Centrum voor Wiskunde en Informatica

REPORT*RAPPORT*

PNA

Probability, Networks and Algorithms

Probability, Networks and Algorithms

MASCOT: metadata for advanced scalable video coding tools
Final report

H.J.A.M. Heijmans, C. Bernard, M. Domanski,
B. Pesquet-Popescu, A. Smolic, P. Schelkens, L. Torres

REPORT PNA-R0307 JUNE 30, 2003

CWI is the National Research Institute for Mathematics and Computer Science. It is sponsored by the Netherlands Organization for Scientific Research (NWO).
CWI is a founding member of ERCIM, the European Research Consortium for Informatics and Mathematics.

CWI's research has a theme-oriented structure and is grouped into four clusters. Listed below are the names of the clusters and in parentheses their acronyms.

Probability, Networks and Algorithms (PNA)

Software Engineering (SEN)

Modelling, Analysis and Simulation (MAS)

Information Systems (INS)

P.O. Box 94079, 1090 GB Amsterdam (NL)
Kruislaan 413, 1098 SJ Amsterdam (NL)
Telephone +31 20 592 9333
Telefax +31 20 592 4199

ISSN 1386-3711

MASCOT:
Metadata for Advanced Scalable Video Coding Tools
Final Report

Henk J.A.M. Heijmans, *Coordinator*
CWI
P.O. Box 94079, 1090 GB Amsterdam, The Netherlands

Christophe Bernard
Centre de Morphologie Mathematique, Ecole des Mines de Paris
35, rue Saint-Honore, F 77305 Fontainebleau, France

Marek Domanski
Poznan University of Technology
Institute of Electronics and Telecommunications
ul.Piotrowo 3a, 60-965, Poznan, Poland

Beatrice Pesquet-Popescu
Dept. Traitement du Signal et des Images
Ecole Nationale Supérieure des Télécommunications
46, rue Barrault, 75013 Paris, France

Aljoscha Smolic
Image Processing Department, Heinrich-Hertz-Institut
Einsteinufer 37, D-10587 Berlin, Germany

Peter Schelkens
Vrije Universiteit Brussel
Department of Electronics & Information Processing (ETRO)
Pleinlaan 2, B-1050 Brussel, Belgium

Luis Torres
Universitat Politècnica de Catalunya
Campus Nord, Modulo D5 Jordi Girona, 1-3 08034 Barcelona, Spain

ABSTRACT
The goal of the MASCOT project was to develop new video coding schemes and tools that provide both an increased coding efficiency as well as extended scalability features compared to technology that was available at the beginning of the project. Towards that goal the following tools would be used:

- metadata-based coding tools;
- new spatiotemporal decompositions;
- new prediction schemes.

Although the initial goal was to develop one single codec architecture that was able to combine all new coding tools that were foreseen when the project was formulated, it became clear that this would limit the selection

of the new tools. Therefore the consortium decided to develop two codec frameworks within the project, a standard hybrid DCT-based codec and a 3D wavelet-based codec, which together are able to accommodate all tools developed during the course of the project.

1998 ACM Computing Classification System: E.4, I.4, I.2.10.
Keywords and Phrases: Video coding and compression, scalability, metadata, wavelet, motion estimation and compensation, MPEG, H.26x.
Note: This is the final report about the EU project MASCOT: Metadata for Advanced Scalable Video Coding Tools.

Mascot

IST–2000-26467
Metadata for Advanced Scalable Video Coding Tools

Final Report

(Deliverable D5.2)

Report Version:	**1.0**
Report Preparation Date:	**12-05-2003**
Classification:	**Public**
Contract Start Date:	01-05-2001
Duration:	24 months
Project Coordinator:	Henk Heijmans, CWI, Amsterdam
Partners:	1: CWI, Amsterdam
	2: ENSMP-CMM, Paris
	3: ARMINES-CMM, Paris
	4: HHI, Berlin
	5: PUT, Poznan
	6: PRF, Paris (till 01-Sep-02)
	7: GET-ENST, Paris
	8: UPC, Barcelona
	9: VUB, Brussels

Project funded by the European Community under the "Information Society Technologies" Programme (1998-2002)

Table of contents

1 Executive Summary

The continuing growth of multimedia applications leads to a great expansion of video transmission over heterogeneous channels such as internet, mobile nets and in-home digital networks. This raises new challenging problems related to varying transport conditions (e.g., bandwidth, error rate) and receiver capabilities (CPU, display). However the improvement in compression efficiency between MPEG-2 and MPEG-4 is not significant and new techniques are required to overcome this limitation and to enable a quick and easy access to large multimedia data repositories. The main goal of MASCOT is to bring a breakthrough in multimedia data coding and access through the exploitation of two innovative techniques, the design of new decompositions and the exploitation of metadata to improve coding efficiency.

The notion of scalability is the expected functionality to introduce a high degree of flexibility in coding/decoding systems, and therefore there is an increasing need for intrinsically scalable video coding schemes providing fully progressive bitstreams. A major objective of the MASCOT project is to fill this need by exploiting novel wavelet decomposition methods and more efficient prediction techniques. A second major objective is to to explore how video sequence encoders might exploit available metadata information to improve the compression efficiency.

Hence the goal of the MASCOT project was to develop new video coding schemes and tools that provide both an increased coding efficiency as well as extended scalability features compared to technology that was available at the beginning of the project. Towards that goal the following tools would be used:

- metadata-based coding tools;
- new spatiotemporal decompositions;
- new prediction schemes.

Although the initial goal was to develop one single codec architecture that was able to combine all new coding tools that were foreseen when the project was formulated, it became clear that this would limit the selection of the new tools. Therefore the consortium decided to develop two codec frameworks within the project, a standard hybrid DCT-based codec and a 3D wavelet codec, which together are able to accommodate all tools developed during the course of the project.

Available codecs have been used as starting point for both. For the hybrid DCT-based codec, the codec developed by the Joint Video Team (JVT), henceforth referred to as H.264/AVC codec, has been used as the baseline codec. The second baseline codec architecture used in MASCOT is a 3D wavelet codec. Initially this was a codec provided by Philips Research France (PRF), who also coordinated the integration of tools from other partners. Unfortunately, PRF withdrew from the project on September 1, 2002. The consortium then switched to another wavelet codec platform, which was publicly available and used within the MPEG community. We will henceforth refer to this codec as the *Woods codec*, after its originator John W. Woods. In parallel to the Woods codec, the consortium also investigated in the first project year the possibility of utilizing an architecture exploiting in-band or wavelet-domain motion estimation and compensation. Due to the fragmentation of the wavelet research activities induced by this additional architecture and the international consensus that existed at that time in relation to the spatial domain approach used by the Woods codec, the consortium decided additionally to focus all research effort on the latter.

These two selected architectures (H.264/AVC and Woods) have the clear advantage that MASCOT results can be easily compared with results from the original codecs and that improvements can be contributed directly. The main objective of the project was then to improve the performance of these two baseline codecs by replacing certain tools or adding additional ones. For the hybrid DCT-based codec these concern among others:

- metadata-based tools such as
  - texture synthesis
  - face encoding
  - video transitions
  - selection of reference frames
  - rate control
  - video segment shuffling
- scalability tools

- long-term global motion compensation (LTGMC)

For the 3D wavelet codec, the following tools have been explored:

- new temporal decompositions, including
  - 5/3 motion-compensated temporal lifting
  - sliding window implementations of the latter
- new spatial decompositions, including
  - morphological wavelets
  - adaptive wavelets
  - no-overshoot wavelets
- wavelet-based motion estimation
- new motion vector encoding techniques
- in-band wavelet video coding technologies.

Both MASCOT codecs have been tested extensively with a test setup similar to MPEG core experiments. Significant improvements compared to both baseline codecs have been achieved. Since the two integrated MASCOT codecs directly extend and improve the state-of-the-art in video coding, they can be regarded to be among the most advanced video codecs available in the world, which is the major outcome of the MASCOT project.

Finally the results obtained within the MASCOT project have been demonstrated to a broad public at the Picture Coding Symposium held on April 23-25 in Saint Malo, France, which is one of the major world wide conferences on image and video coding. An exhibition was organized showing the improvements achieved in MASCOT.

Two main conclusions of MASCOT project:

1. **Metadata can be useful for video coding**
2. **Scalability is an important issue for the future**

Regarding the second conclusion it must be observed that further exploration of both hybrid DCT-based and 3D wavelet-based architectures is necessary to ensure competitive rate-distortion behavior compared to non-scalable MPEG-4 video coding technology and to allow for the selection of a suitable architecture.

The first conclusion is justified by the results summarized in Section 4.3, and the second conclusion is supported by the results in Section 4.4, as well as by the latest developments in the MPEG community.

We conclude this summary with a table listing all the tools investigated in MASCOT, their usefulness, and a measure of risk for their further development.

Tools developed in MASCOT	Codec	Rating	Risk	Reference
Texture (metadata)	Hybrid	Very successful	Low	Section 3.3.2
Faces (metadata)	Hybrid	Successful	Average	Section 3.3.3
Video transitions (metadata)	Hybrid	Very successful	Low	Section 3.3.5
Dissolve transitions (metadata and prediction)	Not integrated	Promising	High	Section 3.3.6
Reference frames (metadata)	Hybrid	Successful	Average	Section 3.3.7
Rate control (metadata)	Hybrid	Successful	Low	Section 3.3.8
Mosaics and key-regions for coding (metadata)	Hybrid	Not integrated	High	
Video segment shuffling (metadata)	Hybrid	Successful	Average	Section 3.3.9
Temporal 5/3 lifting	Wavelet	Very successful	Average	Section 3.4.2
Sliding window implementation of temporal lifting	Wavelet	Very successful	Low	Section 3.4.2
Morphological wavelets for spatial decompositions	Wavelet	Promising	Low	Section 3.4.3
Adaptive wavelets for spatial decompositions	Wavelet	Promising	Low	Section 3.4.3
No-overshoot wavelets for spatial decompositions	Wavelet	Not successful (for video compression)		Section 3.4.3
GoF size selection	Wavelet	Successful	Average	-
Scalability for hybrid codec	Hybrid	Very successful	Average	Section 3.4.4
Wavelet-based motion estimation	Wavelet	Very successful	Average	Section 3.5.2
Motion estimation and compensation in wavelet domain	Not integrated	Very successful	High	Section 3.5.3
Motion vector coding	wavelet	Very successful	Low	Section 3.5.4
Long term global motion compensation	hybrid	Very successful	Low	Section 3.5.5
Long term motion prediction using graph matching	Not integrated	Not successful		Section 3.5.6

2 Project Objectives

2.1 Motivation

The continuing growth of multimedia applications leads to a great expansion of video transmission over heterogeneous channels such as internet, mobile nets and in-home digital networks. This raises new challenging problems related to varying transport conditions (e.g., bandwidth, error rate) and receiver capabilities (CPU, display). Internet is expected to become the major future carrier for all sorts of audio-visual information and data. Consequently, requirements for bandwidth availability and quick and easy access to large multimedia databases will be more and more stringent. To a large extent, this concerns video data. However the improvement in compression efficiency between MPEG-2 and MPEG-4 is not significant and new techniques are required to overcome this limitation and to enable a quick and easy access to large multimedia data repositories. In this context, the main goal of MASCOT is to bring a breakthrough in multimedia data coding and access through the exploitation of two innovative techniques, the design of new decompositions and the exploitation of metadata to improve coding efficiency.

2.2 New decompositions enabling scalable compression

The notion of scalability is the expected functionality to introduce a high degree of flexibility in coding/decoding systems. For all currently available interactive multimedia applications however, which are very demanding in terms of video quality and coding efficiency, the cost as well as the limited performances of scalability obtained in the current standards remain unacceptable. That is why there is a need for intrinsically scalable video coding schemes providing fully progressive bitstreams. A major objective of the MASCOT project was to fill this need by exploiting novel wavelet decomposition methods and more efficient prediction techniques.

2.3 Using metadata to improve coding

It can be expected that in the future, a very large amount of audio-visual documents will be indexed and that metadata information will be rather easy to create. As a result, in many circumstances, audio-visual material will be available together with the metadata describing its content. Therefore, future image and video sequence encoders will be able to use the metadata information in order to improve their efficiency or to optimise their strategy. One of the main objectives of MASCOT is to demonstrate the validity of this approach and to develop an efficient compression scheme exploiting metadata information. Metadata here refers to the indexing information that may be available to support search, query and browse functionalities, e.g. in MPEG-7 or SMPTE metadata standards.

In some situations, existing metadata can help in order to select encoding parameters that are suited to encode particular video sequence. For instance, the rate control tool selects the best IPB GoP structure taking into account the MPEG-7 motion activity metadata that describe the amount of motion that is present in the video sequence. In some other situations, in order to fully exploit the metadata information, the strategy of the encoder will have to severely change the bitstream syntax and semantics of the bitstream. In the later case, the encoder will not be able to decode the sequence unless the metadata information is also available. The video shot transition coding can also be included in this scenario. In this case, the inter frame prediction of the standard hybrid codec is modified so metadata information of shot transition is exploited in order to improve the prediction step.

2.4 MASCOT objectives

Therefore, the main objectives of the MASCOT project are:

- To provide a breakthrough in the domain of video compression and to improve the quality of the reconstructed video at low bit-rates by using metadata during the encoding and decoding steps, exploiting recently developed concepts in nonlinear, i.e., morphological and adaptive wavelet decompositions, and by the development and optimisation of advanced and dedicated prediction schemes.
- To develop intrinsically scalable compression schemes, which fulfill the requirements of multimedia applications, by:
  - covering a wide range of bit-rates;

  - yielding high compression ratios to guarantee transmission of large quantities of video data over one channel;
  - providing a high level of bitstream embeddedness to enable the adaptation of the compressed video data to a variety of networks (with different bandwidths, error rates, etc.) and receivers (characterised by different capabilities in terms of display size, CPU, and memory).
- To provide Europe with a leadership in new video compression techniques by building up a strong patent portfolio in this strategic domain, and contributing to the awareness of the project achievements by publications in renowned journals and communications to key conferences and exhibitions.
- To contribute to standardisation committees like ITU-T and MPEG in order to bring in MASCOT's developments in scalability, and to follow the activities of MPEG-7 (and MPEG-21) and JPEG-2000.

3 Methodologies

3.1 Introduction

In this section we describe all tools that have been developed within the context of the MASCOT project, both for the hybrid DCT-based codec and for the 3D wavelet codec[1]. In Section 4 we summarise the results that have been obtained for these tools.

3.2 Two MASCOT codecs

The goal of the MASCOT project was to develop new video coding schemes and tools that provide an increased coding efficiency and extended scalability features compared to available technology. This should be achieved by new metadata-based coding tools, new spatiotemporal decompositions and new prediction schemes. In the beginning of the project it was intended to build a single codec architecture that should combine all the new coding tools. During the first months and meetings it turned out that the integration of the substantially different new coding approaches in a single new framework could not be achieved since some of the approaches were inherently incompatible.

Most of the ideas in metadata-based coding were strongly related to classical hybrid video coding schemes such as H.264/AVC. For instance the *Selection of Reference Frames* and *Long-term GMC* tools can only be used in connection with the multi-frame prediction of H.264/AVC. Also the *Rate Control*, *Texture*, and *Transitions* tools directly target a classical hybrid video coding framework and the *Scalability* tool directly extends the H.264/AVC architecture and provides new functionality. On the other hand the ideas on 3D wavelet coding and prediction could of course only be investigated within the framework of a 3D wavelet codec. These two research directions (improvement and extension of classical hybrid coding and 3D wavelet coding) are the most important trends in video coding, which is reflected e.g. in the related activity in MPEG.

As a consequence the consortium decided to follow both directions and to use two codec frameworks within the project, a standard hybrid DCT-based and a 3D wavelet codec, which can accommodate all new tools. Available codecs have been used as starting point for both. The main task of the project was then to improve the performance of these baseline codecs that could be measured directly in comparison to the state-of-the-art[2].

MASCOT has developed a metadata-based architecture which uses a standard hybrid codec along with new tools developed within the project. The H.264/AVC codec developed within the Joined Video Team (JVT) has been used as baseline. The H.264/AVC codec is a joined effort of ISO/MPEG and the ITU-T/VCEG to develop a new video coding standard (MPEG-4 part 10 and ITU-T H.264), which represents the state-of-the-art in video coding. The new H.264/AVC codec is very similar to existing standard video codecs such as MPEG-2 or H.263. The impressive performance gain of the H.264/AVC codec of up to 50% bitrate reduction compared to MPEG-4 Advanced Simple Profile mainly comes from an intelligent optimization and combination of existing concepts.

3.3 Metadata-based tools

3.3.1 General

One of the basic objectives of MASCOT has been to use metadata information to improve coding efficiency. During the initial phase of the project, descriptors and description schemes (DS's) included in the MPEG-7 and SMPTE standards were analyzed and their potential use for encoding was investigated.

[1] In this report we shall often use the terminology “3D wavelet codec”. But in the literature one also finds the terminology “2D+t wavelet codec” or “3D subband codec”.

[2] However, it should be noted that the partner (VUB) carrying out the in-band wavelet video research continued its research effort on this type of architectures, utilizing external funding, and succeeded in combining this expertise with its expertise gained with the research activities on the Woods architecture. This resulted in an MPEG proposal, co-authored by Philips USA, that proposes an MCTF-based in-band wavelet video codec that additionally allows plugging in an hybrid DCT-based codec to encode the low-pass subbands (being the base layer). This proposal comes also forward to the requirements imposed by the MPEG-4’s Scalable Video Coding Ad Hoc Group stating that the base layer of a scalable MPEG-4 codec should be backward compatible with the hybrid DCT-based schemes.

The results of this study have been presented in Deliverable 2.1. The main conclusions of the study are that the following metadata can be used to improve coding efficiency:

Feature type	**Descriptor name**	**Application to coding**
Color	Dominant color Scalable color Color layout Color structure GoF/GoP Color	• These descriptors provide rough information about the color distribution of the pixels. Although, the description gives an approximation of the pixel probability density function, it is not precise enough to allow any improvement of the entropy coding. • However, they can be used to search and retrieve images or shots. The classical motion compensation problem in video coding can be formulated as a search and retrieval problem. With this approach, any low-level descriptor useful to search and retrieve visual content can be used.
Texture	Homogeneous texture Texture browsing Edge histogram	• Texture descriptors can also be used for the motion compensation problem formulated as a search and retrieval problem. Some textured areas can be omitted during the encoding process and synthesized at the receiver end. In this case, the texture descriptor is mainly used to classify and select the textures that can be synthesized. In this context, the Edge Histogram descriptor has been used. • These descriptors are also appropriate to tune the quantization parameters. The main idea of the approach is to save some bits for specific textured areas.
Motion	Camera motion Motion trajectory Parametric motion Motion activity	• These descriptors are quite low level and are therefore easy to use for compression. Motion trajectory or parametric motion can be used to construct mosaics. Parametric motion is also used to allow motion compensation of the synthesized texture. • Finally, motion information is a very important feature to design an efficient bit assignment strategy. • The Motion Activity is also used to define the filtering structure and the GOF size of the 3D wavelet encoder.
Face	Face recognition	• This descriptor defines a subspace appropriate for face identification and recognition. Although, it specifies a transform that could be used for face encoding, it has been proven that the direct use of MPEG-7 descriptors is not useful for encoding. The main reason for this conclusion is that the descriptor does not allow to represent a face with sufficient quality.

Beside the low-level descriptors, the content description area of MPEG-7 also allows the creation of Tables of contents and Indexes. They can be considered as high-level description of the structure of the content. This information can be used to improve the compression efficiency. For instance, similarity between shots or frames can be extracted from the structural description of the content to reorganize the sequence to be coded. Finally, MPEG-7 includes a specific tool to describe the sequence structure at the level of shot transition: Analytic Transition DS. This information has been used to improve the compression of frames belonging to a transition.

Fig. 1 shows a high-level block diagram of the MASCOT hybrid encoder. As an illustration, this example also shows how the texture tool has been integrated into the H.264/AVC framework.

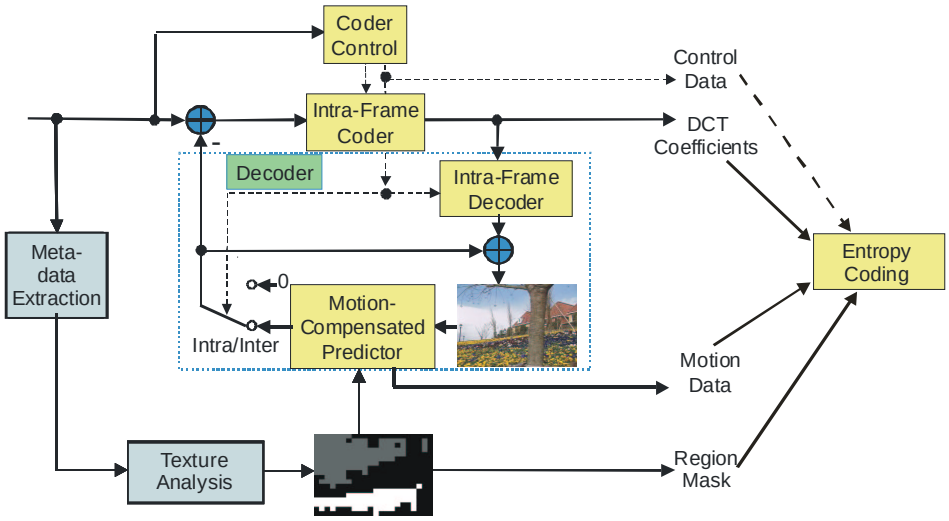


Figure 1: Block diagram of MASCOT hybrid encoder using texture metadata.

The new metadata-based coding tools (the new prediction tools also) developed within MASCOT have been integrated into the H.264/AVC codec. Some of the metadata-based coding tools only influence the encoder while others also influence the decoder and therefore need to be transmitted. Derived data like region assignment masks may also need to be transmitted. This results in a trade-off between the additional bitrate for the overhead and the improved coding efficiency. The H.264/AVC bitstream syntax provides hooks to attach metadata to the video bitstream. Some of the new coding tools leave the H.264/AVC bitstream syntax unchanged; others will require modifications to the syntax. In any way the result has been an improved H.264/AVC codec, which is the best available hybrid video codec.

In the following a summary of the main metadata-based video coding techniques developed in the MASCOT project are presented.

3.3.2 Texture

In this new approach to video coding, the video scene is being classified into subjectively relevant and irrelevant texture regions. Relevant textures are those where details matter and which should be reproduced accurately. Irrelevant textures can be seen as image content with less important subjective details and for which an approximate reproduction can be chosen. This idea is applied to video coding using a texture analyzer and a texture synthesizer. The analyzer identifies subjectively irrelevant texture regions, generates masks for the decoder as well as side information for the synthesizer. The synthesizer replaces the original irrelevant textures by inserting synthetic textures in the identified regions. The approach has been integrated into the MASCOT hybrid codec. Bit-rate savings between 10% and 23% and between 5% and 18.1% compared to the H.264/AVC video codec are shown for a semi-automatic and for the current implementation of an automatic method respectively given similar subjective quality.

3.3.3 Face coding of video sequences

The goal of this tool is to improve the coding efficiency of human faces with respect to other existing coding schemes. The functionality that has been developed is of interest in video conference applications, mobile telephony with video transmission, internet applications where face transmission at low speed is of importance, etc. It is important to emphasize that the coding scheme that has been developed is only suitable for face coding. It is also clear that, prior to the encoding process, the face needs to be detected. The scheme assumes that the face is previously detected before the actual encoding process. The face detection process is out of the proposed codec architecture.

The face encoding scheme is based on the well-known eigenspace concepts used in face recognition systems, which have been modified to cope with the video compression application. The main differences from other proposed schemes are a new way to adapt the eigenspace to take into account the different poses, expressions and lighting conditions of the faces, and the combined use of the H.264/AVC codec to encode the face updates. As an added value to the encoded images, the MPEG-7 descriptors of the image faces have been found and added to the bit stream to allow search and browsing functionalities. Acceptable results have been found for bit-rates around 2.5 kbits/s.

During the MASCOT project it was shown that the MPEG-7 descriptors are not suited to improve coding. This is due to the fact that the MPEG-7 descriptors are focused on detection and recognition. Therefore the reconstructed image quality that can be achieved by combining the descriptor eigenspace and a set of parameters is very poor. During the project, a new technique using a combined adaptive PCA and a hybrid H.264/AVC coding approach was developed and proved to be very successful.

3.3.4 Mosaics and key-regions for coding

Mosaics images, also called sprites, are representations of the background information of a video scene aligned into the same spatial reference. The global motion of the sequence is extracted in order to segment background and foreground objects from the sequence. Recent video coding standards like MPEG-4 can exploit mosaic images to improve the coding efficiency. A mosaic of the background information of a scene is used as reference when coding background blocks of the current image. This mosaic or sprite is made also available for the decoder. The bit-rate increase due to the creation of the mosaic image is compensated by the prediction improvement in the coding process. Metadata descriptors also use sprites to display the background information of a shot. This is useful for quick browsing of video sequences. Metadata descriptors do not restrict the use of mosaics for background regions but foreground regions can also be represented. Figure 2 shows an example of a metadata description for a video shot. The background information is easily presented using a mosaic image. Foreground regions may not be rigid so using a unique mosaic image is not effective. In the case of foreground regions a more complex model is needed to represent different poses or changes of foreground regions along the scene. In the example of the figure, foreground regions are represented using three image models. A texture image with the foreground texture information, an appearance image with a weighted mask of the pixels that appear on the region and a contour image with the shape information of the foreground object.

