

OPTIMUM SCALING OPERATOR SELECTION IN SCALABLE VIDEO CODING

Emrah Akyol¹, A.Murat Tekalp^{1,2}, and M.Reha Civanlar¹

¹ College of Engineering, Koc University, Istanbul, Turkey

² Department of Electrical and Computer Engineering, University of Rochester, Rochester, NY 14627

ABSTRACT

Scalable video coders provide different options, such as temporal, spatial and SNR scalability, where each option results in different kinds and/or levels of visual distortion at the lower scales depending on the content and bitrate. We observe that in most cases a single scalability option does not fit the whole video content well, and the scalability operator should be varied for different temporal segments depending on the content of the segment. In this work, assuming the video is temporally segmented by some content analysis scheme, we propose a method to choose the visually best scaling option that results in minimum visual distortion among temporal, spatial and SNR scalability operators for each temporal segment of soccer videos. We employ four component metrics to quantify artifacts caused by bitrate reduction, spatial size reduction and temporal subsampling, which are a flatness measure, a blockiness measure, a blurriness measure, and a temporal distortion (jerkiness) measure. We then define the best scaling operator for each video segment as the one with the minimum distortion score which is given by a linear combination of these four component measures. Coefficients of this linear combination are tuned to content type using a training procedure. Two subjective tests have been performed to validate the proposed distortion measures and procedure for optimal selection of scalability operators for soccer videos.

1. INTRODUCTION

Recently, scalable video coding has gained renewed interest since it has been shown [1] that it can achieve compression efficiency that is comparable to that of predictive video coding standards such as H.264 [2]. Fully scalable video codecs enable flexible adaptation of video bitrate to time-varying channels through signal-to-noise ratio (SNR), temporal, and, spatial scalability. On the other hand, determination of the visually optimum scalability operator for a given video content at a specific bitrate is an open problem since different scaling operations result in different types of distortion that can not be measured meaningfully using the mean square error or PSNR metric.

The problem of optimum scalability operator selection has been addressed in [3,4,5,6]. In [3], the authors define a ratio of the spatial and temporal information, and compare this metric with a threshold to determine the best operator among the spatial and temporal scalability operators. In [4], a system for automatically selecting the optimal frame rate for MPEG-4 Fine Granular Scalability (FGS) coded video based on the PSNR measure is proposed, such that when the PSNR decreases below some threshold, the frame rate is lowered. In [5], the optimal trade-off between SNR and temporal scalabilities in scalable video coding is addressed, where some content based features are used to choose between the temporal and SNR operators with the assumption that spatial scalability is never preferred as the best operator. A similar approach is followed in [6], where content based features are selected for classification of FGS modes in MPEG-4 according to an objective distortion metric proposed in [7].

In this work we propose a content adaptive scalable video coding framework where each temporal segment is scaled by an optimum scaling operator with respect to a distortion metric which is the linear combination of some flatness, blurriness, blockiness and jerkiness measures. We propose that the optimal combination of the scalability operators depends on the content (shot type) and bitrate. Since a strong relationship exists between the choice of scaling operator and the content, content adaptive streaming of scalable coded video yields significant improvement in the visual quality compared to a fixed scaling option. We performed two subjective tests to train/test our optimal scaling option selection algorithm and test the performance of adapting the scaling operator to shot type and content. Although our analysis is performed on soccer videos, the proposed method can be extended to other types of video content. The paper is organized as follows: we discuss distortion metrics in Section 2. In Section 3, we present the choice of scalability operators (SNR, temporal, spatial and their combinations) and the problem formulation. In Section 4, the subjective tests are explained in detail, and experimental results are presented. Conclusions are presented in Section 5.

2. VIDEO QUALITY MEASURES

It is well-known that different scalability options yield different types of distortions. For example, SNR scalability results in blockiness and flatness when the base layer is at lower bitrates, spatial scalability results in blurriness, and temporal scalability results in jerkiness. Because PSNR measure cannot capture all these distortions or distinguish between them [8], we employ four separate measures to quantify these distortions, which are described below.

