Combining the proposed MVV stream architecture ORMVVT with network architecture MMTS, thirdly, an optimization problem is formulated with the goal of offloading data rate for small cells under the QoS constraint. Then, a novel scheme OBTP is proposed not only to realize small cell offloading but to optimize an MVV traffic placement.

4. Future Research Roadmap

Although, there has a been a thorough literature on video streaming in ICN/CCN and NDN, but according to our knowledge, there are still lots of challenges that need to be researched and investigated before proposing NDN as a replacement of current architecture for multimedia, especially. Following are different challenges involved in transferring video as NDOs. Different architectures face distinct challenges while dispatching video from publisher to consumer. We have tried to put an overall insight of the challenges confronted by NDN and 5G.

4.1 Delivery of NDOs

As a video is composed of a number of frames depending upon the video encoding/compression standards. Architectures, using NDN mechanism, have to make sure that the order of the NDOs remains same when a consumer receives them. In architectures, such as NDN, a consumer can put one interest for a chunk of an NDO at a time. If the chunk against that interest is not provided in time, the consumer issues, interests for the same chunk unless or until it is received. This leads in other challenges such as increased time for NDO reception and a number of distant delays.

4.2 Various Delays

For video streaming, the chunks of video are listed in a consecutive order in which they are being requested by consumer. While the consumer needs them within a definite delay for playback. Usually known as playout or target delay can vary according to the type of video transmission, i.e. on-demand or real-time. The reasons behind the playout delay are not only the visible frailties such as network load or congestion, but also the request generation rate of consumers. To provide a smooth playback, for instance, a consumer must be able to know the time at which the content against an interest will be received. If the time exceeds from normal time (the playout delay) then interest can be resent for that chunk or consumer can move on sending interest for the next chunk assuming the chunk is no longer available. In that case, receiver-driven schemes such as an NDN could fail for real-time communications in which consumer requests one chunk at a time. Routing and forwarding operations can be done quickly by performing specific processes, such as use of bloom filters, though, introducing approaches that can handle live-streaming of video is also viable. In various scenarios, the generation of multiple interests at the same time is adequate. However, the challenge of handling those multiple interests without creating congestion is still unanswered.

4.3 Size of data structures and impact of caching

Some researchers have presented in their work a support to the argument that the performance of video streaming can be improved with sophisticated caching strategies. Also, some of the recent researches exhibited that caching the NDOs for video streaming are not essential as they may not be asked again in-network before their expiration. Although in NDN, the introduction of different data structures used by caching strategies has helped a lot to solve the issues of scalability, availability, and fault-tolerance up to some extent, but still the chunking mechanism in some architectures such as NDN can pose a severe issue of the huge size of data structure for video streaming. Also, contrary to the first argument, the number of NDOs that can be asked run-time over the Internet can reach up to 10^8, which is a huge number for a content router with small cache size.

4.4 Effects of forwarding and routing strategies

From request generation to NDO delivery, every process depends upon forwarding and routing strategies. In NDN, most of the architectures use reverse-path or reverse route strategy in which the NDO is replied back by publisher over the path on which the interest for that particular NDO is received. Various facets of NDN are affected by the forwarding and routing strategies that include scalability, network latency, data structures' related delays and many more. Also, the need of delicate forwarding and routing strategies will increase in future with communication standards such as 5G or implementation of NDN in conjunction of other network types such as or sensor networks.

References

[1] Zhou Su, and Qichao Xu, "Content distribution over content centric mobile social networks in 5G," in IEEE Communications Magazine, no. 6, pp. 66-72, 2015.
[2] Gerhard Wunder, et al. "5GNOW: non-orthogonal, asynchronous waveforms for future mobile applications." in IEEE Communications Magazine, no. 2, pp. 97-105, 2014.
[3] V. Jacobson, et al. "Networking named content," in Proc. of ACM CoNEXT, pp. 1-12, 2009.
[4] 5G-PPP, "Key Performance Indicators", available at https://5gppp.eu/kpis/.
[5] B. Han, X. Wang, N. Choi, T. Kwon, and Y. Choi, "AMVS-NDN: Adaptive mobile video streaming and sharing in wireless named data networking," in IEEE INFOCOM WKSHPS, pp. 375 – 380, April 2013.
[6] Z. L. Kulinsky D, et al, "Video Streaming over Named Data Networking," in IEEE COMSOC MMTC E-Letter, vol. 8, no. 4, 2013.
[7] Llorca, J., Guan, K., Atkinson, G., et al. "Energy benefit of distributed in-network processing for personalized media service delivery," in IEEE ICC, pp. 2901-2906, 2012.
[8] Llorca, J., Guan, K., Atkinson, G., et al. "Energy efficient delivery of immersive video centric services", in IEEE INFOCOM, pp. 1656-1664, 2012.
[9] Jin, Y., Wen, Y. and Hu, H., "Minimizing monetary cost via cloud clone migration in multi-screen cloud social TV system", in IEEE GLOBECOM, pp. 1747-1752, 2013.
[10] Modrzejewski, R., Chiaraviglio, L., Tahiri, I., et al. "Energy efficient content distribution in an ISP network", in IEEE GLOBECOM, pp. 2859-2865, 2013.
[11] Quanxin Zhao, et al. "Multimedia Traffic Placement under 5G radio access techniques in indoor environments." in IEEE ICC, pp. 3891-3896, 2015.

