Efficient Scalable Coding of Video Summaries using Dynamic GOP Structures

Lino Ferreira
ESTG, Instituto Politécnico de Leiria
Instituto de Telecomunicações
Leiria, Portugal
Email:lino.ferreira@ipleiria.pt

Luís Cruz
DEEC, Universidade de Coimbra
Instituto de Telecomunicações
Coimbra, Portugal
Email:lcruz@deec.uc.pt

Pedro Assuncao
ESTG, Instituto Politécnico de Leiria
Instituto de Telecomunicações
Leiria, Portugal
Email:pedro.assuncao@ipleiria.pt

***Abstract*—A method to efficiently encode an arbitrary video summary with temporal scalability and dynamic Group of Pictures (GOP) structures is proposed in this paper. The video summary is encoded as the base layer of a Scalable Video Coding (SVC) bitstream using a novel approach, which matches the summary frames available in temporal segments onto corresponding dynamic GOP structures. An algorithm is devised to compute variable GOP sizes along with an efficient independent prediction structure for the summary. The results show that the proposed method can be used to encode arbitrary video summaries with increased efficiency in the temporal base layer and negligible loss of R-D performance in the whole scalable sequence.**

***Index Terms*—Video summarization, SVC, H.264/MPEG-4 SVC**

I. Introduction

Nowadays, there is a great diversity of multimedia content used in services and applications, which require efficient and flexible management tools for different purposes, such as adaptation, indexing, searching, and browsing. Video summarization can be viewed as special case of content adaptation, which provides high level of flexibility to access the most relevant information contained in a video sequence. A video summary is basically a short version of the whole sequence, i.e., either a subset of key frames or a set of shorter video subsequences, chosen as essential to represent the most important chunks of the original video content according to predefined criteria. Summarization can be based on descriptors such as color, motion, audio, or other type [1]. For instance, a basic functionality provided by a video summary is to allow users to get a rough idea of the content without having to watch the whole sequence (e.g., browsing, fast forwarding, etc). Another useful application of summarization is to provide content adaptation in constrained communication environments where bandwidth, storage capacity, decoding power or visualization time is rather limited.

There are two main types of video summaries, namely those based on *key frames* and those comprised of *video skims* [2]. A summary based on *key frames* is a set of isolated frames selected from the relevant video shots of the original video. This type of summary is static, since the *key frames* do not include the temporal evolution of the sequence. *Video skims* are usually built by extracting the most relevant temporal segments from the source sequence. This type of summary is dynamic and includes information about the local temporal evolution of the sequence.

A straightforward representation method for video summaries is to generate and encode two independent streams, one with the summary and another one with the original sequence. However, such approach is highly redundant and does not benefit from the most recent advances in efficient coding techniques, namely scalable video coding [3]. Using single layer AVC encoding is also a difficult option, because some additional signaling, not compliant with standard decoders and would be necessary to extract a video summary from such a stream. If temporal scalability is used, then the video summary can be independently decoded from the bitstream while the whole sequence is still available without losing coding efficiency. However, scalable encoding of a video summary is not simple because of its inherent variable temporal rate. The usual regular GOP structures are extremely rigid to efficiently accommodate variable temporal rates and flexible prediction structures to allow independent coding of the video summary.

The problem of scalable coding of video summary has been recently addressed in [4], [5] and [6]. In [4] the authors propose a hierarchical frame selection scheme which considers semantic relevance in video sequences at different levels from compressed wavelet-based scalable video. In [5] a method to generate video summaries from scalable video streams based on motion information is presented, while in [6] the authors propose to partition a video summary in summarization units related by the prediction structure and independently decodable. In these approaches the full scalable stream is transmitted and video summary is generated at the user side.

This paper proposes a rather different approach where a video summary generated before encoding is the base layer of a temporally scalable stream whilst the remaining frames are encoded in the upper layers. This is achieved by using a novel dynamic GOP size selection algorithm, capable of mapping the highly variable temporal structure of the video summary to a new prediction structure of scalable video coders. The coding approach followed in this paper is independent from both the methods used to generate video summaries and the type of summary, i.e., either *key frames* or *video skims*. After the summarization process, the video frames of the summary are identified and the respective temporal indices are given to

the proposed scalable encoder which dynamically computes the GOP size based on a coding efficiency criterion given by a function which accounts for the mean squared error (MSE) between the frames of the summary. Note that the video summary frames are implicitly identified as they form the base layer of a temporally scalable stream.

