

technicolor



**A/V CODING: RESEARCH CHALLENGES AND
OPPORTUNITIES**

WHITEPAPER

CONTENTS

INTRODUCTION	3
GOAL OF THIS PAPER.....	3
STATE OF THE ART - COMPRESSION STANDARDS.....	3
STATE OF THE ART - COMPRESSION TECHNOLOGIES	4
RESEARCH CHALLENGES	7
FUTURE IMAGE AND VIDEO FORMATS	7
FUTURE AUDIO FORMATS	7
MULTIMODAL CONTENT	8
VIDEO PRODUCTION	8
DEALING WITH HETEROGENEITY AND ADAPTATION	8
COMPRESSION AND BANDWIDTH.....	9
COMPLEXITY.....	10
DELIVERY AND NETWORK ARCHITECTURES	10
HUMAN PERCEPTION	11
STANDARDS OPPORTUNITIES AND OPEN-SOURCE COMPETITION	12
A/V STANDARDS.....	12
CONCLUSIONS	14
BIBLIOGRAPHY	15
GLOSSARY	18

INTRODUCTION

Goal of this Paper

Audio and video coding are key technologies impacting all components of the media chain. Content creators, broadcasters, network operators, set-top box manufacturers, display manufacturers, and end-users, are all concerned by the compression of data and the way the signals are transmitted. For instance everybody knows about MP3, MP4 or Flash, these technologies are everywhere today in our lives.

As for other technology areas, the history of A/V compression is correlated to technology breakthroughs and new market opportunities: digital signal compression, Internet and MP3 for audio, MPEG-2 for TV, or MPEG-4 AVC for HDTV. The transition from one period to the next always raised questions about the interest in continuing research, about the technology limits, and about the advances of other domains and their consequences on content coding. Today we are at such transition and there is a need to understand the landscape and answer questions such as: Which coding format best suits a given application and why? Are new coding schemes needed? And if so, for which applications?

In addition, other alternatives that did not exist in the past, or with a lower impact, raise more

fundamental questions such as why to develop codecs based on standards while open-source solutions could be available (standardized codecs versus proprietary ones versus open source). Similarly, it may not be clear that compression is as critical as in the past with ever-increasing network bandwidth. Instead, the future might be in new coding schemes for growing media content such as multi-view 3D, Ultra-HD, or mixed natural/synthetic content.

Considering all those aspects, there are many challenges around A/V coding, and so, many opportunities. The objective of this report is to help understand which of these challenges are most critical. After a short state-of-the-art analysis on compression technologies, the technical challenges are discussed in the Research Challenges section. The next section is dedicated to the standards opportunities and the open-source alternatives. Finally, some conclusions and perspectives are proposed.

State of the Art - Compression Standards

Since the end of the 80's, numerous industry companies, in association with academic research actors, have been developing efficient technologies responding to a large number of

commercial applications in the area of digital audio and video coding. Standards such as MPEG-1 for PC and CD-ROMs storage (inc. mp3), MPEG-2 for broadcast radio/TV and DVD storage, MPEG-4 (part 2) for multimedia distribution, H.263 for videophone and teleconferencing, H.264/AVC joint ITU/ISO VCEG MPEG standard (Wiegand, Sullivan, Bjøntegaar, & Luthra, July 2003) for HDTV broadcast and Blu-ray Discs, and its extensions MVC (Smolic A. , Introduction to Multiview Video Coding, Jan 2008) for stereo 3DTV content or SVC (Kouadio, Clare, Noblet, & Bottreau, 2008) for adaptive delivery of content, have been issued.

Today new initiatives have been initiated to develop new standards: MPEG and ITU-T are launching a new project named High Efficiency Video Coding (HVC (Sub-groups, Feb. 2009)) aiming at doubling the compression rate in comparison with H.264/AVC for high-to-very-high resolution. Other projects related to advanced 3D video coding or 3D graphics coding are also under consideration in the same organizations. For audio signals, MPEG has launched an effort to standardize high-efficiency Unified Speech and Audio Coding (USAC). Those initiatives are discussed in the remainder of this document.

In parallel, the field of digital still images encoding has seen the emergence of many standards such as JPEG (Pennebaker & Mitchell, 1993) and JPEG2000 (Taubman & Marcellin, 2002), mainly for satellite images, medical images, or photography applications. The JPEG family has been extended for professional video applications to take care of high quality content storage and archiving, and for cinema applications as defined by the Digital Cinema Initiative (DCI).

Those standards have driven the market up to the 2000's, but more recently some competition from proprietary or open source solutions has appeared:

- Microsoft HD-Photo, standardized as JPEG-XR (Srinivasan, Tu, Regunathan, & Sullivan, Sept. 2007)
- Microsoft WM-9, standardized as SMPTE/VC-1 (VC-1, 2006)
- BBC Dirac, standardized as SMPTE/VC-2 (VC-2, 2009)
- AVID DNxHD, standardized as SMPTE/VC-3 (VC-3, 2008)
- AVS, Audio and Video Coding Standard Workgroup of China
- DivX (originally based on MPEG-4 part2)
- Ogg Theora and Vorbis (respectively Video and Audio open source from the Xiph.Org Foundation)

Further competition comes from Adobe Flash (based on On2 VP6 technology) for internet applications, or Google that turns On2's VP8 coding solution into an open source project

(Google bought On2 in February 2010).

Similarly, a number of proprietary audio codecs have been very successful, in particular in providing multi-channel audio signals in association with video streams, e.g., Dolby Pro Logic, Dolby Digital and DTS.

As shown by this list, the compression domain has been, up to now, mainly driven by standards initiatives. Even the proprietary attempts ended to a standardization body to have a chance of being used (see Microsoft WM-9 solution). ***But the success of a standard depends on many factors, on one hand to its efficiency and on the other hand on its ability to fulfill a market need at the time the standard is developed or produced.*** For instance MPEG-4 (part 2) was not a success because compression gain was around 30% over MPEG-2 and no new application was identified. On the opposite H.264/AVC offered 50% gain over MPEG-2 at the time HDTV was starting with a significant additional bandwidth requirement.

