Multi-Reference Video Coding Using Stillness Detection

Di Chen[*], *Zoe Liu*[†], *Yaowu Xu*[†], *Fengqing Zhu*[*], *Edward Delp*[*]
[†] *Google, Inc., 1600 Amphitheatre Parkway, Mountain View, CA, USA 94043.*
[*] *School of Electrical and Computer Engineering, Purdue University, West Lafayette, Indiana, USA 47907.*

Abstract

Encoders of AOM/AV1 codec consider an input video sequence as succession of frames grouped in Golden-Frame (GF) groups. The coding structure of a GF group is fixed with a given GF group size. In the current AOM/AV1 encoder, video frames are coded using a hierarchical, multilayer coding structure within one GF group. It has been observed that the use of multilayer coding structure may result in worse coding performance if the GF group presents consistent stillness across its frames. This paper proposes a new approach that adaptively designs the Golden-Frame (GF) group coding structure through the use of stillness detection. Our new approach hence develops an automatic stillness detection scheme using three metrics extracted from each GF group. It then differentiates those GF groups of stillness from other non-still GF groups and uses different GF coding structures accordingly. Experimental result demonstrates a consistent coding gain using the new approach.

Introduction

The AOM/AV1 codec [1] is an open source, royalty-free video codec developed by a consortium of major technology companies called Alliance for Open Media (AOM) which is jointly founded by Google. It followed the VP9 codec [2, 3], a video codec designed specifically for media on the web by Google WebM Project [4]. The AOM/AV1 codec introduced several new features and coding tools such as switchable loop-restoration [5], global and locally warped motion compensation [6], and variable block-size overlapped block motion compensation [7]. The AOM/AV1 is expected to achieve generational improvement in coding efficiency over VP9.

Current AOM/AV1 codec divides the source video frames into Golden-Frame (GF) groups. The length of each GF group, i.e. the GF group interval, may vary according to the video's spatial or temporal characteristics and other encoder configurations, such as the key frame interval at request for the sake of random access or error resilience. The coding structure of each GF group is based on their interval length and the selection of reference frames buffered for the coding of other frames. The coding structure determines the encoding order of each individual frame within one GF group.

In the current implementation of the AOM/AV1 encoder, a GF group may have a length between 4 to 16 frames. Various GF coding structures may be designed depending on the encoder's decision on the construction of the reference frame buffer, as shown in Figure 1a and Figure 1b. The `extra-ALTREF_FRAME`s and the `BWDREF_FRAME`s introduce hierarchical coding structure to the GF groups [8]. The VP9 codec uses three references for motion compensation, namely `LAST_FRAME`, `GOLDEN_FRAME` and `ALTREF_FRAME`.

`GOLDEN_FRAME` is the intra prediction frame. `LAST_FRAME` is the forward reference frame. `ALTREF_FRAME` is the backward reference frame selected from a distant future frame. It is the last frame of each GF group. A new coding tool is adopted by AV1 that extends the number of reference frames by adding `LAST2_FRAME`, `LAST3_FRAME`, `extra-ALTREF_FRAME` and `BWDREF_FRAME`. `LAST2_FRAME` and `LAST3_FRAME` are similar to `LAST_FRAME`. `extra-ALTREF_FRAME` and `BWDREF_FRAME` are backward reference frames in a relatively shorter distance. The main difference is that `BWDREF_FRAME` does not apply temporal filtering. The hierarchical coding structure in Figure 1a may greatly improve the coding efficiency due to its multi-layer, multi-backward reference design.

The current AOM/AV1 encoder uses the coding structure shown in Figure 1a for all the GF groups. However, a comparison of the compression performance with `extra-ALTREF_FRAME` and `BWDREF_FRAME` enabled and disabled showed that the coding efficiency for some test videos was actually worse when these two reference frames were enabled. This means that the multilayer coding structure does not always have better coding efficiency for all the GF groups. One such example is the GF groups with stillness feature. In this paper, we propose a new approach that adaptively designs the Golden-Frame (GF) group coding structure through the use of stillness detection. A set of metrics are designed to determine whether the frames in a GF group is of little motion. Little work has been done that investigates the use of difference coding structures depending on video content. In [9], an adaptive video coding control scheme is proposed that suggests using more P- and B-frame while the temporal correlation among the frames in a group of pictures (GOP) are high. A method for using different GOP size based on video content is presented in [10].

