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

R-LETTER

Vol. 4, No. 3, June 2013

IEEE COMMUNICATIONS SOCIETY

CONTENTS

Message from the Review Board

Introduction

The Review Board is glad to report the number of distinguished article nominations by MMTC Interest Groups – QoEIG (3), ASPIG (3), VAIG (2) and IMVIMEC IG (3), for the April and June issue. The IGs' contribution has broadened the scope of high quality review papers recommended to the MMTC community. I would take this opportunity to thank a number of Review Board colleagues who have completed a total of 10 tasks (nomination, review and editorial combined) in the last 4 issues: Koichi Adachi, Xiaoli Chu, Ing. Carl James Debono, Guillaume Lavoue, Vladan Velisavljevic and Jun Zhou. This effort requires personal sacrifice given their busy schedules.

In this issue, the two distinguished articles focus on a continuous challenging research topic: optimize the trade-off between rendered quality and communication constrains for visual data. In particular, the authors discuss effective scanning and perceptually driven compression of images/videos. It is interesting to note that a lot of efforts are devoted to integrating 3D depth sensing and multi-view technology with traditional image/video analysis.

Distinguished Category

With the growing popularity of the applications that use large amounts of visual data, image and video coding is an active and dynamic field. Coverage of both image and video compression yields a unique, self-contained reference, appropriate for all related professions. Image and Video Compression for Multimedia Engineering builds a basis for future study, research, and development. Advanced technologies have increased demands for visual information and higher quality video frames, as with 3-D movies, games, and HDTV. This taxes the available technologies and creates a gap between the huge amount of visual data required for multimedia applications and the still-limited hardware capabilities. Image and Video Compression for Multimedia Engineering bridges the gap with concise, authoritative information on video and image coding.

The first paper, published in *IEEE Transactions on image processing 2012* (vol. 21, no. 4, April 2012), evaluates classic scanning strategies of modern bitplane image codecs using several theoretical-practical mechanisms conceived from rate–distortion theory. **The second paper**, published in *IEEE Transactions on Circuits and Systems for Video Technology* (vol. 21, no. 6, June 2011), proposes a H.264/AVC-based perceptual video codec integrating a ST-JND model to perceptually modulate the quantization steps for each DCT coefficient.

Regular Category

While compression techniques for image/video have been extensively studied in the literature, efficient compression and evaluation of multi-view and multi-modular data, including 3D depth data, still faces a sharp learning curve.

The first paper, published in the *IEEE Transactions on Circuits and Systems for Video Technology*, introduces a non-linear filtering technique to compress depth map. The **second paper**, from *IEEE Information Theory and Applications Workshop*, switches the focus from video codec to speech codec and proposes a perceptually based distortion measure for wideband speech. With increasing 3D data integrated into video streams, the **third paper**, published in *IEEE Transactions on Multimedia*, proposes an efficient fine-granular scalable coding algorithm to compress 3D mesh sequences. The **forth paper**, is from *IEEE Transactions on Image Processing*. Instead of matching 3D geometry, the authors analyze a cluster of 2D views using hypergraph for 3D object recognition and retrieval. The **fifth paper** is published in *IEEE Transactions on Multimedia* and proposes a real-time algorithm for reconstruction of moving 3D objects captured from multiple depth sensors.

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

IEEE ComSoc MMTC R-Letter

Evaluates classic scanning strategies of modern bitplane image codecs

A short review for "Scanning order strategies for bitplane image coding"

Edited by Weiyi Zhang

Francesc Aulí-Llinàs, and Michael W. Marcellin (FIEEE),"Scanning order strategies for bitplane image coding", IEEE Transactions on Image Processing, Vol. 21, No. 4, April 2012.

During the last 20 years there have been, roughly speaking, four generations of wavelet-based image codecs. The first generation included resolution scalability. The second generation introduced quality progressivity and enhanced coding performance through the use of probability models and context based arithmetic coding. The third generation achieved simultaneous scalability in terms of resolution, quality, and spatial location, which triggered new possibilities for the interactive transmission of images. The latest generation of image codecs has provided highly specialized systems that incorporate novel features for specific applications and scenarios.

There are myriad techniques and mechanisms employed by the coding systems and compression standards introduced in the last two decades, but at the core of almost all of them lies a bitplane coding engine. Bitplane coding is a quality progressive coding strategy that can exploit the psycho-visual redundancy of imagery. Its popularity hails from its competitive coding performance and its use of binary representations for wavelet coefficients, which are very convenient for current hardware architectures. Bitplane coding is currently the prevailing technology for lossy (or lossy-to-lossless) image compression. Key in bitplane coding technology is the scanning order, which determines the sequence in which coefficients are visited, the coding passes employed to code each bitplane, and the order in which the coding passes are sorted. Despite the great interest in bitplane coding strategies, they have been generally envisaged from theoretical or experimental insights, without employing rate-distortion models to aid in their conception. The first contribution of this paper is a theoretical-practical framework for the evaluation and the formation of bitplane coding engines. The underpinnings of this framework arise via the latest advances in rate-distortion theory. More precisely, the framework employs distortion estimators to yield high precision approximations of distortion decreases produced when coefficients are coded progressively. It also employs two modeling approaches for high-order statistics of symbols emitted by the coding engine. One of these models is based on the significance state of coefficients, motivated by most recent systems and standards. The other model employs the local average of coefficients, which is a recently introduced concept that permits insightful evaluations of the scanning order. The theoretical-practical framework results in three additional key contributions. First, it is employed to carry out a retrospective analysis of bitplane coding technology. This analysis explains - from a theoretical point of view - the excellent coding performance achieved by bitplane coding engines. It also justifies the use of multiple coding passes to code each bitplane, and the optimal order in which the coding passes should be sorted. Second, the framework is employed to evaluate state-of-the-art systems and new theoretical and practical strategies. This analysis reveals the degree of optimality that current systems achieve. Furthermore, it establishes upper bounds on the coding performance and computational throughput that bitplane coding technology can achieve. Third, the framework is employed to discern the mechanisms of coding engines that are essential from those that are not. This provides perspective on the selection of techniques for effective bitplane coding strategies. Analyses of coding performance, computational throughput, and scalability are provided. These features are commonly employed to appraise modern image coding systems. In addition to the extensive figures provided in the manuscript, supplementary figures and video sequences have been developed to illustrate the coding of a code block via five different scanning orders. These supplementary materials provide significant insight into the underlying processes of each scanning order. They can be found in the "Scanning orders" section at http://www.deic.uab.es/_francesc/. Conclusions drawn from the above analyses are summarized and nine key aspects are presented to guide the design of future coding systems. The formulation of these key aspects allows their inclusion in a wide variety of codecs designed to meet differing specific requirements. To demonstrate their usefulness, they are employed in the design of a coding engine with primary requirement to reduce the computational costs of the JPEG2000 standard without sacrificing any of its features. The proposed coding engine is tested on natural gray-scale and color images, hyper-spectral remote sensing images, and 3D medical images. The resulting compression performance is slightly better than that of JPEG2000 while achieving speedups from 1.2 to 1.6 depending on the image. To facilitate comparison with other techniques and codecs, source code is freely available at the webpage mentioned above.