The paper is organized as follows. Section II briefly describes the metrics associated with video summaries, section III presents the proposed method to compute dynamic GOPs and section IV describes the experimental procedure and discusses the results. Finally section V concludes the paper.

II. Video Summary Associated Metrics

In the context of this work, the most relevant metrics are the temporal rate $R(S)$ and temporal distortion $D(S)$ of a video summary S, given by (1) and (2), respectively. $R(S)$ is the ratio between the number of frames m belonging to the video summary S and the total number of frames of original sequence n, i.e.,

$$R(S) = m/n \tag{1}$$

The temporal distortion $D(S)$ of summary S is defined as the average frame distortion between the original sequence and the one reconstructed from the summary,

$$D(S) = 1/n \sum_{k=0}^{n-1} d(f_k, f_k^{'}) \tag{2}$$

where f_k is a frame in the original sequence and $f_k^{'}$ is the temporally co-located frame in the reconstructed sequence. Frames $f_k^{'}$ in the reconstructed sequence, are either frames from the video summary, i.e., the same as the original, or copies of the last frame in the video summary, i.e., the same frame is repeated along the time where summary frames do not exist. Such reconstruction method expands the video summary by filling the gaps in the temporal domain through a frame replication process (i.e. performing zero-order interpolation). Other reconstruction methods could use more sophisticated techniques by using, for instance, motion compensated interpolation [7]. The frame distortion $d(f_j, f_k)$ is measured between frames f_j and f_k as MSE, i.e.,

$$d(f_j, f_k)_{MSE} = \frac{1}{h * w} \sum_{y=0}^{h-1} \sum_{x=0}^{w-1} (f_j(x,y) - f_k(x,y))^2 \tag{3}$$

The video summarization process used in this work is defined as a temporal rate-distortion optimization problem where the objective is to find the subset of images of the original video that provides its best representation within a given temporal rate budget R_{max}, i.e., without using more than $m = R_{max} * n$ frames. Given the temporal rate constraint R_{max}, the optimal video summary S^* is the one that minimizes the summarization distortion [8], given by,

$$S^* = \arg\min_S D(S), s.t. R(S) \leq R_{max} \tag{4}$$

Note that the scalable coding method proposed in the following sections does not depend on the process used to generate the video summary.

III. Dynamic GOP Size and Prediction Structure

A. Dynamic GOP size selection

In general, the GOP structure used in scalable video coding is fixed and regular over time in order to provide a hierarchical coding structure [3]. In this type of GOP structure the number and type of frames (either P or B) are predefined as encoder configuration parameters. The I frames determine the GOP boundaries. All type of frames are allowed in the temporal base layer, i.e. I, B and P frames, while P and B frames are encoded only in the upper layers following a regular structure over the whole sequence. Note, however, that such regular GOP structure is not mandatory to be compliant with the SVC extension to the standard. Since a video summary does not have a regular frame rate, rather than using a fixed GOP structure, it is better to use GOPs of variable size according to the frames available in the video summary.

In order to match the variable temporal rate of a video summary to a GOP structure, the total number of B and P frames within a GOP must also be variable. Therefore in a dynamic GOP structure, the number of B frames between I or P frames and number of P frames between I frames are variable, depending on the video summary frames.

Essential motivation of constructing dynamic GOP structure is to achieve not only temporal scalability but also higher coding efficiency. Similar frames in the same GOP will help to improve bit saving whilst dissimilar frames cannot be efficiently encoded using temporal prediction. Thus, the dissimilar frames are strong candidates to be coded as I frame at the GOP boundaries. If the most dissimilar frames within a limited set of summary frames (i.e., the maximum allowed GOP size) are selected for the GOP boundaries, then good coding efficiency is expected because all the remaining frames are the most similar ones, which is in favour of efficient temporal prediction.