Method

GF Group Stillness

A GF group may be constructed to contain consistent characteristics to differentiate itself from other GF groups. For instance, some GF group may present stillness across its successive frames, and other may present a zoom-in / zoom-out motion across the entire GF group. We examined the coding efficiency and the stillness feature of each GF group and found that when stillness is present in one GF group, the use of multilayer coding structure as shown in Figure 1a may produce worse coding performance, as opposed to that generated by the one layer structure in Figure 1b.

Automatic GF Group Stillness Detection

An automatic stillness detection of the GF groups is proposed in this paper which allows the GF groups to choose adaptively between two coding structure as shown in Figure 1a and
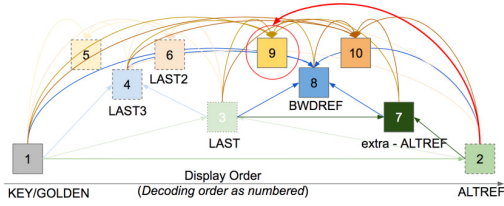
(a) GF Group Coding Structure Using Multilayer
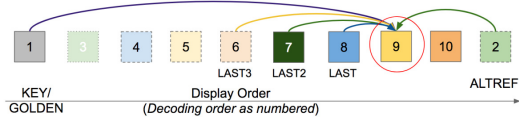
(b) GF Group Coding Structure Using One Layer

Figure 1. *GF Group Coding Structures*

Figure 1b. Three metrics are extracted from the GF group during the first coding pass of AOM/AV1 to determine the GF group stillness. The first coding pass of AOM/AV1 conducts a fast block matching with integer-pixel accuracy and use only one reference frame, the previous frame. Some motion vector and motion compensation information are collected during the first coding pass. Our proposed stillness detection method uses this information to extract three metrics as described below which requires small amount of computation. It then identifies the thresholds and derives the criteria to classify GF groups into two categories: GF groups of stillness and GF groups of non-stillness. The thresholds are obtained by collecting statistics of the three metrics from GF groups of eight low resolution (*cif*) test videos. We manually labeled the stillness or non-stillness of the GF groups. Figure 2 shows the histograms and the thresholds of the three metrics. We intentionally included some test videos that contain GF groups of "stillness-like" characteristics in the non-stillness class because they are more likely to be misclassified as GF group of stillness. The GF group with "stillness-like" characteristics shows either very slow motion or static background with small moving objects. We obtained three criteria which are jointly applied to automatically detect stillness. Finally, the GF group is coded using the workflow given in Figure 3.

Stillness Detection Metrics:

1. **zero_motion_acumulator**: Minimum of the per-frame percentage of zero-motion inter blocks within one GF group:

$$zero\_motion\_accumulator = MIN(pcnt\_zero\_motion_{Fi} \mid Fi \in S) \tag{1}$$
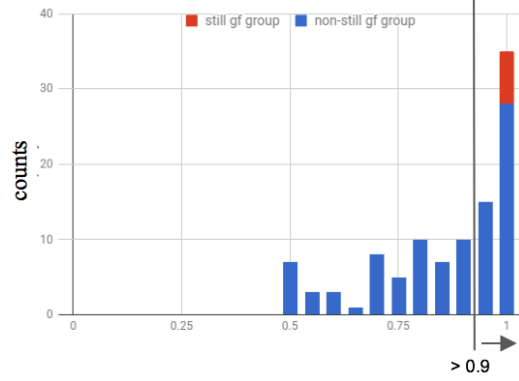
where
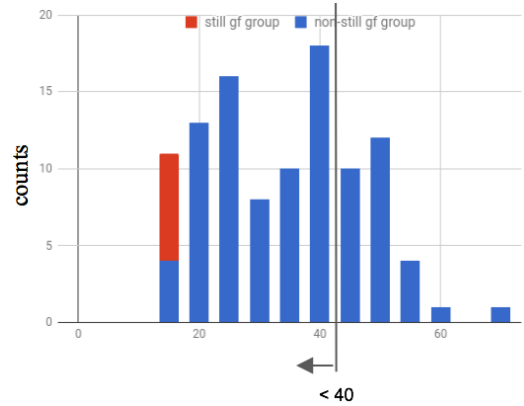$S = \{Fi \mid i = 1, 2, ..., gf\_group\_interval\}$, the set of frames in the GF group
$gf\_group\_interval$: number of frames in the GF group
$pcnt\_zero\_motion$: percentage of the zero-motion inter blocks out of all the inter blocks
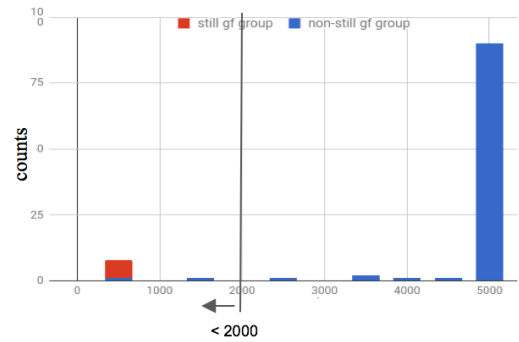
2. **avg_pixel_error**: Average of per-pixel sum of squared

(a) *zero_motion_accumulator*

(b) *avg_pixel_error*

(c) *avg_error_stdev*

Figure 2. *Thresholds for Metrics*

errors (SSE) within one GF group:

$$avg\_pixel\_error =$$
$$MEAN(frame\_sse_{Fi}/number\ of\ pixels\ per\ frame \mid Fi \in S) \tag{2}$$

where
$frame\_sse_{Fi}$ is the SSE of frame Fi

3. **avg_error_stdev**: First calculate the standard deviation of the block-wise SSEs for each frame, where block SSEs are obtained from zero-motion prediction; then obtain the mean value
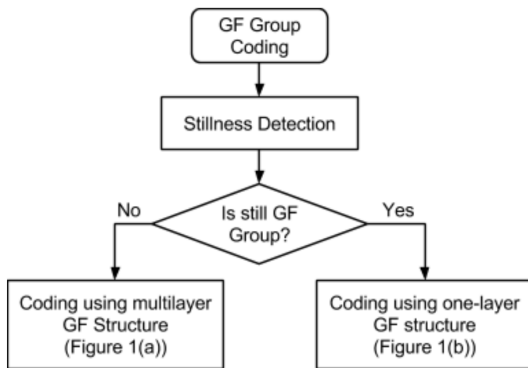
Figure 3. *GF Group Coding With Stillness Detection*

of the standard deviations of all the frames in one GF group:

$$avg\_error\_stdev = MEAN(STDEV_{Fi}(block\_sse_{(0,0)}) \mid Fi \in S) \quad (3)$$

where
$block\_sse_{(0,0)}$ is the block-wise SSEs obtained from zero-motion prediction
$STDEV_{Fi}$ is the standard deviation of the block-wise SSEs of frame Fi

We use the above three metrics to differentiate those GF groups of stillness features from other GF groups, subject to the criteria in Table 1.

Table 1 Criteria for GF group stillness detection

Stillness Detection Metrics	Stillness Detection Criteria (Identified as GF group of stillness)
zero_motion_accumulator	>0.9
avg_pixel_error	<40
avg_error_stdev	<2000

Adaptive GF Group Structure Design

Once a GF group is categorized as a GF group of stillness, no `extra-ALTREF_FRAME` or `BWDREF_FRAME` is used in the single layer coding structure as shown in Figure 1b. The single layer coding structure still has multiple reference frames employed for the coding of one video frame. `LAST_FRAME`, `LAST2_FRAME`, `LAST3_FRAME` and `GOLDEN_FRAME` are used as forward prediction reference and `ALTREF_FRAME` is used as backward prediction reference. If a GF group is categorized as non-still GF group, we will further leverage the use of `BWDREF_FRAME` and `extra-ALTREF_FRAME` to help improve the coding performance.

