IVCA: INTER-RELATION-AWARE VIDEO COMPLEXITY ANALYZER

Junqi Liao† *Yao Li*† *Zhuoyuan Li* *Li Li*⋆ *Dong Liu*

University of Science and Technology of China

ABSTRACT

To meet the real-time analysis requirements of video streaming applications, we propose an inter-relation-aware video complexity analyzer (IVCA) as an extension to VCA. The IVCA addresses the limitation of VCA by considering inter-frame relations, namely motion and reference structure. First, we enhance the accuracy of temporal features by introducing feature-domain motion estimation into the IVCA. Next, drawing inspiration from the hierarchical reference structure in codecs, we design layer-aware weights to adjust the majorities of frame complexity in different layers. Additionally, we expand the scope of temporal features by considering frames that be referred to, rather than relying solely on the previous frame. Experimental results show the significant improvement in complexity estimation accuracy achieved by IVCA, with minimal time complexity increase.

***Index Terms*—** Video complexity, Inter-frame relation, Video streaming, Video coding

1. INTRODUCTION

In the context of growing video content demand, optimizing encoding parameters for videos with different content complexity is essential to ensure seamless and high-quality video streaming. A practical approach is to extract relevant complexity features to perform complexity estimation and adjust the encoding parameters.

There are two complexity estimation approaches: coding-result-based and feature-based. Former achieves high accuracy but requires high computational complexity, unsuitable for real-time scenarios [1, 2, 3]. Latter estimates complexity based on video's spatial/temporal features [4, 5, 6, 7], with lower complexity but challenges for accuracy improvement.

Feature-based methods are popular for real-time scenarios. ITU-T recommendations [8] propose using spatial perceptual information (SI) and temporal perceptual information (TI) scores to assess spatial and temporal complexity. The Video Complexity Analyzer (VCA) [4] achieves a good balance between accuracy and complexity by extracting average texture energy (E) and gradient of texture energy (h) as complexity features. However, VCA does not sufficiently consider motion and reference structure in complexity estimation.

This paper presents IVCA, an inter-relation-aware video complexity analyzer built upon VCA. IVCA incorporates motion and reference structure considerations for enhanced complexity estimation. We introduce feature-domain motion estimation (ME) to improve the accuracy of temporal features. Furthermore, we design layer-aware weights to account for frame complexity variations across different layers, considering the reference structure. Additionally, we calculate temporal features based on the reference structure. Experimental results demonstrate that IVCA achieves improved accuracy compared to VCA, with minimal increase in time complexity.

In this paper we will present the contribution performed by the iVC1 and iVC2 teams in the "Grand Challenge on Video Complexity", part of the IEEE International Conference in Image Processing (ICIP) 2024.

2. BACKGROUND

IVCA builds upon VCA, a complexity analyzer known for its accuracy and low time complexity. To provide context for IVCA, we briefly review VCA. In VCA, a DCT-based energy function is used to evaluate the block-wise texture of each frame, defined as follows:

$$H_{p,k} = \sum_{i=0}^{w-1} \sum_{j=0}^{w-1} e^{\left|\left(\frac{ij}{w^2}\right)^2 - 1\right|} \left|DCT\left(i,j\right)\right|, \tag{1}$$

where k represents the block index in the p^{th} frame, $w \times w$ denotes the block size, and $DCT(i,j)$ represents the DCT component at position (i,j). Based on the calculated energy, the spatial feature E can be calculated as:

$$E = \sum_{k=0}^{B-1} \frac{H_{p,k}}{C \cdot w^2}, \tag{2}$$

where B represents the number of blocks per frame. The temporal feature h can be calculated as:

$$h = \sum_{k=0}^{B-1} \frac{SAD\left(H_{p,k}, H_{p-1,k}\right)}{C \cdot w^2}. \tag{3}$$

Then, the frame-level complexity C is calculated by:

$$C = \sum_{i=0}^{N-1} h_i + \sum_{j=0}^{M-1} E_j, \tag{4}$$

where h_i and E_j are the temporal feature of frame i and spatial feature of frame j, respectively. N and M are the number of inter-coded frames and intra-coded frames, respectively.

3. METHODS

3.1. Feature-domain motion estimation

To improve complexity analysis, excluding simple motion between adjacent frames is crucial, as codecs can efficiently

† Junqi Liao and Yao Li contribute equally to this work.

⋆ Corresponding author: Li Li.