Syed Hassan Ahmed did his Bachelors in Computer Science from Kohat University of Science and Technology (KUST), Kohat, Pakistan. Later on, he joined School of Computer Science and Engineering, Kyungpook National University, Korea, where he completed his Masters in Computer Engineering. Currently he is pursuing his PhD in Computer Engineering at Monet Lab, KNU, Korea. Also, he has been a visiting researcher at the Georgia Institute of Technology, Atlanta, USA in 2015. Since 2012, he has published over 50 International Journal and Conference papers in the multiple topics of wireless communications. Along with several book chapters, he also authored 2 Springer brief books. He is also an active IEEE/ACM member and serving several reputed conferences and journals as a TPC and Reviewer respectively. In year 2014-15, he won the successive Gold and Top Contributor awards in the 2nd and 3rd KNU workshop for future researchers, South Korea. His research interests include WSN, Underwater WSN, Cyber Physical Systems, VANETs and Information Centric Networks in Vehicular Communications.

Safdar Hussain Bouk was born in Larkana, Pakistan in 1977. He received the B.S. degree in Computer Systems from Mehran University of Engineering and Technology, Jamshoro, Pakistan, in 2001 and M.S. and Ph.D. in Engineering from the Department of Information and Computer Science, Keio University, Yokohama, Japan in 2007 and 2010, respectively. Currently he is a working as a Postdoctoral Fellow at Kyungpook National University, Daegu, Korea. His research interests include wireless ad-hoc, sensor networks, underwater sensor networks, and information centric networks.

Houbing Song received the Ph.D. Degree in electrical engineering from the University of Virginia, Charlottesville, VA, in August 2012. In August 2012, he joined the Department of Electrical and Computer Engineering, West Virginia University, Montgomery, WV, where he is currently an Assistant Professor and the founding director of both the West Virginia Center of Excellence for Cyber-Physical Systems (WVCECPS) sponsored by the West Virginia Higher Education Policy Commission, and the Security and Optimization for Networked Globe Laboratory (SONG Lab, www.SONGLab.us). His research interests lie in the areas of cyber-physical systems, internet of things, big data analytics, cybersecurity and privacy, and communications and networking.

RtpExtSteg: A Practical VoIP Network Steganography Exploiting RTP Extension Header

Sunil Koirala[1], Andrew H. Sung[2], Honggang Wang[3], Bernardete Ribeiro[4] and Qingzhong Liu[1]*
[1]Department of Computer Science, Sam Houston State University, USA
[2]School of Computing, University of Southern Mississippi, USA
[3]Dept. of Electrical and Computer Engineering, University of Massachusetts Dartmouth, USA
[4]Department of Informatics Engineering, University of Coimbra, Portugal
[1]{sxk033; liu}@shsu.edu; [2]andrew.sung@usm.edu; [3]hwang1@umassd.edu; [4]bribeiro@dei.uc.pt
**correspondence*

1. Introduction

Enormous amounts of digital multimedia data and network packets are being created and communicated around the world continuously. While they have become indispensable in our daily life, digital data can be easily manipulated for malicious or criminal intent, creating rising concern for realistic threats to our society and posing serious challenges to digital forensics and network security.

Steganography, the ancient Greek term for "secret writing" in various forms, has been revived for the Internet to evolve with the advance of digital techniques. One advantage of steganography over cryptography is that the secret message disappears in the cover media which normally would not arouse suspicion by itself. In contrast, plainly visible encrypted data – no matter how unbreakable – are immediately recognizable and attract attention. By taking this advantage of security through obscurity, steganography or hiding covert contents in multimedia is becoming an increasingly popular tool for cyberattacks and cybercrime.

While most of existing steganographic systems and algorithms utilize multimedia data as carriers [9, 10], network steganography, as an emerging subfield in network security, was first coined in a steganography on TCP/IP network wherein the concept of HICCUPS (Hidden Communication System for Corrupted Network) and the network packets are modified. Checksums are tampered with the usage of frames for covert communication [1]. LACK (Lost Audio Packets Steganography) is a steganography concept by hiding data in lost audio packets. In this concept, some audio packets from a VoIP stream are intentionally delayed. Inside the payload of intentionally delayed packets, the hidden information for an unware receiver is invisible [3]. PadSteg (Padding Steganography) makes use of two or more protocols from TCP/IP stack to perform network steganography [7]. RSTEG (Retransmission Steganography) is an inter protocol hybrid network steganography method based on retransmission of transmitted packets. Instead of sending the data that is to be retransmitted, some steganographic approach is used to send steganograms inside such retransmitted packets [5]. In reference [2], the ideas for embedding data on Stream Control Transmission Protocol (SCTP) packets were discussed. StegSuggest utilizes traffic generated by google suggest in order to enable hidden communication. When google suggest provides new words/letters as suggestion in google search, those inserted words are tampered to carry bits of steganograms [6]. TranSteg (Transcoding Steganography) is based on the idea of compression for limiting the size of a data. TransSteg is the overt data that is compressed to make space for covert message. If a voice stream is to be transferred, then a codec is chosen that will result in similar voice quality but smaller voice payload size then the actual codec. When voice stream is transcoded, the original voice payload size is not tampered so that the change of codec is not noticed. After placing the transcoded voice payload, there exists some unused space which is used to send hidden data [8]. SteganRTP exploits embedding secret data with audio data cover-medium [4].

