

Multimodal Stereoscopic Movie Summarization Conforming to Narrative Characteristics

Ioannis Mademlis, Anastasios Tefas, Nikos Nikolaidis, Ioannis Pitas

Introduction

•The abstracted content to be included in the target summary can be represented as: A train porally ordered collection of selected video frames, i.e., <i>key frames</i> 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 The dostracted end near mainly using a video frame clustering approach. Shot selection is an important step in most movie summarization pipelines, able to drastically reduce the number of key-frames. This work presents a complete, state-of-the-art algorithmic pipeline for skimming stereoscopic 3D movies, by considering video, sound and disparity modalitics, as well as lilm narrative properties. The summarization process. Steps The proposed of the following steps (novel contributions highlighted in bold): Step 1: Shot cut detection Step 2: Stereoscopic video frame description per shot Step 4: Monochroma key-frame filtering (decination) by key-frame selection across movie Step 7: Monochroma key-frame filtering (decination) by key-frame selection across movie Step 6: Monochroma key-frame filtering (decination) by key-frame selection across movie Step 6: Monochroma key-frame filtering (decination) by key-frame select	• <i>Video summarization</i> : generating concise and non-redundant versions of a video, through the identification of its most representative and salient content.	 The For
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 approach. •Stereoscopic 3D video conveys per-frame scene geometry information through the binocular disparity mange thumel. •Shot selection is an important step in most movie summarization pipelines, able to drastically reduce the number of key-frames. •High-level narative properties provide a semantically meaningful structure that can guide the summarization process. •High-level narative properties provide a semantically meaningful structure that can guide the summarization process. •High-level arative properties provide a semantically meaningful structure that can guide the summarization process. •High-level arative properties provide a set esception protection. •A final skim post-processing step eliminates stereoscopic quality defects (Depth Jump Cuts) induced by the summarization process. •High-level arative properties. •Mean 2000 Step 4: Shot cut detection *Step 5: Stereoscopic video frame descript		
 A video skim, i.e., a clip composed of concatenated video segments that each one temporally extends around a previously selected key-frame A video skimmarization 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 approach. 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. His 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 linal skim post-processing step eliminates steroscopic quality defects (Depth Jump Cuts) induced by the summarization process. Step 2: Shot cut detection Step 4: Monochrome key-frame filtering (decimation) by key-frame selection across movie Step 4: Monochrome key-frame filtering (decimation) by key-frame selection across movie Step 7: Audio-assisted key-segment temporal extension Step 9: Derivation of key-segment temporal extension Step 9: Derivation of disorienting editing effects Step 1: Depth Jump Cut climination Step 1: Depth Jump Cut climination Step 1: Depth Jump Cut climination Store sequently, for each shot, a low-level, multi-channet, simple feature description frame description in video summarization. The provides a compact descriptio on statistical inorga properties, in a global and various l		
 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 approach. Stereoscopic 3D video conveys per-frame scene geometry information through the binocular disparity image channel. Storeoscopic 3D video conveys per-frame scene geometry information through the binocular disparity image of key-frames. The field-vel natrative properties provide a semantically meaningful structure that can guide the summarization process. This work presents a complete, state-of-fhe-art algorithmic pipeline for skimming stereoscopic 3D movies, by considering video, sound and disparity modalifies, as well as film natrative properties. A final skim post-processing step eliminates stereoscopic quality defects (Depth Jump Cuts) induced by the summarization process. Step 1: Shot cut detection Step 2: Steroscopic video frame description per shot Step 3: Key-frame selection per shot, via frame clustering Step 4: Monchrome key-frame filtering (decimation) by key-frame selection across movie Step 4: Skim construction by concatenating remaining key-segments Step 4: Derivation of key-segments from filtered key-frame set Step 4: Dubindual tky-frame filtering (Beelmation) by key-frame selection Step 4: Derivation of disorienting editing effects Step 1: Depth Jump Cut elimination Step 1: Depth Jump Cut elimination Step 1: Depth Jump Cut elimination Step 2: Nultimodal Shot Pruning for marrative-based shot selection Step 1: Depth Jump Cut elimi	2. A video skim, i.e., a clip composed of concatenated video segments that each one temporally extends	
 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. Stereoscopie 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 video, sound and disparity image channel. 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. Step 2: Sterooscopic video frame description per shot Step 3: Key-frame selection per shot, via frame clustering Step 4: Monochrome key-frame filtering (decimation) by key-frame selection across movie Step 5: Rodundant key-frame filtering (decimation) by key-frame selection across movie Step 6: Dirivation of key-segments from filtered key-frame set Step 7: Molio-assisted key-segment temporal extension Step 9: Multimodal Shot Pruning for narrative-based shot selection Step 1: Depth Jump Cut elimination Step 1: Depth	•Movie summarization is a special case with wide applicability. It is facilitated by the existence of shot cuts, which	
methods operating on stereoscopic videos have been presented and are mainly using a video frame clustering aptroach. Stereoscopic 3D video conveys per-frame scene geometry information through the binocular disparity image channel. Storeoscopic 3D video conveys per-frame scene geometry information through the binocular disparity image channel. Storeoscopic 3D video conveys per-frame scene geometry information through the binocular disparity image channels. Storeoscopic 3D video conveys per-frame scene geometry information through the binocular disparity image channels are processed. This between the 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. Step 5: Shot cut detection Step 2: Stereoscopic video frame description per shot Step 4: Monochrome key-frame filtering (decimation) by key-frame selection across movie Step 5: Redundant key-frame filtering (decimation) by key-frame selection across movie Step 5: Redundant key-frame filtering (decimation) by key-frame selection across movie Step 6: Sim construction by concatenating remaining key-segments Step 7: Audio-assisted key-segment temporal extension Step 8: Skim construction by concatenating remaining key-segments Step 11: Depth Jump Cut elimination Mear Step 11: Depth Jump Cut elimination Final Step 12: Depth Jump Cut elimination Final Step 2: Step 4: Nonourdary 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, Find) Di semployed oproduce a feature vector per frame. FMoD replaces global color histograms (GCHs), typically employed for frame descrip	•Within each shot, one or more key-frames can be selected by utilizing several image modalities.	
 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. Step 5: 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 (decimation) by key-frame selection across movie Step 5: Redundant key-segment from filtered key-frame set Step 7: Audio-assisted key-segment temporal extension Step 9: Multimodal Shot Pruning for narrative-based shot selection 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 cacles, 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 descriptor derived from Grown 	methods operating on stereoscopic videos have been presented and are mainly using a video frame clustering	• The mor