Experimental Results

We tested the proposed method using two standard video test sets with various resolutions and spatial/temporal characteristics,

as shown in Table 2. More specifically, the set of *lowres* includes 40 videos of *cif* resolution, and the set of *midres* includes 30 videos of 480p and 360p resolution. Each video is coded with a single `GOLDEN_FRAME` and a set of target bitrates. For quality metrics we use the arithmetic average of the frame PSNR and SSIM [11]. To compare RD curves obtained by the base AV1 codec and our proposed method, we use the BDRATE metric [12]. Experimental results demonstrated the advantage of the proposed approach. The Google test set of *lowres* has two video clips that contain detected still GF groups (*pamplet_cif* and *bowing_cif*) and test set *midres* has one (*snow_mnt*). As shown in Table 3, by applying the proposed approach, the BDRATE of video clips that contains GF groups of stillness has decreased by approximately 1%. The classification results of the proposed automatic stillness detector contains no misclassification case in the videos from these two test video sets. There are mainly two reasons that the single layer coding structure has better coding efficiency on the GF groups with stillness feature. One is that the multilayer coding structure in Figure 1a involves more candidate reference frames thus requires more motion information to be transmitted to the decoder. The other reason is that the multilayer coding structure uses an unbalanced bit allocation scheme which is not preferable for GF group of stillness in which the frames are very similar.

Table 2 BDRATE Reduction Using Proposed Method On Google Test Set

test set	BDRATE(PSNR)	BDRATE(SSIM)
test set of *lowres*	-0.063	-0.045
test set of *midres*	-0.026	-0.041

Table 3 BDRATE Reduction Using Proposed Method On Video Clips Contain GF Group Of Stillness

video clip	BDRATE(PSNR)	BDRATE(SSIM)
pamplet_cif	-1.395	-1.076
bowing_cif	-1.118	-0.735
snow_mnt	-0.767	-1.235

Conclusion and Future Work

We proposed an automatic GF group stillness feature detection method. Each GF groups is classified into still GF group and non-still GF group based on three metrics and the encoder adaptively chooses the coding structure based on optimized coding efficiency. Experimental results showed coding gain for videos containing still GF group. We also observed that GF groups containing other features, such as fast zoom-out and high motion, may also benefit from the single layer coding structure.

References

[1] http://www.aomedia.org/.

[2] D. Mukherjee, J. Bankoski, R. S. Bultje, A. Grange, J. Han, J. Koleszar, P. Wilkins, and Y. Xu, "The latest open-source video codec vp9 - an overview and preliminary results," *Proceedings of*

the IEEE Picture Coding Symposium, pp. 390–393, December 2013, San Jose, CA.

[3] ——, "A technical overview of vp9 - the latest open-source video codec," *SMPTE Annual Techical Conference & Exhibition*, October 2013.

[4] http://www.webmproject.org/.

[5] D. Mukherjee, S. Li, Y. Chen, A. Anis, S. Parker, and J. Bankoski, "A switchable loop-restoration with side-information framework for the emerging av1 video codec," *Proceedings of the IEEE International Conference on Image Processing*, pp. 265–269, September 2017, Beijing, China.

[6] S. Parker, Y. Chen, D. Barker, P. de Rivaz, and D. Mukherjee, "Global and locally adaptive warped motion compensation in video compression," *Proceedings of the IEEE International Conference on Image Processing*, pp. 275–279, September 2017, Beijing, China.

[7] Y. Chen and D. Mukherjee, "Variable block-size overlapped block motion compensation in the next generation open-source video codec," *Proceedings of the IEEE International Conference on Image Processing*, pp. 938–942, September 2017, Beijing, China.