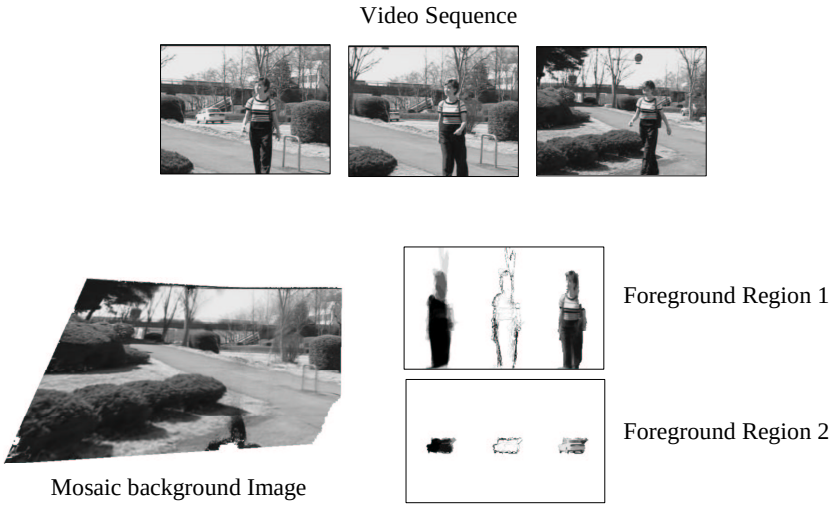


Figure 2: Metadata descriptor for shot browsing

It was intended to use this descriptor also to improve the video coding of foreground objects. If, for instance, a specific object has a considerable presence in a video sequence, the metadata model can be used to improve its prediction along the overall sequence. The coding scheme of this approach is therefore very similar to the sprite coding of MPEG-4 but with the possibility of using various sprite images or foreground models to predict future appearances of these objects through the scene. Coding of video using mosaics was discontinued in the framework of the MASCOT project.

3.3.5 Video shot transitions coding

Transitions between video shots are very common in standard video. Metadata describing video transition effects can be used to improve the inter-frame prediction when coding that specific transition. For instance, in the case of dissolves, prediction of previous or past frames can be modified taking into account the information of the shot transition metadata. The codec can then make use of this extra information to improve the coding efficiency (in terms of rate-distortion efficiency). It can also be modified so that the data already present in the metadata stream is not duplicated in the coded stream.

The proposed method to code transitions using metadata has the following two steps.

- In the first step, the standard IPB structure of current hybrid codecs is modified such that only B frames exists within the transition. To do so, the first P-frame in the transition is moved to the end frame of the transition marked by the metadata. Using only B-frames in the transition has the advantage that current frames can be predicted using two references both at the beginning and at the end of the transition.
- In the second step, the standard interpolative mode of B-frame is changed such that it can use different weights for previous and past references (instead of fixed 0.5 weights for each reference).

Results have shown that using metadata for the coding of video transitions increase the coding efficiency. In the case of short transitions (up to 6 or 7 frames) the bit-rate gains can be up to 80%. In the case of long gradual transitions the bit-rate saving are smaller due to the need of including references within the transitions. In the later case, bit-rate savings of 15% are expected when coding video transitions using metadata.

During the duration of the MASCOT projects, several improvements of this technique were developed, implemented and tested. The strategy was migrated from the Momusys MPEG-4 framework to the H.264/AVC hybrid codec framework. For the case of long gradual transitions, is was noted that using only one reference at the beginning and at the end of the transition was not enough for the prediction step and therefore a new technique was developed in order to include several references (B-frames) inside the transition. This new strategy improved the results when using transition metadata for long transitions.

3.3.6 Dissolve transition encoding

A dedicated strategy was developed to encode efficiently video sequences around dissolve transitions. The motivation of the strategy was that the motion prediction model underlying all video coding approaches, either hybrid of 3D MCTF-based, is a single motion displacement model. Each patch or block of a frame is moving along a single motion in time.
This model is not correct any more during dissolve transitions, where the video sequence is a transparent superposition of 2 sequences each with its own motion:

- one fading-out sequence
- one fading-in sequence.

A sequence analysis framework was devised to extend a wavelet-based single motion estimation scheme in order to

- perform a double motion estimation to obtain two motion maps v_1 and v_2;
- use the multiple motion information to separate the components of the video sequence by velocity: the sequence is $I=I_1+I_2$ where I_1 is moving along v_1 and I_2 along v_2.

Once the two superimposed sequences separated, each sequence with its own motion is encoded separately. Gain is expected from this encoding strategy in dissolve transitions because the underlying model more accurately reflects the video sequence evolution in time, as a single motion model does.

First complexity estimates showed that the approach was computationally very expensive, and was unfeasible in practice if performed on a whole video sequence. Metadata on video transitions are thus used to point frames in time segments on which a dissolve transition occurs, which a small proportion of the whole video sequence. For other frames, classical single motion-based compensation is done.

The approach has been described in deliverable D2.2. A first implementation to estimate everywhere two motion vector maps led to very unstable results. Inaccurate estimates of the flow maps jeopardizing a stable separation of the sequence I into sequences I_1 and I_2, it turned out that substantial changes of the estimation method were required to increase its accuracy. Therefore, further development of this tool was abandoned. However, we believe that under some usage scenarios, such a tool can be useful even if applied to transitions only, because it will prevent the decoded image quality to drop sharply at such dissolves when decoding at constant bitrate. When decoding a bitstream of variable bitrate, the tool is definitely less useful.

3.3.7 Long term selection of reference frames

Normal video sequences have a high degree of temporal redundancy. Standard hybrid coders exploit this fact by using past frames as references for coding the current frame. Once a good candidate has been selected as reference, only the difference between the reference and the current frame is encoded and

written in the bitstream. A simple translational motion model is also used to cope with the internal motion of the video sequence. Therefore, the final codec must send the difference between the current block and the reference block plus the motion vector information for each block to be coded in the current image.

It is natural to think that the closer frame in time will be the most similar to the frame being coded and for that reason current hybrid coders use the closest P- or I-frame in time as a reference for coding the current frame. For each block being coded in the current frame, a motion estimation algorithm finds the best reference block only in the closest reference frame. Therefore, in current hybrid coders, all blocks of the current frame being coded share the same reference image (which is always the closest P- or I-frame in time).

Metadata can be introduced in long-term temporal prediction to improve the overall coding efficiency. It can be used to perform a pre-selection of possible *N* candidates for reference frames. As metadata has been designed for search and retrieval capabilities, the search space of possible references can be increased without severe penalty in the computational cost. Metadata similarity can be used to pre-select *N* frames (to act as reference frames) among a high number of possible candidates frames. Moreover, as the same number of references frames *N* is used, there is no increase in the bit-rate associated to the transmission of the reference frame information. This ordering and pre-selection of previous frames to create the long term buffer can be easily created on the decoder side providing the metadata is also accessible to the decoder. In that case, no extra information is needed.

Metadata can also help the searching of possible reference frames. Shot descriptors can be used so that reference frames are only searched in shots that have similar content. Moreover, metadata about collections of documents can point the encoder to similar content of other sequences. For example, a user may record the news everyday at 7pm. Metadata in the decoder site can inform the encoder that several past bitstreams are already stored and available to the encoder. The codec can make use of those as possible reference for coding the current sequence.

There are several candidates that may improve long-term temporal prediction. The selected metadata should be simple and easy to compute. It also needs to include a similarity criterion so that frames with similar contents can be recognized and included in the buffer of possible reference frames. The MPEG-7 Color Layout descriptor is a valid candidate for this task. This descriptor specifies a spatial distribution of colors for high-speed retrieval and browsing. It targets not only image-to-image matching and video-clip to video-clip matching, but also layout-based retrieval for color, such as sketch-to-image matching. This descriptor can be applied either for the whole image or any (arbitrary shaped) part of the image. In our case, the use of this descriptor is to make a pre-selection of similar frames to fill the reference frame buffer in long-term temporal prediction.

The long term selection of reference frames technique can be entirely included in the framework of the MASCOT project. The coding schema was developed and integrated during the duration of the project. At the initial stage several metadata candidates were studied in order to select one metadata that was simple, easy to compute and that provided similarity measures to pre-select past references. Using the MPEG-7 color layout descriptor as metadata has shown results up to 12% bit-rate reductions with respect to the standard H.264/AVC (version 2.1) codec. At the final stage of the project, the technique was integrated into a common MASCOT codec. Finally, several enhancements to the technique were also implemented. For instance, the color layout metadata was re-computed from previous coded frames instead of using the external metadata. This resulted in better coding efficiency when coding at low bitrates, as in that case the encoded frames were of significantly worse quality than the original references.

3.3.8 Rate control using metadata

The selection of encoding parameters in standard video codecs is a non trivial problem. Therefore, current video codecs use fixed encoding parameters such as GoP structures or quality factors when coding video sequences. The adaptation of these parameters for different video sequences is a difficult problem as changing for instance the quality factor Q of one frame may change the distortion of future frames that use this frame as reference. At the same time, changing the frame type (between I,P or B) affects also to future frames that are going to be encoded. Metadata has been proven useful to improve this step and helps in the decision of choosing among different frame types or between different quantizer levels. Motion descriptors, such as the MPEG-7 *MotionActivityDS* descriptor, indicate the presence of high or low internal motion in the video sequence. This has been used to select different IBP structures according

to the amount of motion present in the bitstream. The selection strategy is based on the assumption that sequences with very high internal motion are better encoded using a low number of B-frames. When internal motion is rather low, the reference frames (previous and past) used to predict the current B-frame are very similar and the B-frame can be predicted with almost no error. In the presence of high internal motion, the references can change considerably and the quality of the predicted B-frame drops. Motion descriptors can inform the coder whether to use more B-frames when less motion is present in the sequence.

Results show that motion information can be used to take apart (more B-frames between P-frames) or join (more P references in the GoP structure) frame references. Bitrate savings up to 5% are expected when using the rate control tool on standard sequences. This technique can also be included as a new outcome of the MASCOT project.

3.3.9 Video segment shuffling

Standard hybrid codecs exploit the temporal redundancy of video sequences by performing a motion estimation between the current image being coded and a reference (or various previous references such as H26L or JVT codecs). B-frame types also use references that are situated in the future. The use of B-frames has been proved to generally increase the coding efficiency. Therefore, the coding frame order and the display order need to be changed in order to be able to use previous and past references. In that case, the ordering is only change a by a small amount of frames (the distance between P-frames). This distance is usually small for two reasons. The internal motion of the sequence makes inefficient to have references too far in time and, for real time applications, the delay introduced in the system cannot be very high.

In the case of non-real-time applications (such as stored video, etc.) it is then reasonable to think that the display order of video frames does not imply that it is the best possible order for coding. Therefore, the main question to this metadata tool is: will the re-ordering of video frames prior the encoding process help to improve the coding efficiency? Obviously, finding the best possible re-ordering is a difficult task. Again, existing metadata can help the decision of the coding order. Metadata, such as the MPEG-7 *SegmentDS* metadata descriptor, makes a relationship between different segments of an image. A hierarchical structure can be constructed that makes relationships between regions of an initial partition of the image. This hierarchical tree representation describes the image content in a similar way the Table of Contents and the index describe the contents of a book.

This metadata can also be applied to Video Segments. In this case, the entire video sequence is analyzed and a hierarchical structure is created from an initial partition of the video sequence. This initial partition can be composed from single frames, groups of N frames or extracted video shots. Initially, all the initial partition is considered and a merging criterion is defined to group frames that (even not closer in time) match a certain similarity criterion. One by one, all partitions are merged with the most similar one (in terms of the criterion chosen). The final hierarchical tree structure keeps information of all merges that have been done until the entire video sequence is merged.

Final results have shown relevant bitrate savings in sequences that present scattered, short and similar shots. This is common in TV programs like interviews or shows that switch over several camera perspectives. In these cases, bitrates savings up to 8% are expected compared to standard H.264/AVC for the same visual quality. The scenario where metadata information must be send together with the content has also been studied. In this case, the extra bitrate needed to send the shuffled information is negligible compared to the bitrate savings (less that 1% of the bitrate savings). This technique can also be included as a direct outcome of the MASCOT project.

3.4 Spatio-temporal decompositions

3.4.1 General

The aim of the workpackage dealing with spatio-temporal decompositions was to develop truly innovative schemes and algorithms in the field of spatio-temporal decompositions. The results that have been obtained as a results of the MASCOT project, have been embedded into new architectures for combined scalability modes (both in the predictive and in the MCTF[3] framework), and have also led to new paradigms for building spatial and temporal decompositions.

[3] Motion-compensated temporal filtering

The 3D wavelet codec developed by MASCOT consortium is described in Figure 3 and its different modules are described in the next sections.

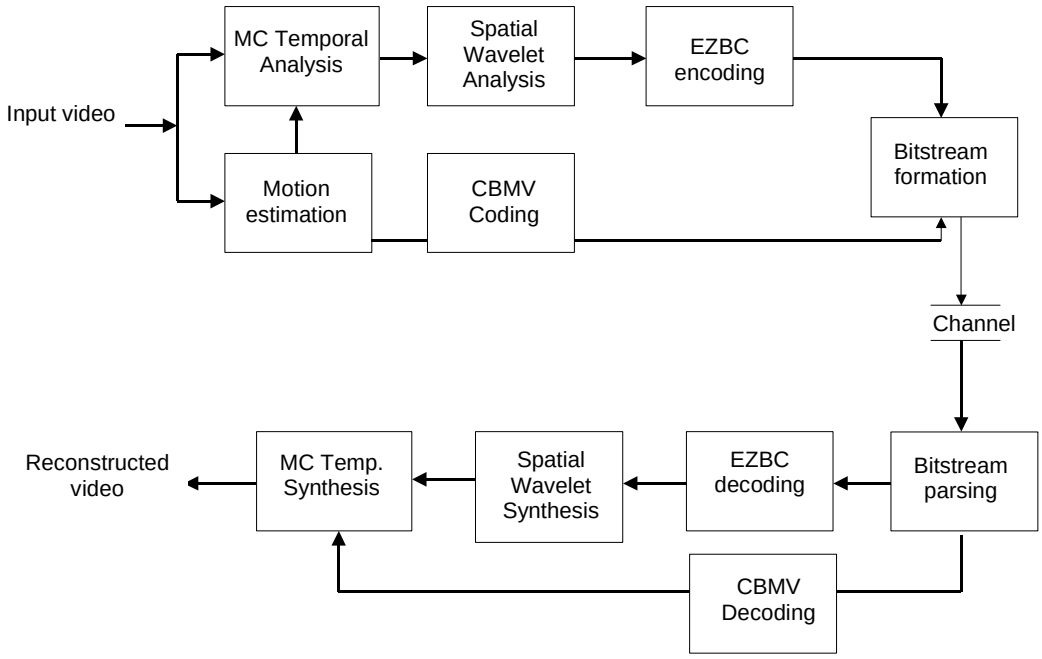


Figure 3: Motion-compensated 3D wavelet video codec developed in MASCOT.

3.4.2 Temporal decompositions

We have highlighted the benefits of a lifting formulation of the motion-compensated subband temporal decomposition (briefly called “temporal decomposition”) entering a 3D wavelet video coding scheme; for a reminder on the lifting implementation, see Figure 4 below. This idea implies that the prediction and update operators involved in such a scheme are intrinsically non-linear and we have shown how to improve the performances of existing Haar filters and also how to extend in this framework the usual Haar filters to longer ones. The advantage of this new temporal filtering paradigm is the increased flexibility in subband motion-compensated temporal filtering, through bidirectional motion compensated prediction-update operators (B-like frames), adaptive filtering, new subband structures, etc. This extension raises however another problem: a double number of motion vector fields. Taking advantage of the multiscale structure, we have proposed solutions to this problem, by exploiting the redundancies between hierarchical motion vector fields. A "sliding window" approach was proposed for implementing the temporal decomposition, leading to reduced memory requirements and avoiding GoF border artifacts, which otherwise decrease significantly the coding efficiency.

3.4.3 Spatial decompositions

The frames resulting after the temporal wavelet decomposition described in the previous subsection are decomposed by a 2D spatial wavelet. The default decomposition is the separable wavelet resulting from the tensor product of two one-dimensional biorthogonal linear 9/7 wavelets. One of the objectives of the MASCOT project was to replace such linear wavelets by nonlinear ones. We have considered three different classes of nonlinear wavelets, morphological wavelets, adaptive (update) wavelets, and “no-overshoot” wavelets. All three classes have been described in great detail in Deliverable D3.1. We will present a brief summary below.

Both the morphological and the adaptive wavelets are based on the lifting scheme, with one important difference: the morphological wavelets use the classical lifting scheme depicted in Figure 4, whereas the adaptive wavelet comprises an adaptive update lifting step as shown in Figure 5, followed by a fixed (i.e., non-adaptive) lifting step.

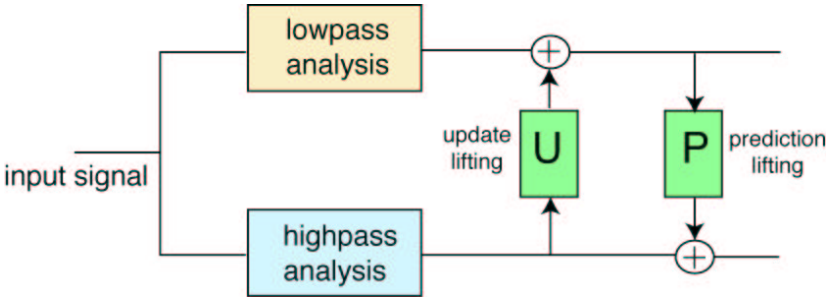


Figure 4: *The classical lifting scheme comprising a prediction and an update step.*

The lifting scheme introduced by Sweldens in 1997, provides a useful and flexible tool for the construction of new wavelet transforms from existing ones. The lifting scheme can be applied from various perspectives. It can be used to increase the number of vanishing moments or to improve or add other properties of a given wavelet, it can be used to design wavelets on complex geometrical surfaces, or to build wavelet decompositions using nonlinear filters, such as rank-order or morphological filters. In the MASCOT project we have concentrated on three different families of morphological wavelets which have all been described in detail in Deliverable D3.1. The first family can be interpreted as the morphological variant of the linear Haar wavelet and uses a direct construction without the lifting scheme. The second family is based on median-type operations, in the sense that both the prediction and the update step in Figure 4 use a median-type filter. The third family also uses the lifting scheme shown in Figure 4, but in this case the prediction and update filters are using a completely different kind of filters. They are based on a very novel approach in mathematical morphology introduced by Heijmans and Keshet in [1]. The basic idea is to provide the real axis with a new partial ordering. This results in a class of morphological filters which are self-dual. The basic filter is called the *cisl*[4] *erosion* and this filter is used in MASCOT as a starting point for a new type of wavelet decomposition. For more details we refer to Deliverable D3.1 or the work by Heijmans and Keshet [1].

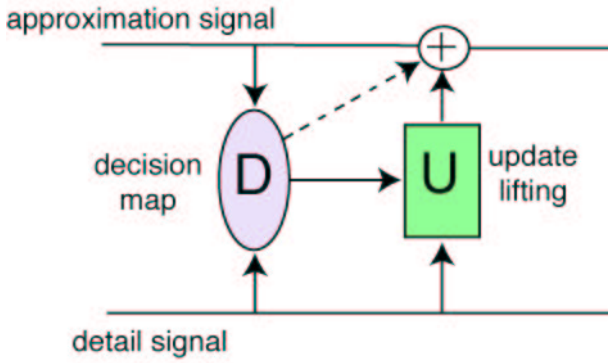


Figure 5: *An adaptive update lifting step comprising the decision map D and an adaptive update filter U.*

The adaptive wavelets that have been used in MASCOT have been introduced by Piella and Heijmans in [3] and explored in much greater detail by Heijmans, Pesquet-Popescu and Piella in [2], mostly within the context of the MASCOT project. The basic idea is captured in Figure 5, which shows a decision map D which yields as output a binary decision $d=0$ or $d=1$. Depending on this output, one uses update U_0 or U_1. This adaptive update step is then followed by a fixed prediction. The important theoretical results that were achieved by Heijmans, Pesquet-Popescu and Piella [2] show that it is possible to design adaptive update scheme which do not require bookkeeping for reconstruction, which, obviously, is of major importance in the context of image and video compression. Some of the theoretical results have been used within the MASCOT project, and some of them are reported in Section 4.4.3.

"No-overshoot" wavelets have been designed in order to provide an alternative way of removing ringing artifacts in decompressed images. The design starts from an original interpolating wavelet transform scheme with a single prediction lifting step, where reconstruction overshoots are canceled with

[4] Here "cisl" stands for "complete inf-semilattice".

continuous clamping operators (the continuity is crucial to guarantee that quantization causes a limited degradation to the reconstructed signal). This interpolatory nonlinear wavelet transform can be derived from various linear wavelet transforms like the Deslauriers-Dubuc schemes of all orders. However, an interpolating wavelet transform is badly conditioned for signal compression, and the scheme was modified with an additional update step. Several strategies for the choice of an adaptive update step were experienced. Also, the scheme was extended for 2-dimensional images.

The resulting transforms were tested for intra-frame coding. Whereas the resulting transform performs substantially better than conventional wavelet transforms for synthetic images (flat areas with sharp contours), it performs worse for natural images with texture and noise. The “no-overshoot” wavelet was also tested for encoding of prediction error images (high pass frames of the motion-compensated temporal transform), which possess more geometrical structure than intra-frames. The results were also disappointing: the distortion was always worse at identical rate than linear wavelets; see D3.1. For this reason, further integration into the wavelet codec was abandoned. It can be noted however, that the underlying “interpolating no-overshoot” wavelet is efficient in reducing artifacts in high-order interpolations, and can be considered as an interesting byproduct of the MASCOT project.

3.4.4 Scalability in hybrid codecs

Generic scheme

Proposed is a scalable coder that consists of two or three motion-compensated coders (Figure 6) that encode a video sequence and produce two or three bitstreams corresponding to two or three different levels of spatial and/or temporal resolution. The generic structure is characterized by mixed spatio-temporal scalability and independent motion estimation and compensation performed in the individual prediction loops. These two features together with other improvements described further are substantial for high efficiency obtained.

The structure of the proposed scalable coders has been proposed and tested for MPEG-2, H.263 and the new H.264/AVC platforms. Nevertheless this approach is also applicable to other hybrid video coders, like MPEG-4.

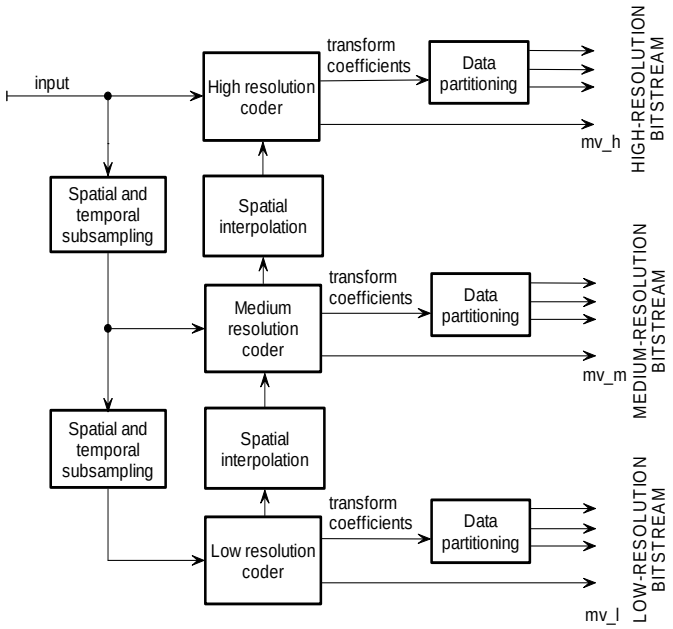


Figure 6: A generic structure of a multi-loop scalable hybrid coder.

A low-resolution macroblock corresponds to a picture area that comprises 4 macroblocks in the middle-resolution bitstream. Therefore the respective motion vectors are different for the particular resolution layers. It means that each coder has its own prediction loop with own motion estimation.

Two-layer coder

The most interesting is the two-layer case (Figure 7).

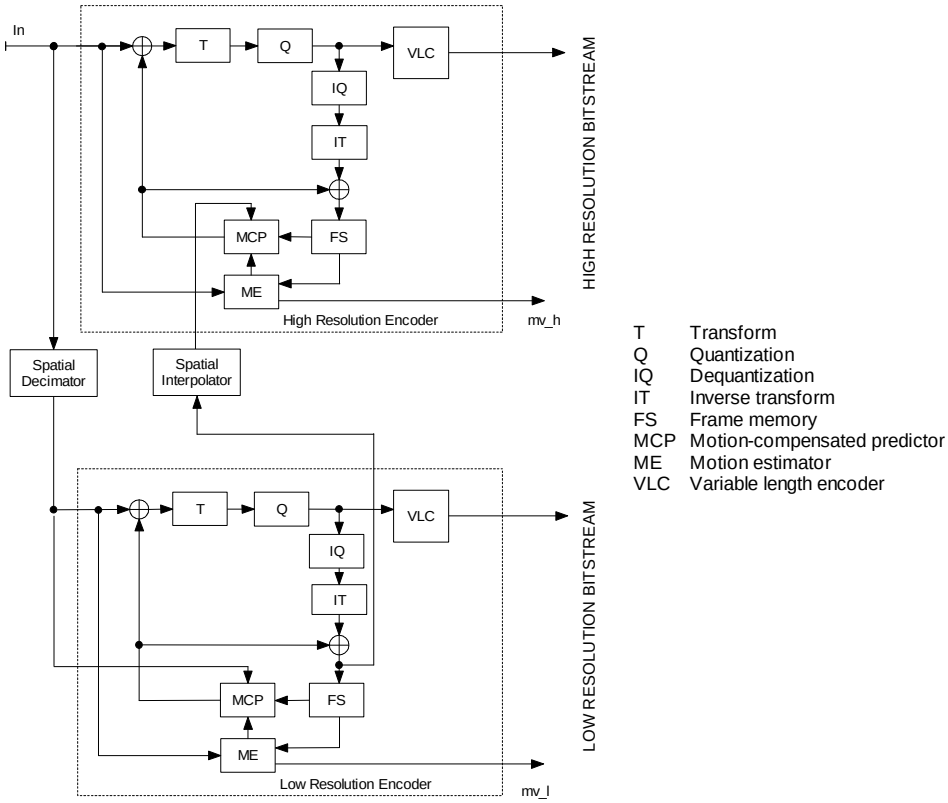


Figure 7: Two-layer scalable coder.