State of the Art - Compression Technologies

During the 90's, research laboratories had been extremely active with creating new technologies for alternative compression technologies. Studies on subband (Taubman & Marcellin, 2002), wavelet-based (Antonini, Barlaud,

Mathieu, & Daubechies, 1992), fractal-based (Wohlberg & De Jager, Dec. 1999), region-based video coding (Cicconi & Nicolas, Jun. 1994), vector quantization (Xu & Kuh, Jul. 1996), perceptual (Jayant, Johnston, & Safranek, Oct. 1993) and model-based audio coding, neural networks (Romaniuk, 1994), and other more recent Wyner-Ziv (Aaron, Setton, & Girod, 2003) coding have been led and are still on going to further improve existing schemes.

The most recent research efforts are focusing on distributed coding (Girod, Aaron, Rane, & Rebollo-Monedero, Jan. 2005), image analysis/synthesis solutions (Ndjiki-Nya, Tobias, & Wiegand, Jul. 2007), long-term motion analysis/representation (Wiegand, Zhang, & Girod, Long-Term Memory Motion-Compensated Prediction, Feb. 1999), depth cue extraction and use (Smolic & Kauff, Jan. 2005), sparse representations (being based on template matching, compressive sensing (Donoho, Apr. 2006), adaptive dictionaries...) (Martin, a, Guillemot, & Thoreau, Sep. 2007), and texture factorizing (extraction of signatures and epitomes) (Wang, Wexle, Ofek, & Hoppe, 2008). More generally, the research on audio and video coding is concentrated on advanced A/V models for improving signal prediction capitalizing on (and cross-fertilizing) advances in domains such as computer vision, image and video analysis, audio-visual perception, and cognitive sciences. Among these research areas, the most promising topics

can be found among the following items:

- **Sparse Representations** (Martin, a, Guillemot, & Thoreau, Sep. 2007) Aiming at finding representations of a signal with a small number of components taken from an over complete dictionary of elementary functions. Pursuit algorithms (matching pursuit MP, orthogonal MP or basis pursuit) and anisotropic waveform representations (bandelets, oriented wavelets) are some examples. A lot of studies are driven on their use in video compression and image denoising.
- **Texture analysis and synthesis** (Ndjiki-Nya, Tobias, & Wiegand, Jul. 2007), where a given texture is approximated (from either parametric description such as the probability density function, or texture representative elements such as patch, epitome...), and sampled to generate similar texture samples (e.g., using Markov Random Fields).
- **Motion analysis and modeling** (Wiegand, Zhang, & Girod, Long-Term Memory Motion-Compensated Prediction, Feb. 1999), to describe the optical flow and/or objects movement in a scene. This field which has been explored for a long time, but that is generating a number of new or improved results, especially regarding long-term motion analysis (tracking, trajectories extraction). These recent

explorations enable to foresee new advances in video compression and scene modeling.

- **Video objects and Scene characterization** (Ebrahimi & Horne, 2000), relevant solutions have been proposed to describe a scene, such as binary trees, edge representation, or object/region segmentation. The concept of layers-based video scene description was proposed for the MPEG-4 part2 standard (Ebrahimi & Horne, 2000) and consists in considering the scene as a composition of multiple planar regions, also called layers. This 2D½ representation is an intermediate step towards explicit 3D scene description.
- **Modeling of audio signals** (Hermusa, Verhelstb, Lemmerlingc, Wambacqa, & Huffelc, Jan. 2005), leading to parametric signal representations that, in turn, can be used to identify and remove redundancy and/or irrelevancy in the compression stage. A simple technique is to model suitable parts of an audio signal by synthetic noise that is shaped in time and frequency in order to be perceptually indistinguishable from the original sound. More sophisticated approaches use sets of sinusoids, models of sound generation, or models of spatial distribution of sound events in multi-channel signals.

- **Rendering and recording of sound fields** (Theile, Oct. 2004), with the aim to reproduce complex spatially distributed sound scenes with a superior quality and immersion as compared to today's multi-channel surround sound. The recording side involves medium-to-large microphone arrays and post processing algorithms. The rendering of complex sound fields requires today many loudspeakers, e.g., with wave-field synthesis in a continuous arrangement all around the listening area. Current research efforts are targeted at reducing the number of required microphones and loudspeakers to a reasonable value.

New video applications emerged in parallel, such as multi-view video, which bring new applications and open the doors to new research. These new applications also bring technical challenges both on the coding format and on the compression of this huge volume of data. Research today explores various format proposals based on some 2D views and additional depth and occlusion information, such as MVD (Multi-View plus Depth), LDV (Layer Depth Video), and some derivation of them, or schemes based on explicit 3D models:

- In the MVD (Smolic, Müller, Dix, Merkle, Kauff, & Wiegand, Oct. 2008) approach, each view is transmitted with its associated depth map. This

format allows obtaining easily the stereoscopic components by discarding the depth information. If the end-user wants to increase/reduce the 3D effect, the depth information is used for synthesizing intermediate views. Even if traditional video compression scheme (from the 2D world) can be used to encode the views and the depth, some opportunities exist for joint texture/depth coding.

- In the LDV (Shade, Gortler, He, & Szeliski, Jul. 1998) approach, only one view is transmitted, with associated depth map, occlusion and occlusion depth map information. With this format, displaying the 3D content on a stereoscopic display requires an intermediate view to be synthesized. The occlusion information improves the rendered quality by filling the holes. Additionally, a second view may be added to the LDV format (LDV-R), which makes the display on basic stereoscopic devices easier
- Some alternatives with more than one or two views are also proposed to cope with autostereoscopic or holographic screens where the number of views will be greater than 20.

Finally, the ultimate goal for encoding 3D video content is the use of 3D models (Balter, Gioia, & Morin, Dec. 2006) for encoding multiple views, for instance, with geometric models (polygon mesh) and texture mapping techniques. These techniques

are addressed in the academic field, but not yet in the relevant standardization groups. There are a number of issues to solve (before comparison to LDV and MVD solutions is even possible), one of them being the 3D model generation from natural views.

The A/V compression area has been mainly driven by standards initiatives and new application requirements. It has also been an important research topic for scientific innovations with involvement from many academics and industrial research laboratories during three decades.