Acknowledgement:

This paper is nominated by the MMTC Image, Video and Mesh Coding (IMVIMEC) Interest Group.

Weiyi Zhang is currently a Research Staff Member of the Network Evolution Research Department at AT&T Labs - Research, Middletown, NJ. Before join AT&T Labs Research, he was an Assistant Professor at the Computer Science Department, North Dakota State University, Fargo, North Dakota, from 2007 to 2010. His research interests include routing, scheduling, and cross-layer design in wireless networks, localization and coverage issues in wireless sensor networks, survivable design and quality-of-service provisioning of communication networks. He has published more than 70 refereed papers in his research areas, including papers in prestigious conferences and journals such as IEEE INFOCOM, ACM MobiHoc, ICDCS, IEEE/ACM Transactions on Networking, ACM Wireless Networks, IEEE Transactions on Vehicular Technology and IEEE Journal on Selected Areas in Communications. He received AT&T Labs Research Excellence Award in 2013, Best Paper Award in 2007 from IEEE Global Communications Conference (GLOBECOM'2007). He has been serving on the technical or executive committee of many internationally reputable conferences, such as IEEE INFOCOM. He was the Finance Chair of IEEE IWQoS'2009, and serves the Student Travel Grant Chair of IEEE INFOCOM'2011.

A Spatio-Temporal JND Model for Enhancing Lossy Video Compression

A short review for "Advanced H.264/AVC-Based Perceptual Video Coding: Architecture, Tools, and Assessment"

Edited by Irene Cheng

Matteo Naccari and Fernando Pereira (FIEEE),"Advanced H.264/AVC-Based Perceptual Video Coding: Architecture, Tools, and Assessment", IEEE Transactions on Circuits and Systems for Video Technology, Vol. 21, No. 6, June 2011.

Human perceptual factors incorporating just noticeable thresholds to eliminate visually redundant data have been considered in many research fields which involve signals and data processing in various dimensions. This paper aims at exploiting certain characteristics of the human visual system (HVS), i.e., the associated spatio-temporal masking mechanism, to enhance lossy video compression efficiency beyond the H.264/AVC standard. This work was delivered when the Joint Collaborative Team on Video Coding (JCTVC) [1] was established with the goal to develop a high efficiency video coding (HEVC) standard. However, the proposed spatio-temporal just noticeable distortion (ST_JND) model continues to have its significance since the proposed techniques, especially the rate-distortion (RD) performance assessment methodology, can be considered for the current HEVC standardization activities as pointed out by the authors in their Conclusion. Thus this review will highlight the human perceptual aspect of the ST_JND model.

Although perceptual criteria are important in evaluating the visual quality of decoded videos, integrating these criteria into quality assessment metrics still poses research challenges due to the masking effects [2] and complex non-linear sensitivity of the HVS towards the quantization distortion introduced in lossy video coding. It is thus worth exploiting these issues to reduce the coding bitrate. This paper proposes using just noticeable distortion (JND) estimation at the decoder to perceptually allocate the available rate so as to avoid redundancy. The advantage of using JND thresholds is to vary the quantization step so that coarser data can be considered in areas with higher JND threshold values and finer otherwise. Accordingly, the authors motivate the following discussions:

1) How the JND models are integrated in a block-based hybrid motion compensated predictive video coding architecture, and
2) How is the rate-distortion (RD) performance of perceptual video codecs assessed in order to avoid the use of the mean square error (MSE) metric, given its widely recognized poor correlation with the perceived quality [3], and to model the variability in the observers when rating video quality.

To address these questions, the paper first explains the relation between JND and spatio-temporal masking, and their impact on visual quality. JND in video signals depends on both spatial and temporal masking effects. Spatial masking arises from three aspects [2]: 1) the transformed frequency representation used; 2) the luminance variations; and 3) the presence of texture patterns. Temporal masking is associated with the motion activity present between two consecutive frames, with distortion less noticeable in areas with fast movements. The proposed ST-JND model accounts for these masking effects and works in the DCT domain over the luminance component of each video frame. A JND threshold is provided for each DCT coefficient and organized in a matrix with the same width and height of the video frame being coded. Interested readers can refer to the original manuscript for technical detail.

The three spatial masking mechanisms considered are 1) Frequency Band Masking, which accounts for the different sensitivity of the human eye to the noise introduced at various spatial frequencies; 2) Luminance Variations Masking, which accounts for the masking mechanism associated to the luminance variations in the various image areas; and 3) Image Pattern Masking, which accounts for the masking mechanism associated to some image patterns. For the temporal masking component, the ST-JND model uses the approach introduced by Wei and Ngan in [4], which outperforms other state-of-the-art approaches. To provide a good approximation of the original video samples, the authors use the predictors specified by the H.264/AVC standard to perform MB predictive coding, and establish three decoder-side estimation criteria: Average Block Luminance Estimation, Integer DCT Coefficients Estimation and ST JND Thresholds Estimation.

