Multimodal Stereoscopic Movie Summarization Conforming to Narrative Characteristics

Ioannis Mademlis, Anastasios Tefas, Nikos Nikolaidis, Ioannis Pitas

Department of Informatics, Aristotle University of Thessaloniki

Department of Electrical and Electronic Engineering, University of Bristol

IEEE Transactions on Image Processing Vol. 25, No. 12, pp. 5828-5840, 2016

Introduction

•*Video summarization*: generating concise and non-redundant versions of a video, through the identification of its most representative and salient content.

•The abstracted content to be included in the target summary can be represented as:

1. A temporally ordered collection of selected video frames, i.e., *key-frames*
2. A video skim, i.e., a clip composed of concatenated video segments that each one temporally extends around a previously selected key-frame

•Movie summarization is a special case with wide applicability. It is facilitated by the existence of shot cuts, which naturally segment the video into a sequence of clearly discernible shots.

•Within each shot, one or more key-frames can be selected by utilizing several image modalities.

•Despite the increased popularity of stereoscopic 3D video content, a very limited number of video summarization methods operating on stereoscopic videos have been presented and are mainly using a video frame clustering approach.

•Stereoscopic 3D video conveys per-frame scene geometry information through the binocular disparity image channel.

•Shot selection is an important step in most movie summarization pipelines, able to drastically reduce the number of key-frames.

•High-level narrative properties provide a semantically meaningful structure that can guide the summarization process.

•This work presents a complete, state-of-the-art algorithmic pipeline for skimming stereoscopic 3D movies, by considering video, sound and disparity modalities, as well as film narrative properties.

•A final skim post-processing step eliminates stereoscopic quality defects (Depth Jump Cuts) induced by the summarization process.

Steps

The proposed pipeline is composed of the following steps (novel contributions highlighted in bold):

•*Step 1*: Shot cut detection

•*Step 2*: **Stereoscopic video frame description per shot**

•*Step 3*: Key-frame selection per shot, via frame clustering

•*Step 4*: Monochrome key-frame filtering

•*Step 5*: Redundant key-frame filtering (decimation) by key-frame selection across movie

•*Step 6*: Derivation of key-segments from filtered key-frame set

•*Step 7*: **Audio-assisted key-segment temporal extension**

•*Step 8*: Skim construction by concatenating remaining key-segments

•*Step 9*: **Multimodal Shot Pruning for narrative-based shot selection**

•*Step 10*: **Elimination of disorienting editing effects**

•*Step 11*: Depth Jump Cut elimination

Steps 1-2

•A proven shot cut/boundary detector is first employed to temporally segment the film [1].

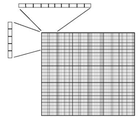
•Subsequently, for each shot, a low-level, multi-channel, simple feature descriptor (Frame Moments Descriptor, FMoD) is employed to produce a feature vector per frame.

•FMoD replaces global color histograms (GCHs), typically employed for frame description in video summarization. It provides a compact description of statistical image properties, in a global and various local scales, while preserving spatial information not available when the frame is summarized by a histogram.

•Luminance, color / hue and binocular disparity image channels are processed. Partial descriptors derived from these channels are concatenated into an aggregate frame descriptor.

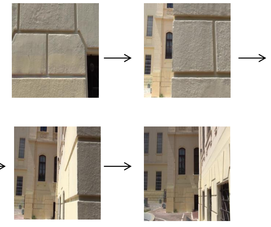
Frame Moments Descriptor (FMoD)

- The frame (M x N) is iteratively partitioned in small blocks (m x n), under a spatial pyramid scheme.
- For each block, profile histograms are computed for the horizontal dimension and the vertical dimension, by averaging pixel values across block columns/rows, respectively.



- The result is an n-dimensional and an m-dimensional vector. Each one is summarized by its first 4 statistical moments.
- The process is repeated multiple times, for larger values of m and n, resulting in different local frame descriptions in different scales.
- The inclusion of disparity-derived information leads to more representative key-frames.

– Example: the camera pans horizontally to the left and, in the middle of the shot, it crosses a wall → the shot is differentiated in disparity (therefore, scene depth), but is homogeneous in luminance and color characteristics



- A single key-frame would suffice if disparity was not taken into account
- Two meaningful key-frames may be found when using disparity information.

Steps 3-8

•Key-frame selection per shot via intra-shot K-Means++ video frame clustering and extraction of cluster medoids.

•Adaptive number of clusters (2 – 5): regulates the number of key-frames per shot, based on an internal clustering evaluation metric (Silhouette Coefficient) within each shot.

•The computed key-frames, derived from all movie shots, are subsequently partitioned in an inter-shot K-Means++ clustering stop with fixed number of clusters (percentage of movie duration).

•All remaining key-frames are temporally extended to key-segments, using p neighboring frames.

•Each key-segment is extended so as to completely include any temporally overlapping speech segment appearances. These are pre-computed using speaker diarization and speaker clustering algorithms. Thus, in the final skim, no speech instance will be abruptly interrupted.

