VIDEO CODING: PREDICTIONS ARE HARD TO MAKE, ESPECIALLY ABOUT THE FUTURE

Antonio Ortega

Signal and Image Processing Institute
Ming Hsieh Department of Electrical Engineering
Viterbi School of Engineering
University of Southern California
antonio.ortega@sipi.usc.edu
http://sipi.usc.edu/˜ortega

ABSTRACT

The goal of this paper is to discuss how current trends in video acquisition, display and applications may impact video compression technology. We start by discussing some of these key trends, and in particular the challenges posed by scaling (increased frame resolutions) and the need for flexible access to complex datasets. Then we provide an overview of the recent evolution of video compression, with a particular focus on recently developed standards, such as H.264/AVC, and identify important factors that have contributed to performance improvements in the last decade. This allows us to identify several areas in which further gains in compression performance may be hard to achieve with current design techniques. Based on this, we propose areas of future research with a common focus on temporal coding tools.

***Index Terms*—** Video coding, RD optimization

1. INTRODUCTION

In spite of numerous predictions of an impending network bandwidth glut or of ever vanishing costs for memory, video compression continues to focus significant R&D activities. In fact, both the over-investment in network bandwidth that accompanied the "dot com" bubble and the truly amazing decreases in the cost of memory have led to more media being transmitted or stored, at even higher resolutions, so that the need for efficient encoding has, if anything, increased. 20 years ago digital video coding was being developed both as an enhancement to phone communications (in fact, serious development of video telephony started even earlier [2]) and as replacement for some of the TV delivery infrastructure. Today video encoding is available in myriad devices from cellphones to game systems or PCs.

In this paper we look back at recent technology developments in video coding and use those to sketch some thoughts about potentially interesting research directions. As should be obvious, we have focused on some specific aspects of recent video technology and do not claim to provide a complete overview of future directions (see [3] for a broader discussion.) In short, what we really address here is not so much the promise of specific coding tools (say, novel directional transforms or distributed video coding), but rather the validity of current "design philosophies", based on using multiple coding tools and rate-distortion (RD) optimization at the encoding, as we seek to apply it to new types of content (from very high resolution video to multi-view or 3D content) As will be seen the paper poses a few questions, but only provides hints about what could lead to some efficient solutions.

Motivation of this paper arises from three observations. First, highly optimized encoding for H.264/AVC produces outstanding RD performance but leads to artifacts that manifest themselves in specific types of content; significant efforts are being spent in improving encoding to address these problems. Second, encoding complexity may now be becoming a more important issue, while perhaps in the past it was considered a somewhat secondary factor; this has led to discussions of the need for new complexity-constrained algorithms. Third, new types of content (3D TV, multi-view video, etc) are being considered and compression tools are being developed that focus mostly on overall coding efficiency; not enough attention is being paid to how users may access these datasets and how well the compressed formats serve typical usage scenarios.

We start by describing in Section 2 how the needs of the field are evolving as a function of general trends in technology. In Section 3 we sketch the key technological developments that have accompanied recent improvements in video compression technology. Section 4 provides some ideas for where important research opportunities may be found based on the above mentioned observations. Section 5 provides some conclusions.

2. TRENDS IN VIDEO COMMUNICATIONS

The cost of pixel capture and display has been dropping dramatically. Video acquisition devices are becoming pervasive, from cellphones, to cameras, to webcams, etc. Available display resolution is ever increasing, from HDTV in the home, to CIF in many cellphones. With the cost of cameras dropping multi-camera systems are being investigated for some new applications (to be discussed below). Progress is also continuing to be made in display technology, starting with increases in display device resolution, improved auto-stereoscopic displays and so on.

The author is currently on sabbatical leave at Tokyo Institute of Technology (Titech). Generous funding from the Japan Society for Promotion of Science (JSPS) and the hospitality of Profs. Koichi Shinoda and Sadaoki Furui at Titech are gratefully acknowledged. A presentation that led to this paper was first made at the Sputnik50 workshop at Ecole Polytechnique Fédérale de Lausanne (EPFL) [1], before being presented in expanded form at the 2007 Picture Coding Symposium of Japan (PCSJ2007). This paper is dedicated to Prof. Martin Vetterli on the occasion of the 50th anniversary of the Sputnik launch.