Recent trends are showing promising opportunities both in terms of technologies and applications, just to mention 3DTV and CGI for instance. But, new alternatives to those standards are coming, especially from the open source community.

Research Challenges

Ten different challenges regarding content coding and compression have been identified and are described in this section. They could be classified according to the following three main issues:

- Support the future audio and video formats with better (3D, resolution, frame-rate, colours, HDR) and richer pictures and sound (multimodal, meta-data, mixed natural and synthetic, interactivity)
- Manage the heterogeneity of content, network and devices in both production and delivery (CDN, Cloud, OTT, P2P)
- Increase compression efficiency to reduce bandwidth requirements, cost, and delay while increasing robustness especially for those future content types

Future Image and Video Formats

Next-generation video formats will probably include more than 2D HDTV or even 3D stereo HDTV to provide the user with better and richer pictures for home and cinema applications. There are already studies on high-resolution images or Ultra-HD (Masayuki, 2008), going even beyond the digital cinema with

8k resolution and high frame-rate (120 fps being already available in display technologies). Multi-view capture for better 3D rendering (and later 3D models) is also discussed (Matusik & Pfister, 2004), especially about the number of views required for acceptable 3D viewing comfort (32 to 64 views might be necessary). We should also add that multi-views and high resolution can be used for panoramic capture (Sun, Foote, Kimber, & Manjunath, 2001).

In addition, the video signal itself is studied to enhance colors, brightness, and luminance through the high dynamic range, wide gamut, and floating-point representations for the video data. All of them with the same goal, i.e., provide images close to reality and to human perception capabilities (viewing angle, accuracy, light, colors...). There are clear opportunities to define new video formats, new file formats, and new coding schemes to cope with those signals.

If Computer-Generated Images (CGI) already integrate these properties and can relatively easily be extended to provide all the above characteristics, we can anticipate similar trend for video (with a longer time frame probably). Therefore, a lot of specifications will be necessary to define the appropriate coding

formats. The different video processes (production, delivery, rendering) will have also to evolve to cope with these new coding formats.

Since synthetic content production is growing each year, and is getting closer to natural images, we expect that more and more graphics, synthetic, and mixed natural/synthetic content will be produced (it is already very popular in movies special effects). The impact on compression must be studied and understood. We cannot accept a loss of quality. At the same time, the production of synthetic images provides much more information than natural capture of a scene, which can be used to design specific and optimized encoding schemes.

Thus there is a clear trend towards enhancing the user experience with rich audio and video content driven by the computer graphics technologies with nearly no limits on the images quality, realism, or artistic rendering and the possible special effects.

Future Audio Formats

A very recent trend in cinematography is stereoscopic 3D presentation. While today's 5.1 surround sound is adequate for 2D home theatre screens, there are questions about whether or not it is enough for

3D cinema and other immersive video formats. We are convinced that high-resolution, truly 3-dimensional surround sound will be required in the midterm in order to match the experience provided by the 3D video. The challenge lies in providing sound objects that are played back in any specific direction (incl. elevation) relative to the listeners. This involves the whole chain from production / post production, potentially even recording, through sound formats and content transmission. Ideally, a future sound format should be able to address various loudspeaker setups that range from conventional stereo to sophisticated wave-field synthesis and 3D loudspeaker setups.

Multimodal Content

The previous sections are about signal fidelity; this one is about information on the content from different sources. It is clear that video processing will benefit from various and different information from the content. As an example the depth of the objects is a significant element, opening room for improved video processing tools. Saliency map or region of interest is another set of useful information. Similarly, alternative shots or multi-camera capture of a scene provide a set of view that can be used in different ways, one of them being to use redundant information as a way to make processes more robust. Such multimodal information offers an opportunity for new production paradigms but requires the

specification of new coding formats.

In a different direction, multimodality can be expressed with meta-data associated to the content. Since meta-data is a generic word that covers too many things, it needs to be detailed. In our case we are talking about production meta-data that describe the content in order to improve subsequent video processing elements (including compression). Depth information and segmentation maps are example of such meta-data. The paragraph on video models and representation below discusses in further details how to encode this meta-data, and how it can be used to improve A/V compression.

Video Production

This is a special case for compression, because the requirements are specific and differ significantly from the distribution area. One of the main challenges of production workflows is the management of the plethora of content formats, being baseband or compressed. It would be therefore desirable to develop a generic coding format, dedicated to production, that is able to support a wide range of video formats and bit-rates (from lossless to lossy), supporting also frame editing, multi-generation, or low delay constraints at an affordable price. The existing JP2K and H.264 standards are a good basis, but none of them is able to fulfill the full range of downstream requirements. However, their respective

scalable and prediction technologies can be re-used to build a new scheme.

Such a new framework would also need to support and manage “production” meta-data, within an appropriate container format (or file format). The container should include a coding format according to the description above, the associated meta-data, and be flexible enough to support the heterogeneity issue described thereafter.

Dealing with Heterogeneity and Adaptation

As a consequence of previous topics, i.e., the plethora of A/V formats and content types, we will face the problem of rendering and adapting the content to different displays, users, networks, and devices. It means that there is a need for a mechanism (e.g., meta-data, scalable compression, transcoding, and/or storage) to support various video formats, various networks, various rendering devices, as well as any possible combination. This mechanism should be as extendable as possible in the future to address new requirements and new formats. The next battle for compression may not be on the compression ratio itself as traditionally, but rather on the capability to address this heterogeneity and compatibility issue. A combination of all techniques (content adaptation, transcoding, repurposing, and interpolation) will probably need

to be deployed to support the heterogeneity and diversity of the environment, which also includes all aspects of resolution, frame rate, tone mapping, color mapping, 2D to 3D conversion, synthetic to natural or vice-versa, zooming, cropping, aspect ratio, and probably some more.

A similar challenge arises for playback of multi-channel audio signals. After stereo and more lately 5-channel surround sound have been dominant for decades, a number of more sophisticated loudspeaker setups have been appearing recently, ending up in formats with 7.1, 10.2, 22.2 specific channels or even in wave-field synthesis with its hundreds of loudspeakers. Moreover, in consumer environments, loudspeakers are rarely positioned according to a standardized setup (even for stereo and surround), but rather where room geometry and furniture allows placing loudspeakers. A flexible coding scheme is required which allows addressing of this heterogeneity of real-world loudspeaker setups with best possible quality.