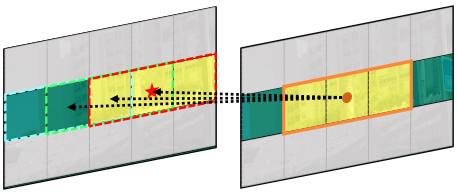

Fig. 1. Illustration of the proposed feature-domain motion estimation in the horizontal direction. Blocks marked with different colors represent feature samples with different energy.

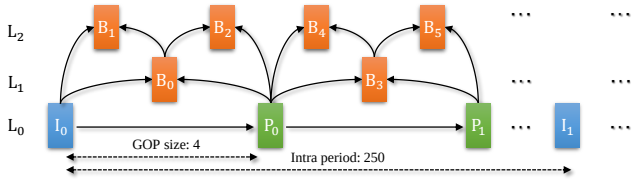

Fig. 2. The reference structure in x264: 3 Layers, GOP Size 4, Intra Period 250.

handle it [9, 10, 11, 12] with few side information bits. Thus, we propose a SAD feature-domain motion estimation method to refine temporal features.

As illustrated in Fig.1, a maximum horizontal feature cosine similarity S_{hor} is firstly calculated between the current feature sample and a set of candidate reference feature samples, with neighboring feature samples also retrieved to perform the calculation:

$$S_{hor} = \max_{j} \left\{ \frac{\sum_{i=0}^{N-1} H_{p,k+\frac{N}{2}-i} H_{p-1,k+\frac{N}{2}-i+j}}{\sqrt{\sum_{i=0}^{N-1} H^2_{p,k+\frac{N}{2}-i}} \sqrt{\sum_{i=0}^{N-1} H^2_{p-1,k+\frac{N}{2}-i+j}}} \right\}, \quad (5)$$

where N stands for the sliding window size for cosine similarity calculation and j stands for the candidate motion offset at the resolution of SAD feature map.

The vertical feature cosine similarity S_{ver} is derived similarly and along with S_{hor} they jointly determine the attenuation factor multiplied to temporal complexity feature h:

$$\mu = \begin{cases} 1 - (S_{hor} + S_{ver}) & (S_{hor} + S_{ver}) \leq 1 \\ 1 - \max(S_{hor}, S_{ver}) & otherwise \end{cases}. \quad (6)$$

H can be quantized and S^2_{hor}, S^2_{ver} may be actually used in implementation to reduce the computation complexity.

3.2. Layer-aware weights scheme

Inspired by the hierarchical reference structure employed in codecs like x264 (as shown in Fig. 2), we propose a layer-aware weights scheme to assign varying weights to frames in different layers. This approach enhances the calculation of sequence-level complexity, resulting in the following improved formula based on (4):

$$C = w_{L_0} \sum_{k=0}^{O-1} h_k + w_{L_1} \sum_{m=0}^{P-1} h_m + w_{L_2} \sum_{n=0}^{Q-1} h_n + w_I \sum_{j=0}^{M-1} E_j, \quad (7)$$

Table 1. Complexity accuracy and speed comparison.

Schemes Applied on VCA			Accuracy	FPS
ME	Weighting	Reference		
			79.15%	48.04
✓			82.88%	48.74
	✓		82.08%	48.04
✓	✓		86.67% (75.70%)	48.74
✓	✓	✓	86.42% (76.95%)	48.64

where O, P, and Q are the number of frames in layer 0, 1, and 2, respectively. w_{L_0}, w_{L_1}, and w_{L_2} are the weights of layer 0, 1, and 2, respectively.

3.3. Reference-based temporal feature

Unlike assuming each frame refers to the previous frame, the actual reference structure follows a hierarchical pattern. Thus, instead of comparing the current frame's DCT energy with the previous frame's DCT energy as in (3), the temporal feature is calculated by comparing the current frame's DCT energy with the potential reference frame's DCT energy using SAD:

$$h = \sum_{k=0}^{C-1} \frac{SAD(H_{p,k}, H_{q,k})}{C \cdot w^2}, \quad (8)$$

where q is the possible reference frame of frame p.

4. RESULTS

The IVCA evaluation involves 50 continuous video clips from the Inter4K dataset [13]. Accuracy is assessed using the Pearson Correlation Coefficient (PCC) against libx264 coding bitrate (medium preset, CRF 26). Processing speed is measured in Frames Per Second (FPS).