Such a coder produces two bitstreams corresponding to two different levels of both spatial and temporal resolution (Figure 7). Each of the coders has its own prediction loop with own motion estimation. Therefore the coder produces a bitstream that consists of four major parts:

- encoded transform coefficients for the low-resolution base layer,
- encoded transform coefficients for the high-resolution enhancement layer,
- motion vectors for the low-resolution base layer (*mv_l*),
- motion vectors for the high-resolution enhancement layer (*mv_h*).

The characteristic feature of this structure is independent motion estimation in both sub-coders resulting in optimum motion vectors estimated for both resolution levels. These motion vectors allow exact motion-compensated prediction in both layers. In the enhancement-layer high-resolution sub-coder additional reference frames can be used for both backward and forward prediction, i.e. interpolated frame from the current low-resolution base-layer frame and linear combinations (averages) of the current interpolated frame and temporal reference. For the latter, independent motion estimation can be performed aiming at estimation of the optimum motion vectors that yield the minimum prediction error for the reference being an average of spatial and temporal references.

Spatial and temporal decomposition

Good performance of spatio-temporal down- and upsampling is critical for good performance of the whole codec. Spatial decimation includes spatial low-pass filtering that prevents spatial aliasing in the base-layer low-resolution sequence. The choice of the filter trades off between high aliasing attenuation and short temporal response. The results of experimental comparisons prove the importance of the careful choice of the decimation-interpolation scheme. The system considered employs edge-adaptive bi-cubic interpolation. The technique is applicable to both luminance and chrominance.

Prediction modes in the high-resolution enhancement-layer sub-coder

Sophisticated intra- and interframe predictions are related to major performance improvements in the H.264/AVC coders. The enhancement-layer sub-coder employs additional prediction modes that exploit the current interpolated base-layer frame as the reference. Other modes exploit averages of temporal

prediction and spatial interpolation as references. These modes are carefully embedded into the mode hierarchy of the H.264/AVC coder thus obtaining the binary codes that correspond to the mode probabilities. The respective mode hierarchy is shown in Table 1.

The choice of the lowest-cost prediction mode plays the key role. The encoding scheme would reduce to simulcast when no interpolated reference macroblocks are used in the enhancement layer. In the other extreme situation, in the enhancement layer, no temporal prediction is used, and only interpolated base-layer frames are used for prediction of the enhancement macroblocks (like in MPEG-4 FGS). The latter situation is very unlikely because of the high efficiency of the H.264/AVC temporal prediction. Nevertheless the extreme situations are related to unsatisfactory coding performance. The spatial interpolation must be very efficient in order to avoid them. Good fidelity of the decimation-interpolation scheme results in reasonable probability that the reference sample block interpolated from the base layer leads to smaller prediction error as compared to the temporal prediction within the enhancement layer.

Frame type	Prediction modes
Intra (I)	1. Spatial interpolation from base layer (16×16 block size). 2. All standard intra prediction modes.
Inter (P)	1. Prediction (forward) from the nearest reference frame. 2. Spatial interpolation from base layer (16×16 - 4×4 block size). 3. Average of two above (1, 2). 4. Temporal prediction modes from other reference frames in the order defined in H.264/AVC specification. 5. All standard intra modes.
Inter (B)	1. Prediction (forward, backward and bidirectional) from the nearest reference frame. 2. Spatial interpolation from base layer (16×16 - 4×4 block size). 3. Average of two above (1, 2). 4. Temporal prediction modes from other reference frames in the order defined in H.264/AVC specification. 5. All standard intra modes.

Table 1: Prediction mode hierarchy.

Fine granularity scalability and sequence structures

Fine granularity may be obtained by use of splitting the data produced on any resolution level. In that way, the bitstream fed into a decoder may be well matched with the throughput available. It means that the decoding process exploits only a part of one bitstream thus suffering from drift. Always, only one of the bitstreams is split, usually the high-resolution one. Therefore only one of the bitstreams received is affected by drift that is related to the reconstruction errors which are accumulating during the process of decoding of the consecutive frames. In this bitstream, inserted are I-frames encoded with respect to interpolated lower-resolution frames. In fact, such frames are encoded similarly as P-frames but with simultaneous interpolated reference frames. The bitstream syntax is that of P-frames. Insertion of such frames does not affect performance so much but bounds propagation of drift errors to groups of pictures (GOPs) (see Figure 8). Moreover, higher percentage of B-frames also decreases the influence of drift. Of course, all layers may exhibit the GOP structure for some reasons as usually in MPEG-2 (Figure 9). The GOP length should be chosen in such a way that drift is acceptable in the worst case of lowest bitrate in the enhancement layer.

In order to improve coding efficiency the prediction scheme from Figure 9 has been proposed. Here, skipped are only B-frames (called BE-frames). Those B-frames that exist in both layers (BR-frames) are also used as reference frames.

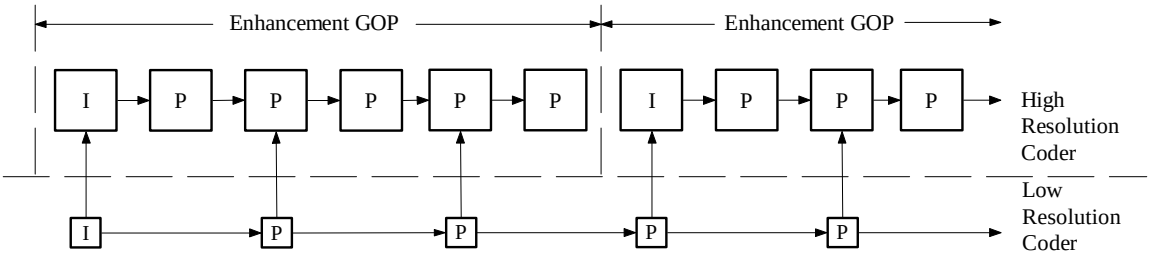


Figure 8: Exemplary structure of a video sequence: No B-frames and no GOP structure in the base layer. There exist GOP structures in the enhancement layer where the I-frames are encoded with respect to the interpolated I- or P-frames from the base layer.

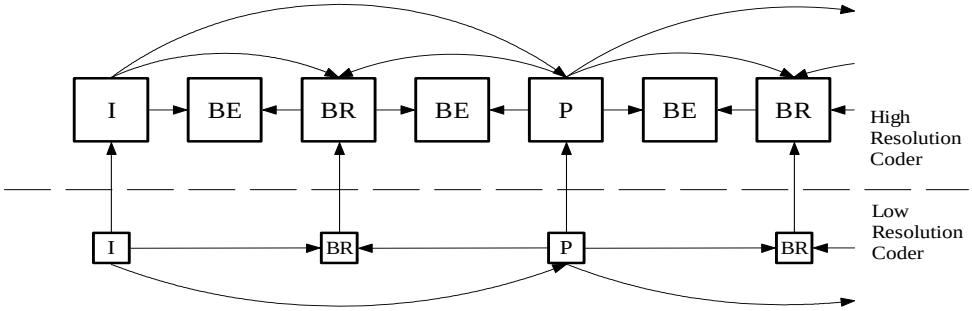


Figure 9: Exemplary structure of a video sequence: Number of B-frames is 75% of the total number of frames. Both the base layer and the enhancement layer divided into GOPs. The BR-frames used as reference frames for the BE-frames in the enhancement layer.

Metadata tools in scalable codecs
Some metadata coding tools developed within MASCOT project are compliant with the scalable coding scheme described above.

Unimportant textures may be synthesized in both base and the enhancement layer. Scalable textures fit particularly well to the scalable codecs. In the simplest case, the synthetic texture parameters are sent in the base layer. The enhancement layer sub-coder exploits these data as well.

In the H.264/AVC scalable coder, efficient choice of the reference frame can be made as described in the Section 3.3.7 The choice is made independently for two layers. The experimental results from Section 3.3.7 are valid for individual layers.

The rate control tool has been also considered in the MASCOT scalable codec. The analysis shown that the tool is applicable for scalable coders.

3.5 Prediction

3.5.1 General

Prediction is a crucial ingredient in video encoding which allows one to take advantage of the temporal redundancy inside an image sequence. The most classical prediction technique is motion-based prediction: in MPEG-2, a blockwise displacement map can be computed at the encoder side that contains the vectors governing displacements between two consecutive video frames. This map can then be transmitted and the decoder can use one frame and the displacement map to estimate the next frame. In such a scenario, prediction pays if the distortion of the predicted frame is low with respect to the coding cost of the displacement map. The purpose of the *Prediction* workpackage was thus to develop, enhance and test new prediction schemes to achieve greater coding efficiency for both MASCOT codecs. Prediction includes the estimation of deformation maps, the definition of a prediction operator, as well as the encoding of the deformation map.

3.5.2 Wavelet-based motion estimation

For hybrid coding schemes using DCT-based spatial redundancy reduction (like MPEG-2 MPEG-4, and H-26x), one always uses blockwise constant displacement maps estimated with block matching. The artifacts occurring at block boundaries are hindering the spatial decorrelation of error frames, because the block boundaries are also block boundaries of the block-wise DCT. For non-DCT-based video codecs however, a block-based motion compensation is not considered optimal.

In the MASCOT project, we first adapted an existing in-house motion estimation technique for video coding purposes. This motion estimation technique also provided an illumination change map to be used as an additional prediction parameter. The adaptation of the motion estimation scheme consisted in adding a flow-smoothing step to control tightly the entropy of the vector map. This modification proved to be useful, as was reported in Deliverable 4.1. However, it did not make the motion estimation and compensation competitive with existing techniques. In fact, several experiments showed that the wavelet-based motion prediction scheme did not provide satisfactory predicted images. The scheme was then rewritten to provide better motion fields. The major feature of the new wavelet-based flow estimation scheme is that the prediction is now explicitly maximized (instead of minimizing subband differences).

The new scheme can be summarised as follows: a multiscale pyramid of each frame is computed, as exemplified below. The pyramid of frames is computed with one filtering and a sequence of filterings and subsampling using an à-trous wavelet transform.

Figure 10: Redundant frame pyramid for motion estimation

Then a piecewise bilinear flow field matching a frame to the next one is found by iterative minimization. First the coarse scale images are matched, the resulting flow maps are refined to provide an initialization for the next iterative flow estimation. The multiscale structure of this flow estimation technique can be considered as a preconditioning step, leading to increased estimation speed. Figure 11 below shows a full refinement step: flow map refinement from scale 2 to scale 1 and iterations at scale 1. The first row displays a frame at the measurement scale taken from the pyramid. The second row displays the current flow map (note: for illustrative purposes, the flow does not correspond to the true motion for the *Paris* sequence, but to a motion where a rectangle is moving to the top right, and the rest of the frame does not move).

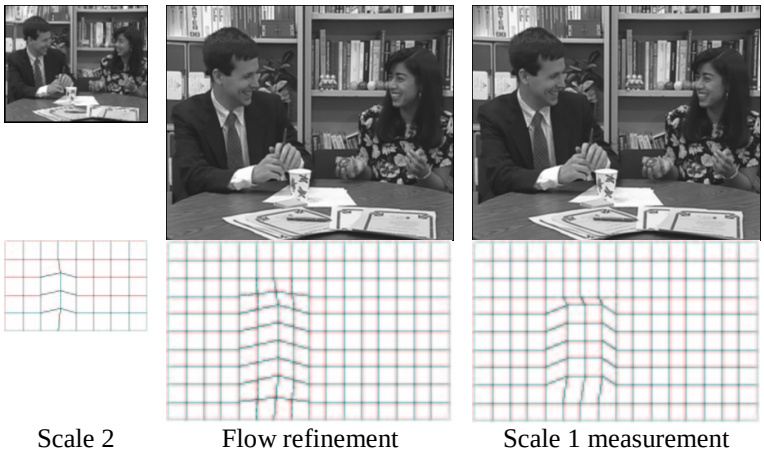


Figure 11: Example of a 2 scale measurement

This second generation motion estimation scheme works well, as is reported in Section 4.5.2 of the present report. It was thus integrated into the MASCOT wavelet codec. An additional prediction parameter was investigated, to see whether it could enhance coding efficiency: illumination. In addition to displacement, a predicted frame $I(t+1)$ was prediction from a frame $I(t)$ using a displacement vector v and an illumination factor L with the following formula:

$$I(t+1,x)=L\cdot I(t,x-v)$$

As was also reported in Deliverable D4.1, illumination change maps did not reduce the bitrate for a given distortion, so their use was abandoned.

3.5.3 Motion estimation and compensation in wavelet space[5]

In the framework of the MASCOT project, wavelet-domain motion estimation and compensation techniques have been developed. The theoretical framework for the transform characteristics used in the proposed methods is presented in [4] [5], while results have been demonstrated in [6] [7] [8]. This work is original in a sense that efficient generations techniques have been proposed for the overcomplete subbands and that it builds further on earlier sparsely proposed architectures for in-band wavelet video coding.

Figure 12 shows an example of a 2-D critically sampled discrete wavelet transform (DWT) of an image. It can be seen that in the wavelet domain, a multiresolution description of the image content is naturally achieved without any redundancy. In addition, due to the fact that the transform basis-functions are of limited region of support, the image structure is preserved, leading to the natural assumption that in-band ME/MC is feasible.

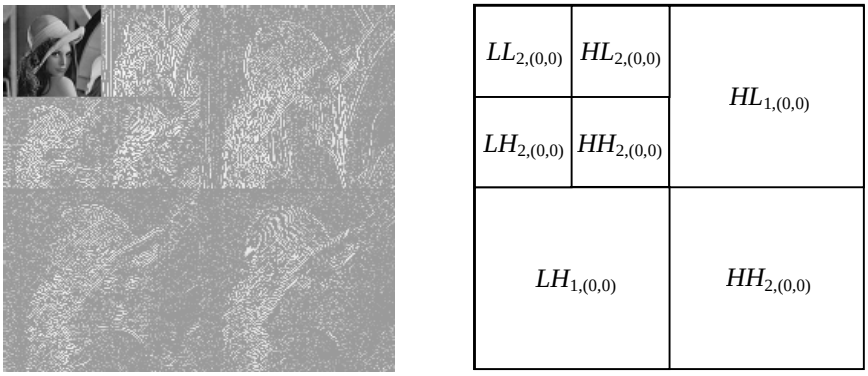


Figure 12: An example of a critically-sampled 2-D decomposition in two levels. The subband content is properly equalized for visualization purposes.

In addition to these findings, a potentially-effective in-band ME/MC technique would possess also the natural side-effect of resolution scalability, since if the motion estimation and compensation procedures are applied separately in a level-by-level fashion, the decoder of every resolution could effectively reconstruct the half-resolution or quarter-resolution frame by simply stopping at a specified decomposition level.

In order to achieve shift invariance to any integer spatial-domain shift, the overcomplete wavelet representation has to be utilized, where all the possible subsampling grids are retained during the transform production. Such a representation is shown in Figure 13 for the 1-D case of an input signal X.

[5] To enable scalable flow encoding in a 3D wavelet codec, techniques originating from VUB were intended to be adapted and used in the MASCOT codec. This work was discontinued because it was found that its use would have been incompatible with other components of the 3D wavelet codec developed by other partners of the project. Utilizing the in-band motion compensation would have led to three different and incompatible MASCOT codecs, and for rationalization, VUB accepted to discontinue its work on the topic in the MASCOT project, and to shift its attention to motion vector encoding in the 3D wavelet codec. This is explained in Deliverable 4.2. Work already done on the topic by VUB was reported on in detail in Deliverable 1.1, Appendix A.

As seen there, for the first decomposition level, a "classical" decomposition occurs which retains the type-0 polyphase components and produces the subbands A_0^1, D_0^1 (low and high-frequency subband respectively). Additionally, a "complementary" decomposition occurs which retains the type-1 polyphase components (complementary grid) and produces the subbands A_1^1, D_1^1. This process iterates for the following levels as seen in Figure 13, where the input in every case is the low-frequency subbands of the previous level. In this way, for each level l an overcomplete representation is produced which has a redundancy of 2^l (and 2^{2l} in 2-D) in comparison to the critically-sampled decomposition of level l, which simply consists of the subbands A_0^l and $D_0^i, i = 1,..,l$.

By using this representation, it has been shown that shift-invariance in the wavelet domain can be attained, meaning that any pixel motion in the spatial domain can be compensated to zero error in the wavelet domain.

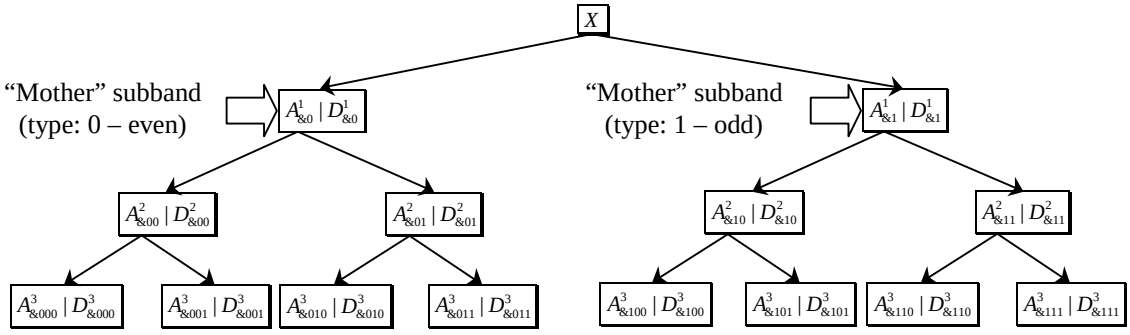


Figure 13: The overcomplete DWT representation of signal X *in 3 decomposition levels. The subband indices (subscripts) are denoted in binary form.*

In coding systems that utilize the DWT, the process of generating the overcomplete transform for the given input requires one additional step before the overcomplete construction itself. This step consists of the inversion of the critically-sampled DWT pyramid to signal X before the application of the process described in Figure 13. Hence this transform is called complete-to-overcomplete DWT (CODWT). It must be noted that this situation is different from the case described before for two reasons:

- The DWT coefficients are usually quantized and this leads to a reconstruction of a signal $\overline{X}$, where in general $\overline{X} \neq X$ due to quantization error.
- When resolution scalability is targeted, for the CODWT that starts from a decomposition in l levels, only a limited set of the subbands D_0^i is utilized, where i is a subset of l. This again leads to the conclusion that the reconstructed $\overline{X}$ will be different from X.

These two differences can be seen as the restrictions imposed when the presented method for shift-invariance in the wavelet-space is applied in video coding systems. Thus, the quantization in this case is performed by the utilized compression system and the CODWT construction under resolution scalability is applied to satisfy the case of a decoder that reconstructs the signal only up to a certain resolution, and hence does not have access to all the subbands.

The instantiation of the motion estimation and compensation process in closed-loop video coding architecture

For every resolution level, block-based motion estimation can be performed in a subband-by-subband manner or in a level-by-level manner. In the first case, motion vectors are produced for each subband, while in the second case one motion-vector field per resolution level is produced. The first approach is illustrated in Figure 14 for one decomposition level, where four different motion vectors are found for the four corresponding blocks located at a position (x,y) in every subband of the current frame. Notice that in this case every motion vector contains information about the in-subband position and phase as well. The predicted frame is simply constructed by compensating all subbands $SUB_{1,(0,0)}$, $SUB=\{LL,LH,HL,HH\}$ of the current frame with the use of the blocks in the overcomplete DWT of the reference frame that are pointed by the motion vectors. Hence, although the search utilizes all the phase components of subbands SUB of the reference, the predicted and the error frames are still critically-sampled.

If the above procedure is extended to a multilevel wavelet decomposition, the result is that an accurate ME is performed, which provides for every resolution level the in-band equivalent of a pixel-accurate spatial-domain ME. Following the rationale of spatial-domain techniques, more complex motion estimation schemes can be envisaged, which include quarter-pixel accurate ME by interpolation in the overcomplete DWT and multiple reference ME.

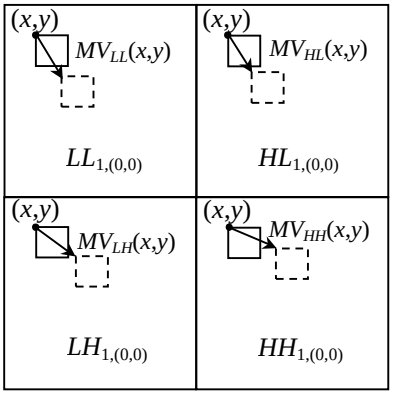


Figure 14: One-level decomposition with the block at position (x,y) *at every subband* $SUB_{1,r,c}$ *and the four corresponding motion vectors* $MV_{SUB}(x,y)$*, with* $SUB=\{LL,LH,HL,HH\}$.

Theoretical analysis of the complete-to-overcomplete discrete wavelet transform

The most challenging process from the combined algorithmic & implementation point of view in an in-band wavelet video codec is the design of the CODWT. It is evident that the functionality of this module is essential for the performance of the full coding scheme and on the other hand the overcomplete construction is a complex and costly procedure for implementation which requires careful handling. As a result, the theoretical properties of the CODWT construction were investigated. This led to a new algorithmic approach for the calculation of the overcomplete DWT of an arbitrary decomposition level *k*, starting from the subbands of the critically-sampled pyramid. This new framework for this calculation is presented in what is called the *prediction-filters* approach [4] [5].

From the theoretical point of view, the new framework is very interesting in the sense that it presents clearly the separate contributions of the subbands of the critically-sampled pyramid in the construction of any subband A_x^k, D_x^k of the overcomplete representation of level *k*. This is very useful under a resolution-scalable construction, since if a reconstruction up to a decomposition level *v* is desired, the high-frequency subbands $D_0^l, l \in [1,2,v], l \in \mathbf{Z}$ can simply be omitted from the proposed calculation scheme [4] [5], since they are not available.

From the implementation point of view, this new theoretical construction can lead in an efficient implementation for the CODWT module for encoders and decoders from two perspectives. The first is that important calculation benefits can be derived when the level-by-level construction is concerned, i.e. under the resolution-scalable mode. The complexity study of the proposed method for the typical case of biorthogonal point-symmetric filter-pairs reveals that in comparison to the conventional approach, the level-by-level derivation of the overcomplete wavelet transform is performed with a reduction of up to 56% in the total multiplication budget. Furthermore, for an $N \times M$ reference frame, the delay occurring with the conventional approach in the calculation of the subbands that correspond to the decomposition level *k* is proportional to $N \times M / 2$, while for the proposed method the delay is proportional to $N \times M / 2^{2k}$, under the same assumptions of system parallelism. These results indicate that, by using the proposed theoretical framework, one achieves scalable system-complexity in comparison to the conventional approach.

The second perspective is that in the decoder side where the motion vector information is a-priori known, the proposed method can provide an optimal tradeoff between memory and computation in the sense that for the compensation process, the actual implementation can balance between the two extremes. The first is the complete construction for the specified subbands of the overcomplete DWT (maximum memory requirements) and the second is a block-wise construction, where only the necessary information for the block-based in-band compensation is calculated by the CODWT in the decoder. This alternative will

require the minimum memory but the calculation budget increases since the specific areas for compensation are recalculated many times if there are many vectors pointing to the same area in the overcomplete DWT.

3.5.4 Motion vector coding

The developed motion vector coding technique can be applied to vectors of arbitrary accuracy, such as the ones produced by optical flow field estimation. The motion vectors are first fitted to the desired interpolation grid and then scaled to integer values. Thereafter, motion vector prediction is performed followed by prediction error coding. In the MASCOT project we have implemented a prediction scheme similar to the one used in the H.263 video coding standard and the prediction error is coded similarly to the way DCT coefficients are coded in the JPEG standard for still-image compression. The latter approach, the combination of spatial prediction schemes and JPEG-alike prediction error coding for the motion vectors, is original. In addition, also more complex alternative approaches for both the prediction schemes (cross-subband and temporal) and the prediction error encoding schemes (wavelet-based coding schemes, look-up tables, etc.) have been investigated.

The general setup of the proposed motion vector encoder is shown in Figure 15. The motion vectors are first quantized to integer values. Thereafter, motion vector prediction is performed, followed by prediction error coding.

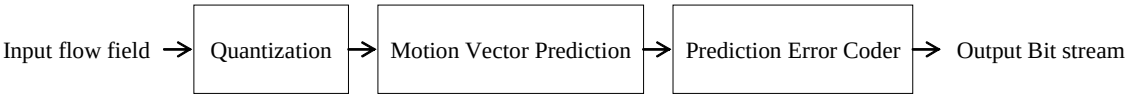


Figure 15: Design of the motion-vector encoding.

The motion vectors are predicted similarly to the motion vector prediction in H.263. Each vector is predicted by taking the median of a number of neighboring vectors. The neighboring vectors that are considered in the default case are shown in Figure 16.

