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

# MMTC Communications – Review

IEEE COMMUNICATIONS SOCIETY

**Vol. 10, No. 3, June 2019**

## TABLE OF CONTENTS

# Message from the Review Board Directors

Welcome to the June 2019 issue of the IEEE ComSoc MMTC Communications – Review.

This issue comprises five reviews that cover multiple facets of multimedia communication research including resource allocation in heterogeneous networks, fine-grained venue discovery from multimedia data, and mobile edge cache placement for adaptive video streaming. Four reviews are briefly introduced below.

The first paper is published in the IEEE Transactions on Multimedia and edited by Dr. Ye Liu. It designs a MultiPath TCP (MPTCP) solution to improve the performance for video transfer specifically in the context of heterogeneous wireless networks and through the understanding of the relationship between the energy consumption and the video quality.

The second paper is published in the IEEE IEEE Transactions on Circuits and Systems for Video Technology and edited by Dr. Jun Zhou. It introduces an interactive system for contour extraction from images. The key novel idea in this paper is to mimic the human sketching process.

The third paper, published in the IEEE Transactions on Multimedia and edited by Dr. Carl James Debono, the authors propose a joint CNN and LSTM architecture that learns to identify multi-view shape descriptors that are used for 3D shape recognition and retrievals.

The forth paper is published in the IEEE Journal of Biomedical and Health Informatics and edited by Dr. Bruno Macchiavello. This work is investigating methods for the better detection of obstructive sleep apnea (OSA). The authors of this work evaluated non-linear entropy-based features, developed a feature selection algorithm, and applied different classification methods in order to provide a minute-by-minute OSA detection.

The fifth paper was published in the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2019 and edited by Dr. Frank Hartung. In this publication the authors discuss the issue of Ambisonics, i.e., full-sphere surround sound, aka 3D audio or 360 audio, for VR. Specifically, the challenge considered is the generation of 360° audio from captured mono or stereo audio plus 360° video.

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

IEEE ComSoc MMTC Communications – Review Directors

Qing Yang
University of North Texas, USA
Email: qing.yang@unt.edu

Roger Zimmermann
National University of Singapore, Singapore
Email: rogerz@comp.nus.edu.sg

Wei Wang
San Diego State University, USA
Email: wwang@mail.sdsu.edu

Zhou Su
Shanghai University, China
Email: zhousu@ieee.org

# Versatile Multipath Video Transfer over Heterogeneous Wireless Networks

*A short review for "Energy-Efficient Multipath TCP for Quality-Guaranteed Video over Heterogeneous Wireless Networks"*

Edited by Ye Liu

*J. Wu. R. Tan and M. Wang, "Energy-Efficient Multipath TCP for Quality-Guaranteed Video over Heterogeneous Wireless Networks," IEEE Transactions on Multimedia, vol. 21, no. 6, June 2019.*

The development of wireless infrastructures, social networks and multimedia applications drive the rapid growth of mobile video traffic. According to the Cisco Visual Networking Index (VNI) [1], video content among the world's mobile data traffic will increase nine times between 2017 and 2022. At that time, around 80 % of mobile data traffic will be video. Unfortunately, the available bandwidth in single wireless network at present becomes a bottleneck to the ever-increasing throughput demand of high definition video stream.

MultiPath Transmission Control Protocol (MPTCP) over heterogeneous wireless network is a promising solution to increase network throughput without significantly changing the existing wireless communication infrastructures, since many kinds of mobile access stations, such as LTE (Long-Term Evolution), HSDPA (High Speed Packet Access) and Wi-Fi, have been widely deployed in the our cities. In addition, mobile terminals in the market have already equipped with multiple different radio interfaces (e.g., 4G, Wi-Fi, Bluetooth). With MultiPath TCP, video server is able to concurrently transmit multimedia traffic over heterogeneous wireless networks to improve throughput. The MultiPath TCP mechanism has been recommended by the Internet Engineering Task Force (IETF) as transport protocol in the communication stack for multihomed terminals [2].

In recent years, many research work have been conducted to improve the performance of MultiPath TCP mechanism in terms of throughput-energy tradeoff [3], robust packet transmission [4], path heterogeneity toleration [5] and so on. However, the relationship between energy consumption and video quality has not been studied in MPTCP-based heterogeneous wireless networks.

