Chrome & YouTube University Research Program
--- 2017 Conference Agenda ---

Location: MTV-GWC2-1-Grand Teton Tech Talk

Live Stream <TODO>

Contact: agrange@google.com

Color Key:

Audio

Video

Geometry

Tuesday, 29th August 2017

Time	Presenter	Institution	Title	Abstract
08:30 - 09:00	**Coffee on Arrival**			
09:00 - 10:00	Prof. Hong-Goo Kang	Yonsei University, South Korea	Generalized Off-line Enhancement Algorithm for Targeted Speech Signals	In this project, we would like to enhance the quality of target speech signals distorted by background noise or interference signals. Under the assumption that the characteristics of the target signals are partially given, we propose a template matching based method where the templates are trained through deep learning based approaches. Since the main application of the developed system would be to enhance pre-recorded speech content publicly available on the Web, there are no strict constraints on causality, computational complexity, and memory usage. We propose a modified recurrent neural network framework that consists of multiple deep learning structures whose weights are adaptively trained to environmental variations. Despite the fact that the proposed structure requires high complexity and a large database for training, it is a reasonable approach for off-line enhancement applications.

10:00 - 11:00	Dr Andrew Hines	Dublin Institute of Technology, Ireland	Streaming VR for Immersion: Quality aspects of Compressed Spatial Audio	Delivering a 360° soundscape that match full sphere visuals is an essential aspect of immersive VR. Multisensory stimulation is important for sustained immersion, with sound being a critical factor. Some VR experts speculate that in VR production, audio is the number one biggest component of believable immersion [1] and it has been shown that even including white noise above no audio improves immersiveness [2]. A user is much more likely to enjoy a VR presentation when it has high quality audio and satisfactory quality video rather than low quality audio and high quality video. Furthermore, VR immersion relies heavily on how the audio is propagated within the scene. Audio cues must occur in the right direction and plane, with the right intensity and need to be rendered in real-time to match your head movements. Recording audio sources to take their position relative to the listener into account, known as ambisonics, was developed in the 1970s [3]. Ambisonics offers a possibility to represent three-dimensional sound. In contrast to existing channel-based methods, ambisonics representation offers the advantage of being independent of a specific loudspeaker set-up. It may also be rendered to set-ups consisting of only few loudspeakers or even to a headphones using binaural rendering. Ambisonics is a full-sphere surround technique that takes into account the azimuth and elevation of sound sources, pinpointing them above and below as well as around the listener[4]. The signal for a given sound source can be represented as a sound field using a spherical decomposition with the B-format standard and scaled to any desired spatial resolution [5]. For example, First Order Ambisonics (FOA) audio is encoded into 4 channels: an omnidirectional gain and 3 dimensional components: forward/backwards, left/right, and distance [6]. Moving to 16 channels (i.e. 3rd order) or Higher Order Ambisonics (HOA) significantly improves the Quality of Experience (QoE). The downside to B-format is the large amount of data and processing power required by HOA to transform a collection of multichannel sound sources into a rendered soundscape. Streaming ambisonic data over the networks requires efficient encoding techniques that would compress the raw audio content in real-time and without compromising QoE. To develop such a codec its performance must be evaluated using formalised quality judgment experiments [7]. In the absence of objective assessment methods, this is done usually with a panel of experienced listeners who evaluate listening audio quality. [8] This work investigates the impact of OPUS 1.2 codec [9] on the quality of 3rd order ambisonic audio encoded at 256kbps. OPUS codec uses lossy compression scheme and reduces the amount of data to be sent by discarding some information regarded as important for spatial audio (i.e. high frequencies and phase). In order to evaluate the impact of OPUS compression on spatial audio quality the ITU-standardized methodology for stereo audio, i.e. ITU-R BS. 1534-3 [10] has been adapted. During listening tests, a panel of experienced listeners rated a selection of 6 audio samples consisting of recorded audio clips from the EBU music database [11] and synthetic audio clips. Specific recorded audio clips have been chosen which are particularly difficult for compression algorithms to deal with [12]. Also synthetic audio clips contained audio signals with a wide range of time-frequency characteristics. In order to evaluate localisation accuracy of compressed audio a number of fixed and dynamic (moving vertically and horizontally) localisations have been chosen. Psychoacoustic research has shown very good localisation precision (0.5° - 1°) in front of a listener and poor (>10°) to the rear of the listener [13]. To account for this, the majority of localisations have been chosen in front of a listener. In addition, the quality and localisation accuracy of hidden reference audio and hidden anchor audio (encoded using FOA) has been evaluated as recommended by the ITU. The collected quality metrics will allow exploration of the relationship between audio quality, localisation accuracy and overall spatial quality for uncompressed audio and OPUS-compressed audio samples. This work will be used to validate and optimise work currently underway to develop a full reference objective spatial audio quality metrics adapted from ViSQOLAudio [14,15]. REFERENCES: [1] Kuhn, Maverick. “Spatial Audio and Immersion”, Brown University, March 2017 [2] Slater, Mel, and Sylvia Wilbur. "A framework for immersive virtual environments (FIVE): Speculations on the role of presence in virtual environments." Presence: Teleoperators and virtual environments 6.6 (1997)