As a consequence of the reduced cost of encoding, storage and bandwidth, professionally produced content is making room for a plethora of video produced by users. Where in the past the number of decoders far exceeded that of decoders, nowadays we see a trend to larger number of encoders. In fact the emergence of coding systems for surveillance leads to a situation where much of the data may never be decoded. Thus, the trade-offs in encoding/decoding complexity are now shifting so as to make low complexity encoding more important. Note that the true cost of encoding can be measured in many ways, and depend on factors other than the specific algorithms (e.g., power consumption will depend on the platform used for encoding and system cost will depend on production volume among other factors). Still, algorithmic complexity is a major factor in encoding cost.

Increases in video resolution often do not lead to changes in the applications themselves. Thus, similar type of content (e.g., films) is distributed in successive generations of video storage devices (such as VCDs, DVDs and emerging high definition systems). However, ongoing research is considering new modalities of video content. For example, immersive environments using very high resolution displays or wearable displays are being considered. To give an idea of the kinds of rates that "true immersion" would require, Tom Holman proposed to calculate "The bit-rate of reality"[4] and indicates that providing such a realistic environment would require "... 340M samples on a sphere, at 100 frames per second, 10bits/sample, double the number for color, double the number for depth [stereo] yields 1.36×10^{12} bits/second." In addition to immersive systems, techniques such as 3D video [5] or free viewpoint TV [6] allow scenes to be viewed from different angles, so that different users will interact in different ways with the same bitstream.

In this paper we consider these trends and ask whether current state of the art design techniques (discussed in Section 3) will continue to be effective. In particular this leads us to pose the following questions, which will be revisited in Section 4: i) Can current encoders perform equally well at higher resolutions? ii) Can encoder complexity be reduced (to support the increasing number of encoders)? iii) How to enable increased decoder flexibility so that individual users can access data in different ways?

3. STATE OF THE ART IN VIDEO ENCODING

3.1. Exploiting temporal redundancy

Video compression technology has focused on exploiting temporal redundancy, while using techniques developed for image coding (or similar to those) in order to exploit spatial redundancy. Thus efficient temporal prediction has always been a key determinant of overall video coding performance. Initial video coding techniques were necessarily limited by computation and memory requirements and so did not exploit temporal redundancy, simply using image coding tools, often low memory ones such as DPCM. A key development was that of block-based motion estimation and compensation tools [7]. These were initially too complex to be useful in practice, but have since then been widely adopted.

It is worth re-emphasizing this obvious point: the temporal dimension is fundamental in defining the efficiency of video coding schemes. One can think of a typical video compression system as using the *past*, i.e., information previously transmitted, as a way to reducing the encoding rate for the *future*, i.e., information yet to be encoded. Most current systems employ closed-loop prediction, so that the encoder includes a decoder, and the encoding order determines which frames are past and future at any given time (and obviously, as in the case of B-pictures or B-slices, the encoding order and the display order could be different). Examples of proposed coding structures include B frames and related ideas [8, 9], extensions of filtering techniques to the temporal dimension [10, 11, 12, 13], prediction from multiple frames [14], and others.

Note, however, that there are other situations where different definitions of future and past apply. For example, distributed source coding techniques [15, 16] can be used for essentially open-loop encoding, so that the encoder no longer needs to replicate the decoder, and indeed frames not seen by the encoder could be used to reduce the encoding rate of current frames. As another example, the encoder may use bit allocation techniques which try to approximate globally optimal bit assignments. In this case, the encoding modes and rates for a given frame may not be decided until some other frames have been analyzed. Thus, future frames (in terms of encoding order) are used to determine how to encode the current frame, even if they are not directly used for prediction.

While significant research effort has been devoted to better exploiting temporal redundancy in video, the ideas proposed in Section4 will suggest that much remains to be done in considering the temporal dimension in coding.

