Real-Time Storyboard Generation for H.264/AVC Compressed Videos

Pei Dong (dongpei@it.usyd.edu.au) Supervisors: *David Dagan Feng* and *Yong Xia*
School of Information Technologies
FACULTY OF ENGINEERING & INFORMATION TECHNOLOGIES

1. Background

Video summarization is attractive in helping people efficiently grasp the main idea and essential information of visual data [1].

H.264/AVC has become the primary video coding standard adopted in most video applications [2].

2. Related work

Conventional pixel domain methods

- Conduct full video decoding to obtain raw pixels
- Rely on raw pixels in video content analysis

Compressed domain methods

- Motion vector-based methods
- Transform coefficient-based methods
- Hybrid methods

Shortcomings of existing methods

- Most pixel domain approaches are computationally expensive.
- Most existing compressed domain approaches are based on the conventional video compression standards, such as MPEG-1, MPEG-2 and MPEG-4 visual.

5. Conclusions and outlook

Conclusions

- Having proposed an H.264/AVC compressed domain summarization algorithm for generic videos
- Fusing multiple types of information in the compressed domain
- Utilizing both compressed domain information and pixel domain information
- Having achieved real-time storyboard generation

Future work

- Investigating automated and adaptive parameter estimation
- Applying this storyboard generation algorithm to videos from more diversified genres

References

[1] A.G. Money and H.W. Agius, "Video summarization: A conceptual framework and survey of the state of the art," *J. Vis. Commun. Image Represent.*, vol. 19, no. 2, pp. 121-143, 2008.

[2] T. Wiegand, J.-R. Ohm, G.J. Sullivan, W. Han, R. Joshi, T.K. Tan, and K. Ugur, "Special section on the joint call for proposals on high efficiency video coding (HEVC) standardization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 12, pp. 1661-1666, 2010.

[3] Z. Liu, Y. Lu, and Z. Zhang, "Real-time spatiotemporal segmentation of video objects in the H.264 compressed domain," *J. Vis. Commun. Image Represent.*, vol. 18, no. 3, pp. 275-290, 2007.

[4] J. Ren, J. Jiang, and J. Chen, "Shot boundary detection in MPEG videos using local and global indicators," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 19, no. 8, pp. 1234-1238, 2009.

Acknowledgement

This research was supported by the Australian Research Council (ARC) grants.

3. The proposed algorithm

Framework

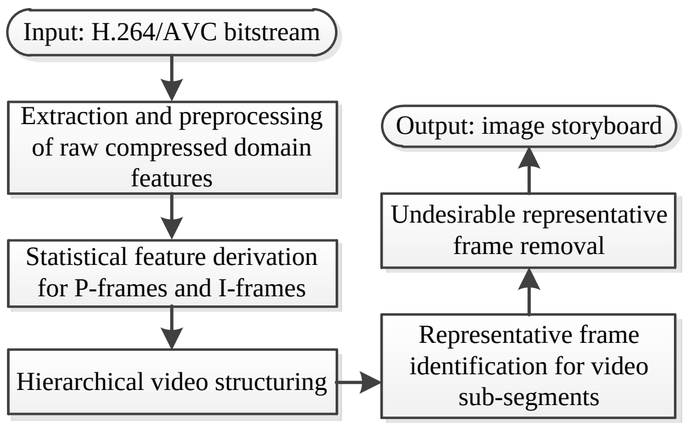


Fig.1. The framework of the proposed algorithm

Main approach

- Extract compressed domain features from H.264/AVC baseline profile bitstream through partial decoding: 1) *motion vectors*, 2) *transform coefficients*, 3) *prediction modes*, and 4) *macroblock bit consumption*.
- Represent video frames by compressed domain features
  - P-frame (predicted frame) features
  - I-frame (intra-coded frame) features
- Analyze the video structure temporally
- Identify representative frames to form a candidate storyboard
- Prune redundant frames

Advantages

- Small amount of video decoding is required.
- The proposed algorithm can be applied to compressed videos in different genres.
- The real-time processing of H.264/AVC compressed videos is achieved.

Visual content representation in H.264/AVC compressed domain

- Extract and preprocess raw information [3]:
  - For each 4×4 block: motion vectors, DCT coefficients, and prediction modes
  - For each 16×16 macroblock: the number of bits consumed
- P-frame feature derivation: compute the statistics and saliency percentages of the extracted P-frame raw information.
- I-frame feature derivation: combine, for each given I-frame, the features of P-frames within a temporal window.

Hierarchical video structuring and representative frame identification

- Two-level temporal segmentation of the compressed video [4]
  - Locate the impulses (possible shot transitions) in the curve of intra mode percentage feature p_{intra}. (See Fig. 1.)
  - Adaptively decide whether an obtained shot should be further divided into a number of sub-segments.
- Keyframe identification
  - Compute the mean I-frame feature of each sub-segment.
  - Select a representative frame (with the most similar feature to the mean) for each sub-segment.

