METADATA-BASED CODING TOOLS FOR HYBRID VIDEO CODECS

Javier R. Hidalgo and Philippe Salembier

Universitat Politècnica de Catalunya, Barcelona, Spain.
{jrh,philippe}@gps.tsc.upc.es

ABSTRACT

Recent initiatives such as MPEG-7 and SMPTE Metadata Dictionary have defined the syntax and semantics of metadata that can be used to describe and index multimedia content. This type of metadata is used in various applications such as indexing, querying and browsing, etc. Within this framework, it can be expected that, in the near future, *metadata* databases will be available for most existing video content. Although the primary goal of metadata is for content retrieval, filtering and browsing, it may also be used to improve the compression efficiency of video codecs. Future image and video encoders will be able to access the metadata and use it to improve their encoding strategy. In this paper, new coding techniques using metadata are presented.

1. INTRODUCTION

It can be expected that, in the next years, a large amount of audio-visual documents will be indexed and if not, metadata will be rather easy to create. *In this paper, the word **metadata** refers to information that has been generated to describe the content with the objective to search, query, filter and browse* (e.g. MPEG-7 [4] or SMPTE [7] metadata standards). In many circumstances then, audio-visual material will be available together with the metadata describing its content. As a consequence, encoders will be able to access to this information and use it in order to improve their efficiency or to optimize their strategy.

Metadata-based encoders are aware of the syntax and semantics of metadata descriptions related to the video content. They optimize their efficiency and modify their strategy taking advantage of the metadata already available. In this framework, the key point is to know whether metadata created for indexing purpose can help in the encoding process. For example, one may ask whether the knowledge of either low-level indexing description such color histogram, texture descriptors or high-level description such as Table Of Contents or Indexes can improve the rate-distortion performances of codecs. In this paper, three different examples are presented. They try to show how metadata can be used for video coding.

Metadata can be used in various ways. In some situations, the encoder will make use of the metadata to simply optimize its encoding strategy and the resulting bitstream will still be compatible with metadata unaware decoders. That is, the decoder will not need any extra information to decode the received bitstream. However, in other situations, in order to fully exploit the metadata, the encoder will have to severely modify its encoding strategy and the bitstream compatibility with classical decoders will be broken. In that situations, the decoder will also need the metadata in order to correctly extract the video content from the bitstream.

Fortunately, there are scenarios where metadata will be available at both the encoder and the decoder sides at no cost. Consider for instance a scenario where a user is browsing a database of movies from a video content provider. The user is able to browse the archive, search and query for different movies. During this initial search phase, the user has the opportunity to get a local copy of the metadata. Once the user has selected a movie and before the downloading starts, both encoder (at the sever side) and decoder (at the user side) have the common knowledge of the metadata used in the browsing, query and searching steps. This extra information can then be used by both the encoder and the decoder to improve the coding efficiency and no transmission of metadata has to be done with the transmission of the actual content.

In other situations, the metadata will need to be sent together with the content. In that case, a careful bit assignment strategy between content and metadata will have to be designed. For example, rate-distortion algorithms may be applied in order to efficiently allocate a given bit rate between data (video content) and metadata (descriptors of the content) which opens a wide area for future research. Note that the availability of metadata at the decoder side and the necessity to transmit it depends mainly on the application and not so much on the type of metadata nor on the tool that uses this metadata. This paper is mainly focused on describing some possible tools that use metadata to improve the encoding efficiency. The proposed techniques can be included in various applications; some of them may require

The authors would like to thank the support from the European Commission and in particular, the MASCOT FET Project (IST-2000-26467).

the encoding of metadata with the content while other applications may not.

The structure of the paper is as follows. The following section introduces the techniques developed to improve video coding using metadata. Section 2.1 explains the coding of video transitions in hybrid codecs. Section 2.2 details the improvement of long term temporal prediction using low-level color descriptors. Section 2.3 describes how metadata can help in the bit rate allocation of encoders. Finally, conclusions are drawn on section 3.

2. METADATA-BASED CODING TOOLS

This section presents three different metadata-based techniques. They are based on current hybrid codecs such as MPEG-4, H.26L or JVT/AVC [5, 2]. The first tool makes use of existing metadata (MPEG-7 VideoEditing Description Scheme for example) that describe the scene transition of a video sequence. The inter-frame prediction during the transition can exploit the modeling of the transition performed by the metadata descriptor. For instance, the codec may use the knowledge that previous frames get darker (or brighter) in a fade-out (or fade-in). Also, geometric transformations can be applied to reference frames inside the transitions. In the case of dissolves, the interpolative mode of B-Frames can be modified to use weights of the transition model.

The second proposed technique shows how long term temporal prediction can be improved using existing metadata. Long term prediction introduced the idea to use N references to predict the blocks of the current image instead of using only one reference (usually the previous one). It will be shown how simple metadata such as color descriptors (for example, the MPEG-7 ColorLayout Descriptor) can make a pre-selection of N reference frames.

