VERY LOW BITRATE VIDEO CODING AND MPEG-4: STILL A GOOD RELATION

Fernando Pereira

Instituto Superior Técnico - Instituto de Telecomunicações
Av. Rovisco Pais, 1096 Lisboa Codex, PORTUGAL
Telephone: + 351 1 8418460; Fax: + 351 1 8418472
E-mail: eferbper@beta.ist.utl.pt

ABSTRACT

MPEG-4 emerged recently as an important development in the field of audio-visual coding aiming at establishing the first content-based audio-visual coding standard. This paper intends to analyse the current relation between MPEG-4 and very low bitrate video coding and corresponding applications, notably by considering the MPEG-4 objectives, functionalities and recent technical developments related to video coding.

1 INTRODUCTION

In November 1992, a new work item proposal for very low bitrate audio-visual (AV) coding was presented in the context of ISO/IEC JTC1/SC29/WG11, well known as MPEG (Moving Pictures Experts Group). The scope of the new work item was described as "*the development of international standards for generic audio-visual coding systems at very low bitrates (up to tens of kilobits/second)*" [1]. The main motivations for the starting of the new work item were basically the prevision that the industry would need very low bitrate video coding algorithms in a few years, the recognition that mobile communications had became one of the most important telecommunication markets and researchers on AV communications had already identified some interesting mobile AV applications, the appearance in the market of videotelephones for the PSTN, and the fact that very low bitrates seemed to be the last bitrate range where no significant standardisation efforts had been made, thus justifying the growing interest of the AV coding research community. The applications being addressed ranged from mobile or PSTN videotelephony and multimedia electronic mail to remote sensing, electronic newspaper, interactive multimedia databases or games [1]. At the same time, the MPEG Ad-Hoc Group on Very-Low Bitrate Audio-Visual Coding, with participation of members from the CCITT Video Coding Experts Group, identified the need for two solutions for the very low bitrate AV coding problem [2]:

- A short term solution (2 years) primarily addressing videotelephone applications on PSTN, LANs and mobile networks (a CCITT H.261 like solution was expected).
- A far term solution (approximately 5 years) providing a generic AV coding system, with non annoying visual quality, using a set of new concepts which had meanwhile to be identified and developed.

Since the near term solution was mainly intended for communication applications, it was decided it should be handled by the CCITT (now ITU-T). The far term solution should produce a generic AV coding system and would be handled by ISO with liaison to CCITT [2].

In September 1993, the MPEG AOE (Applications and Operational Environments) group met for the first time. This group had the main task to identify the applications and requirements related to the far term very low bitrate solution to be developed in the context of ISO/MPEG. In this period, the short-term hybrid solution being developed within the ITU-T SG 15 LBC (Low Bitrate Coding) group started to produce results which were largely accepted as close to the saturation performance of DCT-based hybrid coding schemes.

The turning point of the MPEG-4 work happened in July 1994, at the Grimstad meeting, when members were faced with the need to broaden the objectives of MPEG-4 which could no more be based on a pure compression gain target coming from new coding approaches such as region-based, analysis-synthesis, fractals or any other, since very few people believed in a compression improvement sufficient to justify (alone) a new standard (beside the LBC, latter H.263, standard). The AOE group started then an in-depth analysis of the AV world trends and concluded that the emerging MPEG-4 AV coding standard should support new ways of communication, access and manipulation of digital AV information (notably content-based), offering a common AV solution to the various worlds converging in this interactive AV terminal, hopefully the future MPEG-4 terminal. MPEG-4 should take into account the new expectations and requirements coming from the convergence of the TV/film entertainment, computing and telecommunications worlds.

Since, more than MPEG-1&2, MPEG-4 is appearing in a period of quickly changing conditions, the new standard was considered to need a structure that can cope with the rapidly evolving relevant technologies, providing a high degree of flexibility, extensibility and thus time-resistant through the integration of new technological developments [3].