3.2. Basic technologies

In looking back to the early 90s and to the first widely adopted compression standard, MPEG-1, one can see that the essential coding system architectures essentially remain unchanged. Most current systems used in practice, representing millions of encoders and even greater number of decoders, make use of block-based motion estimation and compensation, followed by a block image transform, quantization and entropy coding. The discrete cosine transform (DCT) is used in every single DVD player as well as in millions of digital cameras, and indeed even when the transform has changed, as in H.264/AVC, techniques for encoding based on block transforms remain very similar to initially proposed ones [17]. Block-based motion estimation was proposed an early stage [7] and continues to be dominant. While it is clear that "true" motion in the scene is neither translational nor block-wise constant, the block-based nature of these algorithms can be very useful computationally. Recent developments have not fundamentally changed the nature of motion estimation, with the block-based approach remaining dominant, but extended to encompass different block sizes, pixel accuracies and prediction modes.

A major role in the development of video coding has been played by standards, such as MPEG-2 or H.264/AVC. All these standards are built on the assumption that the standard specifies the behavior of the decoder. Thus, a standard compliant bit-stream leads to unambiguous decoding, where the only flexibility allowed to the decoder is that of skipping back and forth in the bitstream, i.e., random access is enabled. Standardizing the decoder is needed for interoperability. Also important is the fact that innovation is possible at the encoder, thus letting multiple companies support a standard, while enabling competition in encoding quality, cost and complexity.

3.3. Recent developments

While fundamentally the basics of motion estimation remain the same, what has changed is the number of modes that can be used for encoding. In particular, a major addition to the H.264 standard is that of various block sizes for motion estimation and compensation, not just 16×16 pixels, but also 16×8, 8×16, 8×8, etc. This is clearly useful in that a block-based translational motion model hardly reflects real characteristics encountered in video sequences. By allowing smaller block sizes it is possible to provide a more accurate description of motion, but at the cost of additional bits required to represent the motion field.

The question then becomes how to determine when it is worthwhile to use a smaller block size. Within typical AVC encoders this is done by using rate-distortion (RD) optimization tools [18, 19]. In particular, comparing software used for simulation with the MPEG-2 and AVC standards a striking difference is the availability of Lagrangian based encoding in the latter [20, 21, 22]. While encoder optimization is of course an encoder issue (and thus not codified by the standard), it is important to highlight that comparisons AVC encoders and alternative techniques tend to be based on RD optimized encoding used for AVC.

Finally, two other factors are worth mentioning when explaining the coding gains achieved by AVC. The first one is the inclusion of loop filtering. It is well known that block transforms lead to blocking artifacts, i.e., coding parameters are selected in a blockwise manner and thus when considering pixels at either side of a block boundary there tends to be an increase in the error introduced. When combined with motion estimation, this leads to error increases in the likely case where a block used for matching in a previous frame is not aligned with the block boundaries for encoding. Loop filtering seeks to reduce this error contribution by filtering across prediction boundaries. Second, the standard seeks to entropy code as much of the information being transmitted as possible. This is in particular true for the syntax bits. As can be inferred from the fact that the number of operating modes increases, the number of bits used to describe modes also will increase and thus the percentage of overhead bits increases as well, along with the importance of compressing these well.

4. OPEN QUESTIONS

The current design philosophy followed by the recently developed H.264/AVC standard can be described, at the risk of oversimplifying, as follows: i) A large number of modes of operation is made available, ii) Lagrangian costs are computed for all or some of available mode selections, and iii) The best encoding mode is chosen for each small coding unit (e.g., a macroblock) based on the comparison of Lagrangian costs. Thus, as the number of available modes increases, so does the encoding complexity, as least for approaches that perform an exhaustive search. We now address the questions raised in discussing recent video communication trends.