 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. Step 1: Shot cut detection Step 2: Stereoscopic video frame description per shot Step 4: Monochrome key-frame filtering Step 5: Redundant key-frame filtering (decimation) by key-frame selection across movie Step 5: Redundant key-frame filtering (decimation) by key-frame selection across movie Step 5: Redundant key-frame filtering (decimation) by key-frame selection Step 6: Derivation of key-segments from filtered key-frame set Step 7: Multimodal Shot Pruning for narrative-based shot selection Step 7: Di Elimination of disorienting editing effects Step 1: Depth Jump Cut elimination Step 5: Apply construction by concate a feature vector per frame. Appe apply do produce a feature vector per frame. Appe apply do produce a feature vector per frame. Appe apply of the proved s a compact description of statistical image properties, in a global and various loca acres, while preserving spatial information not savallable when the frame is summarized by a histogram. Ama 	•Stereoscopic 3D video conveys per-frame scene geometry information through the binocular disparity image	dese
 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. Step 5 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 (decimation) by key-frame selection across movie Step 5: Redundant key-frame filtering (decimation) by key-frame selection across movie Step 7: Audio-assisted key-segment temporal extension Step 8: Skim construction by concatenating remaining key-segments 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 loo. Gor -Gor 		– Exa
 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 (decimation) by key-frame selection across movie Step 5: Redundant key-frame filtering (decimation) by key-frame selection across movie Step 7: Audio-assisted key-segment temporal extension Step 7: Audio-assisted key-segment temporal extension 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 con statistical image properties, in a global and various local con statistical image properties, in a global and various local con statistical image properties, in a global and various local con statistical image properties, in a global and various local con statistical image properties, in a global and various local con statistical image properties, in a global and various local con statistical image properties, in a global and various local con statistical image properties, in a global and various		cha
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 •Ada •Step 6: Derivation of key-segments from filtered key-frame set •Step 7: Audio-assisted key-segment temporal extension •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 summarized by a histogram. •FMoD replaces global color histograms (GCHs), typically employed for frame description in video summarized by a histogram. •Luminance, color / hue and binocular disparity image channels are processed. Partial descriptors derived from		
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 5: Redundant key-frame filtering (decimation) by key-frame selection across movie • • Step 6: Derivation of key-segment filtered key-frame set • • Step 7: Audio-assisted key-segment temporal extension • • Step 9: Multimodal Shot Pruning for narrative-based shot selection • • Step 10: Elimination of disorienting editing effects • • Step 11: Depth Jump Cut elimination • • A proven shot cut/boundary detector is first employed to temporally segment the film [1]. • • Step 0: semployed to produce a feature vector per frame. • • FMoD replaces global color histograms (GCHs), typically employed for frame description in video accales, 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 descriptor serived from -	summarization process.	
 Step 1: Shot cut detection Step 2: Stereoscopic video frame description per shot Step 3: Key-frame selection per shot, via frame clustering Step 3: Key-frame selection per shot, via frame clustering Step 4: Monochrome key-frame filtering (decimation) by key-frame selection across movie 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 -Go 	Steps	
•Step 2: Stereoscopic video frame description per shot•Key•Step 3: Key-frame selection per shot, via frame clustering•Key•Step 4: Monochrome key-frame filtering•Key•Step 5: Redundant key-frame filtering (decimation) by key-frame selection across movie•Ada•Step 6: Derivation of key-segments from filtered key-frame set•Ada•Step 7: Audio-assisted key-segment temporal extension•The•Step 8: Skim construction by concatenating remaining key-segments•The•Step 9: Multimodal Shot Pruning for narrative-based shot selection•Ada•Step 10: Elimination of disorienting editing effects•Ada•Step 11: Depth Jump Cut elimination•Eact•Step 11: Depth Jump Cut elimination•Ada•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.•MS•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.•Mai•Luminance, color / hue and binocular disparity image channels are processed. Partial descriptors derived from-Go	The proposed pipeline is composed of the following steps (novel contributions highlighted in bold):	
 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 5: Redundant key-frame filtering (decimation) by key-frame selection across movie Adage value 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 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 -Go 	•Step 1: Shot cut detection	
 Step 4: Monochrome key-frame filtering Step 5: Redundant key-frame filtering (decimation) by key-frame selection across movie Step 5: Redundant key-segment filtering (decimation) by key-frame selection across movie Ada Step 6: Derivation of key-segment temporal extension 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 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 -Go 	•Step 2: Stereoscopic video frame description per shot	
Step 5: Redundant key-frame filtering (decimation) by key-frame selection across movie •Ada Step 5: Redundant key-frame filtering (decimation) by key-frame selection across movie •Ada Step 6: Derivation of key-segments from filtered key-frame set •Ada Step 7: Audio-assisted key-segment temporal extension •Ada Step 8: Skim construction by concatenating remaining key-segments •Mear Step 9: Multimodal Shot Pruning for narrative-based shot selection •All n Step 10: Elimination of disorienting editing effects •Eact Step 11: Depth Jump Cut elimination •Eact 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. •Ms •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. •Mai •Luminance, color / hue and binocular disparity image channels are processed. Partial descriptors derived from -Go	•Step 3: Key-frame selection per shot, via frame clustering	
 Step 5: Redundant key-frame filtering (decimation) by key-frame selection across movie 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 -Go 	•Step 4: Monochrome key-frame filtering	•Key
 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 -Go 	•Step 5: Redundant key-frame filtering (decimation) by key-frame selection across movie	
 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 - Go 	•Step 6: Derivation of key-segments from filtered key-frame set	
 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 - Go 	Step 7: Audio-assisted key-segment temporal extension	
 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 -Go 	•Step 8: Skim construction by concatenating remaining key-segments	
 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 -Go 	•Step 9: Multimodal Shot Pruning for narrative-based shot selection	
 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 -Go 	•Step 10: Elimination of disorienting editing effects	
 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 - Go 	•Step 11: Depth Jump Cut elimination	appe
 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 -Go 	<u>Steps 1-2</u>	•Any
 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 - Go 	•A proven shot cut/boundary detector is first employed to temporally segment the film [1].	
 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 - Go 		•MS add
 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 -Go 	•FMoD replaces global color histograms (GCHs), typically employed for frame description in video	•Mai
•Luminance, color / hue and binocular disparity image channels are processed. Partial descriptors derived from $-Go$		– Dis con
THEN A CHEATHERN ALE PRODUCTION AND THE CONTRACT OF CONTRACT.		– Go