In order to fill above knowledge gap, this paper firstly established theoretical models on effective loss rate, video distortion and energy consumption for concurrent multipath video traffic transfer over heterogeneous wireless networks. Then, the relationship between energy consumption and video quality was formulated and analyzed in detail. Based on this observation, a delay-energy-quality-aware MPTCP scheme was proposed to ensure the high quality video stream is delivered successfully before transmission deadline while minimizing the energy consumption of mobile devices.

The network model studied in this work is that a heterogeneous wireless network integrating many different kinds of access networks between video servers and multihomed devices. Each end-to-end link is independent of another. Three important physical properties of each wireless link are available bandwidth, round trip time and packet loss rate. To understand the energy-quality tradeoff, the Gilbert loss model is used to analyze packet loss pattern, therefore stationary continuous-time Markov chain is able to illustrate the link behavior. Secondly, the adopted video distortion model [6] indicates the quality of real-time video streaming is significantly affected by encoding bitrate, video sequence content and effective loss rate. Thirdly, the ramp, transfer and tail energy are considered for energy consumption of mobile devices[7]. Finally, an overall energy-quality tradeoff model is established based on the above three analytical frameworks. The important findings are: 1) The video quality is proportional to device energy consumption. The higher the video quality received, the more energy consumes; 2) Compared with Wi-Fi, the better video quality can be achieved by delivering more video data over cellular interface at the cost of higher device energy consumption.

The proposed MPTCP system consists four core components: subflow allocator, retransmission controller, parameter control unit and information feedback. In subflow allocator module, the video traffic rate is dynamically reduced based on video quality requirement and frame priority. Different from previous work to drop higher-priority video frame for transmission rate reduction, this work chooses to drop the lower-priority video frames to avoid the decoding failure problem. Utility maximization theory and piecewise linear approximation are introduced to allocate the segments transmitted over different wireless links for guaranteeing the received video quality and lowest energy consumption. A novel energy-delay aware packet retransmission control algorithm is implemented in the retransmission controller to prevent unnecessary retransmissions occurred in conventional retransmission controller.

Extensive semi-physical emulations based on the EXata platform demonstrate the improved performance of the proposed delay-energy-quality-aware multipath TCP solution. Compared with other existing schemes, the proposed solution can achieve the highest quality peak signal-to-noise ratio (PSNR) with minimal energy consumption.

In summary, the proposed versatile multipath TCP transfer scheme is demonstrated to achieve the optimal benefit balance between energy efficiency and streaming quality for video transmission over heterogeneous wireless networks.

**References:**

[1] Cisco, "Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2017-2022," February, 2019.

[2] A. Ford, C. Raiciu, M. Handley, and O. Bonaventure, "TCP extensions for multipath operation with multiple addresses," *No. RFC 6824*, 2013.

[3] Q. Peng, M. Chen, A. Walid, and S. H. Low, "Energy-efficient multipath TCP for mobile devices," in *Proc. ACM MobiHoc*, pp. 257–266, 2014.

[4] Y. Cui, L. Wang, X. Wang, H. Wang and Y. Wang, "FMTCP: A Fountain Code-Based Multipath Transmission Control Protocol," in *IEEE/ACM Transactions on Networking*, vol. 23, no. 2, pp. 465-478, April 2015.

[5] M. Li, A. Lukyanenko, S. Tarkoma, Y. Cui, and A. Ylä-Jääski, "Tolerating path heterogeneity in multipath TCP with bounded receive buffers," in *Computer Networks, 64*, pp.1-14, 2014.

[6] K. Stuhlmuller, N. Farber, M. Link and B. Girod, "Analysis of video transmission over lossy channels," in *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 6, pp. 1012-1032, June 2000.

[7] E. Harjula, O. Kassinen and M. Ylianttila, "Energ y consumption model for mobile devices in 3G an d WLAN networks," *2012 IEEE Consumer Comm unications and Networking Conference (CCNC)*, L as Vegas, NV, pp. 532-537, 2012.

**Ye Liu**, Ph.D, is a researcher with NAU-Lincoln Joint Research Center of Intelligent Engineering, Nanjing Agricultural University. He received M.S. and Ph.D. degrees from Southeast University, Nanjing, China in 2013 and 2018, respectively. He was a visiting scholar at Montana State University, Bozeman, USA from October 2014 to October 2015. He was a visiting PhD student from February 2017 to January 2018 in the Networked Embedded Systems Group at RISE SICS (Swedish Institute of Computer Science). His research interests include wireless sensor networks, energy harvesting systems and mobile crowdsensing. He has published papers in prestigious journals such as IEEE Communications Magazine, IEEE Internet of Things, and ACM Transactions on Embedded Computing System. He was awarded the 1st place of EWSN Dependability Competition in 2019.