4.1. Can encoder RD performance be maintained at high resolutions?

Currently applied RD optimization techniques are not truly optimal (e.g., they ignore temporal dependencies[23] as will be discussed in the next section). Moreover, these techniques are based on mean squared error (MSE), whose limitations are well documented [24]. These lead to encoding problems in the context of highly optimized standards, such as H.264/AVC, and these problems become more obvious as the content resolution increases (e.g., when using HDTV displays). A typical scenario is one where overall MSE has been minimized but it is possible to observe annoying flickering artifacts that may be particularly visible for some types of content [25]. Specifically, when the best Lagrangian cost is chosen on a block per block basis in each frame there is no guarantee that co-located blocks in successive frames will be coded in a similar manner. This leads to artifacts (such as flickering) that manifest themselves only when viewing the video sequence, and not when considering individual frames. Often these errors are clearly visible in very specific type of content, which may not have been included in original test sequences used for the definition of the standard. Thus, these errors in general were uncovered after the standard was completed and when attempts at practical encoder design are undertaken.

Significant progress has been made recently in defining quality metrics that are better at capturing perceptual quality. However, these new metrics (e.g., [26]) are in many cases developed for still images. Comparable results for video, that fully take into consideration the temporal effects, are not as well developed. To be more precise, while there are results in perceptual video quality, most of the work has been devoted to analyzing quality regardless of the encoding used, rather than to developing perceptual tools that can be used as part of the encoding. Developing perceptually oriented tools that incorporate temporal quality criteria is a key challenge for improving the performance of video encoding systems. Note that these kinds of tools are well developed in the context of audio coding [27].

Consider simple tools such as the “quantization matrices” in JPEG, which allow different weights to be applied to each DCT frequency. These are not as sophisticated as audio masking techniques which analyze specific signal segments [27], i.e., they essentially provide absolute masking, rather than relative masking across frequencies. However, they are simple to use and can be easily integrated with RD optimization tools, so that in practice the quality metric being optimized is at least “perceptually aware”. No similar tools exist that can be used for video encoding. We believe that important goal will be to define such simple tools *take into consideration perceptual characteristics in the temporal dimension and can be easily combined with RD optimization techniques.*

4.2. How to reduce encoder complexity?

After the completion of the H.264/AVC numerous authors have investigated techniques for lower complexity mode selection (e.g., [28]). However, other recent work that has instead explored highly complex encoding techniques can provide a more interesting perspective on encoding complexity. Two recent papers [29, 30], have proposed a model for global optimization of encoding choices, given a specific motion compensation structure, for a whole group of pictures. With this model, a quadratic optimization technique allows temporal dependencies to be considered to achieve improved coding performance. Unlike in [23, 31], it is possible to optimize encoding block-wise *and* to do so while taking into account temporal dependencies. Unsurprisingly these novel methods outperform RD optimized H.264/AVC encoding. The gains reported in [30] are below $1dB$ in PSNR, but it is worth noting that these gains are without changing the coding structure, which itself was chosen based on suboptimal minimization of Lagrangian costs (i.e., where modes were individually selected for each block, without taking into consideration what effect this may have on neighboring and future blocks). Thus, potentially, additional gains could be achieved if the motion compensation structure were also part of the optimization. In contrast to [23], in [30] the authors observed that in some cases reducing the rate used for frames close to the “root” of the prediction tree (e.g., initial P frames in a group of pictures) may in fact lead to better overall RD performance.

This result indicates a potential risk in following the current encoder design philosophy. H.264/AVC has introduced numerous additional modes of operation, as compared to previously proposed standards, including more possible motion vectors (e.g., quarter pel resolution vectors), more block sizes, several modes of intra prediction, etc. In theory, an encoder can achieve the optimal operating points by appropriate RD optimization techniques. However there are several problems with this assertion (in addition to the distortion metric itself, as discussed above). Clearly, RD optimization techniques (as widely used in developing the standard) do not provide an optimal solution. These evaluate Lagrangian costs for individual blocks within a slice, and use reasonable approaches to identify operating points for each block. However, they completely ignore the dependencies between choices made at the block level for a frame or slice and the quality of future frames/slices.