Compression and Bandwidth

Traditionally when discussing compression issues, there is a debate recurring after each new standard: *is it still necessary to further compress the data regarding the network capacity evolution?* And the answer is always the same: “yes”, since content bit-rates

and content consumption volumes are increasing at least as fast as the network capacities. In the last 20 years, the raw content bit-rates moved from 166Mbps to 830Mbps for HDTV, to 1.7Gbps for 3DTV, and we are discussing 18Gbps raw data-rates for some PC applications, with perspectives up to 72Gbps. We can also argue that when UGC will increase the quality of the encoded content, and networks will support UGC distribution with reasonable bit-rate, data traffic will drastically increase. However, we can certainly discuss if it means that the next challenge is only on the compression core or if more “intelligence” is required on the way the data is coded and transmitted, i.e., the representation models, the associated meta-data, or adaptation mechanisms.

A similar discussion (for audio, video, and still images) comes up regularly: *are we at the end of capabilities to compress even more efficiently?* The answer has been regularly “no”, because each time, someone came up with a brilliant new scheme that broke the apparent barrier. However, it should be mentioned that either we are reaching the limits of traditional compression schemes, or the increase in complexity of such compression schemes will question the interest in moving from the recent (and costly) H.264 technology to another one. Nevertheless, some room still exists for:

- New A/V formats where some special coding tools could help improving compression (e.g., using meta-data information, extension of existing tools, additional tools). These formats could involve still images, audio and video signals.
- Joint audio/video compression. Historically both domains have been studied separately, but there exists some commonality on the signal processing that can be used to create a single (joint) A/V data representation of the content.
- Advanced audio and video models improving signal prediction for higher compression efficiency
- Technologies from different areas (image synthesis, computer vision, computer graphics, etc.)
- Some special applications with dedicated requirements. For instance some applications could accept a similar content but not exactly the original one, it is what we call the similarity versus difference paradigm.

The community is working towards new paradigms and new approaches to the problem (from a different angle/perspective) that will initiate a series of new opportunities for research and market developments. The reader should refer to the state-of-the-art section above for a description of the promising technologies currently under investigation: distributed coding,

texture analysis/synthesis solutions, long-term motion analysis/representation, depth cue extraction and use, sparse representations (template matching, compressive sensing, adaptive dictionaries...), texture factorizing (extraction of signatures and epitomes), and modeling of A/V signals (as described below).

Video Models and Representation

This is in our view the main challenge for the coming years. As already said, we are probably reaching the limits of traditional compression schemes. The only way to go further is to develop advanced models for the audio and video signals (including 2D and 3D). Similar discussions are already on going in the standardization bodies for multi-view content, but are still limited to 2D plus depth (eventually with additional occlusions). We believe that more complex models are required such as advanced 2D models (scene models, objects) or even real 3D models for multi-view video or graphic contents.

Associated with the modelization is the description of the content with different semantic levels, starting from the low-level signal information (colors, energy, spectral representation), up to high-level information (scene, context, cognition). This includes information available from the previous 2D or 3D models and the specification of the necessary meta-data that will transport this information. This information is then used to

optimize existing coding schemes or to develop new ones.

Besides the compression, we also foresee a large number of new applications of such representations: information search, production/post-production, image processing, and augmented reality applications. For example, video descriptors generated in a mobile phone can be sent to the cloud for analysis and search of similar content, or 3D or 2.5D models of a scene would help integration of graphics or synthetic scenes in a 2D natural video scene.

Complexity

It is important to take into account the complexity of our algorithms and the evolution of the platform architectures. It is worth mentioning for instance the GPU and multi-core platforms. All the recent activities around cloud computing may also have an impact on compression technologies.

Delivery and Network Architectures

Delivery of content is constrained by two aspects. The first one is the end device: it has often limited processing or memory capability (see the paragraph above), and the second one is the network: it is unstable, with packet losses and jitter (at least for IP networks). To cope with the network, different delivery mechanisms can be developed; the most efficient consider jointly the

coding of the content and the network architectures. Adaptive streaming for instance uses different versions of a stream to dynamically select the coding stream suited to the available network bandwidth. Scalable coding is an alternative that eases content management.

When the network behavior (error rate, congestion, delay, etc.) is unstable, or simply variable, some mechanisms are thus required to follow the network variations. Some coding schemes may be more appropriate than others, and some specific technologies could be developed to adapt the stream. An alternative solution is to provide network level mechanisms to cope with network problems (i.e., retransmission or FEC). Similarly, different content delivery architectures (i.e., P2P, OTT, CDN, multicast) will impact the way the content streams are generated and managed. Therefore, understanding the network is essential for the delivery of compressed content, but also to specify efficient coding schemes.

Today there are two topics requiring particular attention: Over-The-Top (OTT) and cloud based delivery. OTT opens the door to new ways to directly deliver the content to the home user, but it is more sensitive to network quality as unmanaged service. Coding schemes that are more robust to network conditions have to be designed; distributed coding for example could be an effective solution. Cloud computing is offering new

ways to distribute content. A clear example is the recent attempt to perform video games using low complexity receivers and high complexity servers in the cloud (OnLive, Games@Large, G-Cluster). This requires the development of coding schemes supporting low delay, low latency, graphics encoding, high quality, low encoding complexity in order to be able to scale-up such services (regarding the number of users).

Video communication has the same kind of latency and robustness requirements, with an additional issue about the complexity on the receiver/transmitter side since it is a bidirectional application. Video communication is mentioned because it could become increasingly important for consumer devices and applications (including STBs and gaming).

Human Perception

We consider the human being the final “sensor” when consuming content. Therefore it is important to retrieve his/her feedback, understand how he/she perceives the content, and how he/she reacts in front of different contents. This is particularly true for 3DTV where the depth perception is not yet completely understood. But even in 2D we still miss a lot of information on the user, where emotion vs. quality of content is difficult to predict. By having this knowledge we expect to be able to create content coding and compression schemes that

maximize the user satisfaction (immersion). One approach is to establish quality of experience models using the well known Mean Opinion Score (MOS) technique, or develop new models, based on experiments with real users in different environments (i.e., home, theaters).