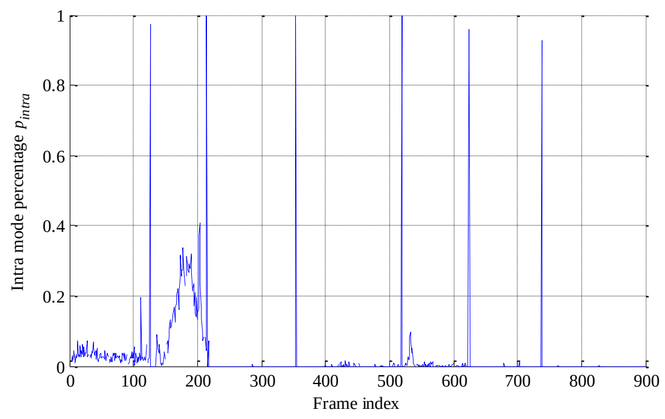


Fig. 2. An example curve of intra mode percentage p_{intra}

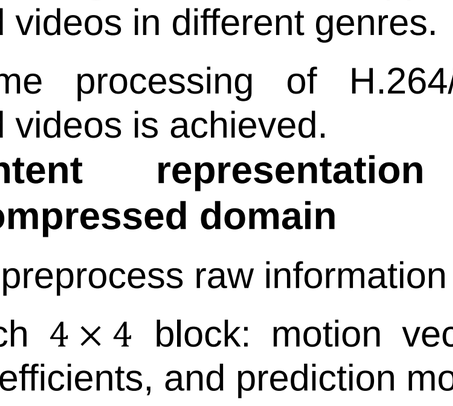
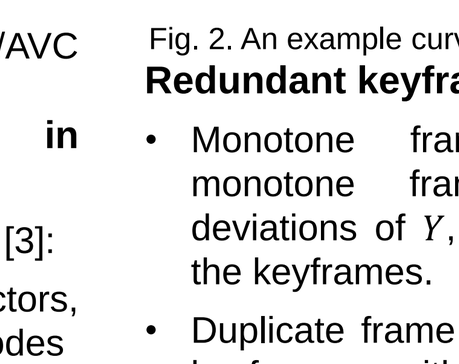
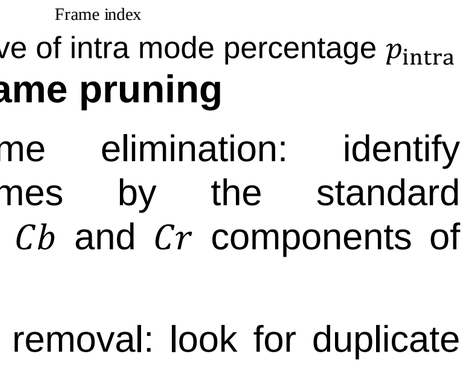
Redundant keyframe pruning

- Monotone frame elimination: identify monotone frames by the standard deviations of Y, Cb and Cr components of the keyframes.
- Duplicate frame removal: look for duplicate keyframes within a search range with updated temporal width.

4. Experimental results

Test sequences

Table. 1. Basic information about the test sequences

	Wildlife	Geology	King's speech
Frame size	352×288	352×240	352×224
Num. of frames	901	1,406	15,000
Sample frames			

Evaluation results (with the metrics of *precision*, *recall* and *F-measure*)

Table. 2. Performances of video structuring and storyboard generation

	WSP *	Wildlife			Geology			King's speech		
		P	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*
Video partitioning	2	**1.000**	**1.000**	**1.000**	**1.000**	0.667	0.800	0.983	**0.772**	0.865
	4	**1.000**	**1.000**	**1.000**	**1.000**	**0.778**	**0.875**	**0.991**	**0.772**	**0.868**
	6	**1.000**	**1.000**	**1.000**	**1.000**	**0.778**	**0.875**	**0.991**	**0.772**	**0.868**
	8	**1.000**	**1.000**	**1.000**	**1.000**	**0.778**	**0.875**	**0.991**	**0.772**	**0.868**
Storyboard generation	2	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**0.644**	0.452	0.531
	4	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	0.629	**0.524**	**0.571**
	6	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	0.615	0.476	0.537
	8	**1.000**	**1.000**	**1.000**	0.857	**1.000**	0.923	0.621	0.488	0.547

* *WSP*: The size parameter of the temporal window.

THIS RESEARCH IS SPONSORED BY

This work has been published in the IEEE International Conference on Multimedia and Expo, ICME 2012, Melbourne, Australia.