The editor would like to direct readers' attention to Section VI in the original manuscript (Proposed RD Performance Assessment Methodology). An objective relative assessment methodology is proposed, which is catered for evaluating the RD performance of perceptual video codec (PVCs). The evaluation is based on full reference to another codec using an objective quality metric while simultaneously addresses the perceptual characteristics in the observers when ranking the video quality. It is interesting to read the authors' elaboration of *variability*, which not only means that "different observers rate differently the same video but also means that different observers may equally rate the same video distorted in different ways." These effects often pose challenges to objective perceptual quality assessment metrics. The authors give an example, where two codecs (*A*, *B*) are considered to encode the same video, with codec *A* as the reference quality. Even though the subjective quality of codec *B* correlates well with the perceived quality, it may happen that the objective quality scores are significantly different leading to the conclusion that codec *B* is drastically worse than codec *A*.

To address this deficiency, the authors apply the works in [5] and [6], where the concept of metric resolving power (MRP) is used. The MRP corresponds to the maximum change in the adopted metric objective scores for which the perceived quality can be considered the same. This is a very important aspect in assessing quality of visual data with full or even partial reference. In this regard, if the difference between the objective scores for a video coded with codecs *A* and *B* (mentioned above) is lower than the MRP, then their qualities will be perceived as identical by a human observer. The use of the MRP to compensate the objective quality scores is the core of the proposed resolving power-compensated (RPC) RD performance assessment methodology.

The RPC methodology includes four phases. The first three define MRP computation and the last phase executes comparison. Observe that the MRP computation requires the availability of the subjective scores collected for relevant video test material [5]. The four phases are: 1) Subjective score statistics computation, 2) MOS estimate from objective scores, 3) MRP computation and 4) MRP objective quality compensation. At the same RPC objective quality, the proposed architecture is able to demonstrate an average bitrate reduction of 30% compared to the H.264/AVC high profile codec and outperforms the perceptual codec proposed in [7], which provides an average bitrate reduction of about 10%.

Acknowledgement:

This paper is nominated by the MMTC Image, Video and Mesh Coding (IMVIMEC) Interest Group.

References:

[1] ISO/IEC JTC1/SC29/WG11, "Joint Call for Proposals on Video Compression Technology," *document N11113*, Kyoto, Japan, January 2010.

[2] H. R. Wu and K. R. Rao, "Digital Video Image Quality and Perceptual Coding," *Boca Raton, FL: CRC Press*, Nov. 2005, p. 640.

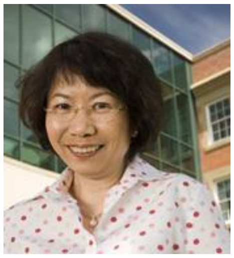
[3] Z. Wang and A. C. Bovik, "Mean square error: Love it or leave it? A new look at signal fidelity measures," *IEEE Signal Process. Mag.*, vol. 26, no. 1, pp. 98–117, Jan. 2009.

[4] Z. Wei and K. N. Ngan, "Spatio-temporal just noticeable distortion profile from grey scale image/video in DCT domain," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 19, no. 3, pp. 337–346, Mar. 2009.

[5] M. H. Brill, J. Lubin, P. Costa, S. Wolf, and J. Pearson, "Accuracy and cross-calibration of video quality metrics: New methods from ATIS/T1A1," *Signal Process.: Image Commun.*, vol. 19, no. 2, pp. 101– 107, Feb. 2004.

[6] M. H. Pinson and S. Wolf, "Techniques for evaluating objective video quality models using overlapping subjective data sets," Inst. Telecommun. Sci., Boulder, CO, NTIA Rep. TR-09-457, Nov. 2008.

[7] Z. Chen and C. Guillemot, "Perceptually-friendly H.264/AVC video coding based on foveated just-noticeable-distortion model," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 6, pp. 806–819, Jun. 2010.



Irene Cheng, SMIEEE is the Scientific Director of the Multimedia Research Centre, and an Adjunct Professor in the Faculty of Science, as well as the Faculty of Medicine & Dentistry, University of Alberta, Canada. She is also a Research Affiliate with the Glenrose Rehabilitation Hospital in Alberta, Canada. She is a Co-Chair of the IEEE SMC Society, Human Perception in Vision, Graphics and Multimedia Technical Committee; was the Chair of the IEEE Northern Canada Section, Engineering in Medicine and Biological Science (EMBS) Chapter (2009-2011), and the Chair of the IEEE Communication Society, Multimedia Technical Committee 3D Processing, Render and Communication (MMTC) Interest Group (2010-2012). She is now the Director of the Review-Letter Editorial Board of MMTC (2012-2014).

Over the last ten years, she has more than 110 international peer-reviewed publications including 2 books and 31 journals. Her research interests include multimedia communication techniques, Quality of Experience (QoE), Levels-of-detail, 3D Graphics Visualization and Perceptual Quality Evaluation. In particular, she introduced applying human perception – Just-Noticeable-Difference – following psychophysical methodology to generate multi-scale 3D models.

Post-filtering for Depth Map Coding

A short review for "Depth Coding Using a Boundary Reconstruction Filter for 3-D Video Systems"

Edited by Gene Cheung

Kwan-Jung Oh, Anthony Vetro, Yo-Sung Ho, "Saliency -cognizant Error Concealment in Loss-corrupted Streaming Video," IEEE Transactions on Circuits and Systems for Video Technology, vol.21, no.3, March, 2011.

