OPTIMAL SELECTION OF ENCODING CONFIGURATION FOR SCALABLE VIDEO CODING[1]

T. Berkin Abanoz and A. Murat Tekalp
College of Engineering, Koç University, 34450 Sarıyer, Istanbul, Turkey

ABSTRACT

It is well-known that the wider the range of extraction points a scalable bitstream supports, the lower the compression efficiency at these extraction points. Moreover, this compression efficiency generally varies according to what combination of scalability types are used to support this range of extraction points as specified by the encoding configuration. Hence, we propose some objective criteria as a measure of coverage, compression efficiency and rate-distortion performance of a configuration, and then present a multiple-objective optimization formulation to select the best encoding configuration for scalable video coding, given a range of bitstreams that must be supported. The method is demonstrated by experimental results.

Index Terms— Scalable video coding, encoding configuration, multiple-objective optimization.

1. INTRODUCTION

Standardization of scalable video coding (SVC), which is often desirable for efficient rate adaptation in video transport over the Internet, is a current work item under the Joint Video Team (JVT) [1]. The reference encoder-decoder, called the Joint Scalable Video Model (JSVM) [2], is based on a scalable extension of the well-established JVT standard H.264/AVC. It provides temporal scalability using of motion-compensated temporal filtering (MCTF) implemented by a lifting framework. For spatial scalability, a combination of motion-compensated prediction and over-sampled pyramid decomposition is employed [3]. SNR scalability is achieved by residual quantization with some modification to the H.264/AVC syntax.

While temporal and spatial scalability modes of SVC allow bitstream extraction at specific rate-distortion points, the FGS mode allows extraction more-or-less over a continuous range of rate-distortion points. The concept of quality layers has been introduced in [4] to allow for rate-distortion optimized bitstream extraction over a range of rates. However, the number and range of extraction points is determined by the encoder configuration, which is often determined prior to encoding in an ad-hoc manner. In order to have flexibility for adaptation of the video rate to a wide range of network conditions, we would like to have as many extraction points within a predetermined operating range of bitrates, which requires definition of several scalability layers. However, the number and type of scalability layers may have significant impact on the compression efficiency, and cannot be easily optimized over a wide range of bitrates.

This paper addresses the problem of selection of the best encoding configuration, namely scalability type and number of layers for each type, and quantization parameter for base FGS layer at each spatial resolution in order to meet some conflicting criteria including maximization of the coverage of extraction points within a predetermined bitrate range, maximization of the rate-distortion performance at each extraction point within the given range, maximization of the average incremental compression efficiency of each layer over all extraction points within the range, and maximization of the maximum picture size represented by the encoder configuration within the given bitrate range. Mathematical definition of these criteria is given in Section 2. We pose a multiple objective optimization (MOO) formulation and provide a solution to this problem in Section 3. Experimental results are presented in Section 4, and conclusions are reached in Section 5.

2. DEFINITION OF OBJECTIVE CRITERIA

In this section, we propose three criteria to quantify an encoding configuration.

2.1 Coverage of a Configuration

Given a target bitrate range, we define total coverage *C* of a scalable bitstream as how much of this range is actually covered by extraction points of this bitstream. Clearly, total coverage is the sum of coverage of the individual layers,

[1] This work has been supported by European Commission within FP6 under Grant 511568 with the acronym 3DTV.

c(*i*), where spatial layers provide single extraction points, and FGS layers provide a range of points. Hence,

$$C = \sum_{i=1}^{L} c(i) \qquad (1)$$

$$c(i) = \begin{cases} 1 & \text{if } i \text{ is a spatial scalability layer} \\ (\text{PSNR}_{\max} - \text{PSNR}_{\min})/0.2 & \text{if } i \text{ is an FGS layer} \end{cases}$$

$\text{PSNR}_{\max}$ and $\text{PSNR}_{\min}$ denote the maximum and minimum PSNR values for an FGS layer calculated after interpolating to 4cif resolution, and *L* is the number of layers within the bitrate range. Additionally, the importance of coverage increase from 25 to 30 is not as important as coverage increase from 10 to 30. For this reason, we propose as objective criteria base three logarithm of the coverage function.

2.2 Efficiency of a Configuration

"Base layer usage" has been proposed as a measure of efficiency of scalable video coding in case of a base layer and one enhancement layer [5]. It measures the efficiency of scalable coding compared to simulcasting at corresponding two rates, given by

$$B = \frac{(R_S + R_B) - R_E}{R_B} = 1 - \frac{R_E - R_S}{R_B} \qquad (2)$$

where R_B, R_E and R_S stand for the base layer rate, total rate of scalable stream (base + enhancement), and the simulcast (non-scalable coding) rate at the same PSNR as that of total scalable rate, respectively. *B* takes values in the range 0-1. When $B \approx 0$, the efficiency of scalable coding is close to simulcasting, and when $B \approx 1$, the efficiency of scalable coding is close to non-scalable coding at the same quality, which is desirable.