The clarification of the MPEG-4 objectives where the concepts of content, interaction and flexibility/extensibility are central, made the relation to any particular bitrate less significant. However, and although MPEG-4 is today said to address both "*fixed broadband and mobile narrowband*" delivery systems [4], lower bitrates still play a primary role in MPEG-4, considering

the applications that will very likely use for first the MPEG-4 standard, such as AV database access, remote monitoring and control, and AV communications and messaging

The current MPEG-4 focus clearly proposes a change in terms of the traditional AV representation architecture in order to provide a set of new AV functionalities closer to the way that users 'relate' to the real world AV information [5]. Since the human beings do not want to interact with abstract entities such as pixels, but rather with meaningful entities that are part of the scene, the concept of content is fundamental to MPEG-4. These content-based functionalities were not supported at all in the context of the available 'pixel-based' video coding standards.

This paper intends to analyse the current role of very low bitrate audio-visual (VLBAV) applications and very low bitrate video coding in the context of MPEG-4. Although MPEG-4 will address both video and audio, this paper will mainly concentrate on video.

2 VLBAV APPLICATIONS AND MPEG-4

The main target of MPEG-4 is thus to provide a new coding standard, supporting new ways of communication, access and manipulation of digital AV information. In this context, MPEG-4 does not want to address any specific application but rather prefers to support as many clusters of functionalities which may be useful for various applications as possible. This functionality-based strategy is best explained through the eight MPEG-4 'new or improved functionalities' - *content-based multimedia data access tools, content-based manipulation and bitstream editing, hybrid natural and synthetic data coding, improved temporal random access, improved coding efficiency, coding of multiple concurrent data streams, robustness in error-prone environments, and content-based scalability* [3]. The eight functionalities came from an assessment of the functionalities that will be useful in future applications, but are not or are not well supported by current coding standards. These functionalities are not all equally important, neither in terms of the technical advances they promise, nor the application possibilities they open. Moreover, they imply rather ambitious goals which will only be fully reached in due time (and provided that the necessary amount of work will be invested by the relevant experts). With the functionality-based approach, MPEG-4 found an identity that can supply an answer to the emerging needs of application fields ranging from interactive AV services, e.g. content-based AV database access, games or AV home editing, and advanced AV communication services, e.g. mobile AV terminals, improved PSTN AV communications or tele-shopping, to remote monitoring and control, e.g. field maintenance or security monitoring.

It is quite clear that many of the applications above mentioned will be provided with (critical) bitrate limitations due to the delivery channel (transmission or storage). In order to find the MPEG-4 feeling about VLBAV applications, we can start to analyse the relation between the MPEG-4 new or improved functionalities and this type of applications [6]. Although these functionalities were not designed keeping in mind any specific bitrate range, it is evident that their provision at very low bitrates will strongly stimulate VLBAV applications.

- **Improved coding efficiency** - This is clearly a functionality useful for VLBAV applications since improved coding efficiency is asked for, to supply an answer to the requests coming from, e.g. mobile network users.
- **Robustness in error-prone environments** - The access to AV applications through channels with severe error conditions, such as some mobile channels, requires sufficient error robustness is added. This functionality, by providing significant quality improvements, will with no doubts stimulate VLBAV applications, notably in mobile environments.
- **Content-based scalability** - The ability to achieve scalability with a fine granularity in content, spatial or temporal resolution, quality and complexity, or any combination of these cases, is a fundamental concept for VLBAV applications since it provides the capacity to adapt the AV representation to the available resources. This flexibility is even more important for low bitrate conditions since the adequate choice of the scalable AV information to transmit is for sure a critical decision which may determine a much higher subjective impact.
- **Improved temporal random access** - The provision of random access efficient methods, within a limited time and with fine resolution, including 'conventional' random access at very low bitrates is the target. No doubts about the clear relation of this functionality to VLBAV applications where limited resources make always random access modes a problem.
- **Content-based manipulation and bitstream editing** - The provision of content-based manipulation and bitstream editing refers to the capability that the user should have to select and manipulate one specific 'object' in the scene/bitstream. This functionality is more related to the syntactic organisation of the information than to any specific bitrate resources. Anyway it is acceptable to think that also some VLBAV applications will benefit from it.
- **Content-based multimedia data access tools** - The possibility to content-based selectively access AV data is for sure an important capability in the context of VLBAV applications since it will allow to optimise the AV information to transmit depending on the available resources.
- **Hybrid natural and synthetic data coding** - The harmonious integration of the natural and synthetic AV worlds is one of the most important MPEG-4 targets. Although this functionality is quite bitrate-independent, the efficient integration of natural and synthetic AV data will for sure make no harm to VLBAV applications.
- **Coding of multiple concurrent data streams** - The efficient coding of multiple concurrent data streams does not seem to specially address very low bitrates; however the provision in the future of services such as mobile virtual reality will also depend on additional developments in this area.