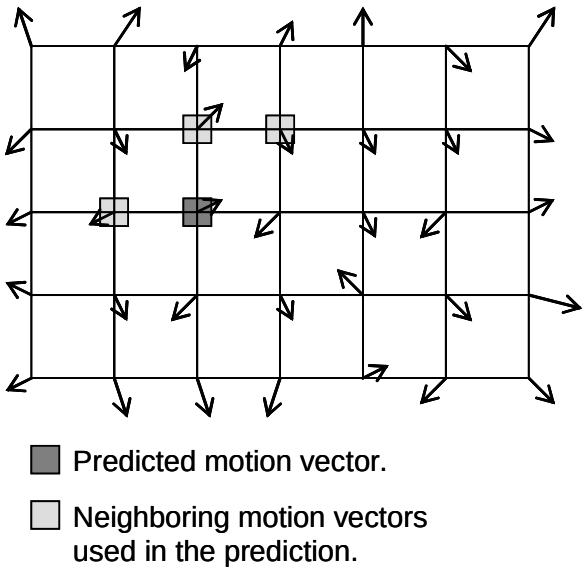


Figure 16: The context-based prediction of a motion vector based on the neighboring vectors.

In the final step, the prediction error vectors obtained after motion vector prediction are entropy coded using adaptive arithmetic coding. The horizontal and vertical components of the prediction errors are always coded separately. In our implementation, only prediction error components belonging to the interval {-511,511} can be represented. However, the implementation can be easily altered to handle larger values should the need arise. The prediction error components are coded similarly to the way DCT coefficients are coded in the JPEG standard for still image compression. For each component value, a symbol representing the interval it belongs to is coded first (Table 2). Thereafter, the offset of the

prediction error component within the interval is coded. A distinction is made between positive and negative components. For positive components, the value that is coded is equal to the prediction error component. For negative components, the algorithm encodes the sum of the prediction error component and the absolute value of the lower bound of the interval it belongs to. It is obvious that no offset is coded for interval 0. One model of the arithmetic coder is used to code the interval-index. For each interval, a different model is used to encode the offset values. The offset value is coded differently for intervals 0 to 4 than for intervals 5 to 9. In the first case the different offset values are directly coded as different symbols of the model. In the second case, the model only allows two symbols 0 and 1, and the offset value is coded in its binary representation.

Interval/value	Symbol
0	0
{-1} ∪ {1}	1
[-3,-2] ∪ [2,3]	2
[-7,-4] ∪ [4,7]	3
[-15,-8] ∪ [8,15]	4
[-31,-16] ∪ [16,31]	5
[-63,-32] ∪ [32,63]	6
[-127,-64] ∪ [64, 127]	7
[-255,-128] ∪ [128 255]	8
[-511,-256] ∪ [256,511]	9

Table 2: The intervals used in the arithmetic encoding of the prediction error of the motion vectors.

3.5.5 Long term global motion compensation

Global motion compensation aims at reducing the bitrate of a video sequence without reducing its visual quality by describing the motion of the scene background with a small number of motion parameters. This feature is included in MPEG-4, but not in H264/AVC because of its limited efficiency. Long term global motion compensation consists in building online a superresolution mosaic by combining different background images to a single reference image using a subpixel flow map of the background.

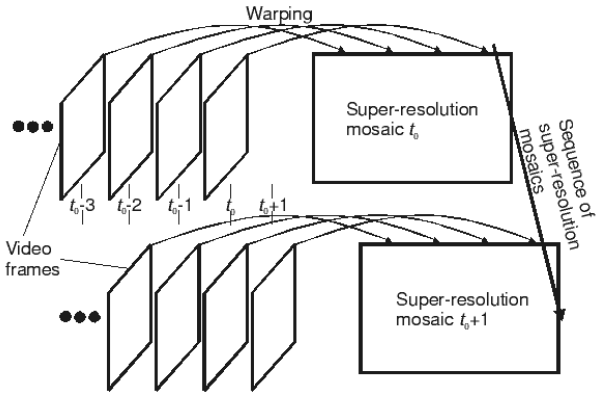


Figure 17: Superresolution mosaic construction by accumulation

The background mosaic is incrementally built by accumulation of pixels of the background images from different frames. The following figure shows the evolution in time of the superresolution mosaic of the background for the *Stefan* sequence.

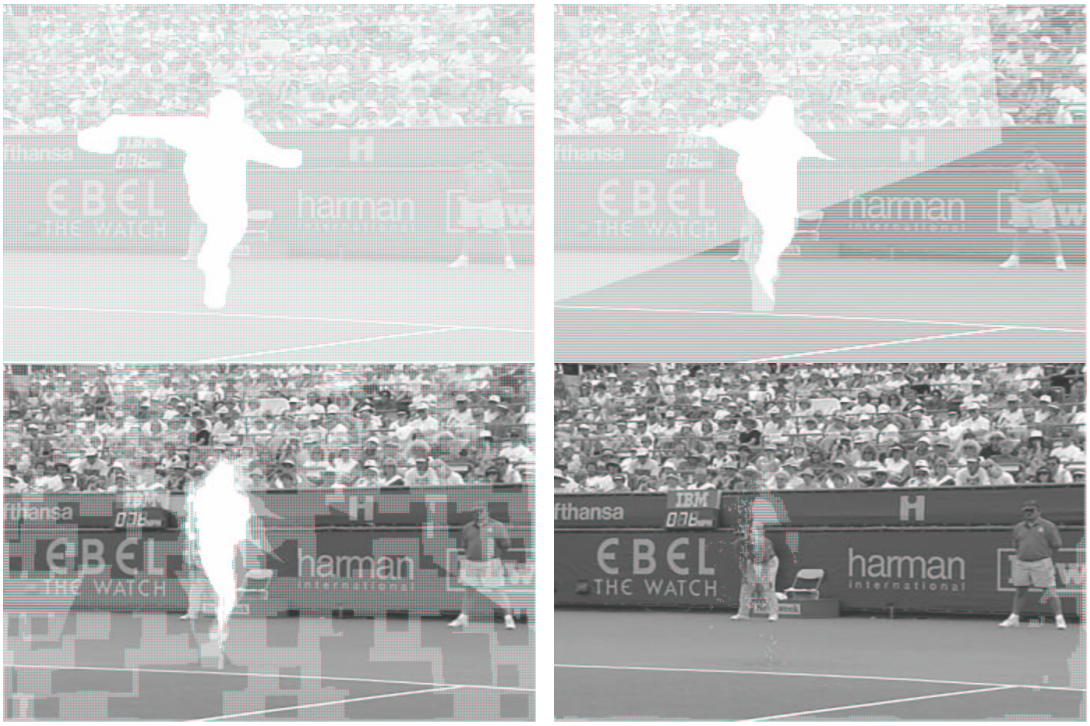


Figure 18: Accumulated superresolution mosaic at frames 1, 5, 15, 49.

As shown in the results section, the sequence encoded with long term GMC often has less aliasing than the original video sequence, and substantial gains in bitrate. The long term global motion compensation is described in substantial detail in Deliverable 4.3.

3.5.6 Long term motion prediction using graph-matching

In this implementation of long term prediction, long term means that a reference frame and a predicted frame are far apart in time, in contrast to the above paragraphs where a long term prediction meant accumulation over several frames of pixels to build a superresolution mosaic of the background.

A graph matching method was developed and described in Deliverable 4.1 to provide a framework for long-term prediction. The purpose of this long-term graph-matching technique was to reduce the bitrate increase occurring at shot transition. The first image of a new shot being not related to the previous video frame, it has to be essentially encoded as an I-frame. The purpose of the present long term motion prediction was to encode this first frame of a new shot in a predictive way, using a past “long term” reference frame possibly several hundred frame before. The test framework included a graph matching technique to select among a number of reference candidates a best reference frame, and also to perform a long range motion estimation between the frame to be encoded and the long term reference frame.

4 Project Results and Achievements

4.1 Introduction

The goal of the MASCOT project was to develop new video coding schemes and tools that provide both an increased coding efficiency as well as extended scalability features compared to technology that was available at the beginning of the project. Towards that goal the following tools would be used:

- metadata-based coding tools;
- new spatiotemporal decompositions;
- new prediction schemes.

The initial goal was to develop one single codec architecture that was able to combine all new coding tools that were foreseen when the project was formulated. During the first six months it became clear, however, that this would limit us in the selection of the new tools. Therefore the consortium decided to use two[6] codec frameworks within the project, a standard hybrid DCT-based codec and a 3D wavelet codec, which together were able to accommodate all tools developed during the course of the project; see also Section 3.2

Available codecs have been used as starting point for both. For the hybrid DCT-based codec, we have used the H.264/AVC codec developed by the Joint Video Team (JVT) as their framework for the hybrid DCT-based codec[7]. The second baseline codec architecture used in MASCOT is a 3D wavelet codec. Initially this was a codec provided by PRF, who provided the software and would also coordinate the integration of tools from other partners. Unfortunately, PRF withdrew from the project on September 1, 2002. The consortium decided to switch to another wavelet codec platform, which was publicly available and used within the MPEG community. We will henceforth refer to this codec as the *Woods codec*, after its originator John W. Woods (Rensselaer Polytechnic Institute). These two choices have the clear advantage that MASCOT results can be directly compared with results from the respective original codecs and that improvements can be directly contributed. The main objective of the project was then to improve the performance of these two baseline codecs by replacing certain tools or adding additional ones.

In the previous chapter we described the new tools that have been developed within the MASCOT. In this chapter we discuss their impact on coding efficiency and functionality. This chapter has been subdivided into four sections according to the four technical workpackages of the MASCOT project. In Section 4.6 we report on the MPEG liaisons within MASCOT and in Section 4.7 we discuss the contribution of this project to the EU policies.

4.2 Results of WP1: Architecture specification, codec integration, test and demonstration

The main objective of this central workpackage was to coordinate the software integration of the MASCOT codecs. As explained in Section 3.2, two different frameworks have been selected as baseline for the tools to be developed in MASCOT.

The first major task in WP1 was then to specify the architecture of the MASCOT codecs, i.e. to describe how all the developed tools will be integrated into the overall framework. The architecture of the MASCOT codecs is described in detail in D1.1. After having specified the architecture, the work in WP1 concentrated on the software integration itself which is described in detail in D1.2.

Then the task was extensive testing of the integrated MASCOT codecs in direct comparison to the state-of-the-art in video coding, which was defined by the two baseline codec frameworks. These tests have been performed similar as MPEG core experiments. A detailed collection of all coding results in MASCOT can be found in D1.3. Significant improvements compared to the baseline codecs have been achieved. Some of these results can be found in the following sections. The descriptions of the developed

[6] VUB also provided a full wavelet video codec, which applies in-band motion compensation. Although the architecture was very promising, it was not clear how the technology could be combined with the architecture provided by PRF. VUB decided to shift attention within MASCOT towards the PRF codec and to continue the development of their own codec in another project.

[7] The JVT is a joint effort of ISO/MPEG and ITU-T/VCEG to develop a new video coding standard (MPEG-4 part 10 and ITU-T H.264), which has to represent the state-of-the-art in video coding.

tools are contained within the deliverables of the other technical workpackages. Since the two integrated MASCOT codecs directly extend and improve the state of the art in video coding, they can be regarded to be among the most advanced video codecs available in the world, which is the major outcome of the MASCOT project.

Concerning the hybrid codec not all of the integrated tools are compatible with each other and some combinations are not useful with regard to coding efficiency. Some of the tools are very specific e.g. to certain types of metadata of content. Also the application of one tool with the respective modification to the standard H.264/AVC framework can decrease the efficiency of other tools or even make it impossible to use some other tools. It is also not always the case that the gain of a useful combination will be additive. These cross-dependencies have been investigated in more detail. Table 3 gives an overview of compatibility of tools (first x/o) and the usefulness of combinations (second x/o).

Faces is a very special tool for face coding that performs some special preprocessing. It is more a standalone application and not compatible with the other tools. *Scalability* adds an enhancement layer to any base layer, which can contain any combination of tools. It is therefore compatible to all other tools (but *Faces*) and the combinations are useful.

The *Video Segment Shuffling* tool can be combined in principle with all other tools except the *Transitions* tool. The *Transitions* tool needs to keep the visual frame order of the transition so it can not be combined with tools that may change the visual order (such as the *Video Segment Shuffling*). For the remaining tools, there should be some cross-influence between them. For instance, the *Selection of Reference Frames* also tries to exploit the temporal redundancy of the bitstream and therefore combining both tools will not obtain pure additive results. The same can be said for combining *Video Segment Shuffling* and *LT-GMC*. The latter tool requires a motion estimation performed on the video sequence. *Video Segment Shuffling* modifies the visual order and therefore the motion estimation step. The combination is useful but the overall gain is not additive.

The *Texture* tool is compatible in principle to all remaining tools. The combination with transitions is useful if the tools are used exclusive, i.e. only *Transitions* during the transition and only *Texture* for the other parts of the sequence. In this way the gains are additive. Applying the *Texture* tool during a transition is not possible due to the inherent nature of texture analysis and synthesis. Combining texture with *Reference Frames* or *Rate Control* is not recommended, since these tools either reorder the frames in the multi-frame buffer or modify the GoP structure, but the *Texture* tools relies on predefined settings of both. The *Texture* and *LT-GMC* tools are very similar in spirit. They recognize regions in the images that are encoded without transmission of prediction errors. They are compatible and the combination is useful, however, the gain is not additive.

Transitions can be combined with all remaining tools in the way it is used only during the transition, while the other tools are used for the remaining frames. The effects are the same as described above for the combination with *Texture*.

The combination of *Reference Frames* and *Rate Control* is useful. Both perform independent optimizations of the encoding. However, there is some cross-influence, so the gains are not completely additive. The combination of *Reference Frames* and *LT-GMC* is not useful, since the efficiency of *LT-GMC* relies on a regular temporal order of the frames in the long-term buffer.

The combination of *Rate Control* and *LT-GMC* is useful, however, the gain is not completely additive, due to the cross-influence.

For the combination of 3 or more tools, the pair wise evaluations above and in Table 3 have to be considered jontly.

	Texture	Scalability	Faces	Transitions	Reference Frames	Rate Control	LT-GMC	Video Segment Shuffling
Texture		x-x	o-o	x-x	x-o	x-o	x-x	x-x
Scalability			o-o	x-x	x-x	x-x	x-x	x-x
Faces				o-o	o-o	o-o	o-o	o-o
Transitions					x-x	x-x	x-x	o-o
Reference Frames						x-x	x-o	x-x
Rate Control							x-x	x-x
LT-GMC								x-x
Video Segment Shuffling								

Table 3: Compatibility of tools.
First entry: possible in principle – Second entry: useful combination
x = yes, o=no

Finally the results obtained within the MASCOT project have been demonstrated to a broad public at the Picture Coding Symposium held on April 23-25 in St. Malo, France, which is one of the major world wide conferences on image and video coding. An exhibition was organized showing the improvements achieved in MASCOT.

4.3 Results of WP2: Metadata-based encoding tools

In the following, the main video coding results obtained using metadata are presented.

4.3.1 Texture

The integrated video codec was tested using well-known test sequences, such as *Flowergarden*, *Canoe* and *Concrete*. The following set-up was used for the H.264/AVC codec: QP=16, 3 B-frames, 1 reference frame, CABAC (entropy coding method), rate distortion optimization, no interlace, 30Hz frame frequency.

The results obtained for *Flowergarden* sequence were of satisfactory subjective quality.

Figure 19 depicts the result obtained for frame #5. The difference signal is nearly zero in the sky region, while it is quite significant in the flower region. However, there is hardly any difference visible when comparing the synthesized and the original sequences. This shows that PSNR is not a suitable measure for this new coding approach, since it is not a classical waveform coding scheme that tries to reconstruct the signal as good as possible. Such a perceptual video codec can only be evaluated subjectively. Thus no PSNR comparisons were done in the experiments.

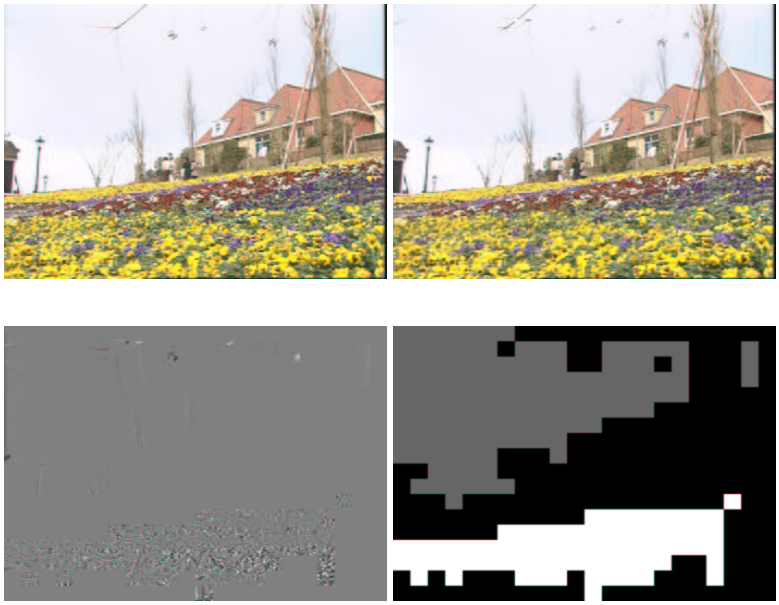


Figure 19: Coding result for Flowergarden*. From left to right and top to bottom: original frame, frame with synthesized texture regions, difference signal (gain factor 3), conservative motion compensated segmentation mask*

Table 4 shows the bitrate savings obtained for each of the test sequences. The visual quality at these bitrate savings was comparable to the quality of the decoded sequences using the standard codec. The largest savings were obtained for the *Concrete* sequence, while the smallest savings were obtained for the *Canoe* sequence. The bitrate savings obtained by using the approach lie between 10% and 23% using manually generated masks and 5% to 18.1% using automatically generated masks.

	Flower-garden	Concrete	Canoe
H.264 JM2.1 [kb/s]	1446.24	2950.32	1958.75
Bitrate savings with manually generated masks[%]	20%	23%	10%
Bitrate savings with automatically generated masks[%]	5.65%	18.1%	5%

Table 4– Bit-rate savings per test sequence

Conclusions

A new approach to video coding using texture analysis and synthesis was presented. Video scenes are classified into relevant and irrelevant texture regions. Irrelevant texture regions are detected by a texture analyser and regenerated by a texture synthesizer. Texture analysis and synthesis were integrated into the reference software of the H.264/AVC video codec. First the system was tested using manually generated masks to evaluate the maximum potential of the approach in terms of bitrate savings. In this case bitrate savings between 10% and 23% were obtained depending on the sequence, whereas using automatically generated masks led to still substantial bitrate savings between 5% and 18.1%.

The interplay between texture analysis and synthesis will be further optimised by integration into a full analysis/synthesis feedback loop. Further a more precise analysis is required to avoid synthesizing texture

regions with low texturization. The bitrate savings may be very low or negative in such cases as the ratio of the number of bits spent for the side information to the number of bits spent for the conventional coding of such textures may be very close to one or greater than one.

4.3.2 Face coding of video sequences

Results are presented and compared to H.264/AVC (JM 2.1 implementation). Two test sequences have been used with an image size of 56x46 pixels (for MPEG-7 compatibility) and 25 frames/sec. Figure 20 shows that for the *Miss America* sequence, the proposed face coder, obtains an efficiency comparable to AVC 2B (IBBPBBPBBP…) and better results than AVC 4B (IBBBBPBBBBP…) or AVC 0B (IPPPPPPP…) for a PSNR in the 29.5-31.5 dB range.

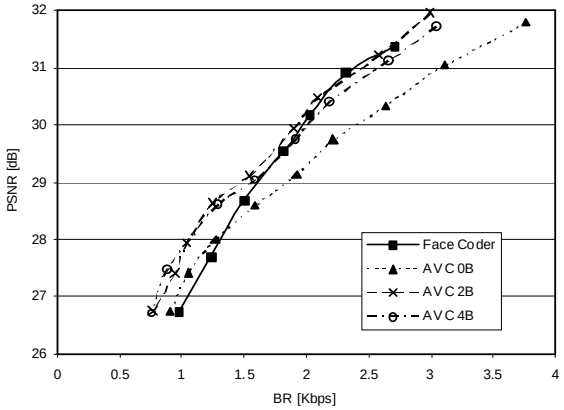


Figure 20: PSNR of coded Miss America *sequence vs. BR*

Table 5 shows the coding results for the *Miss America* sequence at a PSNR of 31 dB.

Esquema	Avg. BR [Kbps]	Avg. PSNR [db]	% A-PCA o B Frames
FACE CODER	2.323	30.90	79.83%
AVC 0B	3.108	31.04	0 %
AVC 2B	2.579	31.22	66.7%
AVC 4B	2.652	31.13	80%

Table 5: Coding results for Miss America *at a PNSR ≈ 31 dB.*

Figure 21 presents visual results for the *Miss America* sequence.

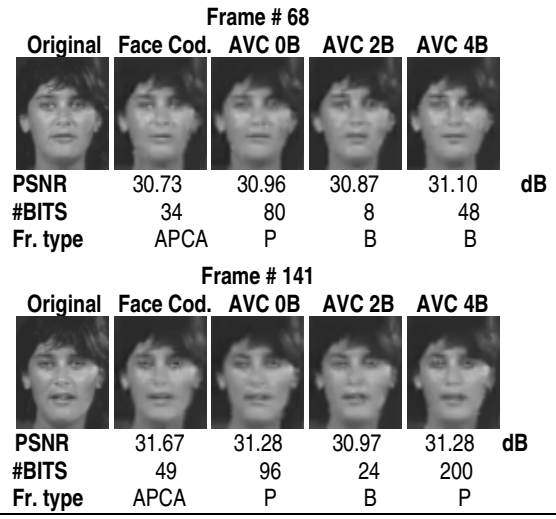


Figure 21: Visual results of the face coder for Miss America *sequence of Table 5.*

Conclusions
It can be generally stated that the MPEG-7 face descriptors are not useful for image and video coding. The provided results confirm this statement. However, they can be embedded in the bit-stream of a new video codec intended for face images to support searching and browsing functionalities.

Regarding the new codec presented the following conclusions can be drawn:

- When the face is located and extracted correctly, better results than H.264/AVC are obtained in the range of 1 – 2.5 kbits/sec. On the contrary, if the face is incorrectly located, the results are similar to AVC 0B (IPPPP…).

- A-PCA frames, those coded using only A-PCA coefficients, offer great bit-rate savings without the limitations of B-frames. Moreover, A-PCA frames do not require motion compensation and have an execution time ten times faster on average. However, a more complete analysis considering all possible optimizations should be done to draw a more definite conclusion about computational complexity.

4.3.3 Mosaics and key-regions for coding

The coding of video using mosaics was discontinued in the framework of the MASCOT project. The expected improvement of coding efficiency of this tool has been difficult to estimate due to the workload implied in the development of a metadata-based coding using mosaics. There is also a need for an automatic mosaic metadata extractor that could be used in the improved codec. Therefore, due to the lack of reliable improvements that this activity could offer, this activity was discontinued.

4.3.4 Video shot transitions coding

The results of this tool have shown promising gains when using metadata transition to improve coding. For the results, two QCIF sequences have been used. Both sequences contain a dissolve transition between the common *Foreman* and *Akiyo*. The dissolves have been manually created so the transition weights are known and can be included in a *VideoEditingDS* metadata descriptor about the transition. The difference between the two sequences is that the first transition lasts 4 frames, while the second transition lasts 40 frames. The total duration of the first test sequence is 13 frames and contains a transition between frames 5 and 8. In order to obtain a fair comparison between default H.264/AVC and H.264/AVC with motion search and compensation with MPEG7 dissolve weights, we started analyzing different GOP structures and selected the optimum for each case. All bitrate results were done using all search modes available and for the following picture qualities: QI=15, QP=16, QB=17.

For short transitions, Figure 22 gives the PSNR/bit-rate plots within the transition using standard H.264/AVC or the metadata enabled codec.

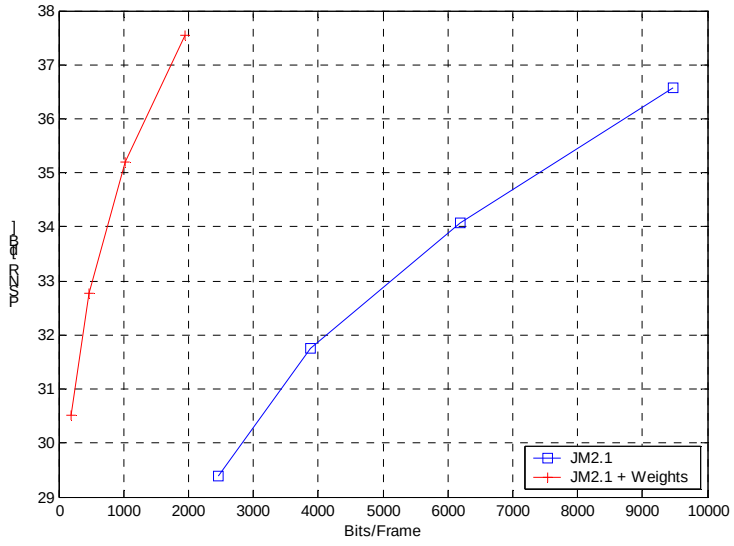


Figure 22: RD curves within the transition using standard H.264/AVC and the metadata enabled codec for a test sequence with short transitions

For long transitions Figure 23 gives the PSNR/bit-rate plots within the transition using standard H.264/AVC or the metadata enabled codec.

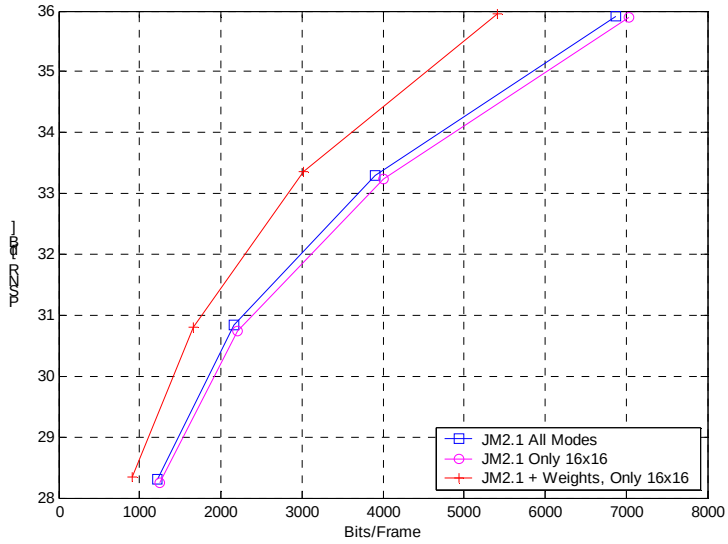


Figure 23: RD curves within transition with or without weights for long transitions.