Department of Informatics, Aristotle University of Thessaloniki

Department of Electrical and Electronic Engineering, University of Bristol

Frame Moments Descriptor (FMoD)

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

e result is an *n*-dimensional and an *m*-dimensional vector. Each one is summarized by its first 4 statistical nents

process is repeated multiple times, for larger values of m and n, resulting in different local frame criptions in different scales.

e inclusion of disparity-derived information leads to more representative key-frames.

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





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

-Typically S >> V

- visible speaker.
- neighboring shots:
- are added

pp. 82-91, 2006

[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 Acknowledgemen

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)

Steps 3-8

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

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

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

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

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

temporally overlapping key-segments are concatenated

Step 9

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

in idea:

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

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

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



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

 $\int 1$, if the *i*-th actor speaks in the *j*-th shot

otherwise

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

where $1 \le i \le V, 1 \le j \le S$.

-Most likely, the basis vector sets for S and F are the standard basis, with one basis vector corresponding to each

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

-Binary matrices are converted to real ones

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

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

–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: $f(\mathbf{c}) = (||\mathbf{S} - (\mathbf{C}^{\mathrm{S}} \mathbf{C}^{\mathrm{S}+})\mathbf{S}||_{\mathrm{F}} + ||\mathbf{F} - (\mathbf{C}^{\mathrm{F}} \mathbf{C}^{\mathrm{F}+})\mathbf{F}||_{\mathrm{F}})^{-1}$

 $- C^{S} / C^{F}$ 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].

Summary Informativeness					Summary Enjoyability					
THOD	Movie1	Movie2	Movie3		METHOD	Movie1	Movie2	Movie3		
oD+MSP	70%	74%	72%		FMoD+MSP	72%	73%	71%		
H, No-MSP	83%	82%	81%		GCH, No-MSP	56%	59%	57%		
	75%	77%	76%		[4]	62%	64%	61%		
References										

Subjective Evaluation

[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,

[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