In this paper, we design and implement a practical VoIP network steganography by exploiting the RTP extension header, called RtpExtSteg.

2. RtpExtSteg (Real-time transport protocol Extension header Steganography)

We begin with a brief description of the VoIP protocols and the basic idea of VoIP steganography.

VoIP protocols and steganography

Voice over IP (VoIP) delivers voice communications and multimedia sessions over Internet Protocol networks and creates large amounts of data exchanging on the Internet. VoIP generally makes use of a number of protocols, of which the most common is a standard called Real-time Transport Protocol (RTP) for transmitting audio and video

packets. Real-Time Transport Control Protocol (RTCP) provides control for an RTP session and allows devices to exchange information about the quality of the media session, including such information as jitter, packet loss and a host of other statistics.

Before audio or video data exchange, call-signaling protocols are employed to find the remote device and to negotiate the means by which media will flow between the two devices. The most popular call-signaling protocols are H.323 and Session Initiation Protocol (SIP). Both can be referred to as "intelligent endpoint protocols" to locate the remote endpoint and to establish media streams between the local and remote device. Complementary to H.323 and SIP, another class of protocol is referred to as "device control protocols", such as H.248 and Media Gateway Control Protocol (MGCP) [12].

In addition to H.323/SIP and H.248/MGCP, some non-standard protocols were designed by various companies wherein Skype has been extremely successful using a proprietary protocol [12].

VoIP technology uses Packet Switching technique for routing the data. Packet switching allows the packet to route through least congested paths. To perform steganography over this VoIP call, some of the RTP stream packets are tampered and the hidden message is embedded inside the packets. The hidden message is embedded among some of the RTP packets flowing to and fro on top of VoIP data.

RTP packet header and hiding options

The RTP header has a minimum size of 12 bytes. After the header, optional header extensions may be present, followed by the RTP payload, the format of which is determined by the particular class of application. Figure 1 shows the format of RTP packet header [11].

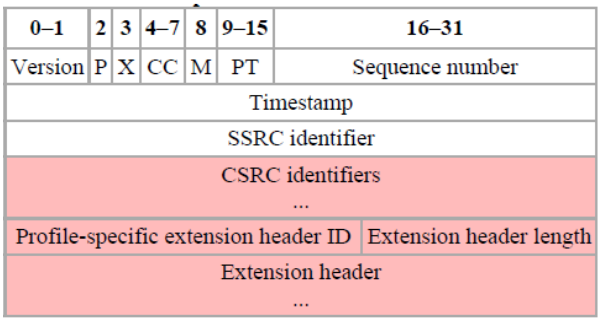


Figure 1. RTP packet header

When these RTP packets are transmitting the VoIP signals, it provides various options to embed the secret information within its header fields. Some of the options available are:

a. *Padding*: This is a single bit field that represents the padding at the end of RTP packet. If padding bit (p) is set, it contains one or more additional padding octets that do not belong to part of payload. This part can be utilized for hiding secret information [8].

b. *Header Extension*: The Header Extension is single bit field that represents if RTP header extension is available or not.

c. *RTP Extension*: This field represents variable length field that contains a 16-bit profile specific identifier and a 16-bit length identifier, followed by variable length extension data. Header Extension (x) and RTP Extension can be combined to hide secret data within it.

d. *Sequence No. and Timestamp Initial Values*: Since the initial values of both these field should be random, it can be used for hiding secret information [6,7].

e. *LSB of Timestamp field*: This least significant bit of timestamp field can also be used in the same way as mentioned above to hide secret information [8].

f. *RTP Payload*: The payload of the RTP header field can also be utilized. The payload can carry any type of data. The original data can be tampered and some steganographic data can be embedded inside it.

RtpExtSteg

RtpExtSteg is a steganography that utilizes the idea of embedding data inside RTP Extension Field. This optional RTP Extension field represents variable length field that contains a 16-bit profile specific identifier and a 16-bit

length identifier, followed by variable length extension data. When a stream of VoIP packets are being transferred, RTP frame carries the audio frame as payload. During this process, RTP extension field can be used to embed covert message. The approach to embed steganography data into RTP packet is based on the fact that RTP standard allows header extension. Header Extension is a RTP header field that defines variable length field that contains a 16-bit profile specific identifier and a 16-bit length identifier, followed by variable length extension data. Any RTP protocol implementation does support RTP header extension, e.g. if software doesn't understand certain header extension then it is just ignored. It means that it is pretty safe to use RTP header extension to transfer steganography data. Another advantage is steganography data will be the part of RTP header, not payload. It means that steganography data can be transferred using RTP stream with any kind of payload.

If header extension is present, the X bit in the mandatory RTP header must be one. A variable-length header extension is appended to the RTP header, following the CSRC list if present or mandatory 12-byte header if not. Figure 2 shows the detailed header extension used in steganography implementation:

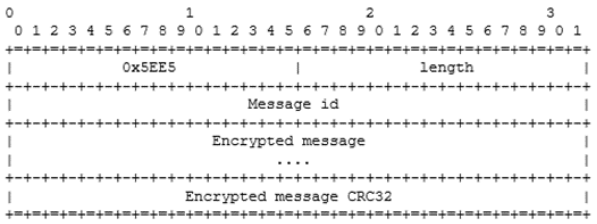


Figure 2: RTP header extension

In Figure 2, the mandatory RTP header is skipped but it is assumed that the X bit of the mandatory RTP header is one. 0x5EE5 is the identifier of our header extension. It's just a random number; any other identifier can be used. The meaning of this identifier is to recognize in our software that it is the steganography header extension. Length is the total length of the following header extension data in 32-bit words. Message id is a 32-bit unique message identifier in the current RTP connection. Message identifier must be a consecutive number starting from 1. Encrypted message is a message text encrypted using AES-128 with pre-defined static encryption key and initialization vector. Because of encryption the encrypted message length is always a multiple of 16, e.g. 16, 32, 48 bytes, etc. Encrypted message CRC32 is a 32-bit CRC of the encrypted message.