We hereby extend this definition to the case of more than one enhancement layers, called incremental efficiency, *b(i)*, of layer *i*. The *b(i)* is defined as in (2), except that the base layer rate is taken as the total rate of all layers up to layer *i*. This is because all those layers, including layer *i*-1, are required for decoding enhancement layer *i*.

Then, we define the overall efficiency, *E*, of a scalable bitstream as the average of incremental efficiency of all layers that fall within bitrate range of interest, given by

$$E = \frac{\sum_{i=1}^{L} b(i)}{L} \qquad (3)$$

where *L* is the number of layers within the bitrate range.

2.3 Rate-Distortion Performance of a Configuration

It is desirable that we have good rate-distortion (RD) performance at all possible extraction points of a scalable bitstream. Hence, we define the overall RD performance of a scalable bitstream as the average of RD performances over all possible extraction points, given by

$$RD = \frac{\sum_{i=1}^{C} [\lambda_i R(i) + D(i)]}{C} \qquad (4)$$

where q_i is the quantization parameter for layer i, $\lambda_i = 0.85 * 2 \wedge ((q_i - 12)/3)$ [7], *R*(*i*) is the bitrate for layer *i* and *D*(*i*) is the SSD distortion for layer *i*, given by

$$D(i) = \frac{255^2 * width * height}{10^{PSNR(i)/10}} \qquad (5)$$

and PSNR(*i*) is the PSNR of layer *i* after it is interpolated to 4cif resolution.

3. THE MULTIPLE-OBJECTIVE OPTIMIZATION FORMULATION

Multiple objective optimization (MOO) was introduced by Pareto for solution of an optimization problem with *P* possibly conflicting objective functions $f_1, f_2, \ldots, f_P$. A solution s^* is called globally Pareto-optimal if any one of the objective function values cannot be improved without degrading other objective values. Then, a Pareto-optimal solution s^* exists if there exists no other feasible solution *s* that satisfies

$$f_p(s) \le f_p(s^*), \quad \forall p \in \{1, \ldots, P\} \qquad (6)$$

with at least one strict inequality. Since different objective functions represent different aspects of the problem, it is difficult to discriminate between these Pareto-optimal points and determine which one is better than the other. The MOO defines a so called *best compromise solution* as the feasible solution that is closest to the utopia point, which is an infeasible solution obtained by minimizing each objective individually [6].

In our problem, there are *P*=4 optimization criteria. In addition to the base three logarithm of coverage, efficiency, and rate-distortion performance criteria, which are defined in Section 2, we also employ maximum picture size as an optimization criterion, since video with the largest size should be preferred if all other parameters were equal. These criteria are optimized with respect to the encoder configuration parameters, which are type (spatial and/or FGS) and number *L* of scalability layers, and the quantization parameter *q* for the base FGS layer at each spatial resolution.

We perform multiple-objective optimization subject to a maximum rate constraint. That is, if there is a scalable bitstream whose total bitrate is greater than the maximum target bitrate, we discard those layers with the minimum bitrate greater than the maximum target bitrate in defining feasible solutions.

Our objective is to find the encoding configuration *j* that strikes the best balance between

- Maximize coverage

$$\max_{j}\left(\log_3(C_j)\right) \quad (7)$$

where C_j is the coverage for bitstream j, given by (1);

- Maximize efficiency

$$\max_{j}\left(E_j\right) \quad (8)$$

where E_j is the overall efficiency of bitstream j, given by (2);

- Minimize rate-distortion performance

$$\min_{j}\left(RD_j\right) \quad (9)$$

where RD_j is the rate distortion for bitstream j given by (4);

- Maximize the maximum picture size

$$\max_{j}\left(\max p_j\right) \quad (10)$$

where max p_j is the maximum picture size for bitstream j.

We assume that all objectives have equal importance; hence, their values will be scaled to range [0,1] as

$$f_{scaled} = \frac{f - f_{\min}}{f_{\max} - f_{\min}} \quad (11)$$

where f is the original objective value prior to scaling.

In order to find the bitstream that strikes the best compromise between our four optimization criteria, we first encode the video with N different choices of encoding configurations. Each one of these encoding configurations is a feasible solution point. The final step is to find the solution point that is closest to the (infeasible) utopia point which is the point (0,1,1,1) after normalization. The utopia point is the one where all optimization criteria are satisfied, in other words, where rate-distortion measure is minimum, and maximum picture size, coverage and coding efficiency are all maximum. To find the closest feasible point, we use Euclidian distance, and calculate the distance for each configuration j as

$$d_j = \sqrt{\left(1 - E_j\right)^2 + \left(1 - \max p_j\right)^2 + \left(1 - \log_3(C_j)\right)^2 + \left(RD_j\right)^2} \quad (12)$$

and we select the configuration with minimum distance $\min_{i}\left(d_i\right)$ as the MOO solution.