Conclusions

Results have shown that using metadata for the coding of video transitions increase the coding efficiency. In the case of short transitions (up to 6 or 7 frames) the bit-rate gains can be up to 80%. In the case of long gradual transitions the bit-rate saving are smaller due to the need of including references within the transitions. In the later case, bit-rate savings of 15% are expected when coding video transitions using metadata.

4.3.5 Long term selection of reference frames

In order to illustrate the results of using metadata in the selection of reference frames, several sequences have been coded and compared using the standard H.264/AVC (version 2.1) and the MASCOT coded. Figure 24 shows the bitrate needed to encode a news sequence extracted from the MPEG-7 database. A normal news sequence usually involves the same presenter intercalated by different shots. It is expected that in that case the metadata help the encoder to select frames from the previous shot where the presenter appears and therefore improve the prediction when coding the current presenter shot.

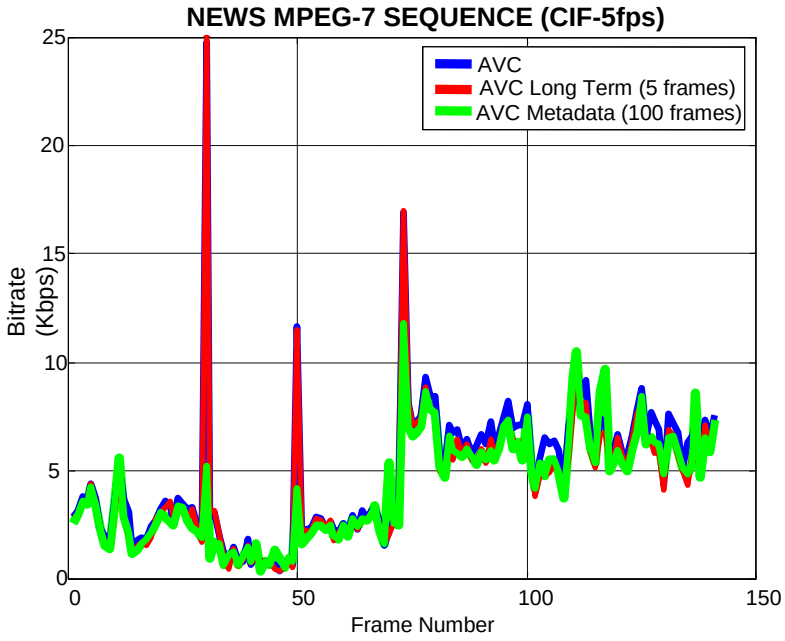


Figure 24: Metadata use in long-term temporal prediction for the MPEG-7 news sequence. The bitrate needed to encode 150 frames is shown. In this analysis, 5 reference frames are pre-selected using 100 previous candidates.

The previous figure shows the results when coding a news sequence from the MPEG-7 database. A pre-selection of 5 reference frames over 100 previous frames is performed using the color layout similarity measure. The blue line shows the bitrate needed when coding the test sequence with H.264/AVC using only one reference per frame. The red line shows the improvement using long-term prediction using 5 previous frames as reference frames (already included in the standard H.264/AVC). Finally the green line shows the bitrate needed color layout metadata to pre-select these 5 reference frames. It can be observed that usually at the beginning of shots, the metadata pre-selection is able to select frames from past similar shots and therefore the overall prediction error is reduced. In this example, simple long term temporal prediction using 5 references frames reduces prediction error by 7.5%. The next figure shows a formal results by indicating the rate-distortion curve of this technique. The same sequence is coded at various bitrates using the standard H.264/AVC codec (in red) and the MASCOT codec (in green). In the example, pre-selection of reference frames using color layout metadata can achieve up to 12.2% reduction in bitrate for the same visual quality.

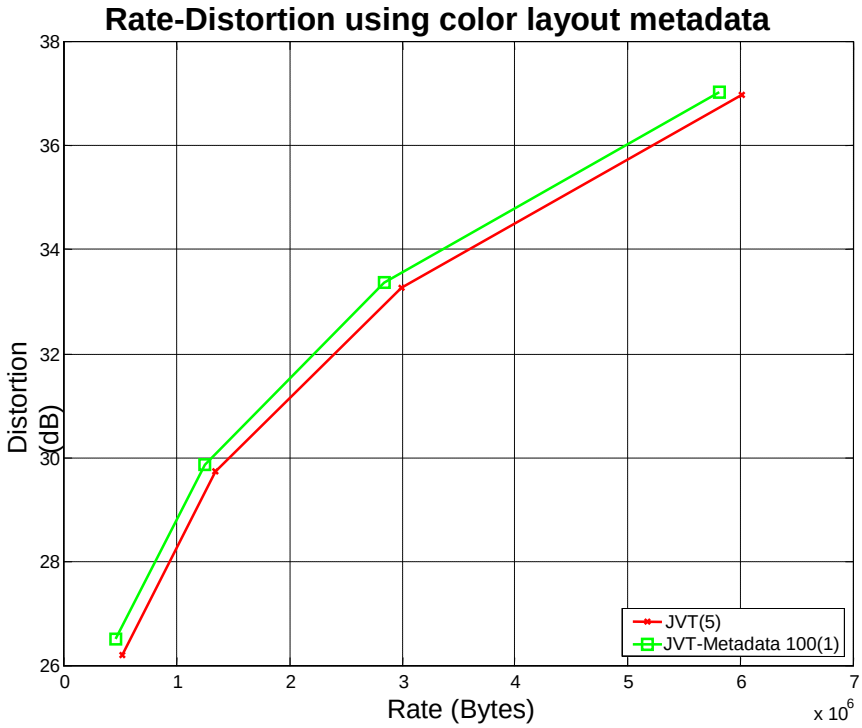


Figure 25: Rate-Distortion comparison of JVT2.1 vs JVT using metadata for the MPEG-7 news sequence. In this example, 4 reference frames are pre-selected using the 100 previous candidates.

Results show that the standard long term temporal prediction can be improved using metadata to pre-select the reference frame buffer. As existing metadata has been created for search and retrieval scenarios, some metadata descriptors are extremely efficient when searching for similar frames. This property can be exploited in order to use the best possible reference candidates instead of using only the previous in time. It can be expected that this coding strategy can obtain up to 10% bit-rate reductions compared to the standard H.264/AVC codec at the same visual quality.

4.3.6 Rate control using metadata

Metadata information is extracted from the content in order to provide indexing, search and retrieval capabilities. It provides information of the content that can also be used to select encoding parameters that are difficult to select otherwise. Included in this technique, motion descriptors can be used to select the IPB structure of standard codecs. Table 6 shows some results using motion descriptors for rate control. The motion activity descriptor includes a simple descriptor related to the amount of motion present between two frames (an integer value between 1 to 5). This descriptor is computed for all five shots of the test sequence (second column shows the mean value for all frames in every shot). Note that values closer to 1 correspond to shots with a small amount of motion while values closer to 5 correspond a large amount of motion present in the shot. The third column shows the mean bitrate (in Kbps) needed to code the shot using the common "2 B-frames between P-frames"' GoP structure in the H.264/AVC codec. The final two columns show the number of B-frames and the bitrate when coding the same shots using the best possible IPB structure (best in the sense of minimum bit-rate for the same quality).

Shot Number	Motion Activity	Bitrate (2 B-frames) Kbps - AVC	Number B-frames (best selection)	Bitrate (best selection) Kbps
1	2.8	61.4	2	61.4
2	4.2	97.4	2	97.4
3	1.1	37.9	3	35.5
4	4.3	11.2	1	10.9
5	4.2	79.8	1	77.9

Table 6: Results for rate control using motion descriptors

It can be seen that, even a simple descriptor such as the motion activity, can be used to have a preliminary idea of the amount of motion of the scene and therefore to control the encoder parameters accordingly. Shot 3 shows for instance how small values in the motion descriptors correspond to low motion. In this case, more B-frames can be introduced in the GoP structure thus reducing the bitrate needed. However, results also show that this simple metadata descriptor can signal the presence of high motion even if the codec is able to motion compensate it (see for instance shot number 4). To compensate this difficulties other metadata descriptors such as the frame displaced difference can be included in the measures in order to obtain more robust estimations of the amount of motion.

The next figures show the rate-distortion curves and relative bitrate saving obtained when coding a news QCIF sequence using the rate control tool. Results show up to 5.6% bitrate savings on this particular sequence. When comparing these results with other GoP structures of the baseline H.264/AVC codec instead of B=2 (such as B=0 or B=1) similar results are obtained as the bitrate is compensated over the entire sequence.

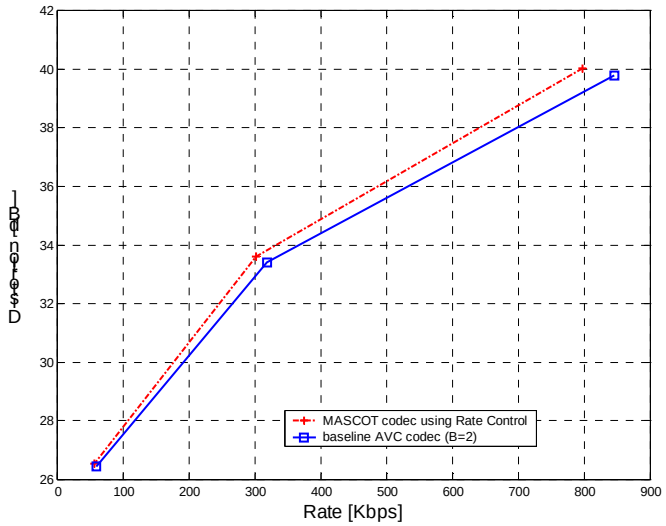


Figure 26: Rate-distortion curves comparing the Rate Control algorithm and the baseline H.264/AVC codec for the sequence News11.

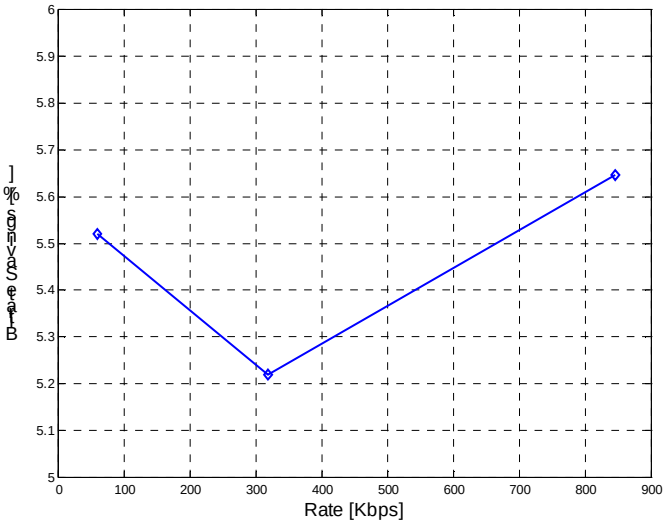


Figure 27: Relative bitrate savings of MASCOT hybrid codec with Rate Control compared to baseline H.264/AVC codec at the same visual quality for the test sequence News11.

Conclusions
Metadata descriptors can carry information that may be useful to select different encoding parameters. Current encoders select frame quality factors or IBP structures based on empiric assumptions about video sequences. Metadata descriptors such as motion activity descriptors or texture descriptors are valid candidates for selecting between a predefined set of IPB structures or quality factors. Bitrate savings up to 5% are expected when using the rate control tool on standard sequences.

4.3.7 Video segment shuffling

Video segment shuffling uses the MPEG-7 *Segment DS* metadata in order to fully exploit the temporal redundancy of video sequences. The sequence to be coded is re-organized so similar frames are grouped in the shuffled sequence. Next figures show the results when coding a sequence using the video segment shuffling tool. Three different test are compared against the H.264/AVC codec (in blue) using video segments of 1 frame, 8 frames and using entire shots. The best results are obtained when shuffling the sequence using complete shots with bitrate savings up to 8.5%.

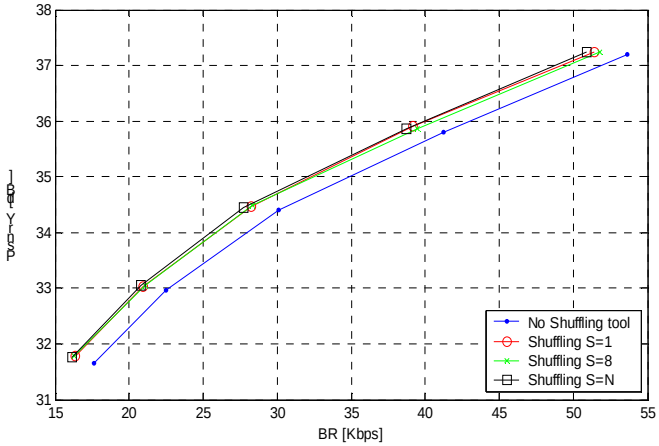


Figure 28: Rate-Distortion curves using the video segment shuffling tool and B=1 (IBPBPBP) GoP structure for the sequence Geri.

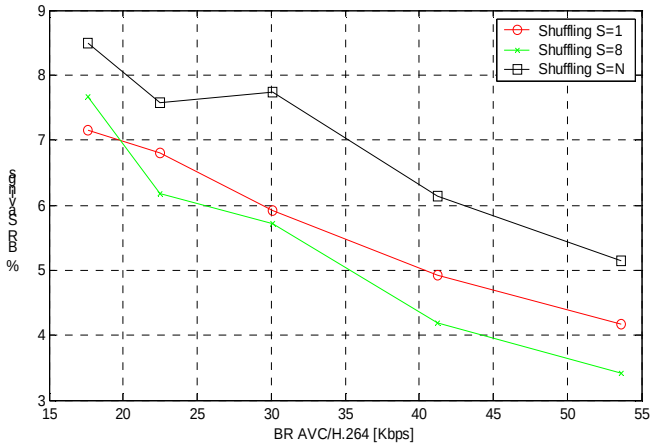


Figure 29: Relative bitrate savings using the video segment shuffling tool and B=1 (IBPBPBP) GoP structure for the sequence Geri.

Conclusions
Segment shuffling achieves relevant bitrate savings in sequences that present scattered, short and similar shots. This is common in TV programs like interviews or shows that switch over several camera perspectives. In these cases, bitrate savings up to 8.5% are expected. In other sequences, the efficiency gain is relatively small or nothing; but segment shuffling could be a convenient way of storing a sequence, i.e. to ease the access through a BPT index.

4.4 Results of WP3: New spatio-temporal decompositions

4.4.1 General

Two different codec architectures have been used as baseline codecs in MASCOT, a hybrid DCT-based codec and a 3D wavelet codec. A detailed description of the architecture of these codecs is given in deliverable D1.1, and the integration of these codecs is described in deliverable D1.2. The partners have developed new coding tools in the framework of these baseline codecs. The tools developed in WP3 are aimed towards an improvent of the performance of the two codecs and to provide extended (scalability) functionalities by exploiting new spatio-temporal decompositions. In Subsections 4.4.2 and 4.4.3 we discuss two tools that have been used for the 3D wavelet codec and in Subsection 4.4.4 a tool for the hybrid DCT-based codec.

4.4.2 Temporal decompositions

Through extensive simulations on classical test sequences in [9] and [15],[16] we have shown the improvement over the reference 3D wavelet video codec (i.e., Woods codec using the Haar temporal filter, hereafter called baseline wavelet codec) brought by longer temporal filters. In the experimental evaluation, the average PSNR was used as an objective evaluation metric. Note that these distortion values are given at different bitrates. Moreover, in our scalable codec, the granularity of the bitrate measure is at the bit level.

First, we have shown that an intelligent strategy for taking into account the redundancy between motion vector fields at different temporal levels may lead to important reductions in the number of bits allocated to motion, and accordingly to improved coding efficiency. Moreover, interesting tradeoffs are possible between algorithmic complexity and coding performances, useful for applications where only a reduced computational load is allowed.

Secondly, we have shown that a "sliding window" implementation of our lifting temporal filtering formalism leads to a better energy concentration and accordingly, to higher coding performances with long temporal filters than with Haar filters. The results of our tests for coding efficiency have shown that:

- Haar temporal filtering without ME performs well at very low bitrates (<150 kbits/s).
- Haar temporal filter with ME performs well at low bitrates (>150 kbits and <350 kbits/s)

5/3 temporal filter with ME seems to perform the best at medium and higher bitrates (>350 kbits/s). In this range of bitrates, it can outperform motion-compensated Haar analysis by more than 2dBs.

In order to illustrate the effectiveness of the proposed approach, we provide in Figure 30 a sample of the coding performance, where a simple Haar motion-compensated temporal multiresolution analysis is compared with the proposed 5/3 one. Significant improvement, of up to 2.5dB at high bitrates, is achieved with the new method.

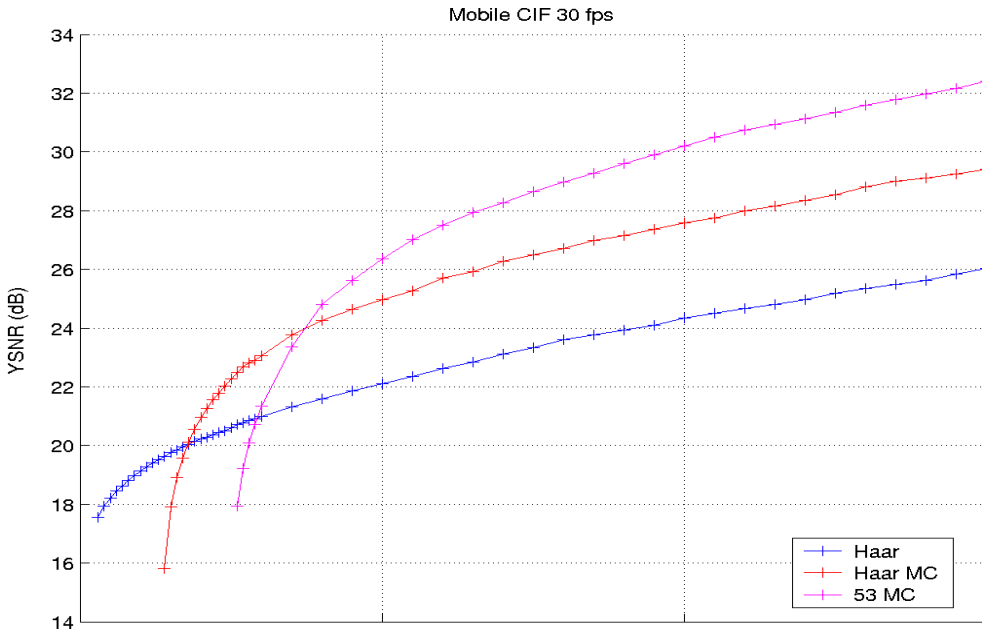


Figure 30: Coding performance comparison between Haar TF, Haar MCTF and 5/3 MCTF.

Moreover, temporal scalability is a key issue for such schemes, which lead to higher energy compaction in the approximation subband, while the temporal details contain less information. Consequently, the quality of the reconstruction obtained by decoding a number of temporal levels smaller than the number of the decomposition levels is better than with classical Haar transform. We have assessed the temporal scalability functionality by decoding from the full bitstream only a part corresponding to a reduced number of temporal levels. We have observed that in the case of Haar-MC temporal filters, the reconstruction at full frame rate based on 3 temporal levels is always of lower quality than that obtained by decoding all temporal levels. In contrast, with 5/3 filters we have obtained a better reconstruction (at the same bitrate) by using less temporal levels, at medium bitrate (≤600kbs). This is remarkable, since it allows complexity scalability for terminals with reduced resources. A second conclusion is that the overall quality of scalability is greatly improved when using 5/3 filters compared to Haar, and this at all bitrates. This is illustrated in Figure 31 for 3 levels decoded over 4 encoded.

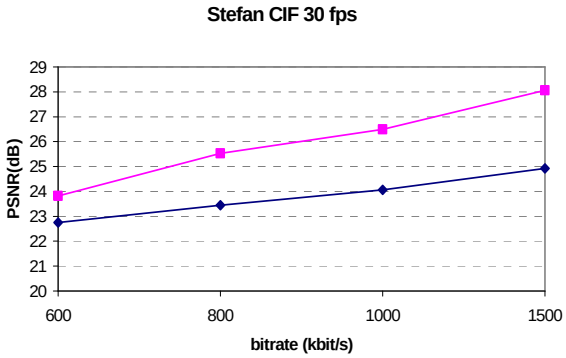


Figure 31: Temporal scalability: average PSNR of reconstructed sequence for the Haar (blue curve) and the 5/3 (pink curve) temporal decomposition.

In particular, it is found that the PSNR-oscillation between odd and even frames is smaller with the 5/3 temporal filters. We have observed situations where the reduced frame rate sequence has higher quality than the full frame rate sequence at the same bitrate (low bitrates). Such a situation can be qualified as "negative cost of scalability" and has shown the importance of a flexible and automatic framework for parameter selection in network adaptation tasks.

4.4.3 Spatial decompositions

In Deliverable D3.1 it has been shown that the statistical properties of the temporally filtered frames (H, HL, HLL, etc) may differ substantially[8]. Therefore, it was decided to apply different wavelets to different frame types. In MASCOT experiments have been carried out for different spatial wavelet sequences, both linear and nonlinear. Such spatial wavelet sequences are described by $n+1$ parameters, 2^n being the size of a group-of-frames (GoF). These parameters describe the type of spatial wavelet which is applied to the subsequent frame types, i.e., H, HL, HLL, etc. Note that the last frame type is always an approximation frame LL…L.

The following spatial wavelet types have been considered:

Type variable	Description
1	classical 9/7-biorthogonal wavelet
2	linear Haar wavelet
3	CISL erosion wavelet
4	adaptive wavelet (initialized with a threshold of 10)
5	morphological Haar wavelet

For example, the parameter sequence 12233 means that frames of type LLLL will be encoded with the 9/7-wavelet , frames of type HLLL and HLL with the linear Haar wavelet , and frames of type HL and H with the CISL erosion wavelet. It was found in Deliverable D3.1 that morphological wavelets perform relatively good for 5/3-MC temporal filtered frames. This is illustrated in Figure 32 where we depict rate distortion curves for wavelet sequences 11111, 12222, and 12233, all applied to *Foreman* CIF at 30 fps.

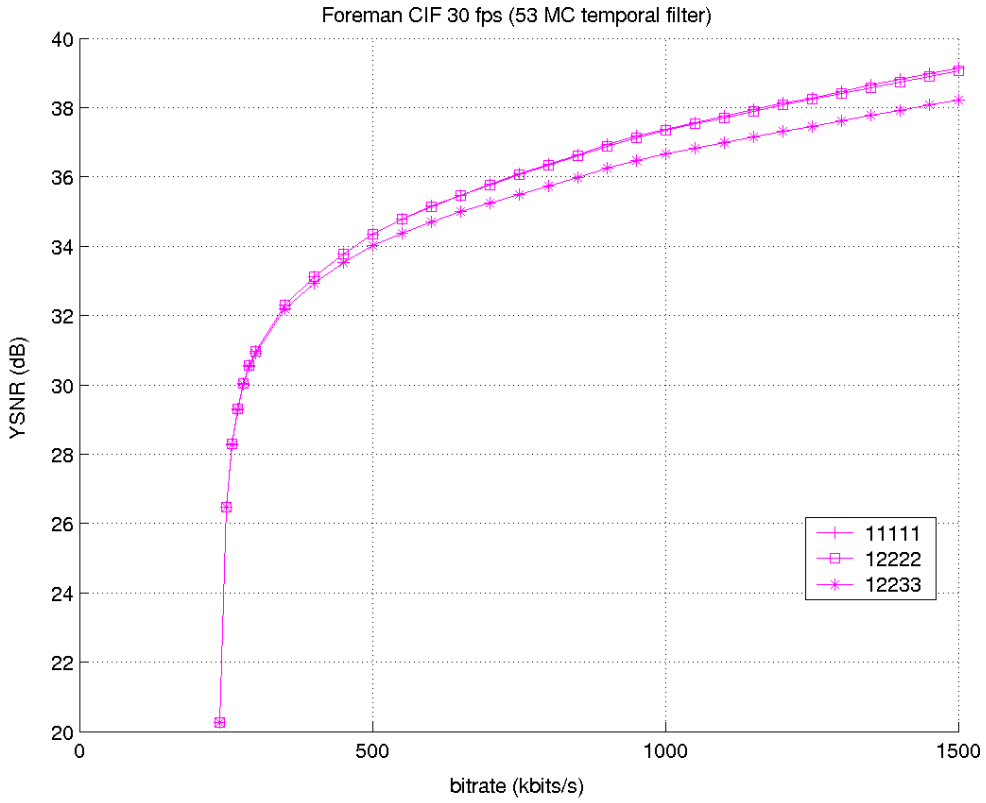


Figure 32: Rate distortion curves for different spatial wavelet sequences applied to Foreman *CIF 30 fps (5/3-MC temporal filter).*

In Deliverable D3.1 it has also been reported that the results, e.g. for the wavelet sequence 12233, for Haar-MC temporal filters are worse. This can possibly be explained by the fact that this spatial wavelet sequence performs better when the temporal decomposition results in strongly decorrelated frames. In this case, the prediction error frames have a stronger "edge-like" appearance than in the case without MC, and

[8] We call H, HL, etc the frame type.

the design of non-linear filters like in the CISL erosion filter, is specific to such cases. However, the spatial filter “12233” seems to give results which are slightly worse than “11111” or “12222”. This may be due, to our opinion, to the “lack of compatibility” between the uniform quantization and the highly non-linear nature of the CISL erosion wavelets. Yet a second cause may be the fact that errors resulting from morphological filters are present in the reconstructed frames as "bright" and "dark" pixels, resulting in a large MSE, and hence low PSNR, even if such errors occur only at very few points.