There is always a risk of losing and corrupting the data during transmission. The following protection can be applied to prevent message loss and corruption:

1. Message id can be used to control message sequence and re-ordering when necessary.
2. CRC32 code can be used to control if message was altered (i.e. corrupted) during transmission.
3. Sending party sends the message 3 times with 1 second interval between sends. If receiving party received the first transmission of the message then consecutive transmissions can be ignored.

3. Implementation

A working system, 'RtpExtSteg' is implemented and built on C# language. It works well for the steganographic purpose. A demonstration of RtpExtSteg is available on YouTube [13]. Our source code and system are available upon request. The system implements the above written logic to deliver the hidden message. As a result, the sample network dump is extracted from the Wireshark software while the communication is being processed. Figure 3 shows the dump of the sample RTP packet with steganography data and the RtpExtSteg extension header identifier is shown in Figure 4 by using Wireshark.

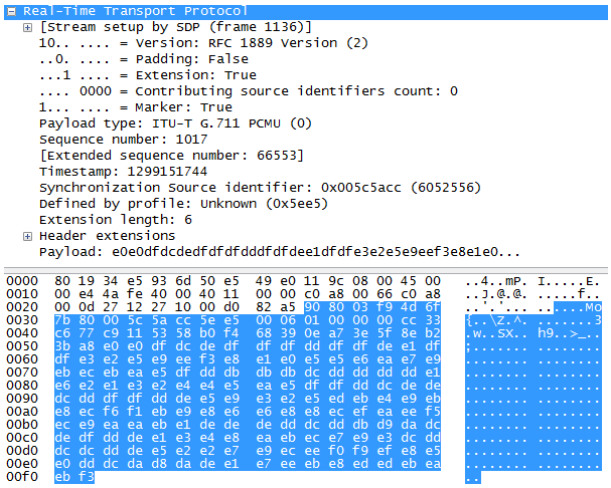


Figure 3. Sample network dump from Wireshark

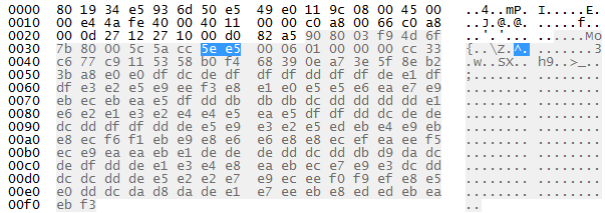


Figure 4. Header extension identifier

Figure 5 shows the extension header length for RtpExtSteg (6 x 32-bit words = 24 bytes):

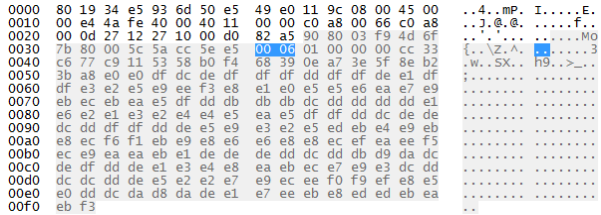


Figure 5. Extension Header Length

Figure 6 shows the 24 bytes of extension header with hidden data: the first 4 bytes 01 00 00 00 are the message id; the next 16 bytes are encrypted message, followed by the 4 bytes CRC32.

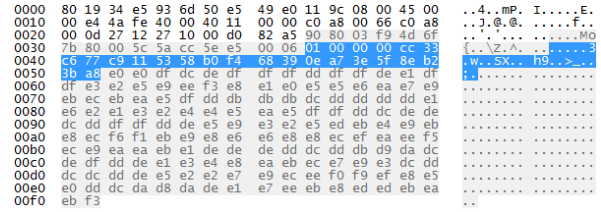


Figure 7. Bytes of Extension Header

4. Conclusion

This paper introduced a pristine concept of steganography for embedding the secret information inside RTP Extension field. We also discussed embedding secret data inside other RTP fields and provided information about the terms used for RtpExtSteg implementation. Our proposed method and system implementation showed the success and effectiveness of RTP network steganography by hiding data in RTP Extension field.

Acknowledgement
The support from National Science Foundation under the award CCF-1318688 is highly appreciated.