A. Flatness Measure

We first find the major edges by the Canny edge detection operator and calculate the local variance at pixels other than the edges. We then set an upper limit on the local variance such that $T(\alpha) = \min(a, t)$ is a thresholding operator, where α is the variance and t is the threshold value. Thresholding serves two purposes: i) measure flatness only in low texture areas where flatness occurs the most, and ii) provides spatial masking of quantization noise in high texture areas. We observe that threshold values in the range 70-80 give the best results so we set $t=75$. The flatness measure is defined as:

$$D_{flat} = \frac{\sum_i [\sigma_{org}^2(i) - \sigma_d^2(i)]}{N}$$

where, N , $\sigma_{org}^2(i)$ and $\sigma_d^2(i)$ denote the number of 4x4 blocks in one frame, variance of 4x4 blocks on the original (reference) and decoded (distorted) frames, respectively.

B. Blockiness Measure

Several blockiness measures exist in the literature to assist PSNR in the evaluation of compression artifacts [9,10] under the assumption that the block boundaries are known a priori. The blockiness metric proposed in [10] is defined as the sum of the differences along straight edges scaled by the texture near that area. When using overlapped block motion compensation and/or variable size blocks, the positions of block boundaries are no longer fixed. To this effect, the straight edges which do not exist in the original frame are treated as block boundaries. The texture near the edge location, which is included to consider spatial masking, is defined as

$$TM_{hor}(i) = \sum_{m=k-1}^3 \sum_{k=1}^L |f(i-mk) - f(i-m+1, k)| + \sum_{m=k-1}^3 \sum_{k=1}^L |f(i+mk) - f(i+m+1, k)|$$

where, L denotes length of the straight edge and f is the frame of interest. We set $L=16$. The horizontal blockiness of the i^{th} straight edge can be defined as

$$Block_{hor}(i) = \frac{B(i)}{1.5 \times TM_{hor}(i) + B(i)} \text{ where,}$$

$$B(i) = \sum_{k=1}^{k=L} |f(i, k) - f(i-1, k)|$$

The cumulative blockiness measure for all horizontal block borders, BM_{hor} , is defined as

$$BM_{hor} = \sum_{i \in \text{All horizontal block boundaries}} Block_{hor}(i)$$

Cumulative blockiness measure for vertical straight edges BM_{vert} can be found similarly. Finally, total blockiness measure is defined as

$$D_{block} = BM_{hor} + BM_{vert}$$

C. Blurriness Measure

In [11] a blur and ringing metric is proposed to measure artifacts in JPEG2000 coded images. Following a similar approach, we find vertical and horizontal major edges by the Canny operator, then find the width change of edges. The sum of the edge width changes in the decoded video is scaled with the sum of the edge widths of the original frame to obtain the blurriness metric.

$$D_{blur} = \frac{\sum_i (Width_d(i) - Width_{org}(i))}{\sum_i Width_{org}(i)}$$

where, $Width_{org}(i)$ and $Width_d(i)$ denote the width of the i^{th} edge on the original(reference) frame and decoded (distorted) frame, respectively.

We take only still regions of frames into consideration as proposed in [12]. The motion detection threshold is selected as 15 [12].

D. Temporal Jerkiness Measure

We use the sum of magnitudes of the integer pixel motion vectors of each 16x16 block for each frame. Since the distortion is to be measured, we look at the difference between the temporal jerkiness of the video in interest and the original video with full frame rate.

$$D_{jerk} = \frac{\sum_i |MV_d(i) - MV_{org}(i)|}{N}$$

where, $MV_{org}(i)$, $MV_d(i)$ and N denote the i^{th} element of the motion vector of the original 16x16 block, motion vector of the 16x16 block of interest and the number of 16x16 blocks in one frame respectively.

E. Interpolation Techniques

In cases where bitrate reduction is achieved by spatial and/or temporal scalability options, the base layer video must be subject to spatial and/or temporal interpolation before proper display. In this work, Daubechies 9/7 inverse filter is used for spatial up-sampling. Zero order

hold interpolation (frame replication) is utilized for temporal interpolation.