The results for the adaptive decompositions are less encouraging at this point. In Figure 33 we show also the rate-distortion curves for the adaptive wavelet sequence 14444 applied to *Mobile* CIF 30 fps, where the spatial frames have been obtained after 5/3-MC temporal filtering. This result suggests that the spatial wavelet sequence corresponding with adaptive update lifting gives the worst results. Although we do not have a valid explanation at this point, the following issues may play a role:

- The optimal threshold for the decision criterion governing the adaptive update filter must be determined for each sequence, at each bitrate, and for each temporal filter. This is an issue that has hardly been addressed in our work up till now; see [15] for some more details.
- The quantization and coding scheme that has been used in the wavelet codec, is not well suited for such highly non-linear filters. More research is necessary in order to develop such coding algorithms.
- As observed before, PSNR is not a "good" measure for judging visual artifacts, especially for non-linear filters such as the ones discussed here.

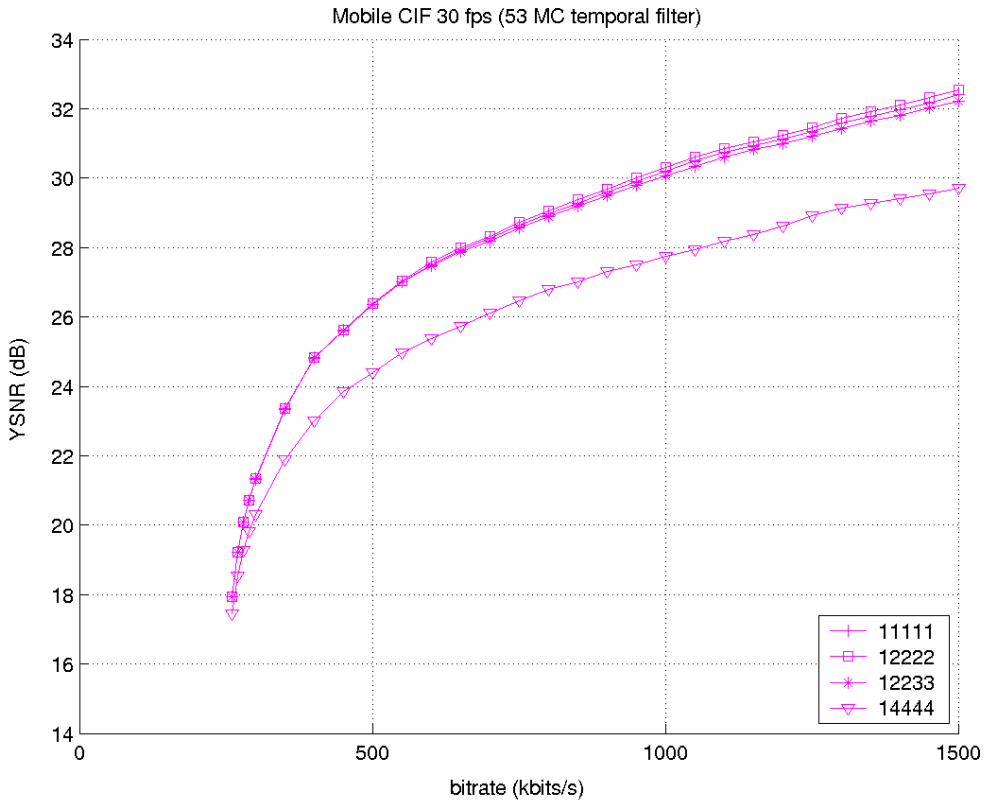


Figure 33: Rate distortion curves for different spatial wavelet sequences applied to Mobile *CIF 30 fps (5/3-MC temporal filter).*

4.4.4 Scalability in hybrid codecs

We refer to Subsection 3.4.4 for a description of our scalability architecture.

The performance of the two-loop structure has been tested for various scenarios. Three basic series of experiments have been performed:

- H.263-based experiments,
- MPEG-2-based experiments,
- H.264/AVC-based experiments.

For the first two experiments, the overall coding performance is summarized in Table 7. The figures are average PSNR values as well as average bitrates for selected test sequences. The values for two-layer bitstreams have been compared to single-layer bitstreams obtained using standard MPEG-2 or H.263 coders with the same options switched on. The values at the output of low-resolution coder are also included in Table 1. The results for the implementation of the H.264/AVC-based scalable codec are shown in Figure 34. The H.264/AVC-based scalable test model has been implemented on the top of standard JVT software version 2.1. Both coder and decoder have been implemented.

In order to test the coding performance of the scalable H.264/AVC codec, a series of experiments have been performed with (352 × 288)-pixel sequences. Horizontal, vertical and temporal subsampling factors have been set to 2 and the video sequence structure was that from Figure 34(a).
In the experiments, the following modes have been switched on:

- CABAC coder,
- ¼-pel motion estimation in both layers,
- all prediction modes.

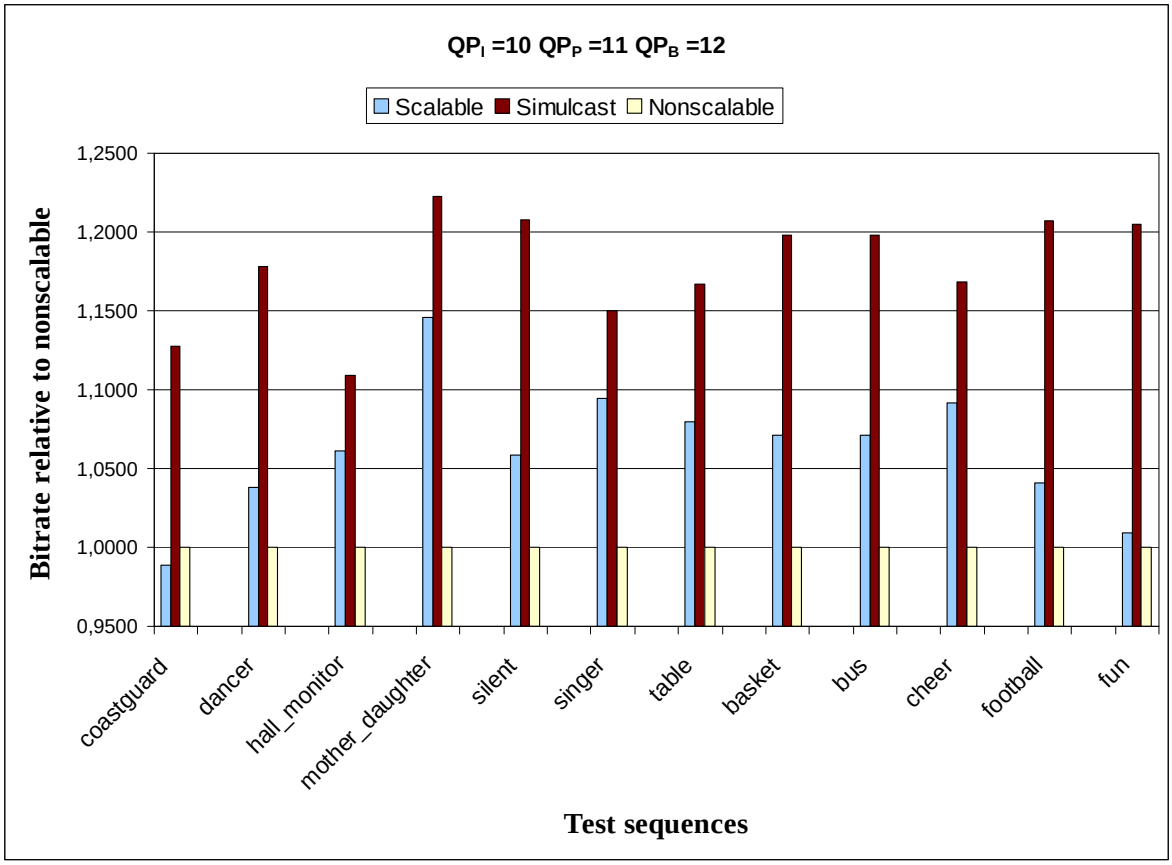
The experiments have been performed for three sets of the quantization parameter values. These values were defined independently for I-frames (QP_I), P-frames (QP_P) and B-frames (QP_B). In the tests, equal values of QP_I , QP_P and QP_B were applied in the base and the enhancement layer, respectively.

In order compare the scalable codec with the nonscalable reference H.264/AVC codec as well as with the simulcast pair of nonscalable H.264/AVC codecs, the experiments have been performed with constant values of QP_I , QP_P and QP_B that imply almost constant quality measured in terms of the *PSNR* factor for the luminance component in a given sequence. Of course, the quality measured for different sequences is different, but for a given video sequence and a given set of QP_I , QP_P and QP_B , the results for scalable, nonscalable and simulcast coding differ mostly less than 0.3 dB and often even less than 0.1 dB. For such conditions, bitrates have been estimated for the scalable coder (*whole scalable coder*), nonscalable coder and simulcast coding. For such test conditions, the approximate bitrate overhead due to scalability was between -1% and 30% of the bitrate for the nonscalable (single-layer) codec. For almost all cases, scalable coder performed better than simulcast coding. Usually scalable coding performance was substantially higher than that of simulcast. Within a scalable coder, the base layer bitrate was about 15% to 22% of the total bitrate produced by a scalable coder for both layers. For some test sequences and some bitrates chosen, the astonishing feature of the results is that the performance of the two-loop coder i.e. scalable coder, is better than that of the reference single-layer coder. Such results have been obtained independently for both series of experiments based on two different coders and two different sequence structures. Similar phenomena are reported also in the references. Fine granularity has been obtained by transmitting only a desired portion of the DCT data from the bitstream of the highest resolution. This can be efficiently done on the basis of bit-planes. Here, in order to have a simple implementation, the partitioning has been simply done by limiting the maximum number of the DCT coefficients transmitted per block (Figure 35). Bit-plane encoding could improve the efficiency slightly.

H.263 - based coder for CIF (352 × 288) sequences		*Football*	*Basket*	*Cheer*	*Fun*	*Bus*
Single-layer coder (H.263)	Bitstream [Kbps]	485,83	448,45	443,64	459,76	443,14
	Average luminance PSNR [dB]	31,38	25,29	25,05	25,69	26,59
Proposed scalable coder	Low resolution layer bitstream [Kbps]	109,17	100,88	104,87	100,62	109,94
	Low resolution layer average PSNR [dB] for luminance	29,17	23,87	23,00	23,71	25,72
	High resolution layer bitstream [Kbps]	350,83	329,97	365,89	327,80	322,65
	Average PSNR [dB] for luminance recovered from both layers	31,36	25,32	25,05	25,67	26,62
Single-layer coder (H.263)	Bitstream [Kbps]	854,33	801,18	667,35	823,62	764,94
	Average luminance PSNR [dB]	33,80	27,39	26,68	28,02	28,45
Proposed scalable coder	Low resolution layer bitstream [Kbps]	184,03	193,59	169,07	196,00	191,75
	Low resolution layer average PSNR [dB] for luminance	31,23	26,00	24,61	26,13	27,44
	High resolution layer bitstream [Kbps]	568,56	587,72	511,44	574,75	546,97
	Average PSNR [dB] for luminance recovered from both layers	33,88	27,37	26,68	28,04	28,49
Single-layer coder (H.263)	Bitstream [Kbps]	1229,74	1193,32	993,22	1234,31	1137,16
	Average luminance PSNR [dB]	35,74	29,07	28,65	29,98	30,18

Proposed scalable coder	Low resolution layer bitstream [Kbps]	261,37	289,09	247,51	291,58	282,86
	Low resolution layer average PSNR [dB] for luminance	32,98	27,89	26,29	28,11	29,33
	High resolution layer bitstream [Kbps]	816,18	878,64	759,43	862,85	820,99
	Average PSNR [dB] for luminance recovered from both layers	35,80	29,07	28,64	29,96	30,17
MPEG-2 - based coder for 4CIF (704 × 576) sequences		*Cheer*	*Flower Garden*	*FunFair*	*Stefan*	*Bus*
Single-layer coder (MPEG-2)	Bitstream [Mbps]	2.99	3.08	3.17	2.96	2.93
	Average luminance PSNR [dB]	28.94	28.19	29.18	31.99	31.45
Proposed scalable coder	Low resolution layer bitstream [Mbps]	0.95	1.04	0.77	0.98	0.99
	Low resolution layer average PSNR [dB] for luminance	28.15	28.78	27.67	32.88	31.86
	Total bitstream [Mbps]	2.91	3.24	3.23	2.99	3.27
	Average PSNR [dB] for luminance recovered from both layers	29.03	28.15	29.21	32.06	31.44
Single-layer coder (MPEG-2)	Bitstream [Mbps]	3.91	3.95	3.91	3.89	3.93
	Average luminance PSNR [dB]	30.66	29.54	30.84	33.84	33.54
Proposed scalable coder	Low resolution layer bitstream [Mbps]	1.26	1.30	1.50	1.27	1.27
	Low resolution layer average PSNR [dB] for luminance	30.43	30.24	32.77	35.73	34.42
	Total bitstream [Mbps]	3.67	4.29	3.80	3.71	4.55
	Average PSNR [dB] for luminance recovered from both layers	30.66	29.52	30.79	33.74	33.59
Single-layer coder (MPEG-2)	Bitstream [Mbps]	5.09	4.85	4.76	4.74	4.87
	Average luminance PSNR [dB]	32.33	30.91	32.17	35.20	34.60
Proposed scalable coder	Low resolution layer bitstream [Mbps]	2.18	2.12	1.66	1.93	2.32
	Low resolution layer average PSNR [dB] for luminance	35.86	36.04	33.75	40.51	38.96
	Total bitstream [Mbps]	5.09	5.45	5.09	5.02	5.7
	Average PSNR [dB] for luminance recovered from both layers	33.11	30.98	32.15	35.14	34.56

Table 7: The two-loop coder performance measured for whole resolution levels. Results obtained for progressive sequences. The H.263 coder without PB-frames, respective scalable coder without GOPs in both layers. The MPEG-2 coder and respective scalable coder with GOP length of 12 frames.



(b)

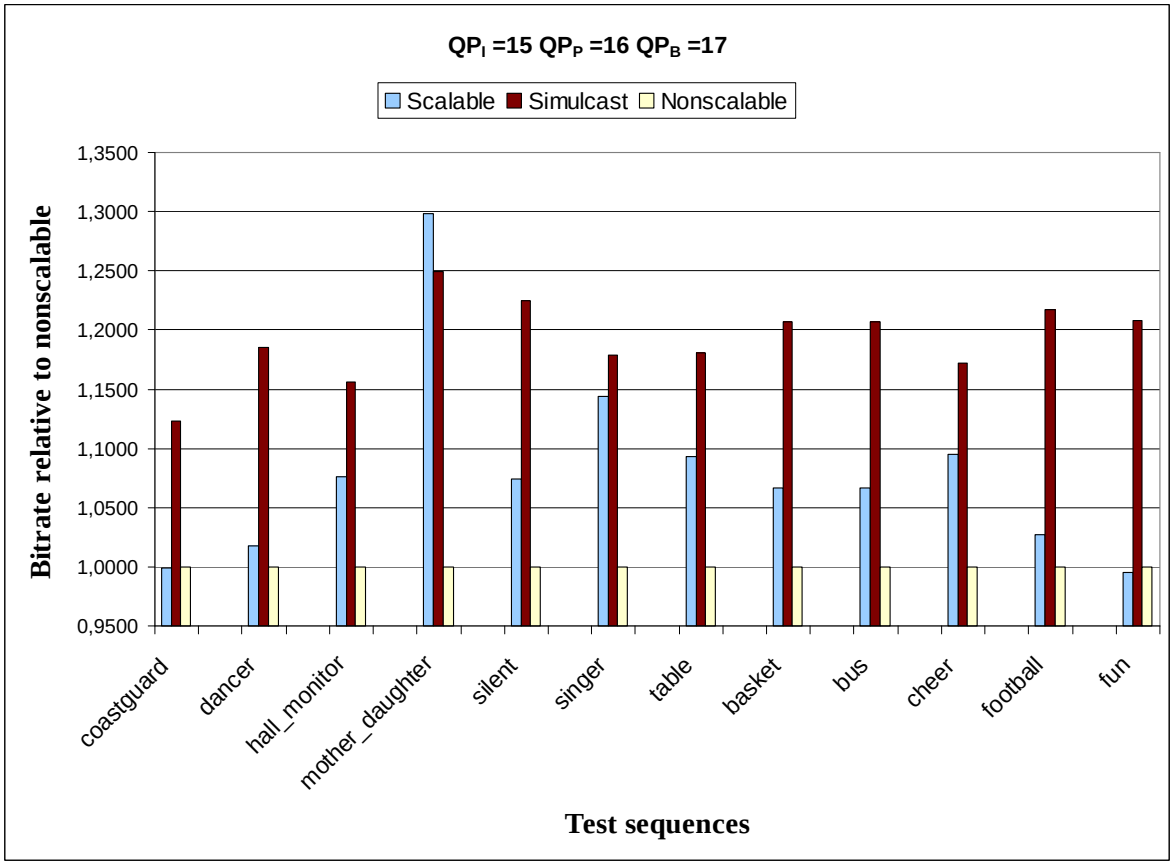


Figure 34: H.264/AVC video codec: approximate bitrate comparison for scalable, nonscalable (single-layer) and simulcast coding.

The comparison for intermediate bitrates proves that also fine-granularity scalability is related to acceptable performance (Figure 35 and Figure 36).

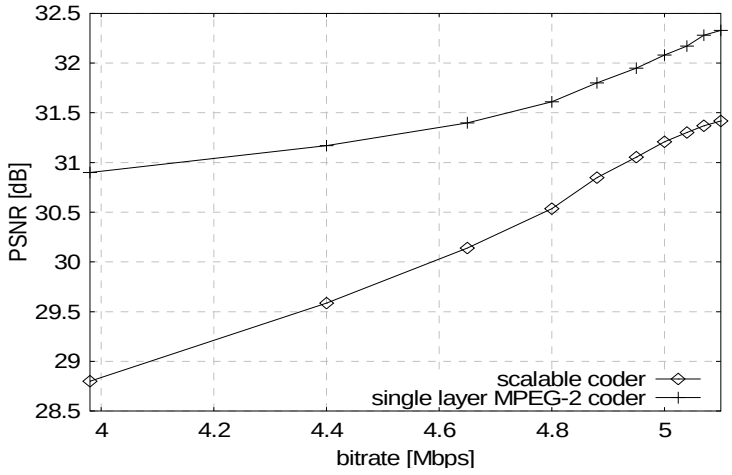


Figure 35: The fine-granularity-scalability in a two-loop MPEG-2-based coder (lower curve) compared to a single layer MPEG-2 coder (upper curve). Test sequence Funfair, *total bitrate 5 Mbps, base layer bitrate about 1.66 Mbps, GOP=12.*

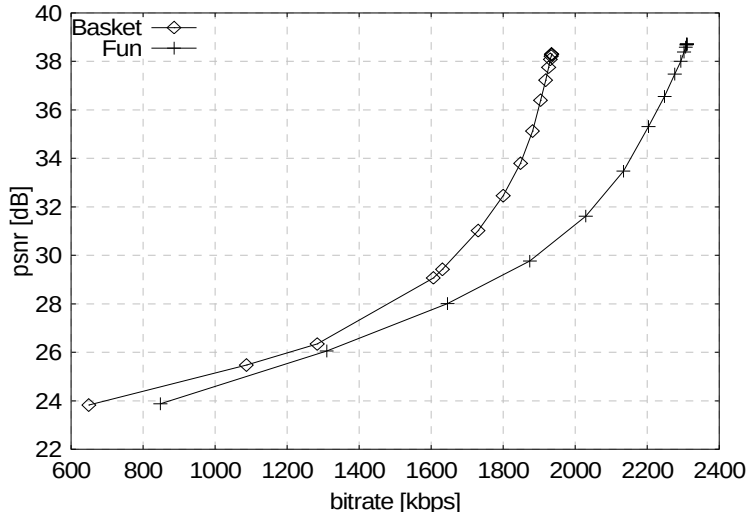


Figure 36: Rate-distortion curves for FGS in the extended H.264/AVC codec: test sequences Fun *and* Basket.

Conclusions

Described is a generic multi-loop coder structure for motion-compensated fine-granularity scalability. The major features are:

- mixed spatio-temporal scalability,
- independent motion estimation for each motion-compensation loop, i.e. for each spatio-temporal resolution layer,
- BR/BE-frame structure,
- improved prediction of BR-frames.

For all layers of a scalable representation, an important feature of the coders is that the bitstream syntax is either fully standard (for H.264/AVC) or only slightly modified (macroblock headers in H.263- and MPEG-2-based scalable coders). The modifications of the bitstream semantics are minor.

4.5 Results of WP4: Prediction

4.5.1 General

In Section 3.5 we have described various new prediction tools that were developed in the context of the MASCOT project. Below, we give a short report on the evaluation of these prediction tools.

4.5.2 Wavelet-based motion estimation

To evaluate the efficiency of the wavelet-based motion estimation technique, the flow field has been estimated between successive pairs of frames in video sequences. The flow has then been used for each pair to predict one frame of the pair from the other one. The results have been compared with those obtained with a full-search block matching providing flow maps of the same encoding complexity. Two parameters determine the complexity of a flow map:

- The density of parameters, which is block size for block matching, being either a vector parameter for each 4x4 block, or a vector parameter for each 8x8 block.
- The accuracy of the parameters, than can be determined for each in pixel precision: results are reported for 2 precisions: half-pixel precision and 1/4-pixel precision.

In the next figures, results are given for *Mobile* and *Paris*. In each figure, the prediction PSNR is given for the wavelet method in 4x4 density (line A), and in 8x8 density (line B), and for the full-search block matching in 4x4 density (line C), and in 8x8 density (line D). First, the *Mobile* sequence:

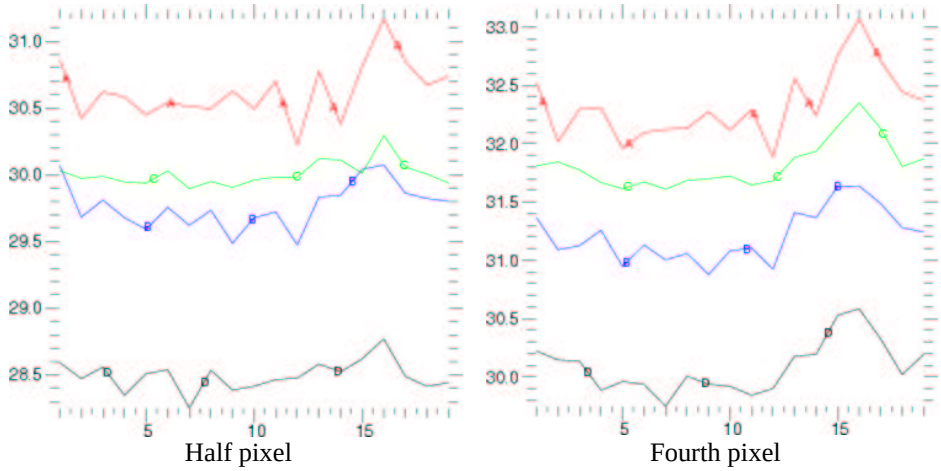


Figure 37: Prediction PSNR comparison for the Mobile sequence (A,B: wavelet motion, C,D: block matching, A,C: 4x4 density, B,D: 8x8 density).

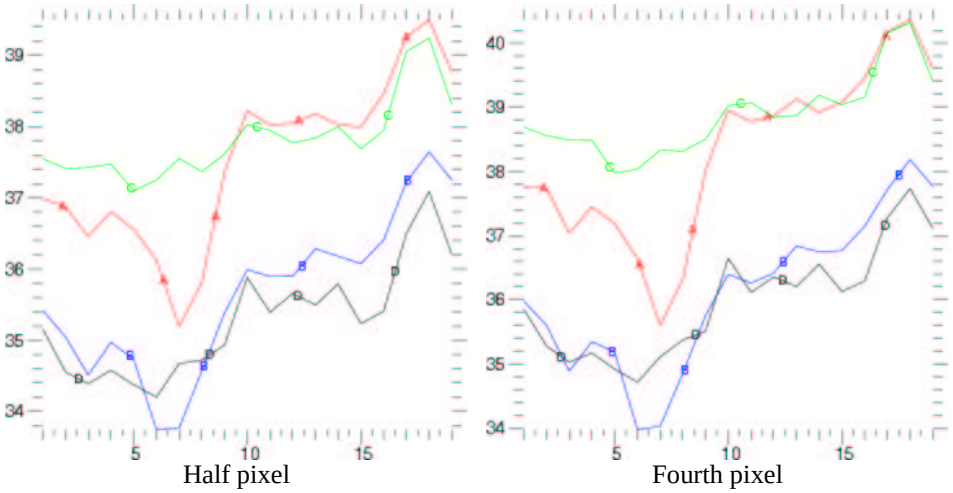


Figure 38: Prediction PSNR comparison for the Paris *sequence (A,B: wavelet motion, C,D: block matching, A,C: 4x4 density, B,D: 8x8 density).*

For the *Mobile* sequence, the wavelet motion estimation and prediction is superior to the classical block-matching. For the *Paris* sequence, the comparison is less clear-cut.