Another dimension of the human perception is the capability to measure the perceived audio/video/image quality. Compression is based on the minimization of a rate-distortion criteria (often called RDO Rate Distortion Optimization), i.e., finding the best trade-off between signal degradation and bandwidth. It needs therefore a criterion to measure this distortion. Today the traditional Mean Square Error (MSE) is predominantly used, but more subjective metrics (i.e., metrics representing more faithfully the end-user’s perceived quality) would help optimize the subjective picture and sound quality.

Similarly monitoring the perceived quality at the user side is a big challenge. This topic often refers to Quality of Experience (QoE) since it includes more than just signal fidelity. Network operators are looking for equipment to provide their customers with a better service. A feedback signal (to the emitter side) can also be used to optimize the transmitted quality of the compressed content stream, or to adapt network level mechanisms.

Subjective evaluation of the video quality with objective metrics is still a research area and no fully satisfactory solutions exist, besides particular restricted cases. In addition, the coming future video formats add other dimensions to this complex issue (depth, colors, etc.). In this context, the work performed by the ITU-T/VQEG group (Video Quality Expert Group) is particularly relevant.

Many technical challenges regarding content coding and compression exist today. Support future audio and video formats with better and richer pictures, management of the heterogeneity of content, network and devices, increased compression efficiency to reduce bandwidth requirements and cost, robustness, or delay, are some of them. All these requirements make it necessary to still invest on new coding schemes. However, the main challenges may not be in the compression core but rather on the coding formats and associated functionalities.

Standards Opportunities and Open-source Competition

In this paper we have highlighted the importance of the standards for the compression area. Standardization is very important to make sure that services and products are conformant and interoperable whatever the product provider or the final user equipments is. The risks of non-standard codecs are the licensing costs (standards rules are clear and defined), the interoperability between different providers that could not be guaranteed, or the need for regular new install or update (in a PC model) and the associated cost for a manufacturer. Just to mention the 2D video codecs, the alternatives to standards are the use of proprietary or open-source codecs such as Microsoft JPEG-XR and VC-1, BBC Dirac, AVID DNxHD, AVS, DivX, ogg theora (open source), and other codecs from Adobe Flash (based on On2 technology). A similar list could be proposed for 3D stereo coding formats (Sensio, Dolby, RealD) and audio coding (Dolby, ogg vorbis).

We have identified several standards initiatives that may impact the next generation of A/V coding formats. The open source announcements and work are also discussed since they are changing the current landscape.

A/V Standards

Regarding video compression there are a significant number of opportunities to push existing standards to the market:

- DVB, SMPTE, Blu-ray and 3GPP bodies discussing adoption of H.264/AVC MVC or SVC extensions for 3DTV or adaptive streaming applications
- WHDI, WiMedia, WiGig, or WiFi Display discussing adoption of the H.264/AVC standard and in particular the professional 4:4:4 profile, and even in the future the MVC and SVC extensions for wireless communication between a device and a display

And looking further, the ISO/IEC MPEG group is currently working on three directions for developing new video and CGI coding standards:

- HVC, next-generation video coding, also wrongly known as H.265, within the Joint Collaborative Team between ISO/IEC MPEG and ITU-T VCEG (JCT-VC). A call was issued in January 2010. The review of the answers was just issued in April 2010, for a standard expected in 2012.
- 3DV is a MPEG ad-hoc group initiative to develop a coding scheme addressing 3DTV multi-view and stereoscopic 3DTV

with baseline adjustment. They are currently trying to achieve good-quality view interpolation and depth estimation as a required preliminary step. The call for proposal is planned for October 2010

- 3DGC is a continuing work targeting coding and compression of 3D graphics data and models. Some specifications have already been published, but the work is still going on for Scalable Complexity 3D mesh compression (SC3DMC), Reconfigurable Graphics Codec (RGC), and IndexedFaceSet (IRS). For the future some ideas are being discussed such as client adaptive streaming of gigantic 3D data and semantic based representation of 3D models and 3D animation.

Regarding audio compression, the MPEG Audio sub-group is currently focusing on the “Unified Speech and Audio Codec” (USAC) which combines concepts from speech codecs and generic audio codecs. The new codec shall provide consistent quality whatever kind of input signal is being processed.

These groups have been the most active and efficient in the past, with the highest impact on the market. They are also considered as the most

important bodies attracting nearly all the main industrial and academics research actors of the audio and video coding field. Their work is thus particularly relevant and needs to be followed by those interested in A/V content encoding and transmission and their applications.

Open Source Initiatives

On the other side, the open source community is pushing for the development of its own standards without any royalties. This question of patent owners, licensing fees, and royalties is the key issue. It leads to the development of open source solutions, but it is also the main reason why they are not widely used today.

Two main topics are particularly relevant to mention and illustrate clearly this issue: HTML5 and the Google codec.

HTML 5 may carry the future of online video, but currently it's bogged down in a debate related to licensing, between supporters of H.264 (standard) and advocates of Ogg Theora (open source). Following Apple, Microsoft recently announced their support to HTML5. Some people are claiming that a video codec cannot be royalty free, because recent coding schemes are all necessary using state-of-the-art coding schemes with still existing IP. The Open source community is arguing that the specifications have been published together with a call for IP declaration and that no

one answers it. This, however, does not mean that there may not be existing IP.

The purchase of On2 by Google, which was finalized earlier this year, is part of a larger campaign by Google to support open source, and therefore royalty free standards in HTML5. The strong position of Google in the internet community makes it a key indicator that royalty free A/V codec may be necessary for internet applications.

Similarly, some companies (in particular Sun Microsystems) supported by academics (Peking University) and the China National Body, have proposed also to start within MPEG a royalty-free standard. It could be an answer of the standardization bodies to proprietary and open-source initiatives.

Standards have been the main drivers of the coding schemes used today. They provided efficient solutions with some guarantees regarding licensing conditions and interoperability issues. They are preparing the next generation of standards for new content types and increased efficiency, providing opportunities for scientific research and competition. At the same time some new models will most probably emerge in the future, such as the open-source initiatives or royalty free technologies.