The total number of mode choices is increasing exponentially, and will continue to increase as frame resolution increases. This is important because of the coupling between modes illustrated by the results in [29, 30]. The risk we are facing is that as the space of possible operating modes becomes larger, the difference between a simple encoder and a highly optimized one may become greater.

In light of this increase in the number of possible mode combinations, and its impact on complexity, it is worth studying the implications of different possible structures for the solution space. Assume very few mode combinations provide excellent coding performance (i.e., there would be a very small subset with good operating performance in the set of possible solutions). This would be clearly inefficient. First, encoding complexity would tend to be high, as the encoder would need to find a small set of solutions. Second, the rate overhead needed for mode description may be in fact be significant, especially if the subset of optimal solutions changes over time. Consider the alternative situation now, where in fact good fast mode selection exist and the subset of optimal solutions does not change much as a function of content. Then, it might be more efficient to design a system with fewer modes and where inefficient mode combinations are ruled out by the system syntax.

Thus, we propose that a challenge for future systems will be to investigate coding structures and syntax that *reduce* the total number of allowable modes, while enabling *larger* numbers of coding tools to be used. What this calls for is a vector oriented approach to mode selection, where certain combinations cannot be selected by the syntax. Thus, for example, if one believes that using multiple reference frames is beneficial as a way to avoid performing subpel interpolation, then perhaps both tools may be not supported jointly by the syntax. Likewise, mode selections for blocks linked via temporal prediction could be similarly restricted. This would again lead to considering the temporal dimension and developing new *spatio-temporal modes, where the basic mode-decision unit is no longer a single macro-block or part of a frame, but information to be encoded across several frames*.

4.3. How to enable increased decoder flexibility?

As mentioned in Section 2, there are a number of applications where viewers are partially accessing some of the data set. This is increasing the need for encoders that allow some decoding flexibility. Taking as an example, a free viewpoint TV scenario, let us say that inputs from multiple cameras have been jointly encoded using tools such as those developed in the context of the multiview video coding (MVC) activity within JVT. Then a user can choose to change the display angle, in effect selecting only one of the views from the bitstream (for the purpose of this discussion it does not matter whether new views are interpolated or only those views already part of the bitstream can be displayed). Because encoding is performed jointly for all the views, in general, blocks from multiple views will need to be available in order to decode a block of interested. If the whole bit-stream is available to the user then this will be at the cost of some additional decoder complexity (i.e., blocks will need to be decoded that will not actually be displayed). Alternatively, if the user is requesting parts of the bitstream, then the overhead will be significant as discussed in [32]. Thus, in this mode of operation (joint encoding across views has been performed, but users wish to view only individual views), the joint encoding itself will lead to performance degradation with respect to simulcast, which would allow access to the each of the views individually [32].

Ideally, the goal the decoder should require just enough data to decode the specific view that has been requested, without requiring a very high bit-rate to be used. As an example, recent work in [33] shows that in a system with feedback, an intra format can be used (in this case JPEG 2000) and redundancy across frames can be exploited by using motion vectors and letting the decoder request only that data needed to update frames of interest, based on what was already received. An alternative approach [34, 35], with a different philosophy, uses distributed source coding to create a single bitstream that can be decoded in several different ways. The basic idea is that an encoder can accommodate decoding based on several different “side information” at the decoder (i.e., previously decoded frames that are used to decode newly transmitted information) by simply ensuring that decoding can proceed in the worst case correlation. These techniques show improved performance with respect to more traditional techniques, such as those used in SP frames [36], which essentially would involve transmitting multiple residual signals (one for each possible scenario of decoding).