# Interactive Contour Generation – Learning from Human Sketching Skills

*A short review for "Interactive Contour Extraction via Sketch-Alike Dense-Validation Optimization"*

Edited by Jun Zhou

*Yongwei Nie, Xu Cao, Ping Li, Qing Zhang, Zhensong Zhang, Guiqing Li, and Hanqiu Sun, " Interactive Contour Extraction via Sketch-Alike Dense-Validation Optimization," IEEE Transactions on Circuits and Systems for Video Technology, Early Access Article, 2019.*

Image contour detection is one of the fundamental tasks in image segmentation, sketching and object detection. Most developed methods aim at automatic edge or contour detection [1], which, however, face the difficulty of finding the best trade-off between keeping main structural information and excluding trivial details of objects or background in an image. This problem can hardly be solved even by the powerful deep learning techniques [2] since the ground truth information may be subjective and not accurate. Therefore, it is natural to explore semi-automatic solutions to address this problem.

The paper written by Nie *et al.* introduces an interactive system for contour extraction. The key novel idea is to mimic the human sketching process by drawing short and dense strokes at the beginning, and then integrating them into the final results. This is an interesting idea which turns out to be very effective since human can provide guidance in ground truth data generation. It not only improves the quality of contour detection, but also minimizes the human efforts, making it potentially useful for many multimedia applications such as movie production.

The user interface of the contour generation system allows users to mark keypoints along the contours of objects. This is the only human input required by the system. The keypoints are then used to generate the initial suboptimal curves around the objects by fitting a Catmull-Rom spline. Given a carefully designed curve-centered coordinate system, whose origin is the beginning of the initial curve, and two axes are defined along and perpendicular to the initial curve, the initial curve can be represented by several local strokes within a certain distance to the target contour. The extraction of the local strokes shall have two properties: "evident" which means the stroke should have larger gradients, and "smooth" which requires the stroke to be rigid enough to follow the characteristic of drawer's strokes. This can be achieved by maximizing an energy function that combines both properties. The optimization can be accelerated by dynamic programming approximation.

Since a single local stroke cannot well fit the target contour, the authors propose to extract densely overlapped local strokes. A multi-scale approach is developed to generate strokes of different lengths that overlap sequentially. These local strokes turn out to be consistent in locations with clear gradients, but do not match well when clear gradient does not existing, e.g. due to similar foreground and background color or highly textured image region. Then a weighted principal component analysis method is adopted to compact the strokes into the global optimal contour. A GPU-based parallel computing is used to speed-up this expensive multi-scale local strong extraction step.

The proposed method is compared with the Megnetic Lasso Tool in Photoshop which follows a similar type of user input. Photoshop seems to require a high number of user input to generate an accurate result, but the proposed method is more robust to few user interactions. The authors also show that the developed system can be used with GrabCut [3] to take advantage of and refine its segmentation results. When compared with several automatic approaches including deep learning based ones [4], the interactive system shows better continuity and crisp edges in contour generation. Although the proposed method has shown superior performance, its drawback is that quite a few parameters need to be tuned in order to achieve sound results. It may limit the adoption of this approach for more general tasks. This problem however, may be solved by adding the parameter setting function as part of the user interface.

In summary, this paper follows an intuitive and brilliant path for multimedia research, i.e. allowing user involvement and learning from human experiences. This is a viable way to the development of robust and useful multimedia systems, especially when performing highly complex contour detection tasks. It is always important to study how human do their work and if possible, embedding such understanding into future research.

**References:**

[1] P. Arbelaez, M. Maire, C. Fowlkes, J. Malik, "Contour detection and hierarchical image segmentation", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 33, no. 5, pp. 898–916, 2011.

[2] W. Shen, X. Wang, Y. Wang, X. Bai, Z. Zhang, "Deepcontour: A deep convolutional feature learned by positive-sharing loss for contour detection", IEEE Conference on Computer Vision and Pattern Recognition, pp. 3982–3991, 2015.