Traditional 2D video enables *passive viewing*; the angle from which the video is captured is exactly the angle from which the video is played back and observed. In contrast, free viewpoint video [1] offers viewers *interactive viewing*, where the angle from which to observe a 3D scene can be freely chosen by an observer. It is quite clear that this interactive aspect—which enhances depth perception in the 3D scene via motion parallax [2]—marks a fundamental departure from traditional 2D video services and a new era in video-based visual experience. It is also clear that to enable such free viewpoint video services, advances in image capture, processing, compression and streaming are necessary.

In particular, for compact representation a new video format call *multi-view video plus depth* (MVD) is proposed in [3] to represent the 3D visual information for free viewpoint video. MVD is composed of texture maps (regular color images) and depth maps (per-pixel distance between the physical objects in the 3D scene and the capturing camera) of the 3D scene of interest as observed from multiple closely spaced viewpoints. Depth maps can be obtained directly from depth sensors [4], or estimated using stereo-matching algorithms using captured texture maps. Armed with texture and depth maps from multiple viewpoints, a novel image as observed from an intermediate virtual viewpoint can be synthesized via *depth-image-based rendering* (DIBR) [5]. Obviously, transmission of multiple texture and depth maps entails a high network cost, and thus efficient compression is necessary. While encoding of multiple texture maps has been studied in the context of *multiview video coding* (MVC) [6], compression of depth maps is relative new. The contribution by Oh et al. is an early work that made a fundamental contribution to this important topic.

Depth maps have unique signal characteristics, such as sharp edges and smooth surfaces away from edges that can be exploited for compression gain when designing new coding tools specifically for depth maps, such as edge-adaptive wavelets [7] or graph-based transform [8]. Instead of proposing new non-standard coding tools, however, Oh et al. proposed to use a boundary reconstruction filter as an in-loop filter during compression to obtain coding gain. In particular, the proposed filter has the effect of removing noise due to signal quantization and recovering true object boundaries from its neighboring pixels—sharp object boundaries are essential towards synthesizing good quality virtual viewpoint images via DIBR. By not disturbing the traditional motion prediction / residual transform coding paradigm, this filtering approach has the advantages of lightweight implementation and easy adaption into video coding standards.

The key idea of the reconstruction filter is the following. It is designed non-linearly by considering three terms: i) occurrence frequency; ii) similarity; and iii) closeness. A cost function composed of the three terms can be written as:

$$J_{recon}(k) = J_F(k) + J_S(k) + J_C(k) \qquad (1)$$

where k represents the depth pixel intensity value. The value that maximizes the above objective for a particular pixel is used to replace the original depth value.

The first term J_F stands for occurrence frequency for each intensity value. It essentially computes how often a particular depth value has occurred in a neighborhood of pixels. The second term J_S reflects how similar a pixel in intensity is to its neighbors. The third term J_C shows how close a given pixel is to its neighbors (in terms of Euclidean distance) of the same intensity value. Oh et al. showed that combing these three terms, on average correct depth values can be restored after lossy signal quantization.

Note that because no averaging of neighboring pixels is performed (unlike typical denoising approaches such as bilateral filter), sharpness of edges can be preserved with no blurring. Experiments show that using this filter as an in-loop filter in the context of H.264 / AVC can result in noticeable depth coding gain as well as higher quality DIBR-synthesized images.

While Oh et al. provided important early groundwork, optimization of depth map coding—taking into consideration how coding artifacts in compressed depth maps affect DIBR-synthesized images—continues to be an active research topic [9, 10]. It is our hope that researchers in all aspects of the 3D visual communications systems (e.g., error-resilient streaming, media interaction) leverage on these works when optimizing the overall system for inter-component performance gain.

Acknowledgement:

This paper is nominated by the MMTC Image, Video and Mesh Coding (MIVIMEC) Interest Group.

References:

[1] M. Tanimoto et al., "Free-Viewpoint TV," in *IEEE Signal Processing Magazine*, January 2011, vol. 28, no.1.

[2] C. Zhang et al., "Improving Depth Perception with Motion Parallax and its Application in Teleconferencing," in *IEEE International Workshop on Multimedia Signal Processing*, Rio de Jeneiro, Brazil, October 2009.

[3] P. Merkle, "Multi-view Video Plus Depth Representation and Coding," in *IEEE Conference on Image Processing*, San Antonio, TX, October 2007.

[4] S. B. Gokturk et al., "A Time-Of-Flight Depth Sensor – System Description, Issues and Solutions," in IEEE CVPR Workshop, Washington, DC, April 2004.

[5] D. Tian et al., "View Synthesis Techniques for 3D Video," in *Applications of Digital Image Processing XXXII, Proceedings of the SPIE*, vol. 7443 (2009), 2009.

[6] *Text of ISO/IEC 14496-10:2008/FDAM 1 Multiview Video Coding*, document w9978, ISO/IEC JTC1/SC29/WG11, October 2008.

[7] M. Maitre et al., "Wavelet-based Joint Estimation and Encoding of Depth-image-based Representations for Free-viewpoint Rendering," *IEEE Transactions on Image Processing*, vol.17, no.6, June, 2008.

[8] G. Shen et al., "Edge-adaptive Transform for Efficient Depth Map Coding," *Picture Coding Symposium*, Nagoya, Japan, December 2010.

[9] G. Cheung et al., "Transform Domain Sparsification of Depth Maps using Iterative Quadratic Programming," *IEEE Conference on Image Processing*, Brussels, Belgium, September 2011.

[10] G. Valenzise et al., "Motion Prediction of Depth Video for Depth-Image-based Rendering using Don't Care Regions," *Picture Coding Symposium*, Krakow, Poland, May 2012.

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 was a

senior researcher in Hewlett-Packard Laboratories Japan, Tokyo, from 2000 till 2009. He is now an associate professor in National Institute of Informatics in Tokyo, Japan. His research interests include 3D visual representation, single- / multiple-view video coding & streaming, and immersive communication. He has served as associate editor for IEEE Transactions on Multimedia from 2007 to 2011 and currently serves as associate editor for DSP Applications Column in IEEE Signal Processing Magazine and APSIPA journal on signal and information processing.