In short, novel datasets are likely to enable more complex access to video information, allowing users to choosing a viewpoint and navigate in almost arbitrary way. Re-thinking how these bitstreams are generated and the flexibility the offer in terms of decoding will be important. In staying with the theme of this paper, in a way this means that “future” and “past” are no longer uniquely determined at the encoder. Instead, the *decoder should be allowed to follow different decoding paths, each one with a different ordering through the data, thus corresponding to different definitions of what is future and what is past for decoding purposes.*

5. CONCLUSIONS

In this paper we have considered recent trends in video communications to evaluate whether current design philosophies can provide sufficiently good performance. We have used recently published work to motivate three areas of research, with a common theme of giving increased importance to the temporal dimension of video coding. In particular, we propose to investigate i) coding tools to take in consideration temporal effects in perceptual quality, ii) spatio-temporal coding modes, and iii) encoding techniques that provide greater flexibility for the user to view the data, with corresponding changes in the decoding order.

6. REFERENCES

[1] “http://www.rle.mit.edu/stir/Sputnik50/,”.

[2] BG Haskell, FW Mounts, and JC Candy, “Interframe coding of videotelephone pictures,” *Proceedings of the IEEE*, vol. 60, no. 7, pp. 792–800, 1972.

[3] GJ Sullivan, J.R. Ohm, A. Ortega, E. Delp, A. Vetro, and M. Barni, “dsp Forum-Future of Video Coding and Transmission,” *IEEE Signal Processing Magazine*, vol. 23, no. 6, pp. 76–82, 2006.

[4] T. Holman, “The bit rate of reality [picture/sound reproduction],” *Intl. Conf. on Consumer Electronics, ICCE 2000*, pp. 398–399, 2000.

[5] W. Matusik and H. Pfister, “3D TV: a scalable system for real-time acquisition, transmission, and autostereoscopic display of dynamic scenes,” *ACM Transactions on Graphics (TOG)*, vol. 23, no. 3, pp. 814–824, 2004.

[6] T. Fujii and M. Tanimoto, “Free viewpoint TV system based on ray-space representation,” in *Proceedings of SPIE*. 2002, vol. 4864, pp. 175–189, SPIE.

[7] T. Koga, K. Iinuma, A. Hirano, Y. Iijima, and T. Ishiguro, “Motion compensated interframe coding for video conferencing,” *Proc. Nat. Telecommun. Conf*, vol. 5, no. 3, pp. 1–5, 1981.

[8] KM Uz, M. Vetterli, and DJ LeGall, “Interpolative multiresolution coding of advance television withcompatible subchannels,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 1, no. 1, pp. 86–99, 1991.

[9] 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, 2007.

[10] G. Karlsson and M. Vetterli, “Three dimensional sub-band coding of video,” in *Intl. Conf. on Acoustics, Speech, and Signal Processing, ICASSP-88.*, 1988, pp. 1100–1103.

[11] J.R. Ohm, “Three-dimensional subband coding with motion compensation,” *IEEE Transactions on Image Processing*, vol. 3, no. 5, pp. 559–571, 1994.

[12] A. Secker and D. Taubman, “Lifting-based invertible motion adaptive transform (LIMAT) framework for highly scalable video compression,” *IEEE Transactions on Image Processing*, vol. 12, no. 12, pp. 1530–1542, 2003.

[13] B. Pesquet-Popescu and V. Bottreau, “Three-dimensional lifting schemes for motion compensated videocompression,” in *Proc. of Intl. Conf. on Acoustics, Speech, and Signal Processing, ICASSP’01*, 2001, vol. 3.

[14] T. Wiegand, X. Zhang, and B. Girod, “Long-term memory motion-compensated prediction,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 9, no. 1, pp. 70–84, 1999.

[15] B. Girod, AM Aaron, S. Rane, and D. Rebollo-Monedero, “Distributed video coding,” *Proceedings of the IEEE*, vol. 93, no. 1, pp. 71–83, 2005.

[16] R. Puri, A. Majumdar, and K. Ramchandran, “PRISM: A Video Coding Paradigm With Motion Estimation at the Decoder,” *IEEE Transactions on Image Processing*, vol. 16, no. 10, pp. 2436–2448, 2007.

[17] W.H. Chen and W. Pratt, “Scene Adaptive Coder,” *IEEE Transactions on Communitcations*, vol. 32, no. 3, pp. 225–232, 1984.