[3] V. Kolmogorov, and A. Blake, "Grabcut: Interactive foreground extraction using iterated graph cuts," ACM transactions on graphics (TOG), vol. 23, no. 3, pp. 309–314, 2004.

[4] Y. Liu, M.-M. Cheng, X. Hu, K. Wang, and X. Bai, "Richer convolutional features for edge detection," IEEE Conference on Computer Vision and Pattern Recognition, pp. 5872– 5881, 2017.

**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 is now an associate professor in the School of Information and Communication Technology in Griffith University. Prior to this appointment, he had been a research fellow in the Australian National University, and a researcher at NICTA, Australia. His research interests are in spectral imaging, pattern recognition, computer vision, and their applications to and environmental informatics and remote sensing.

# 3D Shape Recognition and Retrieval in Multi-view Systems

*A short review for "Learning Multi-View Representation with LSTM for 3-D Shape Recognition and Retrieval"*

Edited by Carl James Debono

*C. Ma, Y. Guo, J. Yang and W. An, "Learning Multi-View Representation with LSTM for 3-D Shape Recognition and Retrieval," IEEE Transactions on Multimedia, vol. 21, vo. 5, pp. 1169 – 1182, May 2019.*

Representing 3D models efficiently poses an important problem in computer vision and related topics. Robust shape recognition and retrieval rely heavily on the accuracy of these models [1 – 2]. Moreover, various applications exploit shape representations, such as augmented and virtual reality, autonomous driving and other applications [3]. Progress in content capturing techniques that allow for 3D and multi-view data capture further demand better techniques to recognize 3D shapes and to retrieve them.

Lately deep neural networks have been applied successfully to a number of computer vision problems. In the 3D shape recognition and retrieval area current methods using convolutional neural networks (CNNs) are based on either volumetric analysis, such as [4], or multi-view based methods, such as [5]. In volumetric based methods, 3D CNNs are applied on the information represented by the 3D volumetric models. On the other hand, 2D CNNs are used in multi-view based methods that learn the 3D shapes from the multiple 2D viewpoints. Therefore, in multi-view systems, techniques that are studied for 2D images and video can be directly applied.

Long short term memory (LSTM) [6] is an implementation of recurrent neural networks (RNNs) that capture long term dependency in a sequence. In the work proposed in the original paper the authors exploit the idea in the time domain and use it in space. In this case each view of the object is treated as a sequence and processed using LSTM. Therefore, any correlation information between the views is captured by the internal memory in LSTM. The LSTM is also computationally efficient reducing resource requirements. This complexity is O(W) for each time step, with W representing the number of weights of the LSTM architecture.

The authors propose a joint CNN and LSTM architecture that learns to identify multi-view shape descriptors that are used for 3D shape recognition and retrieval. The joint architecture is efficient in identifying the 3D shapes from the multi-view descriptors. The CNN is used to extract the low-level features of the images within a view sequence, in this case the different viewpoints of the same scene. These features are then fed as a temporal sequence and aggregated into a shape descriptor through the LSTM. A two-layer bidirectional LSTM architecture is used to look at both the left to right and right to left directions of the views. A sequence voting layer is also employed to take into account the contribution of the hidden states in the LSTM. This gives some improvement in 3D shape recognition and retrieval. Other optimizations are done to improve performance such as the adoption of residual connection [7] in the CNN and a highway network [8] that bridges the CNN and the LSTM.

The original authors evaluated the solution using ModelNet dataset [9] and the ShapeNet Core55 dataset [10]. Their reported results indicate that the solution achieves a performance that is comparable to the state of the art solutions. These comparisons involved a number of methods that use multi-view systems for object recognition. The 3D shape retrieval on the ShapeNet Core 55 dataset were not compared to other results due to the few experiments available using this dataset. The reported accuracy of classification on the testing set was 86.11%. The different modules of the network ere also analyzed on the ModelNet dataset [8]. The solution was also compared to other methods for 3D shape retrieval using the ModelNet dataset [8]. This evaluation was conducted using the area under curve and mean average precision metrics. The reported results on the original author's solution are either similar or

better than the other techniques. Comparison was also conducted with respect to the algorithm used in the SHape Retrieval Contest (SHREC) edition of 2017. Good results were also reported there.