3. PROBLEM STATEMENT AND METHOD

A. Scalability Operators

There are, in general, three types of scalability options: SNR, temporal, and spatial scalability. Scalability operators can also be combined to allow for joint scalability modes. In this work, we allow 6 combinations of scaling operators for each temporal segment, which are:

1. SNR only scalability
2. (Spatial) + SNR scalability
3. (Temporal) + SNR scalability
4. (Spatial + temporal) + SNR scalability
5. (2 level temporal) + SNR scalability
6. (2 level temporal + spatial) + SNR scalability

where, the parenthesis is used to denote the base layer. For example, option 4 denotes one level temporal and one level spatial scaling to form the base layer, such that the base layer has half the original frame rate and half the original spatial resolution.

B. Selection of the Optimum Scalability Operator

We propose to select one of the above scalability operators for each temporal segment to minimize a visual distortion measure (or cost function). In [13], a distortion measure which is a linear combination of some component distortion metrics has been proposed. Following a similar approach, we define a new cost function in the form:

$$D = \alpha_{block} D_{block} + \alpha_{flat} D_{flat} + \alpha_{blur} D_{blur} + \alpha_{jerk} D_{jerk}$$

where, α_{block} , α_{flat} , α_{blur} , and α_{jerk} are the linear combination coefficients for blockiness, flatness, blurriness, and jerkiness measures, respectively. A block diagram of the proposed system is shown in Figure 1, where we employ a fully embedded scalable video coder. Base layers formed by different combinations of scalability operators are then extracted and decoded. Then, the above cost function is evaluated for each combination and the operator that results in the minimum cost function is selected.

We note that the coefficients of the linear combination can be tuned according to content of the shot (shot type). For example, blurriness is more objectionable in close-medium shots; flatness is more disturbing in far shots; and motion jerkiness is more noticeable when there is global camera motion. This is explained in the next section.

C. Determination of the Coefficients

Shot boundary determination and shot type classification can be done automatically depending on the content type,

e.g., for soccer videos [14]. Once all shots are classified, coefficients α_{block} , α_{flat} , α_{blur} and α_{jerk} are computed for each shot type separately by least squares fitting on the results of subjective tests on some training data. In particular, the coefficients are found such that the value of the cost function for some training shots matches subjective visual evaluation scores (see Section 4.C) in the least squares sense. The selection of the training shots is also an important issue since all probable distortion types for that shot type should exist in the training shots.

4. SUBJECTIVE TEST AND RESULTS

We conducted two subjective tests to validate our proposed scalability operator selection method. The dataset obtained from the first test is statistically analyzed to confirm our basic assumptions, such as the best scaling option depends on the bitrate and the shot-type. The second test is conducted for measuring the performance of the proposed method which is trained using the data from the first test.

A. Subjective Test-I

This test is set up with 20 subjects according to ITU-R Recommendation BT.500-10[15], using a three level evaluation scale instead of ten levels. A *Single-stimulus Comparison Scale* is used in the test i.e., assessors viewed six videos generated by the scaling options listed in Section 2.B in random order without seeing the originals. For each “bitrate”–“shot-type” combination, each assessor was asked to rank the six videos according to the three level, Good-Fair-Poor, scale with ties allowed. The video clips used are of 3-5 seconds duration at CIF resolution and contain typical shots from a soccer game. We used four shot types defined based on the camera motion and distance (close-medium and far) as shown in Figure 2. We tested three different bitrates for the base layer: 100kbps, 200kbps and 300kbps. The full rate is kept at 600kbps. At these rates, all shot types other than Shot 3 (close shot with camera pan) are affected by flatness, blurriness and jerkiness distortions; Shot 3 has blockiness instead of flatness as the significant compression artifact.