The following graph reports rate-distortion comparison of the wavelet motion estimation (MME for multiscale motion estimation) and the full-size block matching (FSBM) for identical flow costs. The results have been obtained by integrating the wavelet motion estimation and compensation into the MASCOT 3D wavelet codec based on the Woods wavelet codec. The difference becomes more important at low bitrates, where the wavelet motion estimation based codec has a gain of up to 4 dB in average distortion as compared to full-size block-matching.

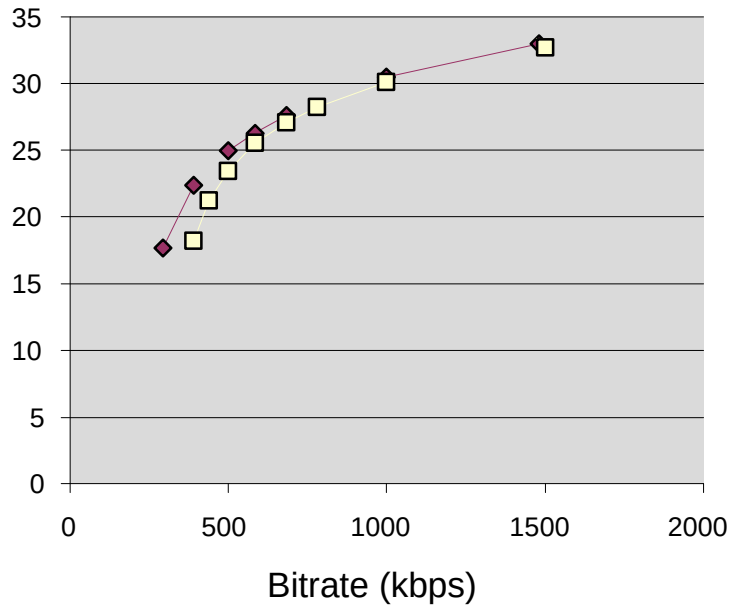


Figure 39: Bitrate comparison for a 3D wavelet codec between wavelet-based motion estimation (MME) and full-search block-matching (FSBM).

4.5.3 Motion-vector coding

In Deliverable D3.1 it has been shown that the proposed motion vector coding achieves superior performance in comparison to the default motion vector coding that exists in the MASCOT 3D wavelet coder. This was demonstrated both for the coding of motion vectors generated by optical-flow field motion estimation and for motion vectors generated by fixed-size block-based motion estimation. Below we present two typical examples of the achieved results. Figure 40 demonstrates the efficiency of the proposed motion-vector coding for the case of block-based motion estimation. The average PSNR achieved with the coder using the context-based motion-vector coding (CBMVC) is constantly superior to the average PSNR achieved using the default motion vector coding scheme. This is especially important for low bitrates. Alternatively, Figure 41 shows the performance of the CBMVC scheme for the coding of vectors generated by optical-flow motion estimation (OFME). Again, in comparison to the coder using the default motion vector coding scheme, the proposed method achieves superior performance.

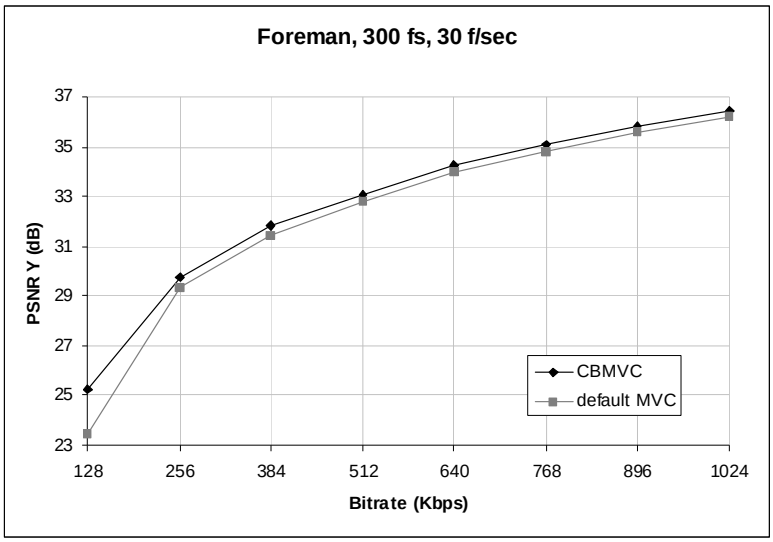


Figure 40: Results for the coding of motion vectors generated by fixed-size block-based motion estimation.

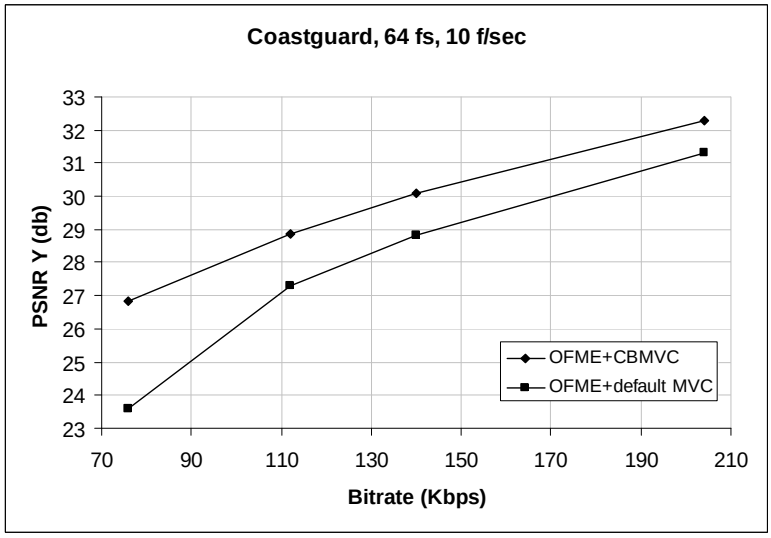


Figure 41: Results for the coding of motion-vectors generated by optical-flow field motion estimation.

4.5.4 Long-term Global Motion Compensation

The results for LT-GMC are first visually very impressive. The following graph compares one frame of the original video sequence, and one frame of the sequence decoded using LT-GMC.

Original sequence Superresolution mosaic

Figure 42: Various close-ups from the Mobile *and* Stefan *sequences. In all cases, the mosaic is visually more pleasant, because of the significant aliasing reduction obtained with LT-GMC.*

As shown in Deliverable 4.3, LT-GMC provides substantial savings over H.264/AVC at equivalent visual distortion. The PSNR is a little lower (38.1 dB vs 39.4 dB) for the LT-GMC version, but PSNR comparisons are a dubious comparison measure since LT-GMC reduces image aliasing.

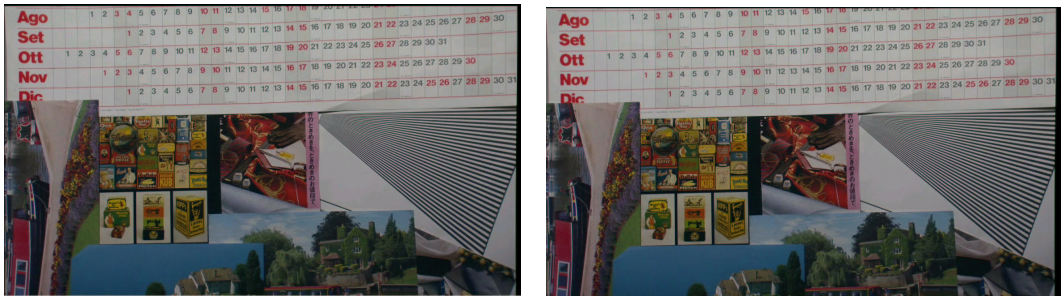


Figure 43 : Decoded H264 frame (left) and with H264 with LT-GMC (right) HDTV format, Spincalendar *sequence.*

The following graph shows the bitrate savings for this sequence provided by the LT-GMC extension as a function of the original sequence bitrate.

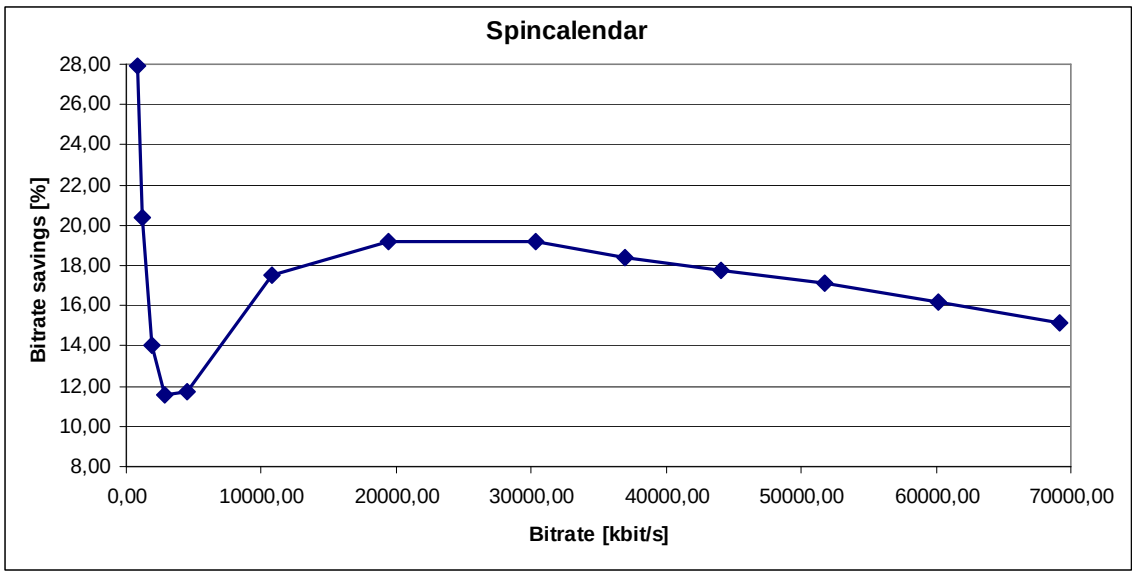


Figure 44: Bitrate gains with LT-GMC at similar visual quality versus H.264/AVC baseline

Figure 45: Decoded H.264/AVC frame (left) and with H264/AVC with LT-GMC (right) HDTV format, City *sequence*

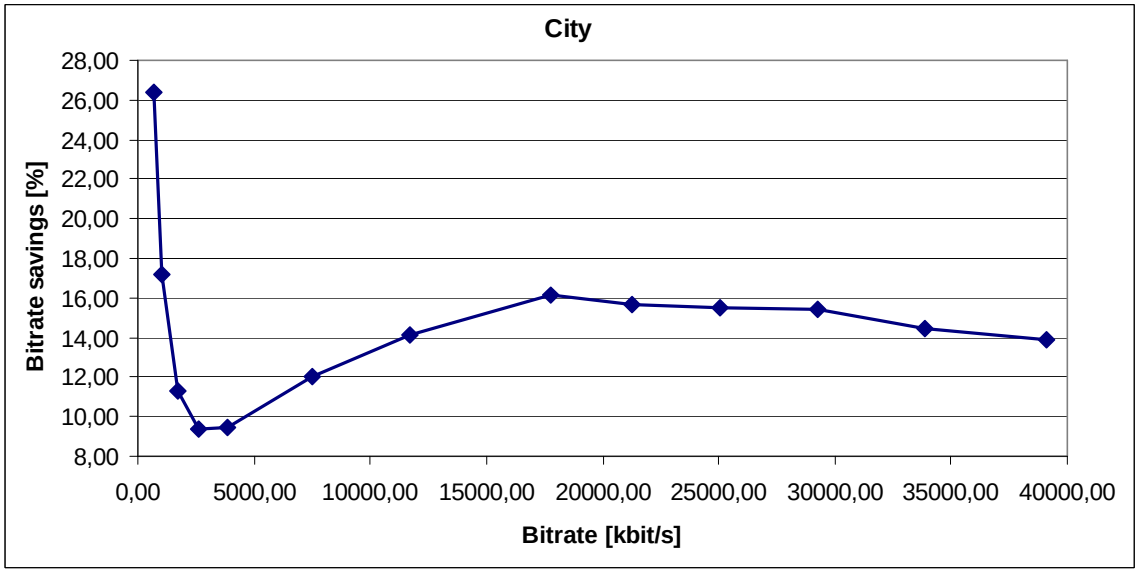


Figure 46: Bitrate gains with LT-GMC at similar visual quality versus the H.264/AVC baseline codec.

To conclude, LT-GMC outperforms standard GMC, and can be included in the H.264/AVC framework very efficiently. This is done by using the rate control of the H.264 codec to decide whether or not a macroblock is predicted with LT-GMC, and on these macroblocks, no prediction error is encoded.

4.5.5 Long term motion prediction using graph matching

The resulting tests are reported in Deliverable 4.1, and suggest that graph matching does not provide a sufficiently accurate motion estimation to produce acceptable gains over a plain I-frame encoding. In addition, the motion estimate obtained with the graph matching method was refined with an additional wavelet motion estimation method, but even in this latter case, simple I-frame encoding is more efficient. For this reason, integration of this tool in the MASCOT codec was abandoned.

4.6 MPEG liaisons

One of the goals of MASCOT was to contribute the results to standardization activities on video coding such as ISO/MPEG and ITU/VCEG. Due to the time schedule it was impossible to contribute to the new standard H.264/AVC, which was recently finalized. However the results obtained in MASCOT already show improvements and it can be expected that various of the MASCOT partners will contribute the developed technology to potential future extensions.

The impact of the MASCOT project on the MPEG activities is clearly reflected by the strongly increased interest of the MPEG committees in wavelet-based video codecs, which, recently lead to the creation of the MPEG Ad-Hoc Groups on *Interframe Wavelets*, and *Scalable Video Coding*. Active participation of the members of the MASCOT consortium in these and other MPEG activities has resulted in several contributions that have been proposed at various MPEG meetings. Recently, these activities have resulted in a "Call for Evidence" on scalable video coding technology that was recently released by MPEG. The technology developed in MASCOT will be proposed in answer to that Call at the MPEG meeting in July 2003 in Norway. This includes 3D wavelet technology as well as scalability based on H.264/AVC.

VUB has delivered input in relation to alternative scalable wavelet video coding architectures, benefiting from in-band motion compensation and utilizing the hybrid and UMCTF frameworks. Some of this work was carried out in collaboration with Philips USA. Additionally, new motion vector coding schemes were proposed that could handle efficiently the motion field data produced by such architectures.

4.7 Contribution to EU policies

Technologies that enable fast and flexible access to multimedia data and that provide better interoperability of different user platforms within heterogeneous networks contribute to the EU objective of building a user-friendly information society. The tools that have been developed within the context of MASCOT help to realize such a society as they will eventually lead to enhanced video compression standards. The MASCOT project has brought together leading players in Europe in the areas of image and video processing and coding, thus creating a consortium with complementary expertise that can only be found at the European level.

For long-term European competitiveness and employment is important to invest in emerging and challenging technologies. This is also one of the goals of the EU FET programme. Video compression is an area where research investments made today can be the key to a sustained scientific and technological lead of Europe in the future. The MASCOT project has been concerned with the development of new tools for video coding which may eventually improve the strategic position of the European enterprises in the area of multimedia technologies. Furthermore, MASCOT has contributed to standardization committees, such as ITU-T SG 16, MPEG and JPEG. As a result, dissemination and use of the project results have become directly available to the industrial community, enabling fast spread and take-up of the results.

The EU Green Paper[9] on the convergence of the telecommunications, media and information technology sectors that appeared in 1997 stresses the enormous potential impact of such technologies on the economy and the live of citizens. An important issue today is to improve the quality of video streams over heterogeneous networks, which is one of the most important means of communications between people, as well as between organisations (including industries and governments) and people. The MASCOT project has developed various tools that can help to solve this issue, and as such the project has contributed to increased and enhanced communications between European citizens.

[9] See http://europa.eu.int/ISPO/convergencegp/97623.html

5 Deliverables and Publications

5.1 Description of Deliverables

5.1.1 Introduction

In this section we briefly describe the MASCOT deliverables that correspond with the four technical workpackages of MASCOT.

5.1.2 Deliverable 1.1 - Specification of architecture of the MASCOT codec

This deliverable specifies the architecture of the MASCOT codecs. It explains the decisions taken concerning the MASCOT codec architectures, defines the roadmap for integration and describes the relation to new standardization activities in MPEG.

Two different codec architectures are used as baseline, a hybrid DCT-based codec and a 3D wavelet codec. The partners developed various new coding tools in the framework of these baseline codecs. The goal is to improve the performance of the reference codecs and to provide extended functionalities. This deliverable gives an overview of the baseline codecs and describes how each tool is going to be integrated.

The deliverable basically consists of two parts. The first part comprises the architecture of the DCT-based codec. First, the basic architecture of the baseline codec, which was then under development in the Joint Video Team (JVT) of ISO/MPEG and ITU/VCEG is described, highlighting improvements and new elements compared to existing hybrid video coding standards. Then the extension of this baseline codec for spatio-temporal scalability with fine granularity is described. Finally, the integration of each coding tool is described in detail. This includes a brief description of the functionality, a detailed description of the integration into the JVT software with block diagrams and a description of the impact on the JVT bitstream syntax and other data transmission.

The second part specifies the architecture of the 3D wavelet codec. It also starts with a description of the baseline framework, which is the 3D wavelet codec developed by PFR. Such a codec is fully scalable by definition. Then the integration of tools into the 3D wavelet codec is described.

A second wavelet codec has been developed within the project by VUB. The consortium decided to concentrate on one wavelet codec framework for further development and integration of new tools and the PRF codec (when PRF left the project this was replaced by the Woods codec) was chosen for this purpose. Nevertheless, a description of the wavelet codec provided by VUB can be found in Annex A of this deliverable, in order to document the successful work.

5.1.3 Deliverable 1.2 - Test model for MASCOT's coding scheme

This deliverable reports the software integration of the MASCOT codecs. Two different codec architectures have been used as baseline, a hybrid DCT-based codec and a 3D wavelet codec. A detailed description of the architecture of these codecs is given in D1.1 which complements this deliverable D1.2. The MASCOT partners have developed new coding tools in the framework of these two baseline codecs. The goal was to improve the performance of the reference codecs and to provide extended functionalities. This deliverable describes how the different tools are integrated and gives a summary of the progress of work in workpackage 1.

5.1.4 Deliverable 1.3 - Report on coding performance

This report contains the results of coding performance evaluation of the two MASCOT codecs, the hybrid DCT-based codec and the 3D wavelet codec. The MASCOT partners have developed various new coding tools in the framework of these two baseline codecs to improve the performance of the reference codecs and to provide extended functionalities.

First, the report describes general issues of performance testing. The experiments and evaluations are carried out in a similar way as MPEG core experiments and fall apart in two main sections. One contains the performance tests of the DCT-based codec with integrated MASCOT tools. Significant gains are reported for the different tools. Since the reference baseline H.264/AVC codec represents the state of the

art in video coding, the integrated MASCOT hybrid DCT-based codec can be regarded as the most advanced DCT-based video codec available in the world. Moreover the MASCOT hybrid DCT-based codec provides fine granularity scalability. The other section contains the performance tests of the 3D wavelet codec. Also in this case significant improvements compared to the state of the art reference are reported. Therefore also this codec can be regarded as one of the most, if not the most, advanced wavelet video codec currently available.

5.1.5 Deliverable 1.4 - Final demonstration

The results obtained within the MASCOT project have been demonstrated to a broad public at the Picture Coding Symposium held on April 23-25 in St. Malo, France, which is one of the major world wide conferences on image and video coding. An exhibition was organized showing the improvements achieved in MASCOT.

5.1.6 Deliverable 2.1 - Analysis and selection of descriptors and description schemes supporting encoding tools and preliminary metadata-based encoding tools and strategies.

It is expected that in the future a large amount of audio-visual documents will be indexed and that metadata information will be rather easy to create. As a result, in many circumstances, audio-visual material will be available together with the metadata describing its content. Therefore, future image and video sequence encoders will be able to use the metadata information in order to improve their efficiency or to optimize their strategy. One of the main objectives of MASCOT was to demonstrate the validity of this approach and to develop an efficient compression scheme exploiting metadata information, i.e., indexing information that may be available to support search, query and browse functionalities (e.g. MPEG-7 or SMPTE metadata standards).

During the initial phase of the MASCOT project, descriptors and description schemes included in the MPEG-7 and SMPTE standards have been analyzed and their potential use for encoding has been assessed. This deliverable reports the consortium's views on the metadata types that may be useful to improve the encoding process. Note, however, that the final report on this issue has been given in Deliverables D2.2 and D2.3.

This report lists a set of Descriptors and Description Schemes that could be used to improve the encoding efficiency. For each Descriptor or Description Scheme, the syntax and semantics is precisely defined. Following the example of the MPEG-7 standard, the syntax is defined by the *XML Schema* language. Also the strategy that could be used to improve the coding efficiency is also described. In some cases, preliminary coding results have been reported.

5.1.7 Deliverable 2.2 - Metadata-based encoding tools

This deliverable defines the two encoding strategies that the project has adopted: a hybrid motion compensated DCT-based scheme and a 3D wavelet scheme. The goal of this report is to highlight the two coding frameworks in which metadata are used. It presents the strategy that is used to improve the coding efficiency using metadata. Furthermore, a number of experimental results are presented. The syntax and semantics of each Descriptor or Description Scheme has been defined precisely in Deliverable 2.1.

5.1.8 Deliverable 2.3 - Metadata-based encoding strategy

This deliverable discusses the metadata encoding strategy for the metadata-based encoding tools developed in the framework of the MASCOT project. The coding strategies developed in the context of MASCOT, have been presented in deliverables D2.1 and D2.2. Final results of these tools have been reported in deliverable D1.3. Metadata-based coding tools employ standard metadata in order to improve the coding efficiency of standards video codecs (such as H.264/AVC). The metadata descriptors have been selected within the two major metadata standards, MPEG-7 and SMPTE. Results in D1.3 have shown promising bitrate gains when coding using metadata-based encoding tools.

Different scenarios can be foreseen when encoding video scenes using metadata-based encoding tools. These scenarios can be classified depending if the metadata descriptors are needed and available in the encoder, the decoder, or both. The first part of this deliverable overviews these scenarios and describes some real applications that illustrate the different scenarios.

The most challenging scenario involves sending the metadata together with the content. Here, metadata is also needed by the decoder but it is not externally available. Therefore, if the encoder wants to use metadata it must be sent to the decoder together with the content. This deliverable reviews all metadata-based coding tools presented in D2.2 and presents several techniques for encoding the metadata that have been used for each tool. An estimate of the bitrate needed for each metadata descriptor will be computed so efficient rate encoding strategies between data and metadata will be developed in the future.

5.1.9 Deliverable 3.1 – New spatio-temporal decompositions

The aim of this report is to present new spatio-temporal decompositions developed in the framework of the MASCOT 3D wavelet codec. This includes all multiscale decompositions (temporal, spatial, motion vectors) as well as methods for coding. The first chapter presents the global 3D wavelet video codec scheme, initially proposed by PRF, that has served as a starting point for the project. Before PRF's withdrawal from the project as of September 1, 2002, several metadata descriptors have been used in conjunction with this codec to prove the advanced coding features provided in MASCOT. The second chapter describes the new motion-compensated temporal wavelet decompositions that have been used in MASCOT. This decomposition is based on the lifting formulation. A main contribution of MASCOT in this context is the generalization of the Haar-type decomposition by a 5/3 decomposition. The problem of motion vector redundancy in these schemes is explained, and solutions are proposed. Furthermore, a new efficient coding algorithm for motion vectors is introduced. A large part of this deliverable is concerned with the discussion of spatial decompositions for the residual frames. It discusses morphological wavelets and introduces a general framework for the design of adaptive wavelets. The most interesting feature of these adaptive schemes is that they do not require any side information to be sent to the decoder. The deliverable also presents an overview of state-of-the-art entropy coding methods for 3D wavelet codecs, with emphasis on an algorithm developed during this project (3D quadtree-limited coding).

5.1.10 Deliverable 3.2 - Impact of new decompositions on video compression

This deliverable can be regarded as a sequel to the second part of Deliverable D1.3 which contains various preliminary coding results for the MASCOT 3D wavelet codec. The aim of Deliverable D3.2 is to present some additional results for this codec, in particular to

- provide evidence for the scalability functionality
- show the efficiency of the scalability tool in the proposed scheme
- to provide new results on multiscale motion estimation and motion vector coding.

It contains results that show the impact of the multiscale motion estimation method on the global coding efficiency and it compares a newly developed motion vector coding strategy with the state-of-the-art method (used in the default Woods codec). Finally it discusses various scalability results in great detail.

5.1.11 Deliverable 4.1 - Alternate prediction parameters in the space domain for better encoding, motion compensation and segmentation, long term prediction

This deliverable reports on some of the work done in MASCOT in order to enhance the prediction tools used in both the hybrid DCT-based codec and the 3D wavelet codec. Two kinds of tools have been studied and enhanced for our purpose: *(i)* dense motion and illumination change field for short-term prediction, and *(ii)* long term prediction tools based on graph matching.

5.1.12 Deliverable 4.2 - Evaluation of new concepts for motion compensated prediction in the wavelet domain – long term memory prediction and NSI filters

The work on in-band motion compensation that was scheduled in the original proposal has been discontinued in order to rationalize the integration of the wavelet video codec. The work has been reported in the *Technology Survey* that has been added as an appendix to Deliverable D3.2.

5.1.13 Deliverable 4.3 - Evaluation, optimisation and development of global motion compensation and sprite coding algorithms, exploiting motion-based metadata

This report discusses the development and evaluation of a new video coding tool called “long-term global motion compensation” (LT-GMC). Since it is a new prediction method the work belongs to WP4, although there are strong relations to WP2.