Algorithms for 3D shape recognition and retrieval are becoming more and more important with the increased use of multi-view and 3D camera systems and their use in various applications. Accuracy and speed of computation need to improve further for such systems to be used in real-time applications.

**References:**


[1] S. Bu, Z. Liu, J. Han, J. Wu, and R. Ji, "Learning high-level feature by deep belief networks for 3-D model retrieval and recognition," *IEEE Trans. Multimedia*, vol. 16, no. 8, pp. 2154–2167, Dec. 2014.

[2] H. Tabia and H. Laga, "Covariance-based descriptors for efficient 3D shape matching, retrieval, and classification," *IEEE Trans. Multimedia*, vol. 17, no. 9, pp. 1591–1603, Sep. 2015.

[3] Y. Guo, M. Bennamoun, F. Sohel, M. Lu, and J. Wan, "3D object recognition in cluttered scenes with local surface features: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 11, pp. 2270–2287, Nov. 2014.

[4] D. Maturana and S. Scherer, "VoxNet: A 3D convolutional neural network for real-time object recognition," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2015, pp. 922–928.

[5] H. Su, S. Maji, E. Kalogerakis, and E. Learned-Miller, "Multi-view convolutional neural networks for 3D shape recognition," in *Proc. IEEE Int.Conf. Comput. Vis.*, 2015, pp. 945–953.

[6] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[8] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Highway networks," 2015, arXiv:1505.00387

[9] Z. Wu *et al*., "3D ShapeNets: A deep representation for volumetric shapes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1912–1920.

[10] M. Savva et al., "SHREC'16 track: Large-scale 3D shape retrieval from ShapeNet Core55," in *Proc. Eurograph. Workshop 3D Object Retrieval*, 2016, pp. 89–98.


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

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

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

# Minute-by-Minute Detection of Obstructive Sleep Apnea and Automatic Measurement of the Apenea-Hypopnea Index

*A short review for "Automatic Detection of Obstructive Sleep Apnea Using Wavelet Transform and Entropy-Based Features From Single-Lead ECG Signal"*

Edited by Bruno Macchiavello

*A. Zarei and B. M. Asl, "Automatic Detection of Obstructive Sleep Apnea Using Wavelet Transform and Entropy-Based Feature From Single-Lead ECG Signal," in IEEE Journal of Biomedical and Health Informatics, vol. 23, no. 3, pp. 1011-1021, May 2019.*

Obstructive sleep apnea (OSA) is a sleep disorder that involves cessation or significant decrease in airflow during sleep [1]. Clinically, there are three types of Sleep Apnea (SA): OSA, Central SA and Mixed SA [2]. OSA is the most common sleep breathing disorder. This type of apnea occurs when your throat muscles intermittently relax and block your airway during sleep. OSA causes excessive daytime drowsiness, neurocognitive deficits, fatigue, depression, and heart stroke [3]. Also, undiagnosed and untreated OSA may lead to a high blood pressure, brain stroke, myocardial infarction, arrhythmias, and ischemia [4]. Even though OSA is detectable, several cases are still not recognized [5]. Currently, gold standard for OSA detection is polysomnogram (PSG). Since this method is time consuming and cost inefficient, practical systems focus on the usage of electrocardiogram (ECG) signals for OSA detection. In 2000, the organizers of Physionet database held a challenge to detect the OSA using a single-lead ECG signal, in order to show the importance of the issue [6]. Previous works, focused on feature extraction from the ECG signal in order to diagnose OSA [7,8,9]. The two common issues found in those works, are the large number of features used, which results in a high computational load; and the lack of computation of the Apnea-Hypopnea Index (AHI).

The authors of this work evaluated non-linear entropy-based features, developed a feature selection algorithm for dimensions reduction and lowering complexity, applied different classification methods in order to verify which one will be more adequate and provided a minute-by-minute OSA detection and AHI measurement for patients in the treatment stage.

The proposed method consists of two main steps. The first step is feature selection. This is done by first per-processing the ECG signal, then use feature extraction and selection in the Wavelet Transform domain. In the second step AHI is calculated by dividing the total number of apnea events by the total number of minutes of actual sleep time, then multiply by 60. Then, considering the calculated AHI, the subjects are marked as apnea or normal.