References

[1] K. Szczypiorski, (2003). HICCUPS: Hidden Communication System for Corrupted Networks, In Proc. of: The Tenth International Multi-Conference on Advanced Computer Systems ACS'2003, pages 31-40. October 22-24, 2003.
[2] W. Frączek, W. Mazurczyk, K. Szczypiorski (2010). Stream Control Transmission Protocol Steganography. Proceedings of 2010 International Conference on Multimedia Information Networking and Security (MINES), pages 829-834.
[3] W. Mazurczyk and K. Szczypiorski, Steganography of VoIP Streams, In: Robert Meersman and Zahir Tari (Eds.): OTM 2008, Part II - Lecture Notes in Computer Science (LNCS) 5332, Springer-Verlag Berlin Heidelberg, Proc. of OnTheMove Federated Conferences and Workshops: The 3rd International Symposium on Information Security (IS'08), Monterrey, Mexico, November 9-14, 2008, pp. 1001-1018.
[4] I)ruid: Real Time Steganography with RTP, http://druid.caughq.org/presentations/Real-time-Steganography-with-RTP.pdf, September 2007.
[5] W. Mazurczyk, M. Smolarczyk, K. Szczypiorski (2011). RSTEG: Retransmission Steganography and Its Detection, Soft Computing - A Fusion of Foundations, Methodologies and Applications - Special issue on Digital Information Forensics, 15(3): 505-515.
[6] P. Białczak, W. Mazurczyk, K. Szczypiorski (2011). Sending Hidden Data via Google Suggest. CoRR abs/1107.4062.
[7] B. Jankowski, W. Mazurczyk, K. Szczypiorski (2013). PadSteg: Introducing Inter-Protocol Steganography. Telecommunication Systems, 52(2):1101-1111.
[8] W. Mazurczyk, P. Szaga, K. Szczypiorski (2014). Using transcoding for hidden communication in IP telephony. Multimedia Tools and Applications, 70(3):2139-2165.
[9] Q. Liu, A.H. Sung, Z. Chen and X. Huang (2011). A JPEG-based statistically invisible steganography. In Proc. 3rd International Conference on Internet Multimedia Computing and Service, pages 78-81.
[10] V.Sedighi, J. Fridrich, R. Cogranne (2015). Content-adaptive pentary steganography using the multivariate generalized Gaussian cover model. Proc. SPIE 9409, Media Watermarking, Security, and Forensics 2015, 94090H, doi: 10.1117/12.2080272.
[11] https://en.wikipedia.org/wiki/Real-time_Transport_Protocol
[12] https://www.packetizer.com/ipmc/papers/understanding_voip/voip_protocols.html
[13] https://www.youtube.com/watch?v=Bjf5blWWKxo

Sunil Koirala1 received his Master Degree in Computer Science from Sam Houston State University in 2015. His research interests include network security and digital forensics.

Andrew H. Sung received his Ph.D. degree in Computer Science from the State University of New York at Stony Brook in 1984. He is currently the Director and a professor of the School of Computing at the University of Southern Mississippi. His research interests include cybersecurity, big data, bioinformatics, and applications of computational intelligence.

Honggang Wang Honggang Wang is an associate professor at UMass Dartmouth. His research interests include Wireless Health, Body Area Networks (BAN), Cyber and Multimedia Security, Mobile Multimedia and Cloud, Wireless Networks and Cyber-physical System, and BIG DATA in mHealth. He has published more than 100 papers in his research areas, including more than 30 publications in prestigious IEEE journals.

Bernardete Ribeiro is a tenured Professor at the Informatics Engineering Department at the University of Coimbra, Portugal. She received a MSc degree in Computer Science and a PhD in Electrical Engineering, specialty of Informatics. Her current research interests include machine learning and pattern recognition and the applications in bioinformatics, multimedia forensics, and financial data analysis. She is member of ACM, IAPR and senior member of IEEE.

Qingzhong Liu received his Ph.D. degree from New Mexico Institute of Mining and Technology in Computer Science, in 2007. He is currently an assistant professor in the Computer Science Department at Sam Houston State University. His current research interests include multimedia forensics, information security, bioinformatics, data analysis, and the applications of computational intelligence.

Video Transmission in 5G Networks: A Cross-Layer Perspective

Jie Tian[1], Haixia Zhang[1], Dalei Wu[2] and Dongfeng Yuan[1]
[1] Shandong provincial key laboratory of wireless communication technologies, Shandong University, Jinan, China, tianjiesdu@gmail.com, {haixia.zhang, dfyuan}@sdu.edu.cn
[2]University of Tennessee at Chattanooga, Chattanooga, TN, USA, dalei-wu@utc.edu

1. Introduction

Cisco forecasts that nearly three-fourths of the world's mobile data traffic will be video by 2019. Mobile video will increase 13-fold between 2014 and 2019, accounting for 72 percent of total mobile data traffic by 2019 [1]. With the explosive growth of mobile video traffic and mobile video-enabled wireless devices (e.g., smartphones), the demands for video services in future 5G networks will also increase dramatically. Therefore, how to satisfy user video service requirements and improve user experience in 5G networks is becoming a major concern. In reality, providing satisfying quality of service (QoS) and quality of experience (QoE) for video services in 5G networks especially in low-latency high-reliability scenarios, e.g., device-to-device (D2D) still faces substantial transmission challenges due to the resource-constrained devices, heterogeneous traffic types, network constraints, complicated link interference, varying channel quality as well as the stringent delay requirement, etc.

Both video transmission quality and QoE not only depends on the video traffic types and video coding rate determined by the video coding parameters at the application layer, but also on the transmission routing, network parameters, scheduling scheme and the radio resource allocation strategies at the multiple lower layers. Therefore, cross-layer design through allowing information exchanges among multiple layers has been thought as a significant potential solution to enhance video transmission quality for the current and future 5G networks [2-3]. Currently, there has been a large amount of work on cross-layer video transmission in wireless network such as our prior work in [4-7]. Readers are referred to [9-11] and references therein for recently comprehensive surveys of this area. Although cross-layer design methods have been widely adopted in literature for video transmission, most of the existing cross-layer solutions focus on investigating different design objectives of different network scenarios through jointly optimizing two or three protocol layers. To the best of our knowledge, there is still lack of a comprehensive analysis and design framework of video communication for future typical low-latency high-reliability 5G network scenarios.