Table 1 shows the effectiveness of the proposed feature-domain motion estimation method and layer-aware weights scheme, resulting in 3.73% and 2.93% accuracy improvements, respectively. Combining these methods (iVC2) achieves a total accuracy improvement of 7.52%. The reference-based temporal feature does not improve performance due to the mild motion and insignificant differences between references on the test dataset. However, it exhibits superiority on a subset of the dataset with intense motion (12 video clips in Inter4K), as indicated by the results in brackets. Thus, when the three contributions are combined (iVC1), IVCA will exhibit better performance on datasets with intense motion, such as the BVI-DVC [14] and USTC-TD [15]. The proposed IVCA schemes have negligible impact on time complexity.

5. CONCLUSION

We propose IVCA as an improvement over the existing VCA, addressing its limitations. IVCA takes into account motion and reference structure, offering a more comprehensive analysis of complexity in video streaming applications. Through the introduction of feature-domain motion estimation, layer-aware weights, and reference-based temporal feature, IVCA achieves improved estimation accuracy while maintaining negligible increases in time complexity compared to VCA.

6. REFERENCES

[1] Li Li, Bin Li, Houqiang Li, and Chang Wen Chen, "$lambda$-domain optimal bit allocation algorithm for high efficiency video coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 1, pp. 130–142, 2016.

[2] Junqi Liao, Li Li, Dong Liu, and Houqiang Li, "Content-adaptive rate-distortion modeling for frame-level rate control in versatile video coding," *IEEE Transactions on Multimedia*, 2024.

[3] Abdul Haseeb, Maria G Martini, Sergio Cicalò, and Velio Tralli, "Rate and distortion modeling for real-time mgs coding and adaptation," in *2012 Wireless Advanced (WiAd)*. IEEE, 2012, pp. 85–89.

[4] Vignesh V Menon, Christian Feldmann, Hadi Amirpour, Mohammad Ghanbari, and Christian Timmerer, "Vca: video complexity analyzer," in *Proceedings of the 13th ACM multimedia systems conference*, 2022, pp. 259–264.

[5] Vignesh V Menon, Prajit T Rajendran, Reza Farahani, Klaus Schoeffmann, and Christian Timmerer, "Video quality assessment with texture information fusion for streaming applications," in *Proceedings of the 3rd Mile-High Video Conference*, 2024, pp. 1–6.

[6] Vignesh V Menon, Christian Feldmann, Klaus Schoeffmann, Mohammed Ghanbari, and Christian Timmerer, "Green video complexity analysis for efficient encoding in adaptive video streaming," in *Proceedings of the First International Workshop on Green Multimedia Systems*, 2023, pp. 16–18.

[7] Hadi Amirpour, Mohammad Ghasempour, Lingfeng Qu, Wassim Hamidouche, and Christian Timmerer, "Evca: Enhanced video complexity analyzer," in *Proceedings of the 15th ACM Multimedia Systems Conference*, 2024, pp. 285–291.

[8] P ITU-T RECOMMENDATION, "Subjective video quality assessment methods for multimedia applications," 1999.

[9] Vassilis E Seferidis and Mohammad Ghanbari, "General approach to block-matching motion estimation," *Optical Engineering*, vol. 32, no. 7, pp. 1464–1474, 1993.

[10] Li Li, Houqiang Li, Dong Liu, Zhu Li, Haitao Yang, Sixin Lin, Huanbang Chen, and Feng Wu, "An efficient four-parameter affine motion model for video coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 8, pp. 1934–1948, 2017.

[11] Yao Li, Zhuoyuan Li, Li Li, Dong Liu, and Houqiang Li, "Global Homography Motion Compensation for Versatile Video Coding," in *2022 IEEE International Conference on Visual Communications and Image Processing (VCIP)*. IEEE, 2022, pp. 1–5.

[12] Zhuoyuan Li, Zikun Yuan, Li Li, Dong Liu, Xiaohu Tang, and Feng Wu, "Object Segmentation-Assisted Inter Prediction for Versatile Video Coding," *arXiv preprint arXiv:2403.11694*, 2024.

[13] Alexandros Stergiou and Ronald Poppe, "Adapool: Exponential adaptive pooling for information-retaining downsampling," *arXiv preprint*, 2021.

[14] Di Ma, Fan Zhang, and David R Bull, "Bvi-dvc: A training database for deep video compression," *IEEE Transactions on Multimedia*, vol. 24, pp. 3847–3858, 2021.

[15] USTC iVC Lab, "Ustc-td dataset," `https://github.com/Junqi98/USTC-TD-dataset` Accessed March 20, 2024.