Conclusions

The A/V coding domain is very challenging because of increasing competition from academic and industrial actors, as well as the open-source community. However, many strong opportunities are in front of us: future audio/video formats (HDR & WG, Ultra-HD, 3D multi-view, mixed natural-synthetic, 3DGI), future standards (HVC, 3DV, 3DGC, spatial audio) and new network delivery architectures (cloud, OTT). More opportunities have also been identified in the area of video processing related to compression: A/V advanced models and descriptors, augmented video, interactive content and video communication, human perception and cognition, multimodal content processing, and cloud-based applications and content. Adjacent technology areas such as the computer vision and computer graphics sciences are also particularly relevant for these activities.

All those different aspects for the short, mid, and long term are based on our feeling that the future of video processing and

coding should consider the following evolutions:

- Move from 2D to 3D workflows: advanced 2D scene models are necessary for the short term, but evolutions to 3D models (inc. natural scenes) and in a longer term to volumetric models, sound fields, light fields/holography should be considered because of their impacts on the content production first, but also on delivery format rendering.
- Convergence towards natural and synthetic content: the frontier between those two types of content will gradually disappear, thus opening the door to the augmented video area and innovations in the field of video processing (with increased user experience and the development of new services).
- Media delivery agnostic to transport and terminals: it means that the coding format and all the delivery chain should be able to auto adapt to the different networks, devices, and users. It would provide content delivery with

opportunities for new services within the home environment.

- Natural and simple human interaction with rich and interactive content at home and theaters. Even if many progresses have been announced for years, there are room for innovations in the production and consumption of new types of content, more immersive and more interactive, but also on the way the users interact with the content.

Thus the A/V/CGI compression domain is at a transition phase from existing and widely used MPEG-2, MP3, and MPEG-4 standards towards the next generation. It is unclear whether traditional standardization bodies will succeed in developing a successful new coding scheme or if an open-source solution will be adopted. However, as described in this paper, there are numerous challenges and consequently numerous opportunities in front of us. It makes this research area particularly exciting, with both competition and potential for innovations.

Bibliography

1. Aaron, A., Setton, E., & Girod, B. (2003). Towards practical Wyner-Ziv coding of video. : ICIP03.
2. Antonini, M., Barlaud, M., Mathieu, P., & Daubechies, I. (1992). Image Coding Using Wavelet Transform. *IEEE Image Processing* , 1 (N° 2, pp.205-220).
3. Balter, R., Gioia, P., & Morin, L. (Dec. 2006). Scalable and Efficient Video Coding Using 3-D Modeling. *IEEE Transactions on Multimedia* , 8 (N° 6, pp. 1147-1155).
4. Cicconi, P., & Nicolas, H. (Jun. 1994). Efficient Region-Based Motion Estimation and Symmetry Oriented Segmentation for Image Sequence Coding. *IEEE Circuits and Systems for Video* , 4 (N° 3, pp. 357-364).
5. Diot, C., & Donnan, G. (17 June 2009). The Thomson Corporate Research Hand-book, Vision, Opportunities, Strategy.
6. Donoho, D. L. (Apr. 2006). Compressed Sensing,, V. 52(4), 1289-1306, 2006. *IEEE Transactions on Information Theory* , 52 (N° 4, pp. 1289 - 1306).
7. Ebrahimi, T., & Horne, C. (2000). MPEG-4 natural video coding - An overview. *Signal Processing: Image Communication* , 15 (pp. 365-385).
8. Francois, E. e. (2010). Spatio-temporal Video Modeling and Applications to Video production and post-production. Modellm@ges ANR proposal.
9. Girod, B., Aaron, A. M., Rane, S., & Rebollo-Monedero, D. (Jan. 2005). Distributed Video Coding. *Proceedings of the IEEE - Special issue on advances in video coding and delivery* , 93 (N° 1, pp 71-83).
10. Gomila, C. a. (17 March 2010). *3D Solutions for Broadcast*.
11. Gomilla, C. (April 2010). *MPEG CfP report*. : Technicolor R&I IMX Steering Committee.
12. Guillotel, P. (June 2009). *A/V Compression Strategy*. Paris: CR Meeting.
13. Hermusa, K., Verhelstb, W., Lemmerlingc, P., Wambacqa, P., & Huffelc, S. V. (Jan. 2005). Perceptual audio modeling with exponentially damped sinusoids. *Signal Processing* , 85 (N° 1, pp.163-176).
14. Jayant, N., Johnston, J., & Safranek, R. (Oct. 1993). Signal Compression Based on Models of Human Perception. *Proceedings of the IEEE* , 81 (N° 10, pp. 1385-1422).
15. Kouadio, A., Clare, M., Noblet, L., & Bottreau, V. (2008). SVC - a highly-scalable version of H.264/AVC. *EBU Technical Review* .
16. Martin, A., a, J.-J. F., Guillemot, n. C., & Thoreau, D. (Sep. 2007). Sparse Representation for Image Prediction. Poznan, Poland: EUSIPCO, 15th European Signal Processing Conference.
17. Mary-Luc, C., Pascal, G., Francois, G., Philippe, G., Francois, L. C., Yves, M., et al. (2010). *Games on Server*. : Technicolor Internal Report.

