Efficient Content-Adaptive Feature-based Shot Detection for HTTP Adaptive Streaming

Paper ID: 2682

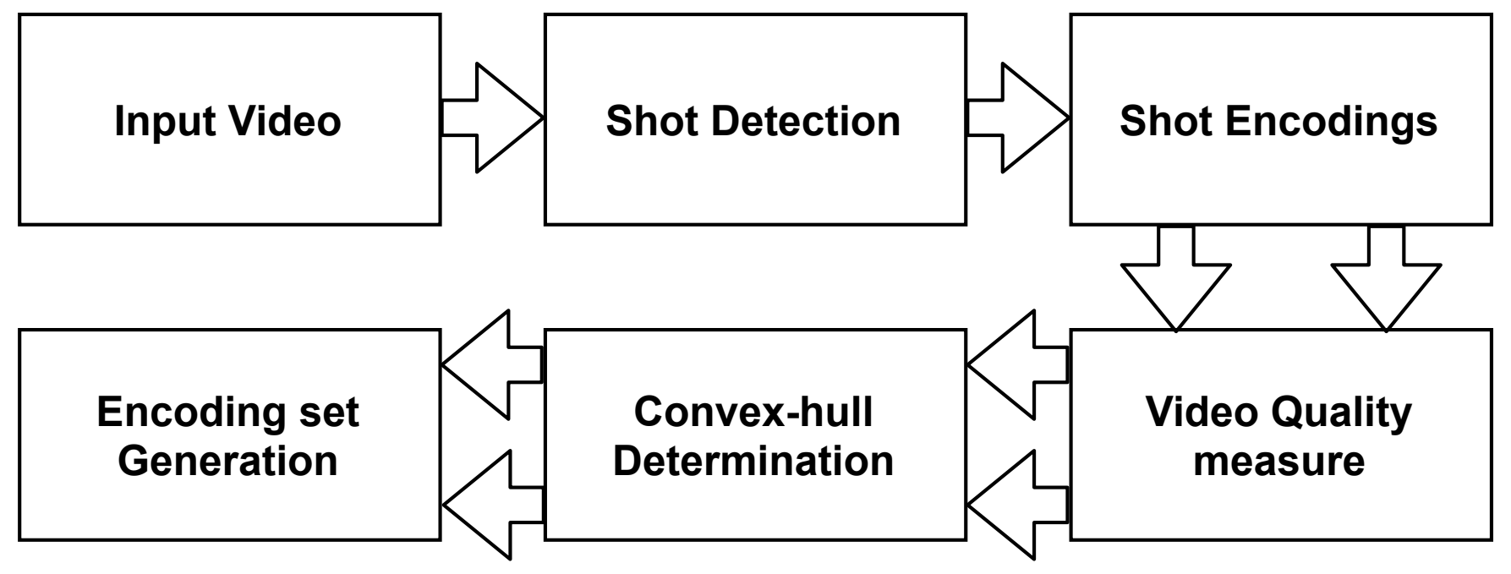
Vignesh V Menon [1] Hadi Amirpour [1] Mohammad Ghanbari [1,2] Christian Timmerer [1]

[1]Christian Doppler Laboratory ATHENA, Alpen-Adria-Universität, Klagenfurt, Austria

[2]School of Computer Science and Electronic Engineering, University of Essex, UK

Introduction

In *HTTP Adaptive Streaming* (HAS), videos are divided into shots and then segments, and each segment is encoded at different bitrates and resolutions referred as *representations* [1].



Input Video

Shot Detection

Shot Encodings

Video Quality measure

Convex-hull Determination

Encoding set Generation

Figure 1. Multi-shot encoding framework for VoD HAS applications [2]

- Each shot is downscaled to a set of spatial resolutions, and all are encoded at a set of bitrates.
- The low-resolution encoded shots are upscaled to the original resolution, and their quality is compared to the original shot.
- Based on these quality measures, a convex hull [3] is formed, and the optimal resolution is selected for each bitrate for each shot.

Hence, precisely detecting shots leads to a higher Quality of Experience (QoE) or bitrate saving.

(a) Frame 255 (b) Frame 262 (c) Frame 269

Figure 2. *snow_mnt* frames 255 to 269 (benchmark algorithm missed this shot transition)

(a) Frame 498 (b) Frame 527 (c) Frame 556

Figure 3. *FoodMarket4* frames 498 to 556 (benchmark algorithm missed this shot transition)

Challenge in shot detection for HAS

Shots can be present in two ways:

- hard shot-cuts, where a frame of one shot is succeeded immediately by a frame from another shot.
- gradual shot transitions, such as dissolve, panning, and zooming, where the changes are accomplished gradually.