The proposed algorithm searches for the best GOP boundaries in the video summary in order to achieve high coding efficiency, both in the base layer (i.e., the video summary) and in the whole sequence (i.e. all layers). To find the GOP boundaries we use the summary frame distortion D_c, defined in (5), where $d(f_c, f_j)$ is the MSE, defined in (3), between the candidate GOP boundary frame f_c and all possible video summary frames f_j within a maximum distance of 32 frames. c is the index of the candidate summary frame and A is the set of ordered summary frame indices, $A = \{l_0, l_1, .., l_{m-1}\}$, such that $l_0, < l_1, .., < l_{m-1}$. Note that $l_0, l_1, .., l_{m-1}$ are defined in the original frame sequence, thus A does not comprise an arithmetic progression.

$$D_c = \sum_{\substack{c-32 \leq j \leq c+32 \\ j \epsilon A}} d(f_c, f_j)_{MSE}; \;\; c = l_0, l_1, .., l_{m-1} \tag{5}$$

D_c is computed for all summary frames and the best upper boundary index for GOP n, l_n^* is given by,

$$l_n^* = \arg\max_c (D_c) \text{ where } l_{n-1}^* < c \leq l_{n-1}^* + 31 \tag{6}$$

l^*_{n-1} is the lower boundary index of GOP n, which is also the best upper boundary index of GOP $n-1$. In the first GOP, the lower boundary is l_0. In this work, the maximum allowed GOP size is 32, since this size provides enough variation headroom for the GOP size and good coding efficiency [3]. The summary is determined before encoding, though some isolated frames can be inserted in the summary during encoding for coping the maximum GOP size. In the case where consecutive summary frames have a temporal distance higher than 32 (e.g. *key frames* or *video skims* summaries), then the algorithm forces the GOP size to take the value of 32 effectively promoting a non-summary frame to a summary frame.

B. Prediction structure in temporal scalable coding

Fig. 1 shows the prediction structure which results from the dynamic GOP size allocation using the method described above. As shown in the Figure, the video summary is encoded in the temporal base layer while the upper layer contains the remaining frames of the sequence. Therefore the full temporal resolution is obtained when both temporal layers are decoded. Since the reference frames of the video summary are all encoded in the temporal layer 0, the summary layer can be extracted and independently decoded from the whole coded stream.The R-D performance of coding a GOP depends not only the coding order but also on the employed reference frames. However, in general, as the temporal interval between each frame and its reference gets shorter, temporal predictive coding becomes more efficient. Therefore, among the available frames in the DPB, we choose the nearest frames of the current one as its forward and backward references. In this work, all frames between the GOP boundary are encoded as B, though they can also be encoded as P type.

In H.264/AVC with scalable extension, the proposed scheme changes the default order of reference pictures in List0 for the P slices or in List0/List1 for B slices. For a correct decoding, the reference frames of the video summary must be signaled to the decoder using Reference Picture List Re-ordering [9].

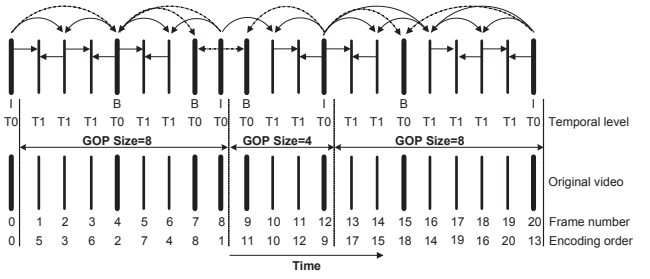

Fig. 1. Example of a prediction structure resulting from dynamic GOP allocation (thicker lines represent video summary frames).

IV. Experimental Results

The proposed method was implemented in the JSVM 8.9 reference software, and the test sequences "Soccer" and "Foreman", QCIF@30Hz were used in the experiments. The main encoding parameters used in the simulations were: *NumberReferenceFrames 2*; *FastSearch*; *Loop Filter on*; *CABAC*; *FRExt no*; *MaxDelay 1200*. The R-D operational points used to obtain the R-D functions were obtained from the set of $QP : \{25, 30, 35, 40, 45\}$. For each sequence, three different video summaries were generated, using the algorithm proposed in [8], with temporal rates $R(S)$ of 25%, 12% and 6%. For comparison, a temporally subsampled version of each sequence, acting as a reference summary, was also encoded as the base layer of a temporally scalable stream at the same R(S), i.e., with the same total number of frames, using the fixed GOP structure of SVC. Thus such reference summary also represents the whole sequence by a subset of frames with the same size and provides the same functionality of being independently decodable.

A. Efficiency of video summary coding

The coding efficiency of the base layer obtained by encoding the video summaries through the proposed dynamic GOP structure was compared with the standard SVC regular GOP structure, again for exactly the same temporal rate $R(S)$. This is actually a different summary because it is comprised of regularly spaced frames. However this is still a fair comparison because the same number of frames is used to represent the whole sequence as a summary.