In this letter, aiming to shed a new light on the analysis and design of the multimedia communication for future 5G networks, we first analyze how decision made by system variables at each protocol layer impacts the quality of video transmission. Then, we further develop a comprehensive cross-layer framework for video transmission. Building on this framework, we propose an interference-aware cross-layer video transmission scheme for future 5G distributed network scenarios such as D2D, vehicle-to-vehicle (V2V) as well as machine-to-machine (M2M) communications. Finally, we conclude this letter.

2. Cross-layer framework for video communication

In this section, we first analyze and discuss the impacts of the key design parameters at each protocol layer on video transmission quality as follows:

Physical layer. The resource allocation strategies, link adaptation scheme, as well as the coding and modulation methods at the physical layer influence the channel capacity and bit error rate of the video transmission.

MAC layer. Medium access control protocol at the MAC layer is responsible for fairly sharing of the wireless medium resource among different users in wireless network. The MAC protocol will directly impact user's access probability and further influence the queue delay, etc. Designing proper MAC protocol can greatly improve the video transmission quality [8].

Network layer. The network layer controls the routing strategy (i.e., path selection) for video packets especially over wireless multi-hop networks. Different routing decisions can significantly impact the end-to-end delay as well as video transmission quality.

Transport layer. Transport layer controls the end-to-end delivery of the video packets [10]. The designed transport control protocol should be able to prevent the network congestion when multiple video streaming sharing the

network especially for the ad hoc network.

Application layer. For different video traffic types, the application layer is in charge of the compression and coding of the video sequences. The different compression techniques can impact the video source distortion and the video transmission quality. In addition, the application layer can also determine the performance metric of video transmission such as video distortion, PSNR, etc. These metrics directly affect the parameter decision at the lower layers.

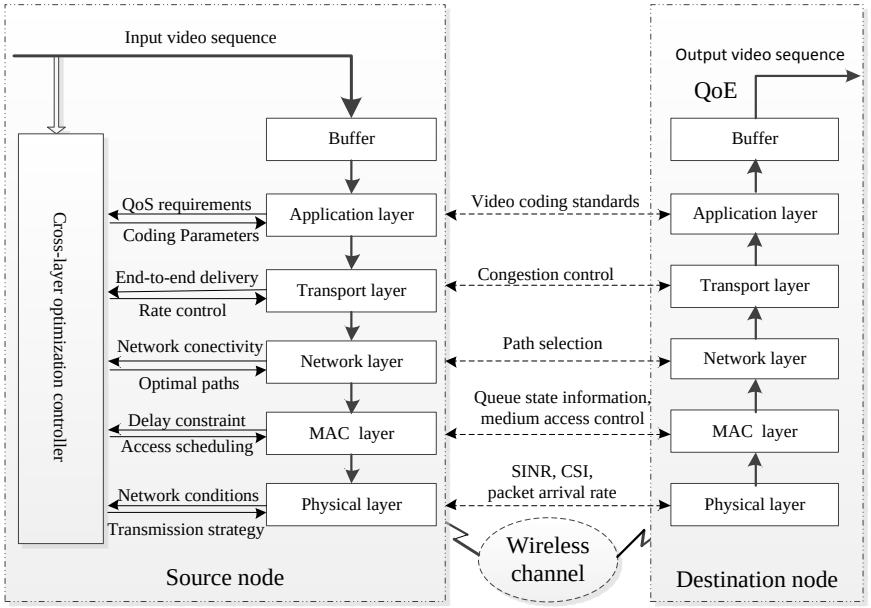


Figure 1. The system diagram of cross-layer video transmission framework

Based on the aforementioned analysis at each protocol layer, it can be seen that video transmission quality and QoE involve different protocol layers and the network parameters from different layers also interact with each other, hence the cross-layer design comes as a natural solution for video transmission.

Next, we present a comprehensive system framework of cross-layer video transmission, as shown in Figure 1. From Figure 1, we can see each layer has one or multiple key design parameters and key technologies which significantly affect the overall design objective. In addition, these system parameters from different protocol layers interact with each other. For instance, the video encoding parameters setting determines the send bit rate and further impacts the rate control at transport layer as well as the queue state at MAC layer. Moreover, both the network conditions and adopted key technologies at the physical layer will impact the link effective capacity and also further influence queue state and delay.

A cross-layer optimization controller was designed to optimize the design parameters from multiple layers [5]. To this end, first, the controller acquires the corresponding network parameter information, such as the video distortion from the application layer and the network conditions from the lower layers. Then, different optimization techniques and strategies could be implemented and deployed at the controller based on its computational capability. Besides the cross-layer optimization, to improve the cross-layer design performance, the controller could also be designed to be able to derive the interactions among and the significance of different design parameters by performing parameter sensitivity analysis on the design objective based on the collected information.