4. EXPERIMENTAL RESULTS

We consider the following problem: Design a scalable bitstream that should simultaneously serve i) a broadband client at about 1.5-3 Mbps, ii) a DSL client at near 256 kbps. To this effect, we encoded two video files (soccer.yuv, harbour.yuv) with N=21 different encoding configurations, that is, with different number of spatial and FGS layers, and with different quantization parameters. The list of feasible configurations is shown in Table 1. The configurations are specified as base spatial layer-followed by the number of FGS layers at that spatial resolution-followed by quantization parameter for the base FGS layer at that spatial resolution. The + sign indicates the next spatial layer specified in the same format. For example, cif-2-38 + 4cif-2-38 denotes that base spatial layer is cif and we have 2 FGS layers for cif resolution, and the quantization parameter for base FGS layer is 38, followed by 4cif resolution with the same parameters.

The values of coverage, base three logarithm of coverage, efficiency, maximum picture size rate-distortion and distance values for each configuration for both of the videos are listed in the Table 1. The distance of each configuration as listed in Table 1 to the utopia point is plotted in Figure 1.

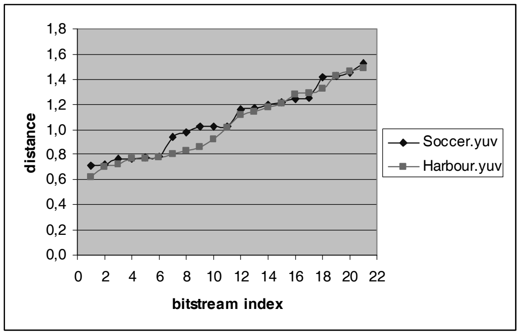

Fig.1. Distance of configurations to the utopia point by rank

5. CONCLUSIONS

Inspection of Table 1 shows that the first five configurations for both videos include the same four configurations, namely cif-1-38 + 4cif-1-38; cif -2-38 + 4cif-2-38; cif-1-40 + 4cif-1-40; cif -2-40 + 4cif-2-40, which indicates that these are the best encoder configurations Their actual rankings are somewhat different, because the two videos contain different amount of spatial and temporal detail.

If we analyze the properties of the first two ranked configurations for "soccer.yuv", we see that their distances are close to each other, but the efficiency of the first configuration is better than that of the second while their coverage and RD values are close to each other.

We note that the best solution may change if the users express different preference (weights) on different criteria.

6. REFERENCES

[1] Joint Video Team of ITU-T VCEG and ISO/IEC MPEG, "Scalable Video Coding – Working Draft 1," *Doc. JVT-N020*, Jan. 2005.

[2] Joint Video Team of ITU-T VCEG and ISO/IEC MPEG, "Joint Scalable Video Model JSVM-6.7," *Doc. JVT-Q202*, Oct. 2005.

[3] H. Schwarz, D. Marpe, T. Schierl, T. Wiegand, "Combined Scalability Support for the Scalable Extension of H.264/AVC," *IEEE Int. Conf. Multimedia & Expo (ICME)*, Amsterdam, The Netherlands, July 2005.

[4] I. Amonou, N. Cammas, S. Kervadec, and S. Pateux, "Optimized rate-distortion extraction with quality layers,"

Proc. IEEE Int. Conf. on Image Processing, Atlanta, Georgia, pp. 173-176, Oct. 2006.

[5] H. Schwarz and T. Wiegand, “Preliminary results for an r-d optimized multi-loop SVC encoder,” JVT Doc. JVT-T080, Klagenfurt, July 2006.

[6] J. Lin, “Multiple Objective Problems: Pareto-Optimal Solutions by Method of Proper Equality Constraints,” IEEE Trans. Automatic Control vol. 21, pp. 641–650, 197

[7] T. Weigand, H. Schwarz, A. Joch, F. Kossentini, G. Sullivan, “Rate_constrained Coder Control and Comparison of Coding Standards,” IEEE Trans. on Circuits and Systems for Video Technology, July 2003.

Table 1. Ranking of several encoding configurations according to their normalized Euclidean distance to the utopia point.