Fig. 2 and 3 show that better efficiency is obtained using the proposed method, compared with that of SVC with fixed GOP size. The difference between the two methods is small for the two sequences with a temporal rate $R(S) = 6\%$,but in the case a temporal rate $R(S) = 12\%$, the PSNR gain for "Foreman" is about 2 to 3dB, while for "Soccer" this gain ranges from 0 to 1.5dB. In the experiments with a temporal rate of 25%, the gains are higher than the $R(S) = 12\%$. In these tests, the PSNR gain for "Foreman" is about 3 to 5dB, while for the "Soccer" this gain ranges from 0 to 2.5dB. This is because the proposed method finds the best GOP size within each interval of 32 frames and B frames find good predictions, whereas in the case of the reference summary with regular GOP structure the GOP size cannot be optimized and the long temporal distance between reference frames makes B frames not useful in temporal layer 0.

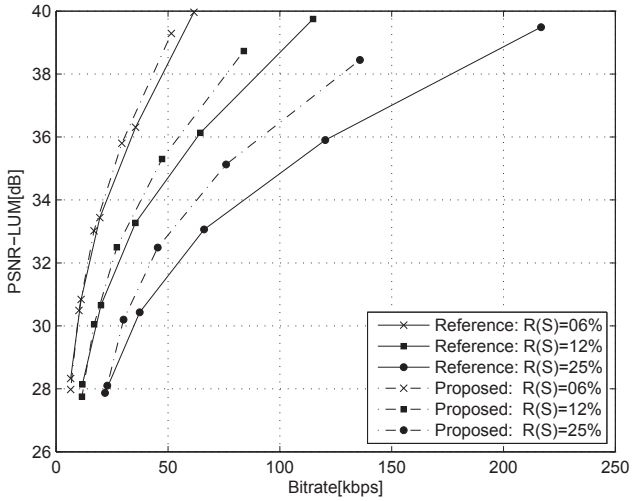

Fig. 2. R-D of "Soccer" summary - temporal base layer (T0).

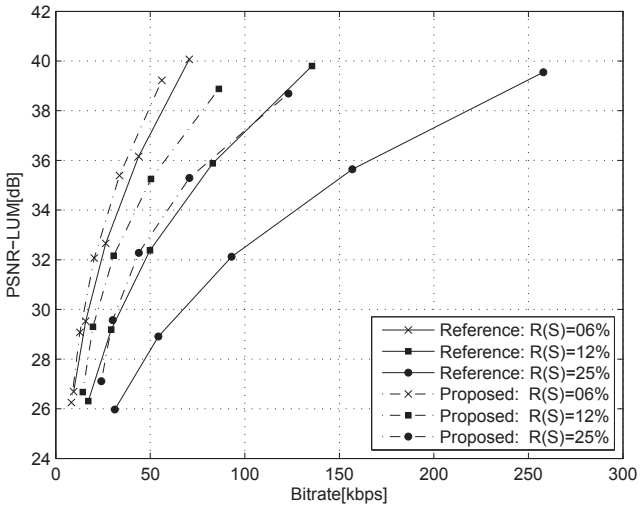

Fig. 3. R-D of "Foreman" summary - temporal base layer (T0).

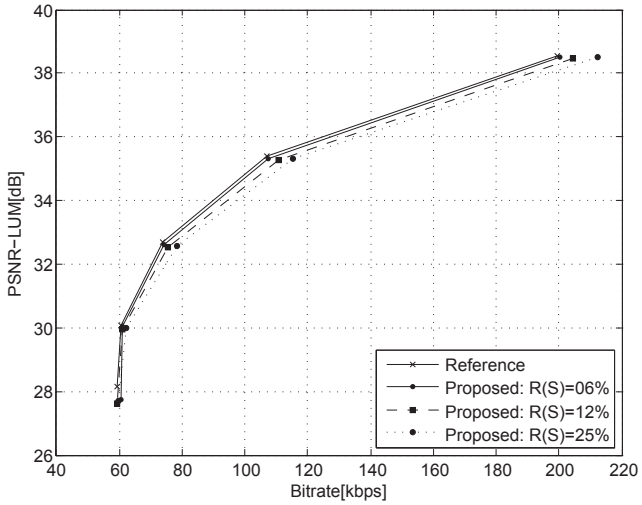

Fig. 5. R-D of "Foreman" - full rate, dynamic GOP in the base layer.