B. Statistical Analysis of the Subjective Test Results

We used statistical tests to answer the following questions:

1. *Is there a statistically significant difference in the assessors choices created by the scaling operator selection? In other words, does the selection of the scalability operator matter?*
2. *If an optimal scaling operator exists, does it change with respect to the shot-type, i.e., is the shot-type a statistically significant factor in ranking scalable coded videos?*

3. *Is the bitrate a significant factor in addition to the scaling option and the shot-type?*
4. *Are there significant clusters in the choices of assessors? Or, is the optimal operator selection subjective?*

To answer the first three questions we applied the Friedman test [16], which evaluates whether a selected test variable, e.g., bitrate, shot-type, etc., can be used to form test result clusters that contain significantly different results as compared to a random clustering. The output of this test, ρ , is the significance level, which represents the probability that a random clustering would yield the same or better groups. A result with ρ less than 0.05 or 0.01 is assumed to be significant in general. The results of the Friedman’s test are:

-Clustering with respect to the scaling option is significant with ρ being almost zero. With this result, scaling operator selection is indeed significant.

-After scaling option clustering, clustering with respect to shot-type is found to be significant with $\rho=0.004$

-In addition to scaling operator and shot-type, bitrate is a significant factor in clustering with significance $\rho=0.001$.

User dependency of the results seemed to be another factor to analyze. We first calculated the correlation of the user’s scores, shown in Figure 3, to see if there is any clustering. We observe two types of user groups: one group prefers higher picture quality over higher frame rate (A type) and the other group prefers higher frame rate (B type). Based on this observation, we clustered subjects into two groups using 2-means clustering. We also determined the significance of the clustering by rank-sum test for each video. The separation of users into two groups is found to be significant at 5% level for 30 videos out of 72 videos coded with different scaling option, bitrate and shot type combinations, and most of the 30 videos that users preferences differ are coded at low bitrates, which leads us to the conclusion that the difference in the users frame rate preferences increases as overall video quality decreases.

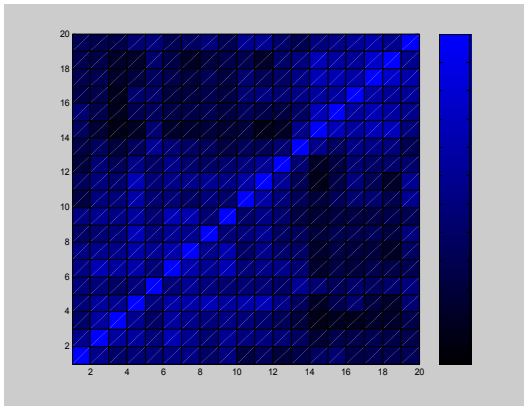


Figure 3: The correlation of user’s scores. A noticeable difference exists between two groups of users.

C. Subjective Distortion Score

The subjective distortion score (SDS) of a video segment is defined as:

$$SDS = \frac{2}{1 + (2 \times S_1 + S_2) / (2 \times S_{\max})}$$

where, S_1 and S_2 are the numbers of “good” and “fair” grades, respectively, and S_{\max} is the number of assessors.

We determine the coefficients of the overall objective cost function by matching it to SDS using the least squares method. In coefficient set calculations, we used 100 kbps video for shots without camera pan (shots 2-4) and 200 kbit video for shots with camera pan (shots 1-3). The coefficient sets computed for the all users, type A users, and type B users are shown in Table-1.

In order to quantify the performance of our method for selection of the optimal scalability option, we computed the Spearman rank correlation between the SDS and the ranking provided by our method as shown in Table 2. It can be seen that our algorithm finds the best or the second best scaling option from the six scaling option choices for most cases.

D. Subjective Test-II

In this test a new test video clip is divided into temporal segments according to the shot-types defined above. Then, for each temporal segment, the optimum scaling operator is determined using our proposed method and the coefficients determined in Test 1. Each temporal segment is coded with the optimum scaling operator. The base layer of the resulting video is compared with that of videos coded with fixed scaling options.

Table 3: The first row shows the percentage of users who preferred the content adaptive scaling option to all fixed scaling options. The second row shows the percentage of testers who preferred the scaling options chosen according to their subjective preferences rather than the average optimal scaling option.