New composite source models and new rate distortion bounds for wideband speech

A short review for "Rate Distortion Performance Bounds for Wideband Speech"

Edited by Weiyi Zhang

Jerry D. Gibson and Ying-Yi Li,"Rate Distortion Performance Bounds for Wideband Speech", Proc. of IEEE Information Theory and Applications Workshop (ITA), 2012, San Diego, CA, pp. 186 - 191, 2012 2011.

Speech codecs based on linear prediction play a significant role in digital cellular, Voice over IP (VoIP), and Voice over Wireless LAN (VoWLAN) applications; however, while speech researchers have been extremely innovative in optimizing speech codecs, meaningful rate distortion bounds for speech sources have not been adequately studied. In particular, it would be of great utility if the host of existing rate distortion theory results could be applied to bounding the performance of practical codecs. Like all rate distortion problems, the two primary challenges are (1) finding good source models for speech, and (2) identifying a distortion measure that is perceptually meaningful, yet computationally tractable. There have been only a few prior research efforts in the last 25 years that have attempted to address various aspects of this problem.

The authors of this paper develop new rate distortion bounds for wideband speech coding based on composite source models for speech and perceptual PESQ-MOS/WPESQ distortion measures. It is shown that these new rate distortion bounds do in fact lower bound the performance of important standardized wideband speech codecs, including, G.718, G.722.1, and AMR-WB. Their approach is to calculate rate distortion bounds for mean squared error (MSE) distortion measures using the classic eigenvalue decomposition and reverse water-filling method for each of the subsource modes of the composite source model, and then use conditional rate distortion theory to calculate the overall rate distortion function for the composite source. Mapping functions are developed to map the rate distortion curves based on MSE to rate distortion curves subject to the perceptually meaningful distortion measures WPESQ. These final rate distortion curves are then compared to the performance of the best known standardized speech codecs based on the code-excited linear prediction paradigm.

There have only been a handful of efforts to calculate rate distortion bounds for speech. However, only few studies [1][2] have considered wideband speech. These works focused on the MSE distortion measure and obtain composite source models for wideband speech by segmentation of the speech into equal order autoregressive subsources. No comparisons to standardized speech codecs are provided since MSE is not a meaningful distortion measure for these codecs. To provide new perspective to the research area, this study first collects essential results from rate distortion theory. A composite source models is developed for speech and used in this paper's rate distortion analyses. Then, the mapping of the distortion measure from MSE to WPESQ is developed and discussed. This work also provided rate distortion bounds based on the WPESQ distortion measures and the composite speech model, where the new rate distortion curves are compared to common, standardized high-performance speech codecs and shown to lower bound the performance of all of the codecs.

In conclusion, this paper develops novel rate distortion bounds for wideband speech sources based on phonetically-motivated composite source models, conditional rate distortion theory, and perceptual wideband PESQ (WPESQ) distortion measures. The new rate distortion curves lower bound the performance of the best known standardized wideband speech codecs. This bound provides significant insight in the design of voice communication system in the future.

Acknowledgement:

This paper is nominated by the MMTC Acoustic and Speech Processing for Communication (ASP) Interest Group.

References:

[1] H. Kalveram and P. Meissner, "Rate Distortion Bounds for Speech Waveforms based on Itakura-Saito-Segmentation," Signal Processing IV: Theories and Applications, EURASIP, 1988.

[2] H. Kalveram and P. Meissner, "Itakura-saito clustering and rate distortion functions for a composite source model of speech," Signal Processing, vol. 18, no. 2, pp. 195 – 216, 1989.

Weiyi Zhang is currently a Research Staff Member of the Network Evolution Research Department at AT&T Labs - Research, Middletown, NJ. Before join AT&T Labs Research, he was an Assistant Professor at the Computer Science Department, North Dakota State University, Fargo, North Dakota, from 2007 to 2010. His research interests include routing, scheduling, and cross-layer design in wireless networks, localization and coverage issues in wireless sensor networks, survivable design and quality-of-service provisioning of communication networks. He has published more than 70 refereed papers in his research areas, including papers in prestigious conferences and journals such as IEEE INFOCOM, ACM MobiHoc, ICDCS, IEEE/ACM Transactions on Networking, ACM Wireless Networks, IEEE Transactions on Vehicular Technology and IEEE Journal on Selected Areas in Communications. He received AT&T Labs Research Excellence Award in 2013, Best Paper Award in 2007 from IEEE Global Communications Conference (GLOBECOM'2007). He has been serving on the technical or executive committee of many internationally reputable conferences, such as IEEE INFOCOM. He was the Finance Chair of IEEE IWQoS'2009, and serves the Student Travel Grant Chair of IEEE INFOCOM'2011.

Dynamic mesh compression with spatio-temporal scalability

A short review for "Efficient Fine-Granular Scalable Coding of 3D Mesh Sequences"

Edited by Guillaume Lavoué

Jae-Kyun. Ahn, Yeong Jun Koh and Chang-Su Kim, " Efficient Fine-Granular Scalable Coding of 3D Mesh Sequences ", IEEE Transactions on Multimedia , vol.15, no.3, pp.485,497, April 2013.

Three-dimensional (3D) meshes, which describe the shapes of 3D objects, are widely used in various applications, such as video games, film production, industrial design, and 3D imaging. Recently, the interest in 3D mesh sequences also has increased because of the rapid growth of 3D TV and 3D films. Various techniques on mesh sequence processing should be developed to address the growing demand. Especially, the compact compression of mesh sequences is essential to transmit and store the huge amount of 3D data efficiently.

In isomorphic mesh sequences, the topology (or connectivity) relationship among vertices is invariant over all frames. Hence, mesh sequence compression encode the topology information once for the first frame, and focus on the compression of the geometry information, i.e., vertex positions, in each frame. Most of the previous related works compress either an entire sequence or an entire frame as a whole to achieve high compression performance [1-3]. It is difficult to employ them in low-latency streaming applications, which demand real-time encoding and decoding. Vertex-based predictive coding is more suitable for low-latency applications, and it supports temporal scalability as well as spatial scalability [4, 5].