It presents a new algorithm for generation of super-resolution video mosaics, which is inspired by video format conversion and spatio-temporal filtering algorithms. Based on this a new prediction tool called

LT-GMC that significantly outperforms standard GMC algorithms is presented. Super-resolution mosaics are used for prediction instead of only the last decoded frame.

The new idea for coding is to use the rate control of the H.264/AVC codec as a recognition tool for global motion macroblocks. If the rate control assigns LT-GMC mode to a macroblock it is very likely that it is only affected by global motion and can be reconstructed visually well using LT-GMC from previously transmitted frames even without transmission of the prediction error. Therefore we do not transmit any texture updates for those macroblocks. In this case the H.264/AVC rate control is used as a recognition or analysis tool for global motion and performs a kind of automatic segmentation of the video content. This is fully in the spirit of MASCOT. Metadata about global motion are used for analysis and recognition of visual content and this information is used for optimized encoding of the video.

The tool is fully automatic and fully integrated into the recursive processing loop of the MASCOT hybrid codec. For selected high-resolution sequences dominated by global motion significant bitrate savings at the same visual quality (note that we don’t do waveform coding that can be measured in PSNR) of 20% and more are reported.

5.2 Table of Deliverables

Nr	Title of Deliverable	Most Recent Version	Delivery month
D1.1	Specification of architecture of the MASCOT codec	5.0	Apr 02
D1.2	Test model for MASCOT's coding scheme	3.0	Jan 03
D1.3	Report on coding performance	5.0	Apr 03
D1.4	Final demonstration	-	23-24 April
D2.1	Analysis and selection of descriptors and description schemes supporting encoding tools and preliminary metadata-based encoding tools and strategies.	3.0	Mar 02
D2.2	Metadata-based encoding tools	3.0	Dec 02
D2.3	Metadata-based encoding strategy	1.0	Apr 03
D3.1	Morphological wavelet decompositions and overcomplete and shift-invariant representations for image and video coding (report)	2.0	Jan 03
D3.2	Impact of new decompositions on video compresion (final report)	1.0	Apr 03
D4.1	Alternate prediction parameters in the space domain for better encoding, motion compensation and segmentation, long term prediction (report)	1.0	Nov 02
D4.2	Evaluation of new concepts for motion compensated prediction in the wavelet domain – long term memory prediction and NSI filters (report)	1.1	Not necessary (Jan 03)
D4.3	Evaluation, optimisation and development of global motion compensation and sprite coding algorithms, exploiting motion-based metadata (report)	2.0	Mar 03
D5.1	Project plan	-	Jul 01
D5.2	Final report (this report)	1.0	Apr 03
D6.1	Quality control measures (section of *Project plan*)	-	Jul 01
D7.1	MASCOT web-site (including *Project Presentation*)	-	Jul 01
D7.2	Dissemination and Use plan	-	Sep 01
D7.3	Technology and Implementation plan	eTIP	May 03

5.3 List of Publications

5.3.1 Publications deriving from MASCOT

This section list publications that are an immediate outcome of the MASCOT project.

1. G.C.K. Abhayaratne, H. Heijmans, B. Pesquet-Popescu and G. Piella,
   Integer to integer adaptive wavelet transforms using adaptive update lifting,
   Accepted for the *Special Conference X-Wavelets of SPIE*, San Diego, 2003
2. G.C.K. Abhayaratne and H. Heijmans,
   A Novel Morphological Subband Decomposition Scheme for 2D+t Wavelet Video Coding,
   Accepted for the *3rd International Symposium on Image and Signal Processing and Analysis, ISPA 2003*, September 18-20, 2003, Rome, Italy
3. Y. Andreopoulos, M. van der Schaar, A. Munteanu, J. Barbarien, P. Schelkens, and J. Cornelis,
   Complete-to-overcomplete discrete wavelet transforms for scalable video coding with MCTF,
   Proceedings of SPIE Visual Communications and Image Processing (VCIP), Lugano, Switzerland, July 8-11, 2003.

4. Y. Andreopoulos, M. van der Schaar, A. Munteanu, J. Barbarien, P. Schelkens, and J. Cornelis, Fully-scalable wavelet video coding using in-band motion compensated temporal filtering, *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Hong Kong, April 6-10, 2003.

5. Y. Andreopoulos, M. van der schaar, A. Munteanu, P. Schelkens, and J. Cornelis, Spatio-temporal-SNR scalable wavelet coding with motion-compensated DCT base-layer architectures, *Proceedings of IEEE International Conference on Image Processing (ICIP)*, Barcelona, Spain, September 14-17, 2003.

6. Y. Andreopoulos, A. Munteanu, G. Van der Auwera, P. Schelkens, and J. Cornelis, Single-rate calculation of overcomplete discrete wavelet transforms, *Signal Processing: Image Communication* (Submitted, December 2002).

7. Y. Andreopoulos, A. Munteanu, G. Van der Auwera, P. Schelkens, and J. Cornelis, Complete-to-overcomplete discrete wavelet transforms: theory and applications, *IEEE Transactions on Signal Processing*, (Submitted, April 2003).

8. J. Barbarien, Y. Andreopoulos, A. Munteanu, P. Schelkens, and J. Cornelis, Motion vector coding for in-band motion compensated temporal filtering, *Proceedings of IEEE International Conference on Image Processing (ICIP)*, Barcelona, Spain, September 14-17, 2003.

9. J. Barbarien, Y. Andreopoulos, A. Munteanu, P. Schelkens, and J. Cornelis, Coding of motion vectors produced by wavelet-domain motion estimation, *Proceedings of IEEE Picture Coding Symposium (PCS)*, Saint Malo, France, pp. Submitted, April 23-25, 2003.

10. J. Barbarien, Y. Andreopoulos, A. Munteanu, P. Schelkens, and J. Cornelis, Coding of motion vectors produced by wavelet-domain motion estimation, *ISO/IEC JTC1/SC29/WG11 (MPEG)*, Awaji island, Japan, MPEG Report M9249, December 7-12, 2002.

11. Błaszak Ł.,Doma□ski M., Ma□kowiak S., Spatio-Temporal Scalability in AVC codecs, ISO/IEC JTC1/SC29/WG11 Doc. MPEG2003/M9469 Pattaya, March 2003.

12. Błaszak Ł., Doma□ski M., Łuczak A., Ma□kowiak S., Spatio-Temporal Scalablity in DCT-based Hybrid Video Coders, ISO/IEC JTC1/SC29/WG11 Doc. MPEG2003/M8672, Klagenfurt, July 2002.

13. Doma□ski M., Błaszak Ł., Ma□kowiak S., AVC Video Coders with Spatial and Temporal Scalability, *Picture Coding Symposium*, 2003, Saint Malo, France, pp. 41-46.

14. Doma□ski M., Ma□kowiak S., Łuczak A., Błaszak Ł., Efficient Hybrid video Coders with spatial and Temporal Scalability, *IEEE International Conference on Multimedia and Expo*, ICME 2002, Lausanne, 26-29 August 2002, pp.133-136.

15. Doma□ski M., Ma□kowiak S., Błaszak Ł., Fine Granularity In Multi-Loop Hybrid Coders With Multi-Layer Scalability, *Proceedings of IEEE Conference on Image Processing ICIP'2002*, Rochester, NY, USA, Sept. 2002, v. III, pp. 742-744.

16. Doma□ski M., Ma□kowiak S., Błaszak Ł., Efficient Structure of Video Coders with Motion-Compensated Fine-Granularity Scalability, *Proceedings of XI European Signal Processing Conference EUSIPCO'2002*, Toulouse, France, September 3-6, 2002, v. II/III , pp. 141-144.

17. Doma□ski M., Ma□kowiak S., Błaszak Ł., Scalable video compression for wireless systems, *URSI X National Symposium of Radio Sciences*, Pozna□, March 2002, pp. 336-340.

18. H. Heijmans, MASCOT- Adaptive and Morphological Wavelets for Scalable Video Coding, *ERCIM News* 48, January 2002.

19. H. Heijmans, B. Pesquet-Popescu and G. Piella, Building nonredundant adaptive wavelets by update lifting, Report PNA-R0212, May 2002, CWI, Amsterdam; submitted to *Appl. Comp. Harm. Anal.*

20. H. Heijmans, G. Piella, B. Pesquet-Popescu
Building adaptive 2D wavelet decompositions by update lifting, *Proceedings of the IEEE International Conference on Image Processing*, Rochester, USA, 2002

21. R.M. Kalluri, M. van der Schaar, B. Pesquet-Popescu,
Differential motion vector coding with application to spatial scalable coding in FGS, *Proceedings of Image and Video Communications and Processing*, vol. SPIE 5022, 21-24 January 2003, Santa Clara, CA

22. A. Munteanu, Y. Andreopoulos, M. van der schaar, P. Schelkens, and J. Cornelis,
Control of the distortion variation in video coding systems based on motion compensated temporal filtering, *Proceedings of IEEE International Conference on Image Processing (ICIP)*, Barcelona, Spain, September 14-17, 2003.

23. B. Pesquet-Popescu. H. Heijmans, G.C.K. Abhayaratne and G. Piella,
Quantization in adaptive wavelet transforms for image compression,
accepted for *ICIP 2003, IEEE International Conference on Image Processing*, Barcelona, Spain, September 2003.

24. B. Pesquet-Popescu, G. Piella and H. Heijmans,
Adaptive update lifting with gradient criteria modeling higher-order differences, *Proc. ICASSP 2002*

25. B. Pesquet-Popescu, G. Piella and H. Heijmans,
Construction of adaptive wavelets using update lifting and quadratic decision criteria,
Proceedings of EUSIPCO, 2002

26. G. Piella, H. Heijmans and B. Pesquet-Popescu,
Adaptive update lifting with a decision rule based on derivative filters,
IEEE Signal Processing Letters 9 (10), p.329-332, 2002

27. G. Piella, H. Heijmans and B. Pesquet-Popescu,
Adaptive wavelet decompositions driven by the gradient norm,
Proceedings of the 3rd IEEE Benelux Signal Processing Symposium, Leuven, Belgium, 2002

28. Roger Piqué and Luis Torres,
Combining Adaptive Principal Component Analysis and a Hybrid Codec Approach for Face Coding in Video Sequences, in *PCS 2003, Picture Coding Symposium* (invited paper), St. Malo, France, April 23-25, 2003.

29. C. Tillier, B. Pesquet-Popescu, Y. Zhan, H. Heijmans,
Scalable Video Compression with Temporal Lifting Using 5/3 Filters, *Picture Coding Symposium*, 23-25 Apr. 2003, St. Malo, France.

30. C. Tillier, B. Pesquet-Popescu, 3D, 3-Band, 3-Tap Temporal Lifting for Scalable Video Coding, accepted for *ICIP 2003, IEEE International Conference on Image Processing*, Barcelona, Spain, September 2003

31. D.S. Turaga, M. van der Schaar, B. Pesquet-Popescu, *Differential motion vector coding in the MCTF framework, MPEG document* M9035, Oct 2002

32. Y.-W. Zhan, M. Picard, B. Pesquet-Popescu and H. Heijmans,
Long temporal filters in lifting schemes for scalable video coding, *Klagenfurt MPEG meeting*, document m8680, July 2002

5.3.2 MASCOT-related publication

This section list publications by MASCOT partners that have benefited from the MASCOT project.

33. Y. Andreopoulos, A. Munteanu, G. Van der Auwera, P. Schelkens, and J. Cornelis,
Wavelet-based Fully-Scalable Video Coding with In-band Prediction, *Proceedings of IEEE Signal Processing Symposium (SPS)*, Leuven, Belgium, pp. 217-220, March 21-22, 2002.

34. Y. Andreopoulos, A. Munteanu, G. Van der Auwera, P. Schelkens, and J. Cornelis, Scalable Wavelet Video-Coding with In-Band Prediction - Implementation and Experimental Results, *Proceedings of IEEE International Conference on Image Processing (ICIP)*, Rochester, New York, USA, Vol. 3, pp. 729-732, September 22-25, 2002.

35. Y. Andreopoulos, A. Munteanu, G. Van der Auwera, P. Schelkens, and J. Cornelis, A new method for complete-to-overcomplete discrete wavelet transforms, *Proceedings of 14th International Conference on Digital Signal Processing (DSP2002)*, Santorini, Greece, Vol. 1, pp. 501-504, July 1-3, 2002.

36. Y. Andreopoulos, K. Masselos, P. Schelkens, G. Lafruit, and J. Cornelis, Cache-misses and Energy-dissipation Results for JPEG2000 Filtering, *Proceedings of 14th International Conference on Digital Signal Processing (DSP2002)*, Santorini, Greece, Vol. 2, pp. 201-209, July 1-3, 2002.

37. Y. Andreopoulos, A. Munteanu, G. Van der Auwera, J. Barbarien, P. Schelkens, and J. Cornelis, Wavelet-based fine granularity scalable video coding with in-band prediction, *ISO/IEC JTC1/SC29/WG11*, Jeju, South-Korea, Report MPEG2002/m7906, March 2002.

38. Y. Andreopoulos, A. Munteanu, P. Schelkens, M. van der Schaar, and J. Cornelis, Control of the distortion variation in motion compensated temporal filtering, *ISO/IEC JTC1/SC29/WG11 (MPEG)*, Awaji island, Japan, MPEG Report M9253, December 7-12, 2002.

39. Y. Andreopoulos, M. van der Schaar, A. Munteanu, J. Barbarien, P. Schelkens, and J. Cornelis, Open-loop, in-band, motion-compensated temporal filtering for objective full-scalability in wavelet coding, *ISO/IEC JTC1/SC29/WG11 (MPEG)*, Shangai, China, Report M9026, October 20-25, 2002.

40. G. Van der Auwera, A. Munteanu, P. Schelkens, and J. Cornelis, Bottom-up motion compensated prediction in the wavelet domain for spatially scalable video coding, *IEE Electronics Letters*, vol. 38, no. 21, pp. 1251-1253, October 2002.

41. G. Van der Auwera, A. Munteanu, P. Schelkens, and J. Cornelis, Scalable Wavelet Video-Coding with In-Band Prediction - The bottom-up overcomplete discrete wavelet transform, *Proceedings of IEEE International Conference on Image Processing (ICIP)*, Rochester, New York, USA, Vol. 3, pp. 725-728, September 22-25, 2002.

42. J. Barbarien, P. Schelkens, and J. Cornelis, Inter-frame Wavelet Video Coding With Mesh-based Motion Compensation, *Proceedings of IEEE Signal Processing Symposium (SPS)*, Leuven, Belgium, pp. 101-104, March 21-22, 2002.

43. V. Bottreau, M. Benetiere, B. Felts, and B. Pesquet-Popescu, A fully scalable 3D subband video codec, *IEEE Conference on Image Processing 2001 (ICIP'01)*, Thessaloniki, Greece

44. J. Cornelis, R. Deklerck, B. Truyen, and P. Schelkens, Selected Topics in Medical Image Processing, *Proceedings of 9th International IEEE/IEE Workshop on Systems, Signals and Image Processing (IWSSIP)*, Manchester, United Kingdom, pp. 1-11, November 7-8, 2002.

45. Gavrilescu, A. Munteanu, P. Schelkens, and J. Cornelis, Embedded multiple description scalar quantizers and wavelet-based quadtree coding for progressive image transmission over unreliable channels, *Proceedings of IEEE International Conference on Image Processing (ICIP)*, Barcelona, Spain, September 14-17, 2002.

46. Javier R. Hidalgo and Philippe Salembier, Metadata-Based Coding Tools for Hybrid Video Codecs, *PCS 2003, Picture Coding Symposium*, St. Malo, France, April 23-25, 2003.

47. S. Ma□kowiak, Scalable Coding of Digital Video, *Doctoral dissertation*, Pozna□University of Technology, Pozna□2002.

48. P. Ndjiki-Nya, B. Makai, A. Smolic, H. Schwarz, and T. Wiegand,
Improved H.264/AVC Coding Using Texture Analysis and Synthesis, *ICIP 2003, IEEE International Conference on Image Processing*, Barcelona, Spain, September 2003.

49. P. Ndjiki-Nya, B. Makai, A. Smolic, H. Schwarz, and T. Wiegand,
Video coding Using Texture Analysis and Synthesis, in *PCS 2003, Picture Coding Symposium*, St. Malo, France, April 23.-25. 2003.

50. G. Piella and H. Heijmans,
Adaptive lifting schemes with perfect reconstruction,
IEEE Transactions on Signal Processing 50, p.1620-1630, 2002

51. G. Piella and H. Heijmans,
An adaptive update lifting scheme with perfect reconstruction,
IEEE Conference on Image Processing 2001 (ICIP'01), Thessaloniki, Greece, pp.190-193

52. Roger Piqué and Luis Torres,
Efficient face coding in video sequences combining adaptative Principal Component Analysis and a Hybrid Codec Approach, in *International Conference on Acoustics Speech and Signal Processing (ICASSP)*, Hong Kong, China, 2003.

53. Smolic and T. Wiegand,
High-resolution video mosaicing, *IEEE Conference on Image Processing 2001 (ICIP'01)*, Thessaloniki, Greece

54. F. Verdichio, I. Andreopoulos, A. Munteanu, J. Barbarien, P. Schelkens, J. Cornelis, and A. Pepino,
Scalable video coding with in-band prediction in the complex wavelet transform, *Proceedings of Advanced Concepts for Intelligent Vision Systems (ACIVS2002)*, Gent, Belgium, pp. 6, September 9-11, 2002.

55. Smolic, Y. Vatis and T. Wiegand,
Long-Term Global Motion Compensation Applying Super-ResolutionMosaics,
Proc. ISCE2002, IEEE International Symposium on Consumer Electronics, Erfurt, Germany, September 24.-26, 2002.

56. Y.-W. Zhan and H. J. A. M. Heijmans,
On the lifting construction of a class of non-separable 2D orthonormal wavelets,
Accepted for the *3rd International Symposium on Image and Signal Processing and Analysis, ISPA 2003*, September 18-20, 2003, Rome, Italy

57. A. Łuczak, M. Doma□ski,
Modeling of Bitstreams Produced by Hybrid Video Coders, *IEEE Symposium on Consumer Electronics ISCE'2002*, September 23-26, 2002 Erfurt, Germany, pp. E69 – E73.

58. A. Łuczak, M. Doma□ski,
Simple Global Model of an MPEG-2 Bitstream, *IEEE International Conference on Multimedia and Expo*, ICME 2002, Lausanne, 26-29 August 2002, pp.645-136.

6 Conclusions

6.1 Conclusions of the project

The MASCOT project was aimed towards the development of new tools in video compression in order to

- achieve high compression ratios
- develop intrinsically scalable compression schemes.

We believe that we have succeeded in achieving both goals, and perhaps even more than that. We have developed two new MASCOT codecs, the *MASCOT DCT-based codec* and the *MASCOT 3D wavelet codec*, which together can accommodate all the tools developed within the context of the project. This concerns metadata-based tools, new spatio-temporal decompositions, and new prediction tools. The latter also includes new coding techniques for motion vectors.

Since the reference baseline H.264/AVC codec represents the state of the art in video coding, the integrated MASCOT DCT-based codec can be regarded as the most advanced DCT-based video codec available in the world. Moreover the MASCOT DCT-based codec provides fine granularity scalability. Also the MASCOT 3D wavelet codec has been provided with significant improvements compared to the state of the art reference 3D wavelet codec of Woods. Therefore also this codec can be regarded as one of the most, if not the most, advanced wavelet video codec currently available. These outstanding achievements could only be realized through the collaborative research in this EU funded project.

Two main conclusions of MASCOT project:

1. **Metadata can be useful for video coding**
2. **Scalability is an important issue for the future**

Regarding the second conclusion it must be observed that further exploration of both hybrid DCT-based and 3D wavelet-based architectures is necessary to ensure competitive rate-distortion behavior compared to non-scalable MPEG-4 video coding technology and to allow for the selection of a suitable architecture.

The first conclusion is justified by the results summarized in Section 4.3, and the second conclusion is supported by the results in Section 4.4, as well as by the latest developments in the MPEG community.

6.2 Future Outlook

As explained before MASCOT members have been very active from the beginning in the new MPEG initiative on scalable video coding. They contributed several technical inputs and participated in the development of requirements, testing methodology, etc. This work resulted in a “Call for Evidence” on scalable video coding technology that was recently released by MPEG. The technology developed in MASCOT will be proposed in answer to that Call at the MPEG meeting in July 2003 in Norway. This includes 3D wavelet technology as well as scalability based on H.264/AVC.

MASCOT was a project in the FET Open Domain action line which is directed towards emerging technologies. In general, projects within this program involve a relatively high risk of (partial) failure, but also hold the promise of innovative technologies. Some of the techniques that have been tested in MASCOT were very new (like the morphological wavelets), and some of them were for the first time developed within the context of this project, like the adaptive wavelets. Both techniques hold a promise for the near future but are still too immature and too little understood to be competitive with tools that have been around for many years.

This project has also proved the validity of the metadata-based coding approach. Several techniques were developed in the framework of this project that showed how metadata descriptors can be used to improve the coding efficiency of current video codecs. It should be noted that in the context of the MASCOT

project, the word metadata refers to data that has been extracted from the content in order to provide indexing capabilities an to allow new functionalities to the content such as search, retrieval, browsing, etc. The results obtained in this project have shown that this extra metadata information can also be employed to optimize the coding strategies of current standard hybrid codecs. To investigate this, metadata standards such as MPEG-7 and SMPTE were explored and several metadata descriptors were selected as good candidates to improve the coding efficiency. Different techniques were developed in order to exploit this new information. For instance, the texture tool and long term selection of reference frames showed significant bitrate savings (up to 20% and 10%) compared to the standard H.264/AVC video codec.

The MASCOT project has created a general framework where new metadata-based coding tools can be developed and analysed. Several of the techniques developed in the project are also opened for further research and development (see table in Section 1). New areas of open research have been discovered through the development of the project such as investigating new techniques to efficiently encode metadata descriptors or finding efficient strategies for rate control between data and metadata.

7 References

7.1 References to Literature

[1] H. Heijmans and R. Keshet, Inf-semilattice approach to self-dual morphology, *Journal on Mathematical Imaging and Vision* 17 (1), p. 55-80, 2002.

[2] H. Heijmans, B. Pesquet-Popescu and G. Piella, Building nonredundant adaptive wavelets by update lifting, *Report PNA-R0212*, May 2002, CWI, Amsterdam (submitted)

[3] G. Piella and H. Heijmans, Adaptive lifting schemes with perfect reconstruction, *IEEE Transactions on Signal Processing* 50, p. 1620-1630, 2002

[4] G. Van der Auwera, A. Munteanu, P. Schelkens, and J. Cornelis, Bottom-up motion compensated prediction in the wavelet domain for spatially scalable video coding, *IEE Electronics Letters*, vol. 38, no. 21, pp. 1251-1253, October 2002.

[5] Y. Andreopoulos, A. Munteanu, G. Van der Auwera, P. Schelkens, and J. Cornelis, A new method for complete-to-overcomplete discrete wavelet transforms, *Proceedings of 14th International Conference on Digital Signal Processing (DSP2002)*, Santorini, Greece, Vol. 1, pp. 501-504, July 1-3, 2002.

[6] Y. Andreopoulos, M. van der Schaar, A. Munteanu, J. Barbarien, P. Schelkens, and J. Cornelis, Fully-scalable wavelet video coding using in-band motion compensated temporal filtering, *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Hong Kong, April 6-10, 2003.

[7] Y. Andreopoulos, M. van der Schaar, A. Munteanu, J. Barbarien, P. Schelkens, and J. Cornelis, Complete-to-overcomplete discrete wavelet transforms for scalable video coding with MCTF, *Proceedings of SPIE Visual Communications and Image Processing (VCIP)*, Lugano, Switzerland, July 8-11, 2003.

[8] Y. Andreopoulos, A. Munteanu, G. Van der Auwera, P. Schelkens, and J. Cornelis, Scalable Wavelet Video-Coding with In-Band Prediction - Implementation and Experimental Results, *Proceedings of IEEE International Conference on Image Processing (ICIP)*, Rochester, New York, USA, Vol. 3, pp. 729-732, September 22-25, 2002.

7.2 References to MASCOT Deliverables

[9] MASCOT Deliverable 1.1, *Specification of Architecture of the MASCOT Codec*, version 5.0, April 2002.

[10] MASCOT Deliverable 1.2, *Test Model for MASCOT's Coding Scheme*, version 3.0, January 2003.

[11] MASCOT Deliverable 1.3, *Report on Coding Performance*, version 5.0, April 2003.

[12] MASCOT Deliverable 2.1, *Analysis and Selection of Descriptors and Description Schemes Supporting Encoding Tools and Preliminary Metadata-Based Encoding Tools and Strategies*, version 3.0, March 2002.

[13] MASCOT Deliverable 2.2, *Metadata-Based Encoding Tools*, version 3.0, December 2002.

[14] MASCOT Deliverable 2.3, *Metadata-Based Encoding Strategy*, version 1.0, April 2003.

[15] MASCOT Deliverable 3.1, *New spatio-temporal decompositions*, version 2.0, January 2003.

[16] MASCOT Deliverable 3.2, *Impact of new decompositions on video compression*, version 1.0, April 2003

[17] MASCOT Deliverable 4.1, *Alternate prediction parameters in the space domain for better encoding, motion compensation and segmentation, long term prediction*, version 1.0, November 2002.

[18] MASCOT Deliverable 4.2, *Evaluation of new concepts for motion compensated prediction in the wavelet domain – long term memory prediction and NSI filters*, version 1.1, January 2003.

[19] MASCOT Deliverable 4.3, *Evaluation, Optimization and Development of Global Motion Compensation and Sprite Coding Algorithms Exploiting Motion-Based Metadata Test*, version 2.0, March 2003.

[20] MASCOT *Technology Survey*, version 1, October 2001.