The detection of gradual changes is very difficult owing to the fact that the criteria used to determine the significance of a change in the visual information between two frames are subjective and hard to be described in a quantitative format.

Proposed Algorithm

Phase 1: Feature extraction: A DCT-based energy function is used to determine the feature of each CTU i of each frame k, $H_k(i)$ defined as:

$$H_k(i) = \sum_{p=1}^{w}\sum_{q=1}^{h} e^{\left|\left(\frac{pq}{wh}\right)^2-1\right|}\left|DCT(p-1,q-1)\right| \tag{1}$$

where w and h are the width and height of the block, and $DCT(p,q)$ is the (p,q)th DCT component when $p+q>2$, and 0 otherwise [4]. The values of $H_k(i)$ are averaged to H_k for each frame, which represents the average energy per frame.

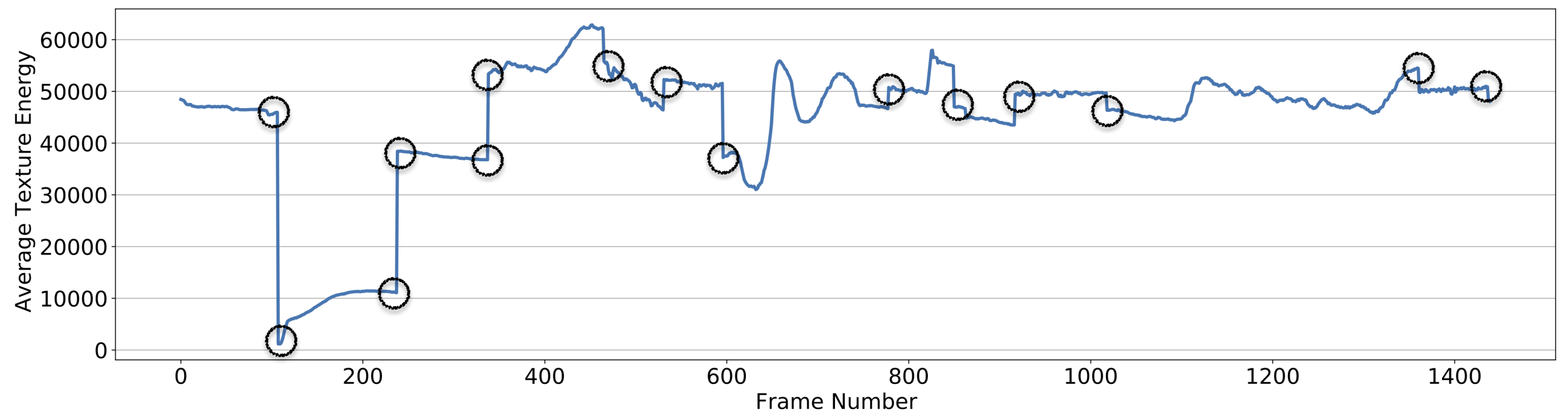


Figure 4. H_k of *Tears_of_Steel* sequence. Black circles denote the regions of shot transitions.

Gradient computation: We define h_k as the Mean Squared Error (MSE) of the CTU level energy values of frame k to that of the previous frame $k-1$, normalized to H_k.

$$h_k = \frac{\sum_{i=1}^{M}(H_k(i)-H_{k-1}(i))^2}{MH_k} \tag{2}$$

where M denotes the number of CTUs in frame k. We define the gradient of h per frame, ϵ given by:

$$\epsilon_k = \frac{h_{k-1}-h_k}{h_{k-1}} \tag{3}$$

Phase 2: Successive elimination:

```
  $T_1$ , $T_2$ : maximum and minimum threshold for $\epsilon_k$
Step 1: while Parsing all video frames do
    if $\epsilon_k > T_1$ then
        $k \leftarrow$ IDR-frame, a new shot.
    else if $\epsilon_k \leq T_2$ then
        $k \leftarrow$ P-frame or B-frame, not a new shot.

  $Q$ : set of frames where $T_1 \geq \epsilon > T_2$
  $q_0, q_{-1}, q_1$: current, previous, and next frame number in the set $Q$
Step 2: while Parsing $Q$ do
    if $q_0 - q_{-1} > fps$ and $q_1 - q_0 > fps$ then
        $q_0 \leftarrow$ IDR-frame, a new shot.
        Eliminate $q_0$ from Q.

Step 3: while Parsing $Q$ do
    if $q_0 - q_{-1} > fps$ and $q_1 - q_0 \leq fps$ then
        compare $\epsilon_{q_0}$ with $\epsilon_q$ when $q$ is from the subset of $Q$ where $q_1 - q_0 \leq fps$
        Frame $q$ with the highest $\epsilon$ value $\leftarrow$ IDR-frame, a new shot.
```