The analysis of the MPEG-4 new or improved functionalities makes quite clear that MPEG-4 will not be a standard addressing (only) very low bitrate AV coding. MPEG-4 will be much more than that. However, it is at the same time more than evident that VLBAV applications will be among the applications that will benefit more from the achievement of the MPEG-4 objectives. The recognition of this fact led again to the establishment of a formal liaison between MPEG and ITU-T, notably through the SG 15 LBC group [7]. The SG 15 LBC group is particularly committed to introduce in MPEG-4 the requirements coming from real-time AV conversational services.

3 TESTING FOR VLBV CODING

One of the new challenges put by MPEG-4 is the request to simultaneously address more than one target/functionality depending on the class of applications it is considered. The

complete specification of the evaluation methodologies for the new MPEG-4 functionalities was, and still is, a new challenge in the framework of standardisation since there was no significant experience for the type of tests needed. Taking into account the functionalities, three types of evaluation may be foreseen: i) conventional subjective tests; ii) task-based tests where the result of a task is evaluated, e.g. through the number of successes, and iii) evaluation by experts, using the description of the functionalities' implementation (when limitations prevent more direct evaluation).

At the MPEG-4 first round of tests, held in Los Angeles, in October 1995, the video bitrates tested ranged from 10 to 1024 kbit/s. The video test material was divided in 5 classes, 3 of which clearly addressing low or very low bitrates: class A - low spatial detail and low amount of movement - 10, 24 and 48 kbit/s; class B - medium spatial detail and low amount of movement or vice-versa - 24, 48 and 112 kbit/s and class E - hybrid natural and synthetic content - 48, 112 and 320 kbit/s. Only conventional subjective tests and evaluation by experts were used [9].

The need to keep the overall tests manageable led to the choice of a representative set of the new or improved MPEG-4 functionalities to be more fully tested and evaluated: content-based scalability, improved compression, and robustness in error-prone environments. It remains to understand if it was just a coincidence that the evaluated functionalities are those more important for VLBAV applications. These tests were pioneer since it was the first time that formal video subjective tests were performed for very low bitrates (starting with 10 kbit/s), it was the first time that subjective content-based tests were performed (the subjects were not only asked to evaluate the global subjective qualities but also the quality of particular objects in the context of a scene) and it was the first time that formal subjective tests were performed with channel corrupted video coded bitstreams - error resilience and error recovery tests. These tests showed the good performance of DCT-based hybrid coding schemes (similar to H.263 or MPEG1), which were easily adapted to work in a content-based representation environment. The test results allowed to realise that MPEG-4 is not fighting against 'conventional' coding technology but rather wants to achieve new functionalities by using the best available technology (independently of how old it is).

The experience of the first MPEG-4 set of tests allows to conclude that testing at very low bitrates is still an open issue. While we are still waiting for valid objective quality measures, notably adequate for coding methodologies which do not try to match the original image pixel by pixel (such as some used at very low bitrates), it became evident that, at very low bitrates, subjective quality evaluation is most of the times a choice between different types of distortion related to different coding approaches. This fact seems to highlight the importance for VLBAV applications of task-based tests, where the success of a relevant task in the context of an application and not a global quality is tested. Since a second round of MPEG-4 tests will be performed in July 1997, it is easy to foresee that we will see, in the near future, new developments in the context of testing for VLBAV applications.

4 VLBV CODING WITH THE MPEG-4 VIDEO VERIFICATION MODEL

Following the first round of MPEG-4 tests, the first MPEG-4 video Verification Model (VM) has been defined at the Munich MPEG meeting, held in January 1996, and it has been updated at the meetings after [8]. The VM is a completely defined encoding and decoding environment such that an experiment performed by multiple independent parties will produce essentially identical results [9]. New tools can be integrated in the VM, substituting other tools, when the corresponding core experiment has shown significant advantages in this integration.