Finally, the last technique presented in the paper consists of a rate-control technique using metadata. Current rate-control algorithms are facing a complex problem as changes in the quality factor or the IPB structure of the GoP affects future coded frames. We will see how motion and texture descriptors (for example, the MPEG-7 MotionActiviy or TextureBrowsing Descriptors) can help in the decision of determining a specific quality factor or selecting a good IPB GoP structure.

2.1. Shot Transitions

Transitions between video shots are common in standard video sequences [1]. Recent metadata standards have proposed some descriptors to organize and describe the evolution of the transition between different shots. For instance, MPEG-7 has a description scheme, called *VideoEditingDS*, that describes a shot transition by specifying the first and the last frames of the transition, as well as a weight function for the modeling of the transition and the fade color (in the case of a fading). This metadata can be used to improve coding inside transitions.

Generally, the visual information during transitions usually depend on both the previous and the past shots. In a hybrid codec framework, the B frame type (frames that can be coded using previous and past references) can be exploited to improve the coding efficiency within the transition. 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 so that only B frames exist within the transition. To do so, the first P frame in the transition is moved to the last 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 so that it can use different weights for previous and past references (instead of the fixed 0.5 weights for each reference). The term $\overline{B}$ will be used as the modified version of B frames able to deal with different weights for previous and past references.

To precisely exploit the weights obtained from the metadata, a new motion estimation block is needed. In current standards such as MPEG-4 and AVC, the interpolative mode is performed using the motion vectors that have been estimated on the previous and past references separately. When using different weights for the previous and the past references, this strategy turns out to be rather inefficient (when the weight associated to a reference is low, the motion estimation is extremely noisy). In that case, a re-calculation of motion vectors using the associated transition metadata weights is needed. As computing a pair of motion vectors using a full search approach is computationally very expensive, an intermediate approach using a 2D-log search algorithm has been used in all results. Re-computing motion vectors using a 2D-log search algorithm has shown better results than using motion vectors estimated separately from the forward and backward reference.

Results such as those reported in Figure 1 have shown promising gains when coding transitions of short duration using metadata. In the case of long gradual transition with high internal motion, the performance of this method drops because having references only at the beginning and the end of the transition (which can be several seconds long) is not enough. To overcome this problem, new references $\overline{B}_r$ can be introduced in the dissolve so that internal $\overline{B}$ frames can

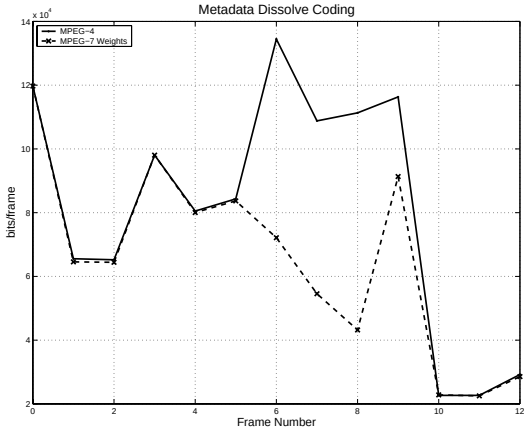

Fig. 1. Bit-rate using MPEG-7 metadata *VideoEditingDS* in the coding of a short transition of 4 frames (between frames 4 and 9). Keeping the same quality, the bitrate drops from 74Kbps (using the standard MPEG-4) to 63.5Kbps using the same codec and transition metadata.

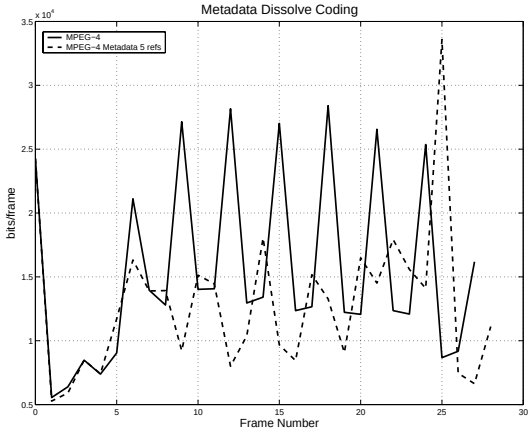

Fig. 2. Bitrate results for a long transition of 25 frames.

be predicted using frame references that are closer in time (and therefore may perform better as predictors).