[18] A. Ortega and K. Ramchandran, “Rate-distortion methods for image and video compression,” *IEEE Signal Processing Magazine*, vol. 15, no. 6, pp. 23–50, 1998.

[19] G.J. Sullivan and T. Wiegand, “Rate-distortion optimization for video compression,” *IEEE Signal Processing Magazine*, vol. 15, no. 6, pp. 74–90, 1998.

[20] H. Everett III, “Generalized Lagrange multiplier method for solving problems of optimum allocation of,” *Oper. Res.*, vol. 11, pp. 399–417, 1963.

[21] Y. Shoham and A. Gersho, “Efficient bit allocation for an arbitrary set of quantizers.,” *IEEE Trans. on Acoust. Speech Signal Processing*, vol. 36, no. 9, pp. 1445–1453, 1988.

[22] T. Wiegand and B. Girod, “Lagrange multiplier selection in hybrid video coder control,” in *Proceedings of Intl. Conf. on Image Proc., ICIP 2001*, 2001, vol. 3.

[23] K. Ramchandran, A. Ortega, and M. Vetterli, “Bit allocation for dependent quantization with applications tomultiresolution and MPEG video coders,” *IEEE Transactions on Image Processing*, vol. 3, no. 5, pp. 533–545, 1994.

[24] B. Girod, “What’s wrong with mean-squared error?,” *Digital images and human vision*, pp. 207–220, 1993.

[25] K. Chono, Y. Senda, and Y. Miyamoto, “Detented Quantization to Suppress Flicker Artifacts in Periodically Inserted Intra-Coded Pictures in H. 264 Video Coding,” in *Proc. IEEE Intl. Conference on Image Processing, ICIP 2006*, 2006, pp. 1713–1716.

[26] Z. Wang and AC Bovik, “A universal image quality index,” *IEEE Signal Processing Letters*, vol. 9, no. 3, pp. 81–84, 2002.

[27] JD Johnston, “Transform coding of audio signals using perceptual noise criteria,” *IEEE Journal on Selected Areas in Communications*, vol. 6, no. 2, pp. 314–323, 1988.

[28] P. Yin, H.Y.C. Tourapis, AM Tourapis, and J. Boyce, “Fast mode decision and motion estimation for JVT/H. 264,” in *Proc. IEEE Intl. Conf. on Image Processing, ICIP 2003.*, 2003.

[29] B. Schumitsch, H. Schwarz, and T. Wiegand, “Optimization of transform coefficient selection and motion vector estimation considering interpicture dependencies in hybrid video coding,” *Proc. SPIE*, vol. 5685, pp. 327–334, 2005.

[30] M. Winken, H. Schwarz, D. Marpe, and T. Wiegand, “Joint Optimization of Transform Coefficients for Hierarchical B Picture Coding in H. 264/AVC,” in *Proc. of ICIP 2007*, 2007.

[31] A. Ortega, K. Ramchandran, and M. Vetterli, “Optimal trellis-based buffered compression and fast approximations,” *IEEE Transactions on Image Processing*, vol. 3, no. 1, pp. 26–40, 1994.

[32] P. Lai, J. H. Kim, A. Ortega, Y. Su, P. Pandit, and C. Gomila, “New rate-distortion metrics for mvc considering cross-view dependency,” ISO/IEC JTC1/SC29/WG11 MPEG Document M13318, Apr 2006.

[33] A. Naman and D. Taubman, “A novel paradigm for optimized scalable video transmission based on jpeg 2000 with motion,” in *Proc. of IEEE Intl. Conf. on Image Proc.*, San Antonio, TX, Sept. 2007.

[34] N.-M. Cheung and A. Ortega, “Distributed source coding application to low-delay free viewpoint switching in multiview video compression,” in *Picture Coding Symposium, PCS 2007*, Lisbon, Portugal, Nov. 2007.

[35] N.-M. Cheung and A. Ortega, “Flexible video decoding: A distributed source coding approach,” in *Proc. of IEEE Workshop on Multimedia Signal Processing, MMSP 2007*, Crete, Greece, Oct 2007.

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