3. Cross-layer design trends for 5G networks.

Future networks will become much denser with more femtocells or picocells as well as device-to-device communications [2]. Along with the network densification, the interference will become one of the most important performance-limiting factors. Therefore, how to design interference-aware adaptive cross-layer video transmission scheme should be envisaged for heterogeneous networks. In addition, in the ultra-dense deployment of future network, many devices could be connected in a distributed manner such as V2V and D2D, then in practice how to

design a distributed cross-layer video transmission scheme is also significant. Moreover, many devices in 5G network are energy-constrained, and the multimedia traffic is delay-constrained [3]. Therefore, one of the major challenges in future 5G networks is how to achieve the optimal design trade-off between energy consumption and delay QoS guarantee for multimedia communications. A delay and energy aware cross-layer scheme could be a potential solution [9].

4. Interference aware cross-layer video transmission

Considering the distributed deployment of 5G low-latency high-reliability scenarios such as D2D and V2V communications, in this section, we propose an interference-aware cross-layer scheme for distributed video transmission under delay constraint over interference-limited distributed networks [8]. We expect the proposed cross-layer scheme could shed a new light on the future multimedia transmission scheme design in 5G networks.

Figure 2 illustrates the main idea of the cross-layer scheme. In this scheme, important system parameters including the distribution of source nodes and the resulting interference, link transmission strategy, queueing delay as well as the video encoding rate are jointly considered to achieve the best received video quality. To achieve this, we first propose a threshold-based link transmission strategy and then develop an interference approximation model.

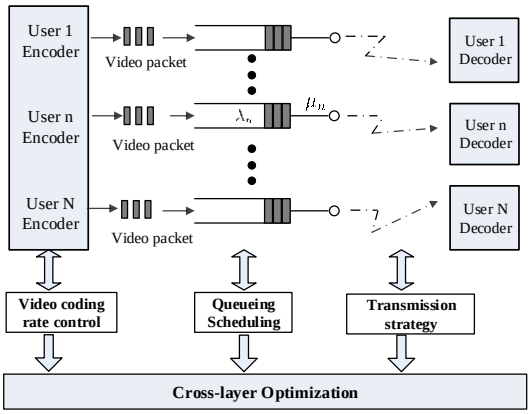


Figure 2. System model of interference-aware cross-layer video transmission scheme [8].

Threshold based video transmission strategy.

In the threshold-based video transmission strategy, each user decides whether to transmit one packet on the selected channel by comparing the channel gain with a given threshold in a given slot. If the channel gain is above the threshold, the packet will be transmitted, otherwise enqueued in its buffer. Hence, the predefined threshold will determine the transmission probability of video packets. In addition, since all users are coupled with each other through interference, the transmission threshold of each user affects all the rest users. Furthermore, either overhigh or overlow transmission threshold will result in a higher packet loss rate, and thus decrease the video transmission quality. Therefore, the transmission threshold is an important design parameter. Moreover, when each user sets its own transmission threshold, the video encoding rate and the channel conditions should be jointly taken into account.

Interference distribution model.

To better analyze the video transmission quality, the stochastic characteristics of interference should be captured. According to the transmission policy, if one channel is selected by multiple users to transmit simultaneously, these users may cause interference to each user. In addition, both users' positions and the transmission thresholds of all users influence the interference distribution. Based on stochastic theory, we propose a log-normal distribution model to characterize the real interference distribution.

Formulated problem and proposed solution.

With the transmission policy and interference distribution model, the queueing delay is analyzed. Finally, the problem of maximizing video transmission quality is formulated as a cross-layer optimization problem by jointly considering network transmission strategy and application performance. To solve this optimization problem, we propose a distributed algorithm based on optimization theory and game theory. Detailed solution is referred to the work in [8].

5. Conclusions

In this letter, we analyzed the impacts of the network parameters at each protocol layer on the video transmission quality. A comprehensive cross-layer video transmission framework was presented. Based on the cross-layer framework, we presented some potential research trends for cross-layer video transmission in 5G networks. Finally, we presented an interference-aware cross-layer video transmission scheme for 5G distributed interference-limited network scenarios.

References

[7] Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update 2014–2019 White Paper, http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white_paper_c11-520862.html.

[8] L. Pierucci, "The quality of experience perspective toward 5G technology," IEEE Wireless Communications, vol. 22, no. 4, pp. 10-16, Aug. 2015.

[9] W. Wang and V. K.N. Lau, "Delay-aware cross-layer design for device-to-device communications in future cellular systems," IEEE Communications Magazine, vol. 52, no. 6, pp. 133-139, Jun.2014.

[10] H. Zhang, Y. Ma, D. Yuan, and H. H. Chen, "Quality-of-Service Driven Power and Sub-Carrier Allocation Policy for Vehicular Communication Networks," IEEE J. Sel. Areas Commun., vol. 29, no. 1, pp. 197-206, Oct. 2011.

[11] D. Wu, S. Ci, H. Luo, and H. Guo, "A theoretical framework for interaction measure and sensitivity analysis in cross-layer design," ACM Trans. Model. Comput. Simul. vol.21, no. 1, pp.1-23, 2010.

[12] Z. Guan, T. Melodia, and D. Yuan, "Jointly Optimal Rate Control and Relay Selection for Cooperative Wireless Video Streaming," IEEE/ACM Trans. Netw, vol. 21, no. 4, pp. 1173-1186, August 2013.

[13] D. Wu, S. Ci, H. Wang, and A. K. Katsaggelos, "Application-Centric Routing for Video Streaming Over Multi-Hop Wireless Networks," IEEE Trans. Circuits Syst. Video Techn. vol.20, no. 12, pp.1721-1734, Dec.2010.