11:00 - 11:30	**Coffee Break**			
11:30 - 12:30	Dr Hamed Sadeghi	University of Wellington, New Zealand	Multi-channel audio compression using deep learning techniques	We report some results on designing new speech codec architectures and tailoring the state-of-the-art image compression and speech enhancement structures to fit our signle-channel audio compression application. So far, the best results have been achieved by using deep 1D convolutional autoencoders, partly inspired by SEGAN (Speech Enhancement Generative Adversarial network) proposed by Pascual et. al. in 2017 for speech enhancement. For example, using deep convolutional layers, competitive quality has been achieved at a bitrate of 4kbs for 16-bit speech signals sampled at 16 KHz. Moreover, we will report the results of previously implemented fully connected and LSTM-based structures inspired by Toderici et. al.'s work proposed for image compression (2015). These are residual multi-step autoencoders (encoder-decoder) comprising fully connected and/or LSTM layers. The results were inferior to that of the deep convolutional ones in terms of quality-compression tradeoff. Finally, we will talk about ideas for generalizing to multichannel signals using 2D convolutions. Depending on the future results, we might also comment on how deep convolutional discriminators might improve the output quality of the mentioned autoencoders in an adversarial training setup. Moreover, we may be able to comment on and report the results of an encoder-decoder structure designed based on the WaveNet structure.
12:30 - 13:30	**Lunch**			
13:30 - 14:30	Prof. Ken Rose	University of California, Santa Barbara	Frequency Domain Singular Value Decomposition for Efficient Higher Order Ambisonics Data Coding	This talk summarizes research work at UCSB over the past year to develop novel compression techniques for encoding higher order ambisonics (HOA) data, to significantly improve both the compression gains and the perceptual quality of reconstructed HOA data. Recent standardization for HOA compression adopted a framework wherein HOA data are decomposed into principal components that are then individually encoded by a standard (MPEG-AAC) audio codec that attempts to optimize a psychoacoustic criterion. A noted shortcoming of this approach is the frequent mismatch in principal components across blocks, and the resulting unnatural transitions in the data when fed to the audio codec introduces significant inefficiencies and degrades the perceptual quality. Instead, we propose a framework where singular value decomposition (SVD) is performed after transformation to the frequency domain via the modified discrete cosine transform (MDCT). This framework not only ensures smooth transition across blocks, but also enables frequency dependent SVD for better energy compaction, instead of defaulting to a "compromise decomposition" for the entire spectrum. Moreover, we introduce a novel perceptual noise substitution technique to compensate for suppressed ambient energy in discarded higher order ambisonics channels, which significantly enhances the perceptual quality of the reconstructed HOA signal.
14:30 - 15:30	Prof. Fengqing Maggie Zhu	Purdue University	Investigation of the VP10/AOMEDIA Video Coding Through The Use Of Texture Analysis And Synthesis	Current video coding standards utilize hybrid coding techniques consisting of 2D transforms and motion compensation techniques to remove spatial and temporal redundancy. Our approach is different in that we will only encode, using VP10/AOMEDIA , areas of a video frame that are "perceptually significant." The "perceptually insignificant" regions will not be encoded. By "perceptually insignificant" pixels we mean regions in the frame that an observer will not notice any difference without observing the original video sequence. The encoder fits a model to perceptually insignificant pixels in the frame and transmits the model parameters to the decoder as side information. The encoder uses the model to reconstruct the pixels. This is referred to as "analysis/synthesis" coding approach. Our goal is to design and develop a novel VP10/AOMedia video coding codec tool through the use of deep learning based texture analysis and synthesis which would potentially achieve a large gain in video coding performance.
15:30 - 16:00	**Coffee Break**			

16:00 - 17:00	Prof. Jie Liang	Simon Fraser University, Canada	Overview of Deep Learning-based Image and Video Compression	In this talk, we first review some recent works on deep learning-based image compression, using technologies such as autoencoder, recurrent neural network (RNN), and generative adversarial network (GAN). These preliminary approaches have already achieved better performance than the well-established image coding standards JPEG and JPEG 2000. We then summarize the applications of deep learning in video compression, including mode decision and compression artifact removal. These results demonstrate the great potentials of deep learning in image and video compression.
18:00 - 21:15	**Dinner - Campo di Bocce, Los Gatos (Bus departs 5:00pm)**			