The pre-processing step is a very simple one, where the ECG signal is divided into frames of 30s of time duration. The frames are then filtered by a Chebyshev type II band-pass filter. The autocorrelation index is computed for each frame. Frames that have an index lower than a certain threshold are discarded. The frames are then transformed to a scale-time domain using the discrete wavelet transform (DWT). The authors used an 8-level DWT, meaning that the transformed signal will have one approximation coefficients sub-band, and 8 high-frequency sub-bands.

Once the signal is in the wavelet transform domain, feature extraction is performed. The authors propose 12 features. The first feature is Approximate Entropy (ApEn), which is and statistical method used for quantization of irregularity of signals. The more ApEn, the more irregularity characteristics of signals, and vice versa. Similar to ApEn, the authors also used Sample Entropy. The third feature is Fuzzy Entropy which measure the irregularity of a signal using exponential function for computation of the similarity degree. The fourth feature is the Correct Condition Entropy, which is a variation of Shanon's conditional entropy adapted for data with low available samples. The fifth one is the actual Shannon Entropy. The next three features are obtained by calculation the graphical correlation of a signal with its shifted version. Those features are the short and log-time standard

deviation and the ration between them. Then, the recurrence plot index, mean absolute deviation and the variance are also used as features. In the end 108 features are computed (12 features for each of the 9 sub-bands). The best set of features is selected by a sequential forward feature selection (SFFS) algorithm and fed into different classifiers.

The classifiers used in this study were: Support Vector Machines (SVM); Least-Square SVM; LIBSVM (all three using RBF, Polynomial or Linear Kernels); Linear Discriminant Analysis; Quadratic Discriminant Analysis; Artificial Neural Networks; K-nearest Neighbor; Naive Bayes Model; Logistic Regression and an Ensemble Model using Gentleboost Method. The dataset used is available online. The ECG signals were sampled at 100 Hz, with 16 bits resolution and modified lead V2 electrode configuration using an overnight PSG recording. The records have variable lengths of 7–10 hours. A specialist has attached the apnea labels (minute-by-minute) using the PSG signals. The authors provide performance of the classifiers in terms of accuracy, sensitivity, specificity, AUC, MCC, F-measure and Kappa coefficient. The SVM classifier with the RBF kernel presented the better results. The proposed method resulted in an accuracy of 92.98% in minute-by-minute classification on the Apnea-ECG database, outperforming other compared methods. As a future work may include a method that include a fusion of different classifiers, OSA detection through other type of signals and application of a deep neural network.

**References:**

[1] T. Young, M. Palta, J. Dempsey, J. Skatrud, S. Weber, and S. Badr, "The occurrence of sleep-disordered breathing among middle-aged adults," New Engl. J. Med., vol. 328, no. 17, pp. 1230–1235, 1993.

[2] M. O. Mendez, A. M. Bianchi, M. Matteucci, S. Cerutti, and T. Penzel, "Sleep apnea screening by autoregressive models from a single ECG lead," IEEE Trans. Biomed. Eng., vol. 56, no. 12, pp. 2838–2850, Dec. 2009.

[3] N. M. Ghahjaverestan et al., "Coupled hidden Markov model-based method for apnea bradycardia detection," IEEE J. Biomed. Health Inform., vol. 20, no. 2, pp. 527–538, Mar. 2016

[4] C. Guilleminault, S. J. Connolly, and R. A. Winkle, "Cardiac arrhythmia and conduction disturbances during sleep in 400 patients with sleep apnea syndrome," Am. J. Cardiol., vol. 52, no. 5, pp. 490–494, 1983

[5] H. D. Nguyen, B. A. Wilkins, Q. Cheng, and B. A. Benjamin, "An online sleep apnea detection method based on recurrence quantification analysis," IEEE J. Biomed. Health Inform., vol. 18, no. 4, pp. 1285–1293, Jul. 2014.

[6] M. Drinnan, J. Allen, P. Langley, and A. Murray, "Detection of sleep apnoea from frequency analysis of heart rate variability," in Proc. Comput. Cardiol. Conf., 2000, pp. 259–262

[7] C. Varon, A. Caicedo, D. Testelmans, B. Buyse, and S. Van Huffel, "A novel algorithm for the automatic detection of sleep apnea from singlelead ECG," IEEE Trans. Biomed. Eng., vol. 62, no. 9, pp. 2269–2278, Sep. 2015

[8] C. Song, K. Liu, X. Zhang, L. Chen, and X. Xian, "An obstructive sleep apnea detection approach using a discriminative hidden Markov model from ECG signals," IEEE Trans. Biomed. Eng., vol. 63, no. 7, pp. 1532– 1542, Jul. 2016

[9] M. Bsoul, H. Minn, and L. Tamil, "Apnea MedAssist: Real-time sleep apnea monitor using single-lead ECG," IEEE Trans. Inf. Technol. Biomed., vol. 15, no. 3, pp. 416–427, May 2011.

**Bruno Macchiavello** is an associate professor at the Department of Computer Science of the University of Brasilia (UnB), Brazil. He received his B. Eng. degree in the Pontifical Catholic University of Peru in 2001, and the M. Sc. and D.Sc. degrees in electrical engineering from the University of Brasilia in 2004 and 2009, respectively. Prior to his current position he helped develop a database system for the Ministry of Transport and Communications in Peru. He is an Editor for the Elsevier Journal Signal Processing: Image Communications. His main research interests include video and image coding, image segmentation, distributed video and source coding, multi-view and 3D video processing. He is currently head of the Graduate Program of Informatics at UnB.

# From 360° Video + Stereo Audio to 360° Audio

*A short review for "Towards Generating Ambisonics Using Audio-Visual Cue For Virtual Reality"*

Edited by Frank Hartung

*Aakanksha Rana, Cagri Ozcinar, Aljosa Smolic, "Towards Generating Ambisonics Using Audio-Visual Cue For Virtual Reality", IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), 2019, pp. 2012-2016.*

Virtual Reality (VR) has been one of the major hypes of the past years, and although it does not yet live up to all initial expectations, it is getting more and more important. Head-mounted VR displays are getting better with less wires, less weight and higher video resolution. Still, capture, encoding, transmission and rendering of video, audio and haptics for VR still pose a number of open research problems.

In the discussed publication, Rana, Ozcinar and Smolic discuss the issue of Ambisonics, i.e., full-sphere surround sound, aka 3D audio or 360 audio, for VR. In contrast to 360° video capture, the capture of corresponding spatial sound is still challenging due to the required sound-field microphones or information about the audio source locations. However, the creation of realistic VR worlds requires the 360° video to be accompanied by spatial audio [1][2]. In the paper, the authors discuss the generation of Ambisonics for 360° video based VR. The concrete challenge is the generation of 360° audio from captured mono or stereo audio plus 360° video.

The approach taken is based on the availability of a dataset of 265 videos of 10 seconds duration each with manually annotated audio source locations. Based on the dataset, deep learning algorithms are trained to estimate and encode the 3D audio source locations. The authors also propose evaluation metrics for the evaluation of the results. They argue that the publication opens up a new area of research investigations in 360° audio-visual analysis for VR.

The basis, as mentioned, is the availability of a newly collected database of 265 annotated 360° videos. The authors have generated them from YouTube where the videos are available without 3D audio information. The authors have annotated the videos to produce a ground-truth providing the audio source directions and locations. The videos cover different types of scenes, such as round-table discussions, presentations, meetings, and documentaries. The authors announce that they shall provide the dataset at https://github.com/V-Sense/360AudioVisual.

Further, the authors have established a workflow pipeline for generating full-sphere surround sound environments from 360° video plus audio. They describe the workflow as a series of four steps or modules that they denote (I) input representations, (II) feature embedding, (III) prediction models, and (IV) Ambisonics encoding.

The first step ("input representations") is a format conversion step where different 360° video formats like equirectangular and cubical are converted.

The second step ("feature embedding") combines deep learning based video and audio embedding methods [3][4] that identify and correlate audio and video features, based on pre-trained VGG-19 and VGG-19 like networks for audio and video classification [5][6]. While such techniques have previously been applied to 2D video, the application to 360° video is novel.

The third step ("prediction models") combines the audio and video feature maps and estimates 3D volumetric maps. This step is also based on a combination of modified (partial) pre-trained networks, followed by a geometrical mapping to the VR geometry. This essential step is unfortunately a little briefly explained and discussed in the publication.

The fourth and last step ("Ambisonics encoding") is a re-formatting of the acquired sound localization map in the special Ambisonic B-format. This step can be regarded as a decomposition of the 360° audio into a number of discrete localized sound sources, followed by a conversion into a sound field described by the four directional channels W, X, Y, Z of the B-format. W is the non-directional sound pressure level, the other three channels are directional

where X is the front-to-back, Y the side-to-side, and Z the up-to-down channel.

In order to evaluate the results, the authors propose two new metrics, named 360 Sound Source Distance (360-SSD) and 360 overlap error (360-OvErr). 360-SSD estimates the Euclidean distance between the center of a predicted audio source and the center of ground truth of the audio source. 360-OvErr is based on the ratio of an intersection of the predicted and ground truth probability volumes. This measure can be seen as a 3D variant of single object localization error.

The authors present results, both in terms of their newly proposed metrics and in terms of visualized qualitative results. Comparison to other work is not easily possible due to the novel nature of the research and the new metrics. However, it can be seen -mainly in the visualized qualitative results- that the method is able to determine 3D/360° sound locations from stereo sound + 360° video. The results seem in any case promising.

The main contribution of this paper is the establishment of a new research area with a novel approach of automatic spatial audio estimation based on audio-visual cues. The collection and publication of a database of 265 videos with annotated ground-truth will hopefully also enable others to continue work in the field. Also, the proposal of suitable metrics will hopefully contribute to better comparability of this and future work.

**References:**

[1] V. Sitzmann, A. Serrano, A. Pavel, M. Agrawala, D. Gutierrez, B. Masia, and G. Wetzstein, “Saliency in VR: How do people explore virtual environments?,” IEEE Transactions on Visualization and Computer Graphics, vol. 24, no. 4, pp. 1633–1642, April 2018.

[2] Martin Naef, Oliver Staadt, Markus Gross, "Spatialized audio rendering for immersive virtual environments," Proceedings of the ACM symposium on Virtual reality software and technology. ACM, 2002.

[3] A. Owens and A. A. Efros, “Audio-visual scene analysis with self-supervised multisensory features,” European Conference on Computer Vision (ECCV), 2018.

[4] T. Yapeng, S. Jing, Bochen L., D. Zhiyao, and X. Chenliang, “Audiovisual event localization in unconstrained videos,” in The European Conference on Computer Vision (ECCV), September 2018

[5] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in International Conference on Machine Learning, 2015.

[6] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. Channing Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. W. Wilson, “CNN architectures for large-scale audio classification,” CoRR, vol. abs/1609.09430, 2016.

**Frank Hartung** is a full professor of multimedia technology at FH Aachen University of Applied Sciences, Aachen, Germany. He received a MSc in electrical engineering from RWTH Aachen University, Germany, and a PhD in Telecommunications from University of Erlangen, Germany. He has been working with Ericsson Research, as a research team leader in Multimedia Technologies, from 1999 to 2011. In 2008, he was a visiting researcher at Stanford University, Palo Alto, USA. In 2015, he was a visiting professor at Namibia University of Science and Technology, Windhoek, Namibia, and in 2016, he was a visiting researcher at Eurecom, Sophia-Antipolis, France. His research interests include media security and forensics, video analysis, networked multimedia, immersive multimedia communication, streaming, and mobile video. He has authored or co-authored more than 60 publications, and is the co-inventor of 25 granted patents.

# Paper Nomination Policy

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

**Nomination Procedure**

Paper nominations have to be emailed to Review Board Directors: Qing Yang (qing.yang@unt.edu), Roger Zimmermann (rogerz@comp.nus.edu.sg), Wei Wang (wwang@mail.sdsu.edu), and Zhou Su (zhousu@ieee.org). The nomination should include the complete reference of the paper, author information, a brief supporting statement (maximum one page) highlighting the contribution, the nominator information, and an electronic copy of the paper, when possible.

**Review Process**

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

**Best Paper Award**

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

## MMTC Communications – Review Editorial Board



## Multimedia Communications Technical Committee Officers

**Chair:** Honggang Wang, University of Massachusetts Dartmouth, USA
**Steering Committee Chair:** Sanjeev Mehrotra, Microsoft Research, US
**Vice Chair – America:** Pradeep K Atrey, University at Albany, State University of New York, USA
**Vice Chair – Asia:** Wanqing Li, University of Wollongong, Australia
**Vice Chair – Europe:** Lingfen Sun, University of Plymouth, UK
**Letters & Member Communications:** Jun Wu, Tongji University, China
**Secretary:** Shaoen Wu, Ball State University, USA
**Standard Liaison:** Guosen Yue, Huawei, USA

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