B. Full temporal resolution coding efficiency

The overall coding efficiency was also evaluated for the full temporal rate, i.e., the whole sequence (all layers) using the proposed method to encode the base layer was compared with a reference using a regular GOP size of 32. Figures 4 and 5 show the results for summaries of different temporal rates $R(S)$. The results show that for $R(S) = 6\%$, the coding efficiency achieved by the proposed method is virtually the same as the reference sequences while for other rates the difference is still negligible. Therefore, using the proposed method to encode summaries in the base layer does not have any noticeable impact in the overall R-D coding performance.

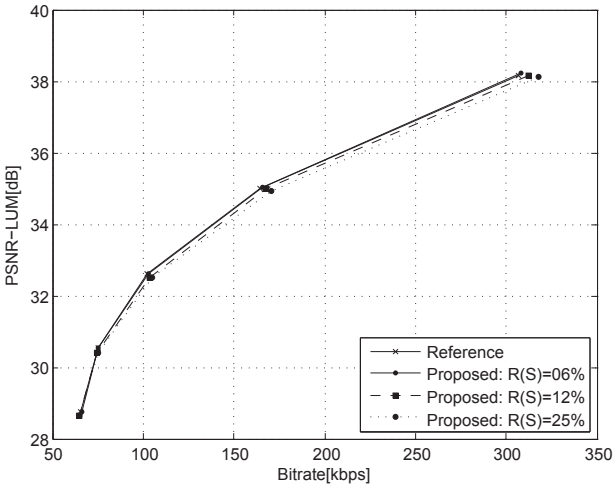

Fig. 4. R-D of "Soccer" - full rate, dynamic GOP in the base layer.

V. Conclusion

A method to encode an arbitrary video summary using dynamic GOP structures in scalable streams was described. The scalable stream obtained is fully compatible with the scalable extension of the H.264/AVC standard. The results show that good coding efficiency is achieved for arbitrary video summaries without compromising the quality of the whole sequence. The proposed method demonstrates that an extra level of flexibility can be achieved by embedding video summaries in scalable streams, which is of practical interest in content adaptation systems and applications.

Acknowledgment

This work was partially supported by Fundação para Ciência e Tecnologia (FCT) of the Portuguese MCTES, under grant SFRH/BD/37510/2007 co-financed by Programa Operacional Ciência e Inovação (POCI 2010) and Fundo Social Europeu (FSE).

References

[1] P. M. Fonseca and F. Pereira, "Automatic video summarization based on mpeg-7 descriptions," *Signal Processing: Image Communication*, vol. 19, no. 8, pp. 685 – 699, 2004.

[2] Y. Li, S.-H. Lee, C.-H. Yeh, and C.-C. Kuo, "Techniques for movie content analysis and skimming: tutorial and overview on video abstraction techniques," *IEEE Signal Processing Magazine*, vol. 23, no. 2, pp. 79 –89, mar. 2006.

[3] H. Schwarz, T. Marpe, and T. Wiegand, "Overview of the scalable video coding extension of the h.264/avc standard," *IEEE Tran. on CSVT,*, vol. 17, no. 9, pp. 1103–1120, Sep. 2007.

[4] J. Bescs, J. Martnez, L. Herranz, and F. Tiburzi, "Content-driven adaptation of on-line vide," *Signal Processing: Image Comm.*, vol. 22, no. 7-8, pp. 651–668, 2007, special Issue on Content-Based Multimedia Indexing and Retrieval.

[5] M. Mrak, J. Calic, and A. Kondoz, "Fast analysis of scalable video for adaptive browsing interfaces," *Comp. Vision and Image Understanding*, vol. 113, no. 3, pp. 425–434, 2009.

[6] L. Herranz and J. Martnez, "An integrated approach to summarization and adaptation using H.264/MPEG-4 SVC," *Signal Processing: Image Comm.*, vol. 24, no. 6, pp. 499 – 509, 2009.

[7] M. Luessi and A. K. Katsaggelos, "Efficient motion compensated frame rate upconversion using multiple interpolations and median filtering," ser. ICIP'09, 2009, pp. 373–376.

[8] G. S. Z. Li, A. K. Katsaggelos and B. Gandhi, "Rate-distortion optimal video summary generation," *IEEE Tran. on Image Processing*, vol. 14, no. 10, pp. 1550–1560, October 2005.

[9] J. Lee and H. Kalva, *The VC-1 and H.264 video compression standards for broadband video services*, S. US, Ed. Berlin, 2008.