The hard shot-cuts characterized by high ϵ_k are detected in Step 1. Step 2 and Step 3 are designed to handle fade-in, fade-outs, and dissolves. In these situations, frames after the gradual shot-cuts are IDR coded, as the subsequent frames shall have a better reference for encoding.

Results

- We used JVET test sequences [5] to validate the algorithms on known sequences and professionally produced UHD HDR cinematic content to validate performance on typical multi-scene content.
- Results are compared against the default shot detection algorithm in x265 [6].
- Metrics used: Accuracy, precision, recall [7], and F-measure [8] to evaluate the performance of the proposed shot detection.
- ”true” event: the frame is a shot boundary.
- ”positive” event: the detection that the frame is a shot boundary.
- Actual shot-cuts: Ground truth, *i.e.*, the number of real shot transitions determined manually.

Table 1. Shot detection results

Video	Actual shot-cuts	Benchmark algorithm				Proposed algorithm			
		Accuracy	Precision	Recall	F-measure	Accuracy	Precision	Recall	F-measure
BigBuckBunny	10	99.88%	100.00%	80.00%	88.89%	**100.00%**	**100.00%**	**100.00%**	**100.00%**
Dinner	4	99.89%	100.00%	75.00%	85.71%	**99.89%**	**100.00%**	**75.00%**	**85.71%**
FoodMarket4	2	99.72%	-	0%	-	**99.86%**	**100.00%**	**50.00%**	**66.67%**
sintel_trailer	14	99.86%	100.00%	85.71%	92.31%	**99.93%**	**100.00%**	**92.86%**	**96.30%**
snow_mnt	3	99.47%	-	0%	-	**99.65%**	**100.00%**	**33.33%**	**50.00%**
Tears_of_Steel	13	99.93%	100.00%	92.31%	96.00 %	**100.00%**	**100.00%**	**100.00%**	**100.00%**
Busy City	11	99.64%	50.00%	18.18%	26.67%	**99.87%**	**100.00%**	**63.64%**	**77.78%**
FunOnTheRiver	12	99.60%	0%	0%	-	**99.80%**	**85.71%**	**50.00%**	**63.16%**

Table 2. Detection rate statistics of the algorithms

Algorithm	TPR	FPR
Benchmark	53.62%	0.03%
Proposed	78.26%	0.01%

Acknowledgment

The financial support of the Austrian Federal Ministry for Digital and Economic Affairs, the National Foundation for Research, Technology, and Development, and the Christian Doppler Research Association is gratefully acknowledged. Christian Doppler Laboratory ATHENA: `https://athena.itec.aau.at/`.

References

[1] A. Bentaleb *et al.*, “A Survey on Bitrate Adaptation Schemes for Streaming Media Over HTTP,” in *IEEE Communications Surveys Tutorials*, vol. 21, no. 1, pp. 562–585, 2019.

[2] V. P. Malladi *et al.*, “MiPSO: Multi-Period Per-Scene Optimization For HTTP Adaptive Streaming,” in *2020 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, 2020.

[3] H. Amirpour *et al.*, ”PSTR: Per-Title Encoding Using Spatio-Temporal Resolutions,” in *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, 2021.

[4] M. King *et al.*, ”A New Energy Function for Segmentation and Compression,” in *2007 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1647-1650, 2007.

[5] J. Boyce *et al.*, “JVET-J1010: JVET common test conditions and software reference configurations,” 07 2018.

[6] VideoLAN, “x265 HEVC Encoder”, https://www.videolan.org/developers/x265.html.

[7] M. Junker *et al.*, ”On the evaluation of document analysis components by recall, precision, and accuracy,” in *1999 International Conference on Document Analysis and Recognition (ICDAR)*, pp. 713-716, 1999.

[8] Y. Sasaki, ”The truth of the F-measure”, 2007. https://www.toyota-ti.ac.jp/Lab/Denshi/COIN/people/yutaka.sasaki/F-measure-YS-26Oct07.pdf.

https://www.athena.itec.aau.at

International Conference on Image Processing (ICIP) 2021, Anchorage, Alaska, USA

vignesh.menon@aau.at