$\overline{B}_r$ frame types are similar to $\overline{B}$ frames but they can act as references for intermediate $\overline{B}$ frames. In that case, the metadata weights used (and referenced to the start and end of the transition) may be re-computed to the new references within the transition. Consider for instance, that two references, named $\overline{B}_r^f$ and $\overline{B}_r^l$, are included in the dissolve. These references use the first and last frames of the transition as references (P^f and P^l respectively). Frames between those two references should use $\overline{B}_r^f$ and $\overline{B}_r^l$ instead of P^f and P^l. The weights used by the metadata-based encoder should then be: $\overline{B} = w_1 \cdot \overline{B}_r^f + w_2 \cdot \overline{B}_r^l$. On the other hand, we have

$$\begin{aligned} \overline{B} &= a \cdot P^f + b \cdot P^l \\ \overline{B}_r^f &= c \cdot P^f + d \cdot P^l \\ \overline{B}_r^l &= e \cdot P^f + f \cdot P^l \end{aligned} \tag{1}$$

where the weights a, b, c, d, e, f are obtained directly from the metadata. So, the final weights w_1 and w_2 that refer to $\overline{B}_r^f$ and $\overline{B}_r^l$ can be computed as:

$$w_1 = \frac{af - be}{cf - de} \qquad w_2 = \frac{ad - bc}{de - cf} \tag{2}$$

Figure 2 shows a result of a large gradual transition. A linear dissolve of 25 frames is coded using the metadata framework. The internal motion of the two shots composing the dissolve is high enough so that middle frames of the dissolve cannot be predicted using only the first and the last frames of the transition. Adding extra references in the dissolve can improve the coding efficiency. In this example, the dotted line shows the results using 3 $\overline{B}_r$ frames inside the dissolve. For the same quality, the bitrate in this example drops from 15.13 Kbps/frame for the standard MPEG-4 codec to 11.82Kpbs/frame using new references inside the transition.

2.2. Long Term Temporal Prediction

Normal video sequences have a high degree of temporal redundancy. Standard hybrid coders exploit this fact by using past (or future) frames as references for coding the current frame. It is natural to think that the closer frame is in time, the most similar it is to the frame being coded. For that reason current hybrid coders use the closest P or I frame in time as a reference for coding the current frame. Studies [8] have shown that using more than one reference frame for encoding can increase the coding efficiency. The best possible reference for a current block being coded may be several frames in the past. In that case, all encoded blocks in a single frame do not share the same reference frame. For each block, it is then necessary to add information about which frame has been used as reference. This implies an increment of the bit-rate. If the number of possible reference frame is limited, results show that the gain in prediction error is, however, greater than the bit-rate needed to encode the reference and so, the final rate-distortion efficiency increases.

Standards like H.26L and the recent JVT have adopted these new ideas into a long-term temporal prediction. A group of N reference frames is created and each block to be coded selects the best reference block among the N possible references. In the same sense as before, the N possible references are generally the N closest P or I frames in the video sequence. In practice, the number N of possible reference frames is limited due to two factors: The computational complexity of performing a motion estimation for all reference frames and the bitrate increase needed to add the

information about the reference frame.

Metadata can be introduced in long term temporal prediction to improve the overall coding efficiency. Metadata can be used to perform a pre-selection of N possible 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. Similarity based on metadata can be used to pre-select N frames (that will be used as reference frames) among a high number of possible candidates frames. Metadata can further help the searching of possible reference frames: for example, shot descriptors can be used so that reference frames are only searched in shots that have similar content. Finally, existing metadata about collections of audio-visual documents can point the encoder to similar content of other sequences. For example, a user may watch and store the news everyday at 7pm. Metadata can inform the encoder that several bitstreams corresponding to past days are already stored and available to the encoder. The codec can make use of those streams as possible reference for coding the current sequence.

In this search and retrieval scenario to select reference frames, many low-level descriptors can be used. Consider as an example the MPEG-7 *ColorLayout* Descriptor. This simple descriptor carries information about the color structure of a single image. The basic idea is to use the similarity criterion of the descriptor to pre-select, among a large number of previous frames, the N closest frames to the one being coded. These "closest" frames are frames with similar color layout descriptors to the frame being coded. Doing this frame pre-selection and taking into account the low-complexity matching criterion of the color layout, the overall computational cost of long term temporal prediction is not severely increased. Moreover, as the same number of references frames N is used, there is no increase in the bitrate 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.