In this work, the authors propose an efficient fine-granular scalable (FGS) coding algorithm to compress 3D mesh sequences in low-latency streaming applications. To support both spatial and temporal scalability, they decompose an input mesh into layers. Then, they perform vertex-wise predictive coding using the decomposed layers. First, they propose a spatial layer decomposition (SLD) algorithm for the FGS mesh coding, which decimates vertices in a mesh frame iteratively to obtain multiple layers. The vertex decimation order affects not only the compression performance but also the visual qualities of intermediate resolution meshes. By exploiting the topological information, and also optionally the geometrical information in the first frame, the proposed SLD algorithm attempts to obtain high quality intermediate resolution meshes and improve the compression performance. Note that since they consider only isomorphic mesh sequences, they perform the spatial layer decomposition only once at the beginning of the compression, and maintain the same spatial layer structure for all frames.

Also, the authors construct temporal layers using the hierarchical prediction structure in the H.264 video coding standard [6]. Specifically, they combine several successive frames as a group of frames (GOF) and encode the frames in the GOF in a hierarchical order. In general, to form T temporal layers, a GOF should include 2^(T-1) successive frames. They encode the 0th frame as I-frame, every 2^(T-1)th frame as P-frame, and the other frames as B-frames [6]. Then, all frames except the 0th frame have their reference frames for the temporal prediction.

For each frame, from the base to the finest spatial layer, the authors predict vertex positions spatially and encode the spatial prediction residuals. Note that the same spatial prediction is performed at the reference frames as well. Since the spatial prediction residuals are similar between the current frame and the reference frames due to high temporal correlations, the spatial residuals at the current frame are further compensated by those at the reference frames. Hence, the authors propose a two-step vertex prediction algorithm: First, they spatially predict each vertex from its neighboring vertices. Second, they temporally predict the spatial prediction residuals at the current frame from those at the reference frames. The spatio-temporal residuals are then quantized and entropy-encoded. To prevent the propagation of quantization errors, the prediction and the reconstruction are performed from the base layer to the finest layer sequentially, and the reconstructed vertices at each layer are used for the prediction at finer layers.

The authors propose an entropy coding scheme, based on the bit-plane coding [7], to transmit bits in the decreasing order of importance and hence improve the rate-distortion performance. Moreover, they observe that, when a vertex has small prediction residuals in the reference frames, it tends to have a small residual also in the current frame. Based on the observation, they construct an efficient context model for the entropy coding.

Experimental results demonstrate that the proposed algorithm consistently outperforms the other algorithms with significant gains on six models. On average, the proposed algorithm reduces the bit rate by 30.1% and 20.4%, as compared with conventional methods in [4] and [5], respectively. In addition to coding efficiency, the proposed algorithm supports much finer-gradual scalability and enables users to control the spatial resolution more flexibly.

The authors' main contribution is to propose a new SLD method, in which the number of spatial layers is as large as the number of vertices. To support high quality rendering of intermediate resolution meshes, the SLD method exploits the topological relations between vertices and also the geometrical relations in the first frame optionally. Moreover, they proposed efficient spatial and temporal predictors for geometry data. In addition, they proposed an efficient context model for the arithmetic coding of prediction residuals.

References :

[1] Z. Karni and C. Gotsman, "Compression of soft-body animation sequences," Computers & Graphics, vol. 28, no. 1, pp. 25–34, 2004.

[2] K. Mamou, T. Zaharia, and F. Prêteux, "A skinning approach for dynamic 3D mesh compression," Computer Animation and Virtual Worlds, vol. 17, no. 3-4, pp. 337–346, 2006.

[3] L. Vássa and V. Skala, "Geometry-driven local neighbourhood based predictors for dynamic mesh compression," Comput. Graph. Forum, vol. 29, no. 6, pp. 1921–1933, 2010.

[4] N. Stefanoski and J. Ostermann, "SPC: Fast and efficient scalable predictive coding of animated meshes," Comput. Graph. Forum, vol. 29, no. 1, pp. 101–116, 2010.

[5] M. O. Bici and G. B. Akar, "Improved prediction methods for scalable predictive animated mesh compression," J. Vis. Commun. Image Rep., vol. 22, no. 7, pp. 577–589, 2011.

[6] H. Schwarz, D. Marpe, and T. Wiegand, "Analysis of hierarchical B pictures and MCTF," in Proc. IEEE ICME, Jul. 2006, pp. 1929–1932.

[7] J.-K. Ahn, D.-Y. Lee, M. Ahn, J. D. K. Kim, C. Kim, and C.-S. Kim, "Progressive compression of 3D triangular meshes using topology-based Karhunen-Loève transform," in Proc. IEEE ICIP, Sep. 2010, pp. 3417–3420.

Guillaume Lavoué received his Ph.D. degree in computer science from the University Claude Bernard, Lyon, France (2005). Since September 2006 he is associate professor at the French engineering university INSA of Lyon, in the LIRIS Laboratory.
He is author or co-author of over 60 publications in international journals and conferences. His research interests include 3D mesh analysis and retrieval, 3D data transmission and streaming (including compression and watermarking), Web 3D, Perception and Human factors for computer graphics.

A Pioneering Breakthrough of 3-D Object Analysis

A short review for "3D Object Retrieval and Recognition with Hypergraph analysis"

Guest Edited by Jun Liu

Yue Gao, Meng Wang, Dacheng Tao, Rongrong Ji and Qionghai Dai, "3D Object Retrieval and Recognition with Hypergraph Analysis," IEEE Transactions on Image Processing, vol.21, no.9, pp. 4290-4303, 2012.

With the development of computer hardware and network, 3-D technology has been widely applied in many applications, such as 3DTV, CAD and medical industry. The increasing 3-D objects aggravate the requirement of effective and efficient 3-D object management technologies. 3-D object retrieval and recognition are important topics and have high marketing values for both the enterprises and individual users. However, the performance is not satisfied for real applications.