	100kbit	200kbit	300kbit
Adaptive scaling performance	%95	%75	%75
Additive gain from bimodal user separation	%20	%5	%5

These results confirm that adaptive scaling provides significant improvement over fixed scaling. The effect of subjective preferences on the scalability operator selection is observed to be somewhat important at the low bitrates

and not important at the higher rates. This result was also observed in the first subjective test.

5. CONCLUSION

In this work we propose a content adaptive scalable video streaming framework, where each temporal segment is coded with the optimum scaling option. Optimum scaling option is determined by a cost function which is a linear combination of different distortion measure such as blurriness, blockiness, flatness and jerkiness. Two subjective tests are performed to find the coefficients of the cost function and test the performance of the proposed system. Statistical significance of the test variables is analyzed. Results clearly show that optimum scaling option changes with the content and content adaptive coding with optimum scaling option results in better visual quality.

6. REFERENCES

[1] MPEG documents, "Registered Responses to the Call for Proposals on Scalable Video Coding," ISO/IEC JTC1/SC29/WG11 MPEG04/M10569.

[2] T. Wiegand, G. Sullivan and A. Luthra, "Draft ITU-T Recommendation and Final Draft International Standard of Joint Video Specification (ITU-T Rec. H.264 | ISO/IEC 14496-10 AVC)," 27 May 2003.

[3] C. Kuhmünch, G. Kühne, C. Schremmer and T. Haenselmann, "A video-scaling algorithm based on human perception for spatio-temporal stimuli," Proc. of SPIE, Multimedia Computing and Networking (MMCN), SPIE Press, Jan. 2001.

[4] R. Kumar Rajendran, M. van der Schaar and S.-F. Chang, "FGS+: Optimizing the Joint Spatio Temporal Video Quality in MPEG-4 Fine Grained Scalable Coding," International Symposium on Circuits and Systems (ISCAS), Phoenix, Arizona, May 2002.

[5] Y. Wang, T. Ng, M. van der Schaar and S.-F. Chang. "Predicting optimal operation of MC-3DSBC multi-dimensional scalable video coding using subjective quality measurement". SPIE Video Comm. and Image Processing (VCIP), San Jose, CA, January 2004.

[6] B.-F. Hung and C.-L. Huang, "Content-based FGS coding mode determination for video streaming over wireless networks", Selected Areas in Communications, IEEE Journal on , Volume: 21 , Issue: 10 , Dec. 2003

[7] S. Wolf and M. H. Pinson, "Spatial-Temporal Distortion Metrics for In-Service Quality Monitoring of Any Digital Video System." Proceedings of SPIE International Symposium on Voice, Video, and Data Communications, Boston, MA, September 11-22, 1999.

[8] B. Girod, "What's wrong with mean-squared error", A.B. Watson (Ed.), Digital Images and Human Vision, MIT Press, Cambridge, MA, 1993, pp. 207-220.

[9] Z. Yu; H.-R. Wu; S. Winkler, and T. Chen, "Vision-model-based impairment metric to evaluate blocking artifacts in digital video," Proc. of the IEEE, vol. 90, no. 1, pp.154 – 169, Jan. 2002.

[10] F. Pan, X. Lin, S. Rahadja, W. Lin, E. Ong, S. Yao, Z. Lu, and X. Yang, "A locally adaptive algorithm for measuring blocking artifacts in images and videos," Signal Processing: Image Communication, vol. 19, no. 6, pp. 499-506, July 2004.

[11] P. Marziliano, F. Dufaux, S. Winkler and T. Ebrahimi., "Perceptual blur and ringing metrics: Application to JPEG2000." in Signal Processing: Image Communication, vol. 19, no. 2, pp. 163-172, Feb. 2004.

[12] A. Webster, C. Jones, M. Pinson, S. Voran, and S. Wolf, "An objective video quality assessment system based on human perception," SPIE Human Vision, Visual Processing, and Digital Display IV, Feb. 1993

[13] A. P. Hekstra, J.G Beerends, D.Ledermann, F.E. Caluwe, S. Kohler, R.H Koenen, S. Rihs, M Ehrsam and D. Schlauss, "PVQM: A perceptual video quality measure", Signal Processing :Image Communication (17) No. 10, Nov. 2002, pp. 781-798.