Soccer.yuv							
Configuration Parameters	**Conf. Effici.**	**Max. Pic. Size**	**Conf. Cov.**	**Log3(Cov.)**	**Conf. RD**	**Distance**	**Rank**
cif-1-38 + 4cif-1-38	0,674	405504	21	2,771	16729169	0,26	1
cif -2-38 + 4cif-2-38	0,619	405504	27	3,000	16550397	0,27	2
cif -2-40 + 4cif-2-40	0,616	405504	32	3,155	18256597	0,31	3
qcif- 0-32 + cif-0-32 + 4cif-2-40	0,624	405504	18	2,631	13359059	0,32	4
cif-1-40 + 4cif-1-40	0,670	405504	21	2,771	20183890	0,37	5
qcif- 0-40 + cif-0-40 + 4cif-2-40	0,570	405504	22	2,814	14772208	0,40	6
qcif-0-32 + cif-1-40 + 4cif-2-40	0,567	405504	26	2,966	17701892	0,43	7
qcif-1-38 + cif-1-38 + 4cif-1-38	0,555	405504	28	3,033	21972394	0,56	8
qcif-1-34 + cif-1-34 + 4cif-1-34	0,590	405504	16	2,524	21915195	0,56	9
qcif-1-40 + cif-1-40 + 4cif-1-40	0,549	405504	27	3,000	25357453	0,68	10
qcif-2-38 + cif-2-38 + 4cif-2-38	0,494	405504	30	3,096	25157722	0,79	11
qcif- 2-40 + cif-2-40 + 4cif-2-40	0,472	405504	39	3,335	24679739	0,82	12
qcif-1-32 + cif-1-32 + 4cif-1-32	0,654	101376	12	2,262	23597170	1,17	13
qcif-2-38 + cif-2-38	0,533	101376	25	2,930	27835771	1,28	14
qcif-0-32 + cif-0-32 + 4cif-0-32	0,371	405504	3	1,000	20820273	1,34	15
qcif-1-38 + cif-1-38	0,618	101376	15	2,465	31578156	1,34	16
qcif-2-40 + cif-2-40	0,516	101376	26	2,966	30784696	1,37	17
qcif-0-34 + cif-0-34 + 4cif-0-34	0,360	405504	3	1,000	23309491	1,39	18
qcif-1-40 + cif-1-40	0,598	101376	14	2,402	35915953	1,48	19
qcif-0-32 + cif-0-32	0,682	101376	2	0,631	27654914	1,55	20
qcif-0-34 + cif-0-34	0,677	101376	2	0,631	30437866	1,60	21
Harbour.yuv							
Configuration Parameters	**Conf. Effici.**	**Max. Pic. Size**	**Conf. Cov.**	**Log3(Cov.)**	**Conf. RD**	**Distance**	**Rank**
cif -2-38 + 4cif-2-38	0,567	405504	24	2,893	35091883	0,27	1
cif-1-40 + 4cif-1-40	0,563	405504	24	2,893	36246351	0,28	2
cif-1-38 + 4cif-1-38	0,538	405504	25	2,930	30209428	0,29	3
qcif- 0-40 + cif-0-40 + 4cif-2-40	0,538	405504	22	2,814	29226375	0,31	4
cif -2-40 + 4cif-2-40	0,520	405504	30	3,096	36216258	0,34	5
qcif-1-38 + cif-1-38 + 4cif-1-38	0,482	405504	28	3,033	45981028	0,50	6
qcif-1-40 + cif-1-40 + 4cif-1-40	0,505	405504	29	3,065	56904820	0,58	7
qcif-0-32 + cif-1-40 + 4cif-2-40	0,384	405504	28	3,033	36018061	0,65	8
qcif- 0-32 + cif-0-32 + 4cif-2-40	0,380	405504	18	2,631	25779420	0,68	9
qcif-2-38 + cif-2-38 + 4cif-2-38	0,443	405504	25	2,930	62524322	0,74	10
qcif-1-34 + cif-1-34 + 4cif-1-34	0,459	405504	15	2,465	61130323	0,75	11
qcif- 2-40 + cif-2-40 + 4cif-2-40	0,421	405504	35	3,236	62316472	0,77	12
qcif-1-32 + cif-1-32 + 4cif-1-32	0,515	101376	11	2,183	60294933	1,24	13
qcif-2-38 + cif-2-38	0,496	101376	22	2,814	68342409	1,25	14
qcif-0-34 + cif-0-34 + 4cif-0-34	0,343	405504	3	1,000	66496759	1,28	15
qcif-1-38 + cif-1-38	0,612	101376	13	2,335	78193351	1,32	16
qcif-2-40 + cif-2-40	0,498	101376	25	2,930	78702638	1,33	17
qcif-1-40 + cif-1-40	0,645	101376	15	2,465	88819162	1,41	18
qcif-0-32 + cif-0-32 + 4cif-0-32	0,227	405504	3	1,000	62118924	1,43	19
qcif-0-34 + cif-0-34	0,597	101376	2	0,631	92271302	1,74	20
qcif-0-32 + cif-0-32	0,445	101376	2	0,631	87305844	1,76	21