18. Masayuki, S. (2008). Super Hi-Vision - Research on a future ultra-HDTV system.
19. Matusik, W., & Pfister, H. (2004). 3D TV: a scalable system for real-time acquisition, transmission, and autostereoscopic display of dynamic scenes. Los Angeles, California, USA: International Conference on Computer Graphics and Interactive Techniques.
20. Ndjiki-Nya, P., T. H., & Wiegand, T. (Jul. 2007). Generic and Robust Video Coding with Texture Analysis and Synthesis. Beijing, China: Proceedings of IEEE International Conference on Multimedia and Expo.
21. Pennebaker, W. B., & Mitchell, J. L. (1993). *JPEG still image data compression standard (3rd ed.)*. Springer.
22. Romaniuk, S. (1994). Applying constructed neural networks to lossless image compression. : ICIP94.
23. Shade, J., Gortler, S., He, L.-w., & Szeliski, R. (Jul. 1998). Layered depth images. Orlando, Florida, USA: Proceedings of the 25th annual conference on Computer graphics and interactive techniques (SIGGRAPH'98).
24. Smolic, A. (Jan 2008). *Introduction to Multiview Video Coding*. Antalya, Turkey: ISO/IEC JTC 1/SC 29/WG 11.
25. Smolic, A., & Kauff, P. (Jan. 2005). Interactive 3-D Video Representation and Coding Technologies. *Proceedings of the IEEE*, 93 (N° 1, pp. 98-110).
26. Smolic, A., Müller, K., Dix, K., Merkle, P., Kauff, P., & Wiegand, T. (Oct. 2008). Intermediate View Interpolation Based On Multiview Video Plus Depth For Advanced 3d Video Systems. San Diego, California, U.S.A.: International Conference on Image Processing (ICIP'08).
27. Srinivasan, S., Tu, C., Regunathan, S. L., & Sullivan, G. J. (Sept. 2007). HD Photo: A New Image Coding Technology for Digital Photography”,. : SPIE Applications of Digital Image Processing XXX.
28. Sub-groups, M. V. (Feb. 2009). *Vision and Requirements for High-Performance Video Coding (HVC)*. Lausanne, Switzerland: ISO/IEC JTC1/SC29/WG11.
29. Sun, X., Foote, J., Kimber, D., & Manjunath, B. S. (2001). Panoramic video capturing and compressed domain virtual camera control. Ottawa, Canada : Proceedings of the ninth ACM international conference on Multimedia.
30. Taubman, D., & Marcellin, M. (2002). *JPEG2000 Image Compression Fundamentals - Standards and Practice*. : The International Series in Engineering and Computer Science.
31. Taubman, D., & Zakhor, A. (Jul. 1994). Orientation adaptive subband coding of images. *IEEE Image Processing*, 3 (N° 4, pp.421-437).
32. Theile, G. (Oct. 2004). Wave Field Synthesis - A Promising Spatial Audio Rendering Concept. Naples, Italy: Proceedings of the 7th International Conference on Digital Audio Effects (DAFx'04).
33. VC-1, S. (2006). Television - VC-1 Compressed Video.
34. VC-2, S. (2009). VC-2 Video Compression.
35. VC-3, S. (2008). VC-3 Picture Compression and Data.

36. Wang, H., Wexle, Y., Ofek, r. E., & Hoppe, H. (2008). Factoring Repeated Content Within and Among Images. Los Angeles, CA, USA: Proceedings ACM Transactions on Graphics (SIGGRAPH'08), 27(3).
37. Wiegand, T., Sullivan, G. J., Bjøntegaar, G., & Luthra, A. (July 2003). Overview of the H.264/AVC Video Coding Standard. *IEEE Transactions on Circuits and Systems for Video Technology* , 13 (N° 7).
38. Wiegand, T., Zhang, X., & Girod, B. (Feb. 1999). Long-Term Memory Motion-Compensated Prediction. *IEEE Transactions on Circuits and Systems for Video Technology* , 9 (N° 1, pp. 70-84).
39. Wohlberg, B., & De Jager, G. (Dec. 1999). A Review of the Fractal Image Coding Literature. *IEEE Image Processing* , 8 (N° 12, pp.1716-1729).
40. Xu, M., & Kuh, A. (Jul. 1996). Image-Coding Using Feature Map Finite-State Vector Quantization. *Signal Processing Letters* , 3 (N° 7, pp. 215-217).

Glossary

A/V: Audio/Video.

MPEG: The Motion Pictures Experts Group is a sub-group of ISO/IEC Joint Technical Committee 1, Subcommittee 29, Working Group 11 (ISO/IEC JTC 1/SC 29/WG 11) - titled as Coding of moving pictures and audio.

JPEG: The Joint Photographic Experts Group is a sub-group of ISO/IEC Joint Technical Committee 1, Subcommittee 29, Working Group 1 (ISO/IEC JTC 1/SC 29/WG 1) - titled as Coding of still pictures.

VCEG: The Video Coding Experts Group is the Question 6 (Visual coding) of Working Party 3 (Media coding) of Study Group 16 (Multimedia coding, systems and applications) of the ITU-T - titled as ITU-T Q.6/SG 16.

SMPTE: The Society of Motion Picture and Television Engineers, is an international professional association of engineers working in the motion imaging industries.

IEC: The International Electrotechnical Commission prepares and publishes International Standards for electrical and electronic technologies.

DVB: The Digital Video Broadcasting group developed international specifications for digital television published by a Joint Technical Committee (JTC) of European Telecommunications Standards Institute (ETSI), European Committee for Electrotechnical Standardization (CENELEC) and European Broadcasting Union (EBU).

ATSC: The Advanced Television Systems Committee (ATSC) developed the ATSC Standards for digital television in the United States, also adopted by Canada, Mexico, South Korea, and recently Honduras.

HDMI: The High-Definition Multimedia Interface is a compact audio/video interface for transmitting uncompressed digital data. It represents a digital alternative to consumer analog standards, such as radio frequency (RF) coaxial cable, composite video, S-Video, SCART, component video, D-Terminal, or VGA.

MVC: MultiView Coding, an extension of the MPEG-4 AVC/H.264 standard for 3DTV applications.

SVC: Scalable Video Coding, an extension of the MPEG-4 AVC/H.264 standard for scalable video coding.

HVC: High efficiency Video Coding, a new initiative within MPEG to develop a new standard for video compression.

JCT-VC: Joint Collaborative Team on Video Coding (JCT-VC) is a group of video coding experts from ITU-T Study Group 16 (VCEG) and ISO/IEC JTC 1/SC 29/WG 11 (MPEG). It was created in 2010 to develop a new generation video coding standard.

Ogg: Ogg is a free, open standard container format maintained by the Xiph.Org Foundation and distributed without licensing fees.

Theora: Theora is a free lossy video compression format. It is developed by the Xiph.Org Foundation and distributed without licensing fees.