•Any temporally overlapping key-segments are concatenated.

Step 9

•MSP is a post-processing shot selection method that incorporates narrative information and does not require additional data (e.g., the script) beyond the film itself.

•Main idea:

– Discard shots based on known temporal speech (audio) and face (visual) appearance segments, thus considering the narrative prominence of each actor.

– Goal: produce a shorter skim (arguably, more enjoyable), by eliminating key-segments contained in the discarded shots.

Multimodal Shot Pruning (MSP)

•Two V x S shot matrices, **S** and **F**, are initially constructed:

–S is the total number of movie shots

–V is the total number of visible speakers, i.e., different actors that speak

–Typically $S >> V$

$$\mathbf{S}_{ij} = \begin{cases} 1, & \text{if the } i\text{-th actor speaks in the } j\text{-th shot,} \\ 0, & \text{otherwise} \end{cases}$$

$$\mathbf{F}_{ij} = \begin{cases} 1, & \text{if the } i\text{-th actor appears in the } j\text{-th shot,} \\ 0, & \text{otherwise.} \end{cases}$$

$$\text{where } 1 \leq i \leq V, 1 \leq j \leq S.$$

–Most likely, the basis vector sets for **S** and **F** are the standard basis, with one basis vector corresponding to each visible speaker.

•**S** and **F** are modified, through a Gaussian expansion process, in order to extend each speech / face appearance to neighboring shots:

–Binary matrices are converted to real ones

–For each $\mathbf{S}_{ij} = 1$ / $\mathbf{F}_{ij} = 1$, a discrete approximation of a Gaussian distribution, with its mean at $\mathbf{S}_{ij}$/ $\mathbf{F}_{ij}$, is locally assigned to the entries of the i-th row around $\mathbf{S}_{ij}$/ $\mathbf{F}_{ij}$

•Shot matrix values derived from different speech / face appearances and corresponding to the same shot matrix entry are added

–Neighboring speech / face appearances are temporally diffused, to achieve rudimentary scene modeling

–Most likely, the basis vector sets of the modified shot matrices include vectors corresponding to the most prominent actors and vectors corresponding to combinations of more and less prominent actors

–Intuition: lead actors can appear alone, while supporting actors appear mainly along with leads

•Approach: Cast the problem as a joint matrix Column Subset Selection Problem (CSSP) on **S** and **F**, where the desired solution is a vector **c** of matrix column indices (corresponding to retained movie shots). Solve it with a genetic algorithm [2], using the following joint-CSSP fitness function: $f(\mathbf{c}) = (\|\mathbf{S} - (\mathbf{C}^S \mathbf{C}^{S+})\mathbf{S}\|_F + \|\mathbf{F} - (\mathbf{C}^F \mathbf{C}^{F+})\mathbf{F}\|_F)^{-1}$

– $\mathbf{C}^S$ / $\mathbf{C}^F$ are sub-matrices of **S** / **F**, respectively, containing only the columns indicated by **c**.

Steps 10-11

•Key-segments contained within the same shot and separated by less than a second of video duration are merged. Too short key-segments are eliminated. Purpose: eliminate abrupt *temporal jump cuts*.

•Visually annoying *depth jump cuts*, i.e., severe mean disparity mismatches among consecutive video frames induces by the skim construction process, are detected and fixed by applying the method in [3].

Subjective Evaluation

Summary Informativeness

METHOD	Movie1	Movie2	Movie3
FMoD+MSP	70%	74%	72%
GCH, No-MSP	**83%**	**82%**	**81%**
[4]	75%	77%	76%

Summary Enjoyability

METHOD	Movie1	Movie2	Movie3
FMoD+MSP	**72%**	**73%**	**71%**
GCH, No-MSP	56%	59%	57%
[4]	62%	64%	61%

References

[1] Z. Cernekova, I. Pitas, and C. Nikou, "Information theory-based shot cut/fade detection and video summarization.," IEEE Transactions on Circuits and Systems for Video Technology, vol. 16, no. 1, pp. 82–91, 2006

[2] P. Kromer, J. Platos, and V. Snasel, "Genetic algorithm for the column subset selection problem," Complex, Intelligent and Software Intensive Systems, pp. 16–22, 2014

[3] S. Delis, I. Mademlis, N. Nikolaidis, and I. Pitas, "Automatic detection of 3D quality defects in stereoscopic videos using binocular disparity," IEEE Transactions on Circuits and Systems for Video Technology, vol. 27, no. 5, pp. 997-991, 2017

[4] N. Doulamis, A. Doulamis, Y.S. Avrithis, K.S. Ntalianis, and S.D. Kollias, "Efficient summarization of stereoscopic video sequences," IEEE Transactions on Circuits and Systems for Video Technology, vol. 10, no. 4, pp. 501517, 2000

Acknowledgement

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement number 287674 (3DTVS)