[8] Z. Liu, D. Mukherjee, W. Lin, P. Wilkins, J. Han, Y. Xu, and J. Bankoski, "Adaptive multireference prediction using a symmetric framework," *Proceedings of the IS&T International Symposium on Electronic Imaging, Visual Information Processing and Communication VIII*, pp. 65–72(8), January 2017, Burlingame, CA.

[9] S. C. Hsia, "An adaptive video coding control scheme for real-time mpeg applications," *EURASIP Journal on Advances in Signal Processing*, vol. 2003, no. 3, p. 161846, Mar 2003. [Online]. Available: https://doi.org/10.1155/S1110865703210040

[10] B. Zatt, M. S. Porto, J. Scharcanski, and S. Bampi, "Gop structure adaptive to the video content for efficient h.264/avc encoding," *Proceedings of the International Conference on Image Processing*, pp. 3053–3056, September 2010.

[11] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, April 2004.

[12] G. Bjntegaard, "Calculation of average psnr differences between rdcurves," *VCEGM33, 13th VCEG meeting*, March 2001, Austin, Texas.

[13] D. Mukherjee, H. Su, J. Bankoski, A. Converse, J. Han, Z. Liu, and Y. Xu, "An overview of video coding tools under consideration for vp10: the successor to vp9," *Proceedings of the SPIE, Applications of Digital Image Processing XXXVIII*, vol. 9599, p. 95991E, September 2015.

[14] M. Paul, W. Lin, C.-T. Lau, and B. Lee, "A long-term reference frame for hierarchical b-picture-based video coding," *IEEE Transaction on Circuits and Systems for Video Technology*, vol. 24, no. 10, pp. 1729–1742, October 2014.

[15] J. Bankoski, J. Koleszar, L. Quillio, J. Salonen, P. Wilkins, and Y. Xu, "Vp8 data format and decoding guide, rfc 6386," http://datatracker.ietf.org/doc/rfc6386/.

Author Biography

Di Chen is a PhD candidate in Video and Image Processing Laboratory (VIPER) at Purdue University, West Lafayette. Her research focuses on video analysis and compression. Currently, she is working on texture segmentation based video compression using convolutional neural networks.

Zoe Liu received her Ph.D. in ECE from Purdue University. She once worked with corporate research institutes, prototyping mobile video conferencing platforms at Bell Labs (Lucent) and Nokia Research Center. Zoe then devoted her effort to the design and development of FaceTime at Apple, Tango Video Call with the startup of TangoMe, and Google Hangouts. She is currently working in the WebM team of Google.

Dr. Yaowu Xu is currently the Tech Lead Manager of the video coding research team at Google. Dr. Xu's education background includes the BS degree in Physics, the MS and PhD degree in Nuclear Engineering from Tsinghua University at Beijing, China. He also holds the MS and PhD degree in Electrical and Computer Engineering from University of Rochester. His current research focuses on advanced algorithms for digital video compression.

Fengqing Zhu is an Assistant Professor of Electrical and Computer Engineering at Purdue University, West Lafayette, IN. Dr. Zhu received her Ph.D. in Electrical and Computer Engineering from Purdue University in 2011. Prior to joining Purdue in 2015, she was a Staff Researcher at Huawei Technologies (USA), where she received a Huawei Certification of Recognition for Core Technology Contribution in 2012. Her research interests include image processing and analysis, video compression, computer vision and computational photography.

Edward J. Delp was born in Cincinnati, Ohio. He is currently The Charles William Harrison Distinguished Professor of Electrical and Computer Engineering and Professor of Biomedical Engineering at Purdue University. His research interests include image and video processing, image analysis, computer vision, image and video compression, multimedia security, medical imaging, multimedia systems, communication and information theory. Dr. Delp is a Life Fellow of the IEEE, a Fellow of the SPIE, a Fellow of IS&T, and a Fellow of the American Institute of Medical and Biological Engineering.