- Vorbis:** Vorbis is a free lossy audio compression format. It is developed by the Xiph.Org Foundation and distributed without licensing fees.
- BD:** Blu-ray Disc is an optical disc storage medium designed to supersede the standard DVD format. Capacity is 25 GB per single layered, and 50 GB per dual layered disc.
- fps:** Frame per second or frame rate of a video.
- 4k, 8k:** Next generation video standard resolution after HDTV. 4K is a standard resolution with approximately 4,000 pixels (4096x2304 pixels with a 16:9 aspect ratio in computer graphics, or 4096 × 1714 with 2.39:1 aspect ratio for Digital cinema). 8K is the double.
- UHD:** Ultra High Definition (UHD), Ultra High Definition Video (UHDV), Ultra High Definition Television (UHDTV), Extreme Definition Video or 8K is an experimental digital video format, currently proposed by NHK of Japan and British Sky Broadcasting. It corresponds to a resolution of 7680 × 4320.
- HDR:** High Dynamic Range imaging is a set of techniques that allow a greater dynamic range of luminances between the lightest and darkest areas of an image. It allows HDR images to more accurately represent the wide range of intensity levels found in real scenes, ranging from direct sunlight to faint starlight.
- H.264:** H.264/AVC/MPEG-4 Part 10 (Advanced Video Coding) is a standard for video compression completed in May 2003. It succeeded to the MPEG-2 and MPEG-4 part 2 video coding standards.
- JP2K:** JPEG 2000 is a wavelet-based image compression standard and coding system published in 2000. It succeeded to the discrete cosine transform-based JPEG standard.
- MP3:** MPEG-1 Audio Layer 3, more commonly referred to as MP3, is a standardized, lossy compression and digital audio encoding format.
- AAC:** Advanced Audio Coding is a standardized, lossy compression and encoding scheme for digital audio. It succeeded to the MP3 format.
- STB:** A set-top box or Set-Top Unit (STU) is a device that connects to a television and an external source of signal, turning the signal into content which is then displayed on the television screen or other display device.
- CGI:** Computer-generated imagery is the application of the field of computer graphics or, more specifically, 3D computer graphics to special effects in films, television programs, commercials, simulators and simulation generally, and printed media.
- 3DCG:** 3D computer graphics are graphics that use a three-dimensional representation of geometric data for the purposes of performing calculations and rendering.
- 3DTV:** 3D television employing techniques of 3D presentation, such as stereoscopic capture, multi-view capture, or 2D plus depth, and a 3D display.
- IC:** An integrated circuit is a miniaturized electronic circuit manufactured in the surface of a thin substrate of semiconductor material.
- UGC:** User-generated content, refers to content, publicly available and produced by end-users.

- DPX:** Digital Picture Exchange is a common file format for digital intermediate and visual effects work. It is an ANSI/SMPTE standard (268M-2003).
- VGA:** Video Graphics Array refers to the 640×480 video resolution.
- QVGA:** Quarter VGA, thus the 320×240 video resolution.
- DCT:** The discrete cosine transform is a sum of cosine functions with different frequencies.
- Metadata:** Metadata is defined as data about data. It is mainly used to describe the content (inc. definition, structure, format, context, properties...).
- OpenGL:** The Open Graphics Library is a standard API for 2D and 3D computer graphics applications.
- CPU:** The Central Processing Unit is the processor that runs the instructions of a software program.
- GPU:** A Graphics Processing Unit is a specialized processor for 3D or 2D graphics rendering.
- NSP:** A network service provider is an organization that provides network access.
- QoE:** Quality of Experience is a subjective measure of a customer's experience. It differs from Quality of Service (QoS), which attempts to objectively measure the service delivered.
- FEC:** Forward Error Correction is a method to recover data lost during transmission. Redundant data are added to the original message (error-correction codes) to detect and correct errors.
- OTT:** Over the Top, describes third party services that are delivered on top of a broadband network without affiliation with the broadband service provider.
- CDN:** A Content Delivery Network is a network architecture where various servers containing copies of data, are placed at various points in a network so as to maximize client bandwidth.
- P2P:** A peer-to-peer network is a distributed network architecture composed of participants that make a portion of their resources (processing power, disk storage or network bandwidth) directly available to other network participants, without central coordination.
- Cloud:** Cloud computing is Internet-based computing, whereby common resources, software, services, information are shared and provided to the user on-demand.
- DLNA:** The Digital Living Network Alliance is a consumer electronics standard that allows home devices to share content with the others within a home network.
- HTML5:** HTML5 is the next major revision of HTML (HyperText Markup Language). It aims to reduce the need for proprietary plug-in-based rich internet application (RIA) technologies such as Adobe Flash, Microsoft Silverlight, and Sun JavaFX.
- MPEG-LA:** MPEG-LA is a firm that licenses patent pools covering essential patents required for use of the MPEG-2, MPEG-4 Visual (Part 2), IEEE 1394, VC-1, ATSC and AVC/H.264 standards. MPEG LA is based in Denver, Colorado, USA.
- MOS:** The Mean Opinion Score is a numerical indication of the perceived quality of media content. It ranges from 1 to 5, where 1 is the lowest and 5 the highest perceived quality.

- MSE: The Mean Square Error is the average of the square of the difference between an estimation and a true value of a signal (i.e., the error).
- RDO: Rate-Distortion Optimization refers to the optimization of the amount of distortion (loss of video quality) against the amount of data rate required to encode the video.

ABOUT TECHNICOLOR

With more than 95 years of experience in entertainment innovation, Technicolor serves an international base of entertainment, software, and gaming customers. The company is a leading provider of production, postproduction, and distribution services to content creators and distributors. Technicolor is one of the world's largest film processors; the largest independent manufacturer and distributor of DVDs (including Blu-ray Disc); and a leading global supplier of set-top boxes and gateways. The company also operates an Intellectual Property and Licensing business unit. For more information: www.technicolor.com .

Technical Contact Information

Any technical questions may be addressed to the following individuals and principal contributors.

- Philippe.Guillotel@technicolor.com
- Christophe.Diot@technicolor.com
- Peter.Jax@technicolor.com
- Marie-Jean.Colaitis@technicolor.com
- Edouard.Francois@technicolor.com
- Cristina.Gomila@technicolor.com
- Xavier.Bonjour@technicolor.com