Existing methods [1,2,3] can be divided into two paradigms: model-based and view-based. Model-based methods require explicit 3-D models, which severely limits the practical applications of model-based methods. With the rapidly increasing cameras in mobile devices, such as iPhone and iPad, it is possible to capture multiple views of real 3-D objects, which has attracted extensive research efforts on view-based 3-D object retrieval and recognition.

View-based methods benefit from the flexibility of multiple views on representing 3-D objects. In this type of methods, each 3-D object is described by a set of views. It is noted that the multiple views of 3-D object not only provide rich information but also bring in difficulties in comparison between 3-D objects. Most of existing works employ direct view-matching to measure the distance between objects, such as the Earth Mover's distance [4] and Hausdorff distance [5]. These methods have been deeply investigated in single image search, while they are not precisely designed to represent the higher order information in view-base 3-D object analysis.

Two main challenges in view-based 3-D object analysis are the representation of multiple views and the relevance exploration. In this work, authors propose an unified framework to handle both of the two issues successfully. This is the first attempt to bring all 3-D object views into one structure for analysis, which has high potential to lead the future view-based 3-D objet analysis research.

Different from existing view-based object analysis methods, each view is regarded as the element of 3-D objects directly, and the relationship among 3-D objects is formulated in a hypergraph structure. Under this framework, each object is denoted by one vertex, and the links among vertices are generated by using view clustering. Multiple hypergraphs are generated by using view clustering with different granularities, and they are further combined for semi-supervised learning. This fusion scheme can reduce the influence of different clustering results and make the method robust. Authors also propose a learning framework to automatically update the weights for multiple hypergraph, which is able to further improve the representation ability of the hypergraph structure on 3D objects.

The merit of this work can be summarized as follows.

1) This work presents an unified framework for 3-D object analysis. This method abandons the traditional bi-view matching scheme and proposes an innovative 3-D object formulation by using hypergraph to effectively exploit the underneath structure among 3-D objects. In this formulation, multiple views are denoted as "features" of 3-D objects. The relationship among 3-D objects is modeled in a hypergraph framework. The relationship among all 3-D objects is estimated in one round, which can explore the underneath structure and connection in the 3-D database. This formulation is groundbreaking in comparison with state-of-the-art methods.

2) This work proposes to jointly learn the weights of individual hypergraphs and the relevance matrix among 3-D objects in one semi-supervised learning (SSL) process. The SSL stage is able to leverage a large amount of unlabeled data and labeled data for 3-D object retrieval and recognition. A hypergraph weight learning procedure automatically updates the weights for hypergraphs to reduce the influence of clustering results with different granularities. This framework has great potentialities on applications in many other tasks.

Generally, this work is a pioneering breakthrough of 3-D object analysis, and the proposed hypergraph formulation for 3-D representation makes remarkable progress in view-based 3-D object retrieval and recognition leading to the new state-of-the-art. The relevance exploration scheme has great capabilities for wide applications in other tasks and significant influence on future research.

References:

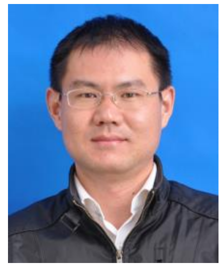
[1] R. Osada, T. Funkhouser, B. Chazelle, and D. Dobkin, “Shape distributions,” ACM Trans. Graphic, vol. 21, no. 4, pp. 807–832, 2002

[2] Y. Gao, M. Wang, Z. Zha, Q. Tian, Q. Dai, and N. Zhang, “Less is more: Efficient 3-D object retrieval with query view selection,” *IEEE Trans. Multimedia*, vol. 11, no. 5, pp. 1007–1018, Oct. 2011.

[3] Y. Gao, J. Tang, R. Hong, S. Yan, Q. Dai, N. Zhang, and T. Chua, “Camera constraint-free view-based 3-D object retrieval,” IEEE Trans. Image Process., vol. 21, no. 4, pp. 2269–2281, Apr. 2012.

[4] Y. Rubner, C. Tomasi, and L. J. Guibas, “The earth mover’s distance as a metric for image retrieval,” *Int. J. Comput. Vis.*, vol. 40, no. 2, pp. 99–121, 2000.

[5] M. J. Atallah, “A linear time algorithm for the Hausdorff distance between convex polygons,” *Inf. Process. Lett.*, vol. 17, no. 4, pp. 207–209, 1983.



Jun Liu received the B.E. and Ph.D degrees in school of remote sensing and information engineering from Wuhan University, China, in 2007 and 2012 respectively. Currently he is the assistant research fellow with the Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Sciences. His research interests include processing, retrieval and information extraction of remote sensing and natural images, such as image fusion and quality assessment, image enhancement, cloud and fog removal, image retrieval, the applications of high resolution remote sensing images and hyper-spectral data, and data mining, especially the association rule mining from image

.

A Multiple-Kinect System for Full 3D Constructions of Humans

A short review for "Real-time, full 3-D reconstruction of moving foreground objects from multiple consumer depth cameras"

Edited by Jun Zhou

Dimitrios S. Alexiadis, Dimitrios Zarpalas, and Petros Daras, "Real-Time, Full 3-D Reconstruction of Moving Foreground Objects From Multiple Consumer Depth Cameras", IEEE Transactions on Multimedia, Vol. 15, No. 2, pages 339-358, 2013.

From the day it was released by Microsoft in 2010, Kinect has attracted much attention in the consumer market. Due to its revolutionary low setting that integrates a RGB camera and a depth sensor, its applications has expanded quickly to applications other than games, such as interactive medical practice [1] and robotics [2]. This is boosted by the release of a non-commercial Kinect software development kit that allows researchers to access and effectively use the data captured by Kinect.