[14] A. Ekin, A. M. Tekalp, and R. Mehrotra, "Automatic soccer video analysis and summarization" IEEE Trans. on Image Proc., vol. 12, no. 7, pp. 796-807, June 2003.

[15] Methodology for the Subjective Assessment of the Quality of Television Pictures, Recommendation ITU-R BT.500-10, ITU Telecom. Standardization Sector of ITU, August 2000.

[16] J. Devore, "Probability and Statistics for Engineering and the Sciences," Duxbury Press, December 1999.

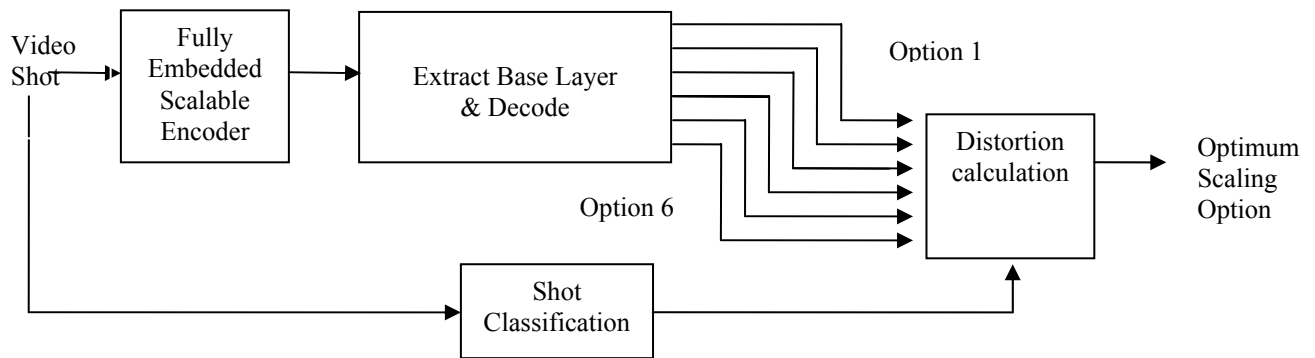


Figure-1: The proposed algorithm of optimal scaling option selection



a) Far-medium shot with camera pan



b) Far shot without camera pan



c) Close shot with camera pan



d) Close-medium shot without camera pan

Figure 2: Four shot types with respect to distance of shots and type of motion.

Table 1: The coefficients of the overall cost function for all users / type A users / type B users, respectively.

	Blurriness	Flatness	Blockiness	Jerkiness
Shot-1	8.436/8.636/ 8.433	0.071/0.080/0.056		0.467/ 0.290/ 0.788
Shot-2	1.22/ 1.30/ 0.85	0.0505/0.0569/0.0435		0.043/ 0.029 /0.069
Shot-3	1.18/ 1.63/ 0.39		0.0129/0.0130/0.0130	0.028 / 0.023 / 0.037
Shot4	3.64/ 4.58/2.36	0.0244/ 0.0247/0.0242		0.046/ 0.009/ 0.114

Table 2: The performance of our optimal operator selection algorithm: The Spearman rank correlation and subjective rank of the option given by our algorithm, respectively.

	Total			Type-A			Type-B		
	100kbit	200kbit	300kbit	100kbit	200kbit	300kbit	100kbit	200kbit	300kbit
Shot-1	0.74/1	0.94/1	0.77/1	0.6/1	0.83/1	0.54/2	0.84/1	0.9/1	1/1
Shot-2	-0.31/3	0.71/1	0.99/1	0.17/3	0.37/1	1/1	0.99/1	0.99/1	1/1
Shot-3	0.43/4	0.77/1	0.49/1	0.5/4	0.93/1	0.6/1	0.77/3	0.79/1	0.37/1
Shot-4	0.86/1	0.94/1	1/1	0.93/1	0.84/2	0.69/2	0.81/2	0.9/1	1/1