[14] J. Tian, H. Zhang, D. Wu, and D. Yuan, "Interference-Aware Cross-layer Design for Distributed Video Transmission in Wireless Networks," IEEE Trans. Circuits Syst. Video Technol., May 2015, in press.

[15] Y. Ye, S. Ci, N. Lin and Y. Qian, "Cross-layer design for delay-and energy-constrained multimedia delivery in mobile terminals," *IEEE Wireless Communications*, vol. 21, no. 4, pp. 62-69, Aug. 2014.

[16] S. Pudlewski, N. Cen, Z. Guan, and T. Melodia, "Video Transmission over Lossy Wireless Networks: A Cross-layer Perspective," IEEE J. Sel. Topics Signal Process., vol. 9, no. 1, pp. 6–22, Feb. 2015.

[17] S. Ci, H. Wang, and D. Wu. "A theoretical framework for quality-aware cross-layer optimized wireless multimedia communications." Advances in Multimedia, vol.2008.

Jie Tian received the BE degree from Shandong Normal University in 2008, and the ME degree from Shandong Normal University in 2011, both in Electrical Engineering in China. She is currently a PhD student at the School of Information Science and Engineering, Shandong University, China. Her current research interests are in cross-layer optimizations for multimedia communication over wireless networks, dynamic resource allocation for heterogeneous network.

Haixia Zhang received the BE degree from the Department of Communication and Information Engineering, Guilin University of Electronic Technology, China, in 2001, and received the MEng and PhD degrees in communication and information systems from the School of Information Science and Engineering, Shandong University, China, in 2004 and 2008. From 2006 to 2008, she was with the Institute for Circuit and Signal Processing, Munich University of Technology as an academic assistant. Currently, she works as full professor at Shandong University. She has been actively participating in many academic events, serving as TPC members, session chairs, and giving invited talks for conferences, and serving as reviewers

for numerous journals. She is the associate editor for the International Journal of Communication Systems. Her current research interests include cognitive radio systems, cooperative (relay) communications, cross-layer design, space–time process techniques, precoding/beamforming, and 5G wireless communications.

Dalei Wu received the B.S. and M. Eng. degrees in electrical engineering from Shandong University, Jinan, China, in 2001 and 2004, respectively, and the Ph.D. degree in computer engineering from the University of Nebraska-Lincoln, Lincoln, NE, USA, in December 2010. From 2011 to 2014 he was a Post-Doctoral Researcher with the Mechatronics Research Lab in the Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA. He was a member of OFDMA Technical Staff with ZTE Corporation, Shenzhen, China from 2004 to 2005. Currently, he is an assistant professor with University of Tennessee at Chattanooga, Chattanooga, TN, USA. His research interests include wireless communications and networking, mobile computing, cyber-physical systems, and complex dynamic system modeling and optimization.

Dongfeng Yuan received the MS degree from the Department of Electrical Engineering, Shandong University, China, 1988, and obtained the PhD degree from the Department of Electrical Engineering, Tsinghua University, China in January 2000. Currently, he is a full professor in the School of Information Science and Engineering, Shandong University, China. From 1993 to 1994, he was with the Electrical and Computer Department at the University of Calgary, Alberta, Canada. He was with the Department of Electrical Engineering in the University of Erlangen, Germany, from 1998 to 1999; with the Department of Electrical Engineering and Computer Science in the University of Michigan, Ann Arbor, USA, from 2001 to 2002; with the Department of Electrical Engineering in Munich University of Technology, Germany, in 2005; and with the Department of Electrical Engineering Heriot-Watt University, UK, in 2006. His current research interests include cognitive radio systems, cooperative (relay) communications, and 5G wireless communications.

MMTC OFFICERS (Term 2014 — 2016)

CHAIR

Yonggang Wen
Nanyang Technological University
Singapore

STEERING COMMITTEE CHAIR

Luigi Atzori
University of Cagliari
Italy

VICE CHAIRS

Khaled El-Maleh (North America)
Qualcomm
USA

Liang Zhou (Asia)
Nanjing University of Posts &Telecommunications
China

Maria G. Martini (Europe)
Kingston University,
UK

Shiwen Mao (Letters & Member Communications)
Auburn University
USA

SECRETARY

Fen Hou
University of Macau, Macao
China

E-LETTER BOARD MEMBERS (Term 2014—2016)

Periklis Chatzimisios	Director	Alexander TEI of Thessaloniki	Greece
Guosen Yue	Co-Director	Futurewei Technologies	USA
Honggang Wang	Co-Director	UMass Dartmouth	USA
Tuncer Baykas	Editor	Medipol University	Turkey
Tasos Dagiuklas	Editor	Hellenic Open University	Greece
Chuan Heng Foh	Editor	University of Surrey	UK
Melike Erol-Kantarci	Editor	Clarkson University	USA
Adlen Ksentini	Editor	University of Rennes 1	France
Kejie Lu	Editor	University of Puerto Rico at Mayagüez	Puerto Rico
Muriel Medard	Editor	Massachusetts Institute of Technology	USA
Nathalie Mitton	Editor	Inria Lille-Nord Europe	France
Zhengang Pan	Editor	China Mobile	China
David Soldani	Editor	Huawei	Germany
Shaoen Wu	Editor	Ball State University	USA
Kan Zheng	Editor	Beijing University of Posts & Telecommunications	China