When applied to computer vision and multimedia research, one of the most important functions that Kinect facilitates is building 3D models. This can be done by fusion depth data captured from different poses of the sensor. For example, KinectFusion [3], the ISMAR'11 best paper award winner system, allows single global surface model be constructed progressively by moving a Kinect around the scene. The current sensor pose is estimated by discovery the relationship between the live depth frame and the global model using an iterative closest point (ICP) algorithm. In such setting, challenges lies in how to efficiently learn a full-view model so that the best alignment between multiple views can be recovered [4].

There are also limited systems that use multiple Kinects to generate 3D models. Tong et al presented a 3D human body models by fusion data from three Kinects [5]. These sensors scan different parts of the body to generate data to be registered under non-rigid deformation. In [6], six Kinects are used to perform data capturing. This method combines separate 3D meshes from Kinects at the rendering stage. However, as the author of this paper pointed out, all these systems are either not fast enough to be real-time, or not produce a single 3D mesh. Therefore, the goal of the work addressed in this paper is to develop the first system that uses multiple Kinects for real-time 3D reconstruction.

The reported system uses four Kinect sensors connected to the same computer with an NVIDIA graphics card. The sensors are horizontally positioned at the vertices of a square, pointing towards the center of the square. Therefore, this system can capture data from four different directions that covers the full 360 degree of view of a region of 2 square meters.

The proposed 3D reconstruction method consists of two parts. The first part is offline calibration, including single Kinect calibration and external calibration. The single Kinect calibration uses a checkerboard to estimate the internal RGB cameras' parameters, while keeping the default setting of the depth camera. The external calibration step generates correspondence between the images from different RGB cameras.

The second part of the method is dedicated to an online real-time 3D construction algorithm. This algorithm commences from an optional registration step to refine the alignment of point cloud from different Kinects. Similar to [3], an ICP framework is adopted, but with an extension to consider the visibility of points in each view into the modelling. At the same time, one mesh is created from each depth map via an initial triangulation step. The generated meshes from two adjacent sensors normally contain large amount of overlapping regions, which requires iterative removal efforts that are very time consuming. To address this problem, the authors have introduced a fast redundant boundary removal algorithm and a coarse fine processing strategy. Finally, a clipping process is employed to smooth the mesh regions at their adjacency boundaries. This step allows small holes be removed and a final seamless 3D mesh.

The authors have done several experiments to show the effectiveness and efficiency of the reported system. The effectiveness is evaluated by the visual quality of the reconstruction results. This was assessed by 17 human raters, who gave higher rating to the proposed system than a traditional separate meshes setting. Furthermore, this method can reconstruct close to 10 frames per second, which is much faster than other full 3D models. It makes this suitable for some real-time applications. The real-time effect of this work comes from both algorithms, i.e., coarse to fine mesh redundancy removal, and the employment of high performance computing, such as CUDA and multi-threading

In summary, the contribution of this paper comes from the implementation of a real-time full 3D reconstruction system. The details on hardware configuration and reconstruction methods have been clearly described in the paper. It shall be easily reproduced for various applications that require tracking of moving objects.

References:

[1] H. Mentis, K. O'Hara, A. Sellen, and Rikin Trivedi, “Interaction Proxemics and Image Use in Neurosurgery”, Proceedings of the ACM Conference on Computer-Human Interaction, pages 927-936, 2012

[2] J. Giles, “Inside the race to hack the Kinect”, New Scientist, Vol. 208, No. 2789, pages 22-23, 2010.

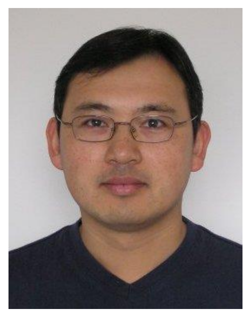
[3] R. Newcombe, S. Izadi, O. Hilliges, D.Molyneaux, D. Kim, A. Davison, P.Kohli, J. Shotton, S. Hodges, and A. Fitzgibbon, “KinectFusion: Real-Time Dense Surface Mapping and Tracking”, in Proceedings of the IEEE International Symposium on Mixed and Augmented Reality, pages 127-136, 2011.

[4] J. Stuckler and S. Behnke, “Model learning and real-time tracking using multi-resolution surfel maps”, In Proc. of the AAAI Conference on Artificial Intelligence, 2012.

[5] J. Tong, J. Zhou, L. Liu, Z. Pan, and H. Yan, “Scanning 3D full human bodies using Kinects”, IEEE Transactions on Visualization and Computer Graphics, Vol. 18, No. 4, pages 643-650, 2012.

[6] A. Maimone and H. Fuchs, “Encumbrance-free telepresence system with real-time 3D capture and display using commodity depth cameras,” in Proceedings of the IEEE International Symposium on Mixed and Augmented Reality, pages 137-146, 2011.

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



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

Paper Nomination Policy

Following the direction of MMTC, the R-Letter platform aims at providing research exchange, which includes examining systems, applications, services and techniques where multiple media are used to deliver results. Multimedia include, but are not restricted to, voice, video, image, music, data and executable code. The scope covers not only the underlying networking systems, but also visual, gesture, signal and other aspects of communication.

Any HIGH QUALITY paper published in Communications Society journals/magazine, MMTC sponsored conferences, IEEE proceedings or other distinguished journals/conferences, within the last two years is eligible for nomination.

Nomination Procedure

Paper nominations have to be emailed to R-Letter Editorial Board Directors:

Irene Cheng (locheng@ualberta.ca),
Weiyi Zhang (maxzhang@research.att.com), and
Christian Timmerer (christian.timmerer@itec.aauat)

The nomination should include the complete reference of the paper, author information, a brief supporting statement (maximum one page) highlighting the contribution, the nominator information, and an electronic copy of the paper when possible.

Review Process

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

R-Letter Best Paper Award

Accepted papers in the R-Letter are eligible for the Best Paper Award competition if they meet the election criteria (set by the MMTC Award Board).

For more details, please refer to http://committees.comsoc.org/mmc/rletters.asp

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