Figure 3 shows the results when coding a news sequence (from the MPEG-7 database). A pre-selection of 4 reference frames over 100 previous frames is performed using the color layout similarity measure. The temporal prediction of frame at time t is based on the frame at time $t-1$ plus the four reference frames obtained with the Color Layout descriptor. The rate-distortion plots shows an increase up to $0.9dB$ when using color layout metadata to pre-select reference frames. If the metadata is not available in the decoder site, it has to be sent together with the data. Results show that the extra information (compressed using the standard gzip) only represents the 9% of the bitrate savings when

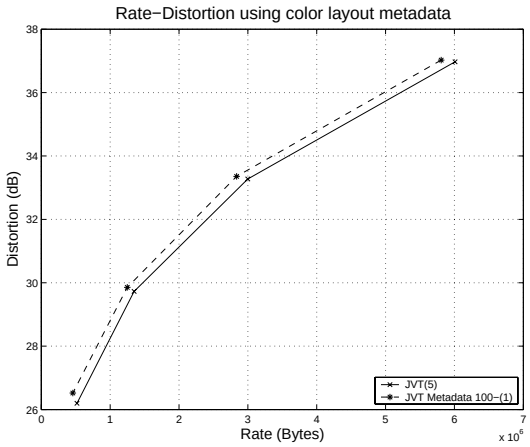

Fig. 3. Rate-distortion of the JVT codec vs. JVT using metadata.

coding QCIF images. This results on less than $0.08dB$ loss on PSNR when the color layout metadata must be sent together with the bitstream.

2.3. Rate Control

The problem of bit allocation given a specific budget consists of distributing the given bitrate over the complete sequence maximizing the rate-distortion quality. In [6], the bit allocation is performed by controlling the quality factor of the DCT quantization process. The bit budget is distributed between all frames of the video segment being coded taking into account the different relations of the coded frames. This relations may be complex as fixing the quality factor of a P frame, for instance, can determine the distortion associated to future P or B frames that use it as reference. The distribution of this quality factor in all frames minimizes the overall distortion of the final coded video scene.

The bitrate may also be reduced by carefully choosing the type of frames (I, P or B frames) between the video scenes [3]. As before, the distortion of the encoded video is minimized under the restriction of a limited bit budget but now, the quality factor is fixed and the type of coded frames is changed to minimize the overall distortion. Again, relationships between coded P and B frames make this a challenging problem as future decisions will depend on current ones.

Metadata can be useful to improve this step and may help 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 can be used to select different IBP structures according to the amount of motion present in the bitstream. This information can help in the decision of choosing dif-

ferent frame types for coding the sequence. 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.

Texture metadata, such as the MPEG-7 *Homogeneous Texture DS* or *TextureBrowsingDS* descriptors, describe the regularity of the texture present in an image. As the standard DCT transformation used for the residual coding behaves better on smooth textures, this information can be used to select the quantizer levels on the DCT coefficients. Areas with detailed textures may need higher quantizer levels than textured areas without details. Texture descriptors can be used to improve this step. Smooth texture can be associated with low Q parameters without any loss of perceptual quality. On the other hand, when texture metadata signals the presence of detailed texture, the Q parameter that controls the quantization and truncation step can be increased to maintain the same distortion quality.

Table 1 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 (the 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 to 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 classical "2 B frames between P frames" GoP structure. 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 bitrate for the same quality).

Shot Number	Motion Activity	Bitrate (2 B-frames)	B frames (best)	Bitrate (best)
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 1. 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.

3. CONCLUSIONS AND FUTURE WORK

A set of metadata based coding tools have been presented. Note here that metadata refers to indexing information that has been generated to describe the content with the objective to search, query, filter and browse. These tools exploit metadata descriptors about the content of the video scene in order to improve the coding strategy and performances. The results reported here show promising gains when enabling metadata based coding into current standards video codecs such as MPEG-4 and JVT/AVC.

Our future research is directed toward investigating metadata descriptors to include in this framework and combine different metadata descriptors into the same tools. For instance, results on Table 1 show that the simple motion activity metadata can signal the presence of high motion even if the codec is able to motion compensate it (see for instance shot number 4). In this case, other metadata descriptors are currently investigated in order to obtain more robust estimations of the amount of motion.

Finally, different type of codecs (such as wavelet or 3D codecs) can also be investigated to see if they can benefit from the metadata-based coding strategy.

4. REFERENCES

[1] John S. Boreczky and Lawrence A. Rowe. Comparison on video shot boundary detection techniques. *SPI Storage and Retreival for Image and Video Databases*, 5(2):122–128, April 1996.

[2] JVT. http://standard.pictel.com/ftp/video-site.

[3] J. Lee and B.W. Dickinson. Rate-distortion optimized frame type selection for mpeg encoding. *CSVT*, 7(3):501–510, June 1997.

[4] B. Manjunath, P. Salembier, and T. Sikora, editors. *Introduction to MPEG-7*. Wiley, 2002.

[5] F. Pereira and T. Ebrahimi, editors. *The MPEG-4 Book*. Prentice Hall, 2002.

[6] K. Ramchandran, A. Ortega, and M. Vetterli. Bit allocation for dependent quantization with applications to multiresolution and mpeg video coders. *IP*, 3(5):533–545, September 1994.

[7] SMPTE Metadata Dictionary RP210.4. http://www.smpte-ra.org/mdd.

[8] T. Wiegand, X. Zhang, and B. Girod. Long-term memory motion-compensated prediction. *CSVT*, 1997.