The representation architecture adopted for the first MPEG-4 video VM (see figure 1) is based on the concept of Video Object Plane (VOP) which "*corresponds to entities in the bitstream that the user can access and manipulate*" [8]. The scene is 'understood' as a composition of VOPs with arbitrary shape but the method to produce the VOPs is not considered in the MPEG-4 VM. This means that the MPEG-4 VM is able to code scenes with more than one VOP, if the scene is, by some means (automatic or manual), previously structured in VOPs; a VOP can be (it is usually) a semantic object in the scene. The VOPs may have different spatial and temporal resolutions and each VOP has assigned a composition order.

Following the results of the MPEG-4 first round of tests, the coding tools used in the VM to code each VOP (VOP coding block in figure 1) are basically those already used in the

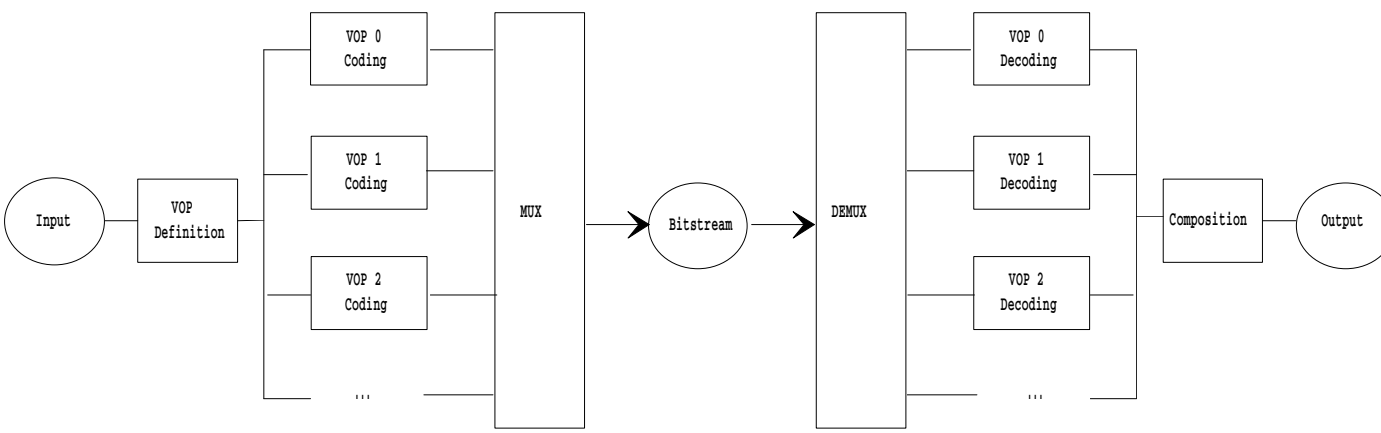

Figure 1 - MPEG-4 video Verification Model encoder and decoder architecture

available video coding standards, notably ITU-T H.263. In terms of texture and motion coding, the basic difference with ITU-T H.263 is the possibility to separate, at the VOP level, the motion and texture information. Three VOP prediction modes are considered: I (intra), P (predicted), and B (interpolated). The VOP arbitrary shapes are coded using a quad-tree scheme, both for binary and grey scale (n-bits) masks. Since some of the macroblocks to be coded may fall over the VOP contour, a padding technique had to be specified.

The MPEG-4 video VM bitstream syntax considers only two layers: the session layer and the VOP layer. The session layer encompasses a given span of time and contains all the video information needed to represent this span of time without reference to information in other session layers. The VOP layer encompasses all the syntactic elements, coding and composition elements, corresponding to a specific VOP in the scene.

The analysis of the current MPEG-4 video VM [8] from the point of view of VLBV coding leads to the following remarks:

- **VOP definition** - This block is at the heart of the MPEG-4 vision since it clearly states the scene is a composition of scene elements which will be individually available for manipulation. For that the relevant scene elements have to be somehow identified (automatically, manually, etc). The criteria and the methods to make this identification are not the main task of MPEG-4 which has rather to provide the means to represent the composition of video elements. The VOP definition block has the important but rather difficult task to organise the scene in terms of content; it is expected this organisation will strongly depend on the application at hand (and the corresponding available resources, either computational or transmission). At very low bitrates, the task to identify the most significant elements in the scene is even more important but also more difficult. Due to the complexity of some images, it seems useful that the VOP definition process is preceded or made in parallel with a *simplification process* which should have the task of simplifying the image by eliminating irrelevant details, noise, etc, in order to easy the VOP definition task not only in terms of computational effort but also to 'clarify' the image content in order its 'understanding' is more easily possible, specially in a very low bitrate context. This simplification task should depend on the target application and corresponding bitrate/quality and it is always a very difficult compromise between what it is a simplification or an elimination of the image information content. Taking into account the 'price' of contour coding, it is obvious that the number of VOPs to be used in very low bitrate applications will have to be limited. Thus an adequate choice of the VOPs together with a certain 'cleaning' of irrelevant and expensive image elements may become a key issue to reach images with higher subjective impact. Finally, it is important to highlight that this type of processing - simplification and VOP definition - will never be defined as mandatory in the context of MPEG-4. This means that any relevant technological developments which may appear in the future for these tasks will be easily integrated in the context of MPEG-4.
- **VOP coding architecture** - The VOP coding architecture refers to the coding tools used to efficiently code each VOP resulting from the VOP definition block; it has a shape and a motion plus texture component. Following the MPEG-4 tests results which have showed that DCT-based hybrid coders still perform very well, a H.263-like coding structure has been adopted. Since H.263 is largely recognised has a very good coding scheme for very low bitrates, there seems to be no doubts that, from the coding efficiency point of view, very low bitrate applications are well supported by the MPEG-4 video VM.
- **Syntactic structure** - One the most important characteristics of the MPEG-4 video VM syntax is the clear coexistence of coding and composition syntactic elements. Until now, syntactic elements have been added to the VM with a close look to the global efficiency. A careful analysis of the current video VM syntax allows to conclude that special attention is being paid to the case of MPEG-4 sessions with a unique VOP. Although this case is very likely to happen at very low bitrates (e.g. for simple videotelephony), we should not confuse very low bitrate applications with one VOP sessions since there are many very low bitrate applications which do require more than one VOP [6]. Although very low bitrates do ask to pay more attention to the VOP definition and to the number of VOPs identified (contours are expensive), the choice to have one or more VOPs is more a consequence of the type of application being provided that of the bitrate itself. In this sense, it may be expected that videotelephone applications may work with one VOP (or two) even for higher bitrates, while interactive database access applications will very likely use more VOPs, even for lower bitrates.

The main conclusion is again in the same direction: although MPEG-4 has not the target to define a very low bitrate AV coding standard but rather to "*establish a universal, efficient coding of different forms of audio-visual data, called audio-visual objects*" [4], VLBAV applications still play a fundamental role within MPEG-4 and the recent technical developments have taken their requirements into account.

ACKNOWLEDGEMENTS

The author would like to acknowledge the support to this work by the Commission of the European Union under the ACTS program - project MOMUSYS AC098 and by the Junta Nacional de Investigação Científica e Tecnológica from Portugal.

REFERENCES

[1] L. Chiariglione, Doc. ISO/IEC JTC1/SC29/WG11 N271, "New work item proposal (NP) for very-low bitrates audio-visual coding", London meeting, November 1992

[2] K. O'Connell, Doc. ISO/IEC JTC1/SC29/WG11 N270, "Project description for very-low bitrate A/V coding", London meeting, November 1992

[3] AOE Group, "Proposal package description (PPD)", Doc. ISO/IEC JTC1/SC29/WG11 N998, Tokyo meeting, July 1995

[4] L. Chiariglione, Doc. ISO/IEC JTC1/SC29/WG11 N1177, "MPEG-4 project description", Munich meeting, January 1996

[5] F.Pereira, "MPEG-4: a new challenge for the representation of audio-visual information", Picture Coding Symposium' 96, Melbourne - Australia, March 1996

[6] F. Pereira, R. Koenen, "Very low bitrate audio-visual applications", in Image Communication Journal, 1996

[7] L. Chiariglione, Doc. ISO/IEC JTC1/SC29/WG11 N1231, "Liaison to ITU-T SG 15 LBC", Florence meeting, March 1996

[8] MPEG Video Group, "MPEG-4 video verification model 2.0", Doc. ISO/IEC JTC1/SC29/WG11 N1260", Florence meeting, March 1996

[9] F. Pereira (editor), "MPEG-4 testing and evaluation procedures document", Doc. ISO/IEC JTC1/SC29/WG11 N999, Tokyo meeting, July 1995