Wednesday, 30th August 2017

Time	Presenter	Institution	Title	Abstract
08:30 - 09:00	**Coffee on Arrival**			
09:00 - 09:30	Prof. Ofer Hadar		Novel Modes and Adaptive Block Scanning Order for Intra Prediction in AV1	The demand for streaming video content is on the rise and growing exponentially. Networks bandwidth is very costly and therefore there is a constant effort to improve video compression rates and enable the sending of reduced data volumes while retaining quality of experience (QoE). One basic feature that utilizes the spatial correlation of pixels for video compression is Intra-Prediction, which determines the codec's compression efficiency. Intra prediction enables significant reduction of the Intra-Frame (I frame) size and, therefore, contributes to efficient exploitation of bandwidth. In this presentation, we propose new Intra-Prediction algorithms that improve the AV1 prediction model and provide better compression ratios. Two (2) types of methods are considered: (1) New scanning order method that maximizes spatial correlation in order to reduce prediction error; and (2) New Intra-Prediction modes implementation in AVI. Modern video coding standards, including AVI codec, utilize fixed scan orders in processing blocks during intra coding. The fixed scan orders typically result in residual blocks with high prediction error mainly in blocks with edges. This means that the fixed scan orders cannot fully exploit the content-adaptive spatial correlations between adjacent blocks, thus the bitrate after compression tends to be large. To reduce the bitrate induced by inaccurate intra prediction, the proposed approach adaptively chooses the scanning order of blocks according to criteria of firstly predicting blocks with maximum number of surrounding, already Inter-Predicted blocks. Using the modified scanning method and the new modes has reduced the MSE by up to five (5) times when compared to conventional TM mode / Raster scan and up to two (2) times when compared to conventional CALIC mode / Raster scan, depending on the image characteristics (which determines the percentage of blocks predicted with Inter-Prediction, which in turn impacts the efficiency of the new scanning method). For the same cases, the PSNR was shown to improve by up to 7.4dB and up to 4 dB, respectively. In the end of the presentation we also present some future directions, including incorporating Deep Learning (DL) algorithm to optimize prediction coefficients, and optimization algorithm to select K Intra Prediction modes out of N modes according to Rate Distortion (RD) optimization.

09:30 - 10:30	Prof. Ken Rose	University of California, Santa Barbara	Optimal Adaptation of Motion-Compensated Prediction and Decorrelating Transforms in Video Coding	This talk summarizes UCSB's research work in the last year on developing new prediction and transform schemes that better exploit spatial and temporal redundancies via data statistics-dependent optimal design, and which achieve significant compression performance gains over conventional approaches. Specifically, we will focus on the following facets: 1) Design of K-mode interpolated motion compensation for the variable-block-size setting, via an interpolated prediction tree structure, wherein a pixel's prediction is generated as a linear combination of predictions based on the four nearest neighboring motion vectors. 2) Initial investigation into optimal transform design for the inter mode residual, which accounts for and circumvents the underlying instability due to quantization error propagation in a closed-loop video coding system, and which shows promise in preliminary experiments, outperforming the default 16 transforms currently in use.
10:30 - 11:00	**Coffee Break**			
11:00 - 12:00	Prof. Antonio Ortega	University of Southern California	Graph-based methods for video compression	In this talk we will discuss recent progress in using graph-based techniques for video coding, with a focus on three main areas: i) learning transforms from data, ii) fast transforms exploiting symmetries, and iii) enhanced motion based predictors.
12:00 - 13:00	**Lunch**			
13:00 - 14:00	Dr Francesca DeSimone, Prof. Pascal Frossard	École Polytechnique Fédérale de Lausanne	Motion estimation on omnidirectional video sequences	Classical motion estimation algorithms are built on the assumption that the video is captured by a perspective camera and that any motion of an object in space can be modelled by block translations in the imaging plane, i.e., the video frame. This model is not accurate any more for panoramic videos, resulting from the map projection of an omnidirectional, i.e., spherical, video to a plane. In this talk we present an extension of block-based motion estimation for omnidirectional videos, based on a camera and translational object motion model that accounts for the spherical geometry of the imaging system and the impact of the map projection. Experimental results demonstrate that significant gains can be achieved with respect to the classical exhaustive block matching algorithm (EBMA) in terms of accuracy of motion prediction.
14:00 - 14:30	Dr Michael Hemmer	Google	Introduction and Overview of Project DRACO	Draco is a library for compressing and decompressing 3D geometric meshes and point clouds, intended to improve the storage and transmission of 3D graphics, and designed and built for compression efficiency and speed. The code supports compressing points, connectivity information, texture coordinates, color information, normals, and any other generic attributes associated with geometry. Applications using 3D graphics can be significantly smaller without compromising visual fidelity. For users, this means apps can now be downloaded faster, 3D graphics in the browser can load quicker, and VR and AR scenes can be transmitted with a fraction of the bandwidth and rendered quickly.
14:30 - 15:00	**Coffee Break**			
15:00 - 16:00	Dr Pierre Alliez	INRIA	Advanced Progressive Geometry Compression	We will present a novel method for the progressive compression of surface triangle meshes with color textures. We focus on genericness, fine grain and flexible perceptual metrics.
16:00 - 17:00	Dr Michael Hemmer	(on behalf of) University of Mainz	Texture atlas generation for progressive mesh compression	<TBD>
17:00	**Close**			