STEREOSCOPIC VIDEO DESCRIPTION FOR KEY-FRAME EXTRACTION IN MOVIE SUMMARIZATION

Ioannis Mademlis, Nikos Nikolaidis and Ioannis Pitas

Department of Informatics, Aristotle University of Thessaloniki, Thessaloniki, Greece
Email: {imademlis,nikolaid,pitas}@aiia.csd.auth.gr

ABSTRACT

A novel, low-level video frame description method is proposed that is able to compactly capture informative image statistics from luminance, color and stereoscopic disparity video data, both in a global and in various local scales. Thus, scene texture, illumination and geometry properties may succinctly be contained within a single frame feature descriptor, which can subsequently be employed as a building block in any key-frame extraction scheme, e.g., shot frame clustering. The computed key-frames are subsequently used to derive a movie summary in the form of a video skim, which is suitably post-processed to reduce stereoscopic video defects that cause visual fatigue and are a by-product of the summarization.

Index Terms— Video Summarization, Stereoscopic Video Description, Bag-of-Features

1. INTRODUCTION

Video summarization aims at generating condensed versions of a video stream through the identification of its most important and pertinent content [1]. The derived video summaries can be subsequently exploited in various applications, like interactive browsing and search systems, thereby offering the user the ability to efficiently view, manage and assess video content [2]. Such methods initially try to select a set of salient video frames, such as shot key-frames that represent the video context. Information is extracted by analysing the available modalities (video, audio or text) for abstracting intuitive semantics, such as objects, events, as well as low-level features from the video stream. The abstracted content that needs to be included in the target summary can be represented as still images (key-frames), a video skim, or by employing graphical and textual descriptions [1]. Due to the inherently subjective nature of the task (there is no such thing as a globally agreed good video summary), evaluation of the success of a summarization method is typically subjective.

Generic video summarization algorithms extract key-frame cues, i.e., sequences of key-frames presented in temporal order [3]. To achieve this, each video frame is first described by low-level image descriptors, such as global color-based, texture-based or shape-based features [2]. Composite descriptors which may additionally consider visual attention attributes have also been employed [3]. In general, the most commonly employed frame descriptors are variants of joint image histograms in the HSV color space [4] [5] [6]. Moreover, dimensionality reduction on such color histograms has been attempted [7], in order to decrease the computational cost of the subsequent summarization steps. In a few cases [8], local image region descriptors have been employed for video description, using the Bag-of-Features representation model [9].

In order to extract key-frames, the frame descriptors are typically processed by unsupervised learning algorithms, i.e., clustering is employed to create frame groups, under the assumption that the camera focuses more on important frames [4]. The number of clusters may be set proportionally to the video length [5]. Subsequently, a set of frames that are closest to each cluster centroid are initially selected as key-frames. Typically, a refinement post-processing stage filters out a percentage of the extracted key-frames and the remaining ones are presented in temporal order to produce a storyboard. In [10] a similarity metric is described that assesses the video frame-by-frame, in order to detect whether each frame should be included in the summary. Frames similar to their previous ones are excluded, while a noise reduction technique based on histograms is applied to exclude homogeneously colored frames (e.g., black frames).

Video skims are series of short video segments concatenated in the correct temporal order, in order to form a shorter version of the original stream that contains the informative content. In video summarization with applications to movie post-production, the state-of-the-art approach exploits content selection techniques and video skimming.

Recently, the rise in popularity of 3D video content has reoriented video analysis research towards the exploitation of scene depth information. In stereoscopic 3D video content derived from filming with stereo camera rigs (matched pairs of cameras), two images of the scene are available for each video frame, taken at the same time from slightly different positions in world space. From every such *stereo-pair*, a *disparity map* may be derived using a disparity estimation algo-

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).

978-0-9928626-3-3/15/$31.00 ©2015 IEEE

rithm. Thus, a binocular disparity value (also called *stereo disparity*) is assigned to each color video pixel, that indicates relative distance from the stereo rig. Obviously, the disparity map corresponding to a video frame is indicative of the imaged scene geometry.

Despite the increased popularity of 3D video content, a very limited number of video summarization methods operating on stereoscopic or multi-view videos have been presented and are mainly using a shot clustering approach. For instance, in [11] object segmentation utilizing both color and disparity-derived relative-depth information is performed. Next, feature vectors are constructed using multi-dimensional fuzzy classification of segment features including size, location, color and relative-depth.

This work presents the Frame Moments Descriptor (FMoD), a novel video frame descriptor developed with the goal of summarizing stereoscopic movies through key-frame extraction and the derivation of stereoscopic video skims. These skims are subsequently post-processed in order to reduce 3D video defects which cause visual fatigue and are a by-product of the summarization.

2. STATISTICAL STEREOSCOPIC VIDEO DESCRIPTION FOR MOVIE SUMMARIZATION

In the proposed approach, the stereoscopic video is assumed to be composed of a temporally ordered sequence V^L of N_f luminance frames and a temporally ordered sequence V^D of N_f corresponding disparity maps, containing pixel values. Each luminance frame and each disparity map can be considered a matrix $\mathbf{V}_i^L \in \mathbb{R}^{M\times N}$ and $\mathbf{V}_i^D \in \mathbb{R}^{M\times N}$, respectively, where $i = 1, \cdots, N_f$. Both V^L and V^D are assumed to have been identically partitioned into non-overlapping shots, e.g., by employing the information-theoretic method described in [12].

Key-frames are automatically extracted per shot, exploiting both luminance and disparity information. The number of key-frames extracted at each shot (K) is adaptive, i.e., it lies between 2 and a user-provided maximum K_{max}, a parameter that regulates the granularity of the shot summarization. Initially, a feature vector is extracted per frame, using a particular feature descriptor. Subsequently, all shot frames are partitioned into K clusters. Finally, the frames closest to the cluster centroids in the feature space, in terms of Euclidean distance, are selected as the resulting key-frames. These are subsequently employed to generate a video skim, containing the selected key-frames from all shots and a temporal neighbourhood of frames around each of them.

The K-Means++ algorithm [13] has proven to be sufficient for clustering. Other clustering algorithms have been tested and shown to provide similar results. A novel feature descriptor, called hereafter Frame Moments Descriptor (FMoD), is used for video frame description. It preserves spatial information not available when an entire frame is summarized by a histogram. It can be used with any type of image modality (luminance, color, disparity etc.). Additionally,

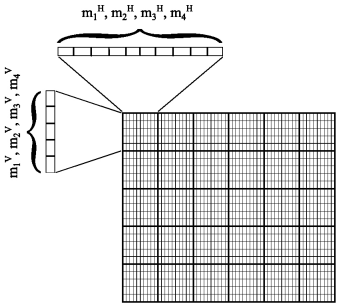

Fig. 1. Statistical summarization of an image block.

a variant of FMoD called Position-Invariant Frame Moments Descriptor (PI-FMoD), is also being described below. It removes most of the spatial information by employing the Bag-of-Features representation model.

The Frame Moments Descriptor operates by partitioning a $M \times N$ frame ($\mathbf{V}_i^L$ or $\mathbf{V}_i^D$) in small blocks of $m \times n$ pixels, where $m < M$ and $n < N$. In each block, one *profile vector* is computed for the horizontal dimension and one for the vertical dimension, by averaging pixel values across block columns / rows, respectively. The result is an n-dimensional and an m-dimensional profile vector. Each of the two vectors is summarized by their first 4 statistical moments (mean, standard deviation, skewness, kurtosis). The resulting 8-dimensional vector $\mathbf{f}_i^v = [m_1^H, m_2^H, m_3^H, m_4^H, m_1^V, m_2^V, m_3^V, m_4^V,]^T$, where v is either L (luminance) or D (disparity), compactly captures the statistical properties of the block, as shown in Figure 1. The process is successively repeated d times, for larger values of m and n. In the last iteration, $m = M$ and $n = N$. Finally, all the 8-dimensional vectors are concatenated into a single feature vector that describes the entire frame. This vector set (one per frame) is used for key-frame extraction.

FMoD feature extraction was implemented recursively, in a top-down manner, with the image region that is currently being statistically summarized at each time, subsequently being partitioned into 4 quadrants. These quadrants serve as input blocks to the 4 recursive function calls of the next step. Thus, the total number of 8-dimensional block vectors that are to be concatenated is given by the sum of the first d terms of a geometric progression:

$$1 \cdot 4^0 + 1 \cdot 4^1 + \cdots + 1 \cdot 4^{d-1} = (4^d - 1)/3 \qquad (1)$$

Therefore, the final FMoD feature vector has $8 \cdot (4^d - 1)/3$ dimensions. It compactly describes the video frame in a global and in various local scales, with local information being more spatially focused for higher values of d.

The Position-Invariant Frame Moments Descriptor (PI-FMoD) is a variant of FMoD that employs an additional step, in order to discard spatial information from the frame description. This step consists in transforming the set of all

8-dimensional block vectors that