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

# *R-LETTER*

Vol. 5, No. 4, August 2014

IEEE COMMUNICATIONS SOCIETY

## CONTENTS

## Message from the Review Board Directors

Starting with the August issue of the MMTC R-Letter we present a new team of directors and we would like to acknowledge the efforts of the review board members. Thus, we have invited the members of the review board to present their own work within this R-Letter.

The scope of the R-Letter is to stimulate research on all aspect of multimedia communication by nominating papers to be included in the regular category. The distinguished category comprises papers nominated by MMTC Interest Groups (IGs).

The **first paper**, published in the *Workshop on IIMC* and *co-authored by Carl James Debono*, describes means for low-complexity disparity estimation for 3D video coding.

The **second paper**, published in the *IEEE Transactions on Image Processing* and *co-authored by Vladan Velisavljević*, presents scalable multiview video coding approach which is adapted to the user behavior.

The **third paper** is *co-authored by Gene Cheung* and has been published within the *IEEE International Conference on Image Processing*. It proposes distributed source coding approach for stream switching without channel codes.

The **forth paper**, published in the *IEEE Transactions on Image Processing* and *co-authored by Rui Shen and Irene Cheng*, targets embedding perceptual criteria in a multi-exposure fusion hierarchy.

The **fifth paper**, published in *IEEE Transactions on Image Processing* and *co-authored by Jun Zhou*, describes a method towards an effective structure analysis for object detection.

The **sixth paper**, *co-authored by Carsten Griwodz* and published within the *ACM Transactions on Multimedia Computing, Communications and Applications*, describes experiences in engineering a video-based system to improve sport analytics for football.

Finally, the **seventh paper** is published in the *Proceedings of the ACM MMSys Conference* and *co-authored by Gwendal Simon*. It presents an optimal set of video representations for adaptive streaming over HTTP.

We would like to thank all the review board members for their time and efforts, specifically Irene Cheng for leading the MMTC review board in the past two years.

**IEEE ComSoc MMTC R-Letter**

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

**Christian Timmerer** is an associate professor at the Institute of Information Technology (ITEC), Alpen-Adria-Universität Klagenfurt, Austria. His research interests include immersive multimedia communication, streaming, adaptation, and Quality of Experience with more than 100 publications in this domain. He was the general chair of WIAMIS'08, QoMEX'13 and has participated in several EC-funded projects, notably DANAE, ENTHRONE, P2P-Next, ALICANTE, QUALINET, and SocialSensor. He also participated in ISO/MPEG work for several years, notably in the area of MPEG-21, MPEG-M, MPEG-V, and DASH/MMT. He received his PhD in 2006 from the Alpen-Adria-Universität Klagenfurt. Follow him on twitter.com/timse7 and subscribe to his blog blog.timmerer.com.

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

**Weiyi Zhang** is currently a Senior 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, Far-

go, 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 80 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.

**Co-Director**: Yan Zhang, Simular, Norwayr

Email: yanzhang@simula.no

**Prof. Yan Zhang** is currently Head of Department, Department of Networks at Simula Research Laboratory, Norway. He is also an Adjunct Associate Professor at the Department of Informatics, University of Oslo, Norway.

He received a PhD degree in School of Electrical & Electronics Engineering, Nanyang Technological University, Singapore. He is an associate editor or on the editorial board of a number of journals. He also serves as the guest editor for IEEE Communications Magazine, IEEE Wireless Communications Magazine, IEEE Transactions on Industrial Informatics, IEEE Systems Journal, and IEEE Intelligent Systems. His current research interests include: communications solutions for reliable and secure cyber-physical systems (e.g., transport, smart grid), machine-to-machine communications, Internet-of-Things, economic approaches for networks performance optimization.

# Low Complexity Disparity Estimation for 3D Video Coding

*A short review for "Low Complexity Disparity Estimation for Immersive 3D Video Transmission"*
*(Edited by Carl James Debono)*

*B. W. Micallef, C. J. Debono, and R. A. Farrugia, "Low Complexity Disparity Estimation for Immersive 3D Video Transmission," in Proc. of the Workshop on IIMC in International Conference on Communications 2013 (ICC'13), June 2013.*

The latest developments in display technology allow for an improved representation of 3D videos. This led to the standardization of the 3D video format [1], which consists of various textures and their per-pixel depth map video information coming from different viewing angles. 3D rendering techniques, such as those of Depth Image Based Rendering [2], can use this format to generate any arbitrary viewpoint, which is essential for an immersive service. Thus, both 3D depth perception [3], through a 3DTV display and free-viewpoint navigation [4], through an auto-stereoscopic 3DTV display, can be obtained at the receiver.

This large amount of data has to be transmitted through a bandwidth-limited channel, where the H.264/AVC's multi-view video and multi-view depth map coding extensions can be used to compress such videos [5]. These add disparity estimation techniques to single-view video coding to remove also inter-viewpoint redundancies [6], making Multi-view Video Coding (MVC) quite computational expensive.

To reduce the complexity of disparity estimation, the authors of this paper exploited the multi-view and the epipolar geometries, together with the transmitted depth map data, to geometrically assist the search for the inter-view correlation between different viewpoints' videos. Thus, disparity estimation will not have to do an intensive and exhaustive search for the optimal disparity match. This drastically reduces the computations and encoding delays. Moreover, they extended this idea to perform an adaptive search area along the epipolar lines; as depth and disparity are directly proportional, reducing further the disparities' search areas. Through simulation results on sequences with different characteristics, the authors demonstrated that the search area can become adaptive and reduced, giving speed-up gains of up to 32 times compared to the full-search estimation and of 4.5 times for the diamond search estimation, with little degradation to the multi-view video quality.

The authors investigated these techniques for the general disparity estimation technique, which can be applied on both Low-latency and Hierarchal bi-Prediction MVC structures. For the success of 3D video, it is crucial to reduce the coding complexities such that the coding delays will be as low as possible, or such that they do not consume a lot of resources or energy. This means that MVC coding time can become closer to single-view video coding whose performances are quite known and manageable with today's technology.

## References:


[1] MPEG & VCEG, "Multi-view video plus depth (MVD) format for advanced 3D video systems," JVT-W100, Apr. 2007.

[2] P. Kauff, N. Atzpadin, C. Fehn, M. Müller, O. Schreer, A. Smolic, and R. Tanger, "Depth map creation and image based rendering for advanced 3DTV services providing interoperability and scalability," Signal Processing: Image Comm. Special Issue on 3DTV, Feb. 2007.

[3] L. Onural, "Television in 3-D: What are the prospects?," Proceedings of IEEE, vol. 95, no. 6, Jun. 2007.

[4] M. Tanimoto, "Overview of free viewpoint television," Signal Processing: Image Communications 21, 2006.

[5] ISO/IEC IS 14496-10, Advanced Video Coding for Generic Audiovisual Service, ITU-T Rec. H.264, Version 18, Apr. 2013.

[6] P. Merkle, A. Smolic, K. Müller, and T. Wiegand, "Efficient prediction structures for multi-view video coding," IEEE Trans. on Circuits and Systems for Video Tech., vol. 17, no. 11, Nov. 2007.

# Scalable Multi-view Video Coding Adapted to User Behavior

*A short review for "User-Action-Driven View and Rate Scalable Multiview Video Coding"*
*(Edited by Vladan Velisavljević)*

*J. Chakareski, V. Velisavljević, V. Stanković, "User-Action-Driven View and Rate Scalable Multiview Video Coding," in IEEE Trans. on Image Processing, Vol. 22, No. 9, pp. 3473-3484, September 2013*

The recent advances in camera technology and associated algorithms have enabled capturing high–quality videos using consumer cameras at a reasonable cost. Using synchronized camera networks covering the same scene or enriching the video signals with depth maps comprising per-pixel distances between captured objects in the scene and the camera leads to a new paradigm called multi-view video. Having received the complex multi-view video signal, the end-users can synthesize novel views of the scene of interest that do not necessarily coincide with the capturing camera viewpoints. Common techniques for view synthesis include depth-image-based-rendering [1] or three-dimensional warping [2] that use videos and depth maps captured at two closest viewpoints as anchors. Such techniques enable a great enhancement in viewers' visual experience allowing for interactive selection of any desired viewpoint within an observation range. They also provide a platform for novel products, such as free viewpoint television [3].

The multi-view video signals are characterized by a huge amount of data required for seamless view synthesis by the end-users. Since transmission resources are generally limited in communication networks, compression of the multi-view videos and depth maps plays an important role. In the standard approach, the available bit budget would be optimally allocated between the videos and depth maps so that the expected distortion of the synthesized views is minimized. However, there are two practical problems in the described scenario. First, the available transmission bit rate that can be utilized is not often precisely known at encoding time. This occurs due to the presence of users with heterogeneous device capabilities and Internet access links. Second, the end users can exhibit a variety of behaviors in terms of interactive view selection. This affects the rate-distortion optimization problem so that there is no a single rate allocation that minimizes the synthesized view distortion for all users.

This paper addresses the two issues by examining how an efficient representation of multi-view videos and depth maps can be developed given the variable rate constraints and user behavior. The proposed method consists of an edge-adaptive wavelet multi-view image transform that is providing a view and rate scalable bit stream. Portions of this bit stream can be extracted and decoded at different bit rates to match the variable rate constraints. At each step of the iterative coding process, the encoder considers either refining the quality of already encoded views or introducing a new viewpoint into the representation. The specific choice of the encoder's action is made based on the rate-distortion efficiency of these two operations. The constructed scalable representation enables adding encoding bits incrementally to the compressed bit stream, while ensuring high quality of synthesized views is delivered. The scalability level can be controlled with an arbitrarily fine granularity in response to the deployment scenario.

In conjunction, the derived codec is user-driven, where the encoded bit stream is dynamically adapted to the anticipated user's view selection. The user behavior is modeled as a discrete-time Markov chain over the state-space spanned by the collection of possible viewpoints. The coding resources are then optimally allocated according to the anticipated actions of the user over a horizon of time instances using the Markov chain. This enables a higher multi-view video quality perceived by the user given the available rate budget and it also reduces the application latency.

The presented simulation results demonstrate the potential of the novel optimized rate allocation method. The quality of the delivered and synthesized views in several multi-view streaming scenarios is superior to those of the same codec with a simple uniform rate allocation and the state-of-the-art H.264 / Scalable Video Coding (SVC) codec [4]. In addition, the novel codec can provide an arbitrarily fine granularity of the encoded bit rate and simultaneous view scalability, unlike H.264/SVC.

In summary, the derived optimization framework for joint view and rate scalable coding of multi-view video content enables the encoder to select a subset of coded views and their encoding rates such that the aggregate distortion over a continuum of synthesized views is minimized. The constructed embedded bit stream delivers optimal performance simultaneously over a discrete set of transmission rates. The method also includes a newly developed user interaction model that characterizes the user's view selection actions as a Markov chain over a discrete state-space. The designed codec is adaptive to various rate con-

straints and to dynamic user behavior. The achieved quality of the synthesized views on the user side exceeds the same obtained by using the same coding without the adaptation or the standard H.264/SVC.

**References:**

[1] H.-Y. Shum, S.-C. Chan, and S. B. Kang, Image-Based Rendering. New York, NY, USA: Springer-Verlag, 2007.

[2] Y. Morvan, D. Farin, and P. H. N. de With, "Multiview depth-image compression using an extended H.264 encoder," in Proc. Adv. Concepts Intell. Vis. Syst., Lect. Notes Comput. Sci., 2007, pp. 675–686.

[3] T. Fujii and M. Tanimoto, "Free viewpoint TV system based on ray-space representation," Proc. SPIE, vol. 4864, pp. 175–189, Jan. 2002.

[4] H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the scalable video coding extension of the H.264/AVC standard," IEEE Trans. Circuits Syst. Video Technol., vol. 17, no. 9, pp. 1103–1120, Sep. 2007.

.

# Distributed Source Coding for Stream Switching w/o Channel Codes

*A short review for "Rate-distortion Optimized Merge Frame using Piecewise Constant Functions"*
*(Edited by Gene Cheung)*

*Wei Dai, Gene Cheung, Ngai-Man Cheung, Antonio Ortega, Oscar Au, "Rate-distortion Optimized Merge Frame using Piecewise Constant Functions," IEEE International Conference on Image Processing, Melbourne, Australia, September, 2013.*

In *interactive video streaming*, a client can in real-time freely choose subsets of a high-dimensional media content for personalized consumption. In response, the server must transmit pre-encoded data that corresponds to the requested media subsets for correct decoding and display at client. Examples of interactive video streaming include switching among streams of the same video encoded at different bitrates for real-time bandwidth adaptation, view-switching among videos capturing the same dynamic 3D scene from different cameras [1], etc. To support interactive video streaming, the technical challenge is to pre-encode a high-dimensional video content efficiently, while providing flexible mechanisms to facilitate stream switching. One simple method is to insert an intra-coded I-frame at each application-required switching point. While I-frames can facilitate stream switching, their large size means frequent insertion is not practical.

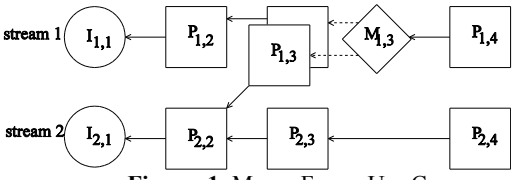


**Figure. 1**: Merge Frame Use Case

Towards a more efficient stream-switching mechanism, *distributed source coding* (DSC) exploits the correlation between the set of possible frames from which a client is switching (called *side information* (SI)) and the target frame for coding gain [2]. Specifically, each code block is first mapped from pixel to transform domain. Then bit-planes of each transform coefficient from all SI frames are compared to the bit-planes of the target frame. The "noisiest" bit-planes among SI frames – ones with the largest deviation from target frame – are then identified, and channel codes strong enough to overcome the worst-case noise are encoded as the DSC frame. At the decoder, any SI frame plus DSC frame can result in an identical reconstruction of the target frame. See Fig. 1 where a *merge* (DSC) frame $M_{1,3}$ is used to overcome the noise in the two SI frames $P_{1,3}$ with respective predictor frames ($P_{1,2}$ and $P_{2,2}$) in two different streams to reconstruct an identical target frame. DSC frame can potentially be much smaller in size than a comparable I-frame [2].

However, DSC frame design employs bit-plane encoding and channel codes, which translate to high computation complexity in both encoder and decoder. In this paper, the authors pursue a new approach to the stream-switching problem based on the concept of "signal merging"—*merging any SI into an identically reconstructed good signal*—using *piecewise constant* (pwc) function as the merge operator. Specifically, the authors propose a new merge mode for a code block, where for the *k*-th transform coefficient in the block, they encode appropriate step size and horizontal shift parameters of a **floor** function at the encoder, so that the resulting **floor** function at the decoder can map corresponding coefficients from any SI frame to the same reconstructed value. The selection of step size and horizontal shift directly affects both the merged signal fidelity and the coding rate; they proposed rate-distortion (RD) optimization procedures to optimize these parameters, as well as the selection of coding modes between intra and merge on a per-block basis. Experimental results show noticeable coding gain over a previous DSC implementation at low- to mid-bitrates at reduced computation complexity.

**References:**

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

[2] N.-M. Cheung, A. Ortega, and G. Cheung, "Distributed source coding techniques for interactive multiview video streaming," in *27th Picture Coding Symposium*, Chicago, IL, May 2009.

# Embedding Perceptual Criteria in a Multi-Exposure Fusion Hierarchy

*A short review for "QoE-Based Multi-Exposure Fusion in Hierarchical Multivariate Gaussian CRF"*
*(Edited by Rui Shen and Irene Cheng)*

*R. Shen, I. Cheng, and A. Basu, "QoE-Based Multi-Exposure Fusion in Hierarchical Multivariate Gaussian CRF", IEEE Transactions on Image Processing, vol. 22, no. 6, pp. 2469-2478, Jun. 2013.*

Natural scene contains a rich amount of visual information, which is beyond the capture range of any image sensor. For a scene with high dynamic range, a commonly adopted solution is to acquire the scene using multiple exposures and to apply image fusion techniques [1], in order to preserve the combined detail in a single image. However, the challenge lies in optimally integrating the perceptual characteristics in these multi-exposure images so that the human eye can view the best visual content, with the ability to see detail that otherwise may be over- or under-exposed.

The authors propose to exploit human perception on local contrast and apply in image fusion, in order to achieve the optimal reproduction of scene detail. This perceived local contrast is a measure of the probability that human eyes correctly discriminate a stimulus with certain contrast from the background. This is accomplished by calculating the perceived local contrast in a multi-scale framework. First, physical luminance contrast is calculated at each scale of the input multi-exposure images. Second, the contrast values are aggregated to form a single contrast score for each region. Third, a transducer function and a psychometric function [2] are applied to the aggregated physical contrast to approximate the nonlinearity of human perception of local contrast. The output from this third step is the perceived local contrast. Each pixel in each input image is associated with a perceived local contrast value. To achieve better color reproduction, the authors also employ a color saturation measure, which complements the perceived local contrast measure.

An important consideration in multi-exposure image fusion is to deliver seamless patch boundaries. Simply taking the measurement from the QoE (quality of experience) calculation above as fusion weight can generate unnatural boundary between image patches. To avoid such artifact, the authors model the fusion weights as a piecewise smooth surface embedded in a multi-dimensional space, named Multivariate Gaussian Conditional Random Field (MGCRF). With the QoE metric defined, the optimal set of fusion weights is equivalent to the maximum *a posteriori* (MAP) configuration of MGCRF. Compared with the Generalized Random Walks (GRW) model introduced in their previous work [3], MGCRF offers a more flexible fusion weight calculation. Varying the boundary vectors in MGCRF can control different appearances of the fused image. For computational efficiency, the authors propose to calculate the fusion weights in a hierarchical version of MGCRF. Test results show that this hierarchical computation improves both speed and memory usage.

Detailed objective and subjective analysis on their multi-exposure fusion algorithm were conducted. Two objective metrics to measure overall edge information reproduction [4] and per-pixel contrast reproduction [5] were deployed. The authors use four criteria in the subjective analysis, i.e., global contrast, detail, color, and overall appearance. The objective and subjective experimental results demonstrate the algorithm's superior performance and advantage over several other state-of-the-art methods.

This algorithm can be extended and applied to fusion problems in general, *e.g.* multi-modality medical image fusion [4]. It is also worth exploring the capability of the MGCRF model in dealing with the backward compatibility, high dynamic range compression and visualization problems studied by the research community.

**References:**


[1] R. S. Blum and Z. Liu, editors. *Multi-Sensor Image Fusion and Its Applications*. CRC Press, 2005.

[2] M. A. García-Pérez and R. Alcalá-Quintana, "The transducer model for contrast detection and discrimination: Formal relations, implications, and an empirical test," *Spatial Vision*, vol. 20, nos. 1–2, pp. 5–43, 2007.

[3] R. Shen, I. Cheng, J. Shi, and A. Basu, "Generalized random walks for fusion of multi-exposure images," *IEEE Transactions on Image Processing*, vol. 20, no. 12, pp. 3634–3646, 2011.

[4] C. S. Xydeas and V. Petrovíc, "Objective image fusion performance measure," *Electronics Letters*, vol. 36, no. 4, pp. 308–309, 2000.

[5] T. O. Aydin, R. Mantiuk, K. Myszkowski, and H.-P. Seidel, "Dynamic range independent image quality assessment," in *ACM SIGGRAPH*, 2008, pp. 1–10.

[6] R. Shen, I. Cheng, and A. Basu, "Cross-scale coefficient selection for volumetric medical image fusion," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 4, pp. 1069–1079, 2013.

# Towards Effective Structure Analysis for Object Detection

*A short review for "Object Detection via Structural Feature Selection and Shape Model"*
*(Edited by Jun Zhou)*

*Huigang Zhang, Xiao Bai, Jun Zhou, Jian Cheng and Huijie Zhao. "Object Detection via Structural Feature Selection and Shape Model", IEEE Transactions on Image Processing, Vol. 22, No. 12, Pages 4984-4995, 2013.*

Object detection is an active research topic in computer vision and pattern recognition, which requires effective object description by developing structural and textural features. Amongst various structure-based features, contour has shown unique properties as it is invariant to color, texture, and brightness changes [1, 2]. Its construction and model development does not rely on large number of training samples either [3].

A disadvantage of contour features is that they are sensitive to background noises. This is normally alleviated by training foreground models using bounding boxes. This, however, is often hindered by the lack of ground truth in training samples. As a generic feature, contour features are also easily matched to irrelevant parts of an image without considering the geometric relationship between different parts of objects.

To address these two problems, in this paper, Zhang et al proposed a method for object detection via structural feature selection and part based shape model. This method uses feature selection strategy to automatically reduce the influence of background noises without need of using bounding boxes in the training data. It then builds a part based model with selected foreground features learned from training data. Finally, object detection is performed by matching each testing image with this built model. This method has demonstrated exceptional performance on the ETHZ shape dataset [4] and INRIA horse dataset [5], which, according to the authors, was the best result that by the date of paper submission.

The technical value of this paper can be summarized in two aspects. The first one is on how to develop a good contour descriptor for efficient object detection. A simple solution is to adopt a local feature voting strategy. However, as the authors pointed out, such strategy is not sufficiently reliable in the matching process. Therefore, the authors proposed a novel contour context descriptor, which combines PAS feature [6] and its context information for shape part description. A codebook is first generated via clustering the extracted PAS features. Then a log-polar coordinate system is established, which centers at the location where each PAS feature is extracted. This allows statistical Bag-of-Words fashion descriptor being constructed.

The second contribution is on how to eliminate the background responses for object boundary model learning. The ideal case is that the background noises can be removed automatically. To achieve this goal, the authors proposed an iteratively matching strategy to select the foreground features. Each feature is initially assigned an equal weight, then Earth Mover's Distance (EMD, [7]) is used to perform pairwise matching between images. Based on the contribution of the feature towards the EMD calculation, these feature weights are updated. The weights for foreground features become larger as the iteration goes on, and those for the background features are gradually reduced, which facilitates explicit feature selection. This is a practical solution to get a prediction model when abundant training information is not available.

The method in this paper has been tested on several datasets, including ETHZ shape dataset, INRIA horse dataset, and PASCAL 2007 dataset. The proposed method has shown comparable performance as several state-of-the-art methods, even though this method does not use any labelled bounding boxes.

Probably the only disadvantage of this method is the computational cost during the training step. It requires pair-wise feature comparison between training images in the EMD method and feature weight update step. However, because the required number of training samples is small, the training time is not a major concern in this case. The proposed method can be applied to object detection tasks, not only efficiently but also effectively.

**References:**

[1] B. Leibe, K. Schindler, N. Cornelis, and L. Van Gool, "Coupled object detection and tracking from static cameras and moving vehicles," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 30, no. 10, pp. 1683–1698, 2008.

[2] J. Shotton, A. Blake, and R. Cipolla, "Multiscale categorical object recognition using contour fragments," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 30, no. 7, pp. 1270–1281, 2008.

[3] Q. Zhu, L. Wang, Y. Wu, and J. Shi, "contour context selection for object detection: A set-to-set contour matching approach," Proceedings of European Conference on Computer Vision, 2008, pp. 774–787.

[4] V. Ferrari, T. Tuytelaars, and L. Van Gool, "Object detection by contour segment networks," Proceedings of European Conference on Computer Vision, 2006, pp. 14–28.

[5] F. Jurie and C. Schmid, "Scale-invariant shape features for recognition of object categories," Proceedings of Computer Vision and Pattern Recognition, vol. 2. 2004, pp. 90–96.

[6] V. Ferrari, F. Jurie, and C. Schmid, "Accurate object detection with deformable shape models learnt from images," Proceedings of Computer Vision and Pattern Recognition, 2007, pp. 1–8.

[7] O. Pele and M. Werman, "Fast and robust earth Mover's distances," Proceedings of Incremental Conference on Computer Vision, 2009, pp. 460–467.

# Experiences in Engineering a Video-based System to Improve Sports Analytics for Soccer

*A short review for "Bagadus: An Integrated Real-Time System for Soccer Analytics"*
*(Edited by Carsten Griwodz)*

*H.K. Stensland, V.R. Gaddam, M Tennøe, E. Helgedagsrud, M. Næss, H. K. Alstad, A. Mortensen, R. Langseth, S. Ljødal, Ø. Landsverk, C. Griwodz, P. Halvorsen, M. Stenhaug, D. Johansen, "Bagadus: An Integrated Real-Time System for Soccer Analytics", ACM Transactions on Multimedia Computing, Communications, and Applications, vol. 10, no. 1s, Jan. 2014.*

International football (or soccer) feeds a large international sports entertainment market, and the largest clubs have fans all over the world. These clubs invest in huge support staff to understand the improvement potential of individual players and the team as a whole. Smaller or less financially endowed football clubs cannot afford such staff, but gaining an understanding of players' and team's performance is no less important for them than for the big players. Various commercial systems aim at providing statistical data to trainer teams [1-2], but they are aimed either at animated real-time output or after-action, offline analysis, and not meant for providing situation-based video feedback to trainers and players.

The paper presents a system named "Bagadus" that was developed in cooperation with a football club in Northern Norway, which combines panoramic video with tracking functions and trainer notes to allow immediate video review of relevant situations. The paper documents the challenges that were overcome and engineering effort required to create an integrated system, which consists of the three elements: video subsystem, tracking subsystem and analytics subsystem. It reports the state of the complete system and a variety of remaining challenges.

The paper provides a review of the state-of-the-art and concludes that integrated solutions for video-supported analytics are not found yet. It is argued that sports analytics will be a game-changer in sports, and Bagadus is introduced as a system that provides real-time video presentation of sport events within an analytics system. The three elements comprising the Bagadus system are introduced as video subsystem, tracking subsystems, and analytics subsystem; it is clarified that the main contribution of the paper lies in the evaluation, selection and integration of existing work (ranging from theoretical algorithms to existing products) into a complete analytics system that operates in real-time.

The paper provides a variety of details into the decision-making within individual components, explaining the choices that might be most valuable for other implementers. Both speed and quality aspects have been considered, and the paper does for example argue for a use of a homography stitcher according to Hartley and Zisserman [3] instead of its earlier panoramic projection after Brown and Lowe [4]. A variety of remaining challenges are also discussed.

The system is developed to achieve real-time performance on a single machine with CPU and GPU, and presents a pipeline whose stages are spread between host and GPU to achieve real-time performance. While several of the pipeline steps are explained in detail, others remain unclear. The use of color corrector and stitcher, for example, is explained in detail, whereas the need for color space conversion and the role of the background extractor remains unclear.

The tracking subsystem is concerned with the reason for selecting a commercial sensor-based tracking system [2] instead of a vision-based system, and explains the position mapping between tracking data and panoramic video. After showing the frontend of the system as provided to the trainer team in Tromsø, a variety of remaining challenges are deferred to the discussion section of the paper.

**References**

[1] Prozone, "Prozone Sports -- Introducing Prozone Performance Analysis Products," 2013. [Online]. Available: http://www.prozonesports.com/subsector/football/.
[2] ZXY, "ZXY Sport Tracking," 2013. [Online]. Available: http://www.zxy.no.
[3] R. Hartley and A. Zisserman, Multiple View Geometry in Computer Vision, 2nd ed. Cambridge University Press, 2004, p. 670.
[4] M. Brown and D. G. Lowe, "Automatic Panoramic Image Stitching using Invariant Features," Int. J. Comput. Vis., vol. 74, no. 1, pp. 59–73, Dec. 2006.

# Video Encoding Parameters for Adaptive Streaming

*A short review for "Optimal Set of Video Representations in Adaptive Streaming"*
*(Edited by Gwendal Simon)*

*Laura Toni, Ramon Aparicio, Alberto Blanc, Gwendal Simon, and Pascal Frossard, "Optimal Set of Video Representations in Adaptive Streaming", Proceedings of ACM MMSys conference, 2014.*

Adaptive Streaming is a technology that has been pushed by content providers to address the heterogeneity of video consumers. Since this technology is now widely adopted, it has become a central object of research in both multimedia and network communities. Studies can be roughly summarized into the following question "given a set of video representations, how to deliver them to end-users in a fair, efficient, cheap manner?"

What I like in this paper is that it takes a new perspective: *what if the set of video representations is not given*?

There are no commonly accepted rules on how to choose the encoding parameters (here resolution and bit-rate) of each video representation. Content providers typically use arbitrary rules of thumb. In particular, content providers often follow manufacturer's recommendations, which are general settings that take neither the nature of the video nor the characteristics of the population into account.

In this paper, the authors study this problem from the optimization standpoint. The optimization problem aims at maximizing the average user satisfaction for a content provider with a catalog of video and a given population of users. Some constraints have to be satisfied: a minimum ratio of end-users must be served, the overall bandwidth required to serve users must be below a given budget, and the total number of representations to encode must also be below a given number.

By using the solutions from this optimization problem, the authors derive two contributions: they measure the performances of recommended representation sets vis-à-vis the optimal set, and they provide guidelines for the selection of encoding parameters.

For a content provider with a large, popular, diverse catalog of videos, the recommended sets are not bad in terms of average Quality of Experience (QoE). However, for a given expected quality, the number of video representations in the recommended sets is almost twice the number of representations in the optimal solutions. In other words, the average QoE is obtained at the price of more video representations, which means more encoders, more storage, more delivery bandwidth in the CDN infrastructure and more complexity in the overall service management.

Content providers with more specific catalog of videos (such as those specialized in e-sport or those targeting mobile phones) should avoid following the recommendations from manufacturers. According to the results shown in the paper, the gains obtained from using specific encoder settings are worth the time spent at deciding them properly.

Some guidelines are extracted from the analysis of the solutions found by the optimization problem solver for a large set of synthetic configurations. Some of these guidelines are obvious, but authors make effort to provide proper measures to support them. These guidelines include:

- How many representations per video? The repartition of representations among videos needs to be content-aware (e.g. more representations for sport than for cartoon).
- How to decide bit-rates for representations in a given resolution? The higher is the resolution, the wider should be the range of rates, with an emphasis on lower rates.

In summary, this paper addresses the topic of adaptive streaming from a new perspective. For live streaming services, one of the most immediate future works is about the design of a strategy to automatically adjust the video encoding parameters to the type of video being delivered and to the population of users. More generally, since transcoding is becoming a commodity service in the cloud, this paper can also be seen as an entry point for studies about adjusting representation sets in order to cope with delivery issues. In the literature related to content delivery, network scientists deliver immutable packets. When it is about adaptive streaming delivery, researchers can play with another dimension, which is the data to be transferred. This paper is a remainder about this often-neglected option.

# 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:

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

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

### Review Process

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 R-Letter Editorial Board



# Multimedia Communications Technical Committee (MMTC) Officers

**Chair**: Yonggang Wen, Singapore
**Steering Committee Chair**: Luigi Atzori, Italy
**Vice Chair – North America**: Khaled El-Maleh, USA
**Vice Chair – Asia**: Liang Zhou, China
**Vice Chair – Europe**: Maria G. Martini, UK
**Vice Chair – Letters**: Shiwen Mao, USA
**Secretary**: Fen Hou, China
**Standard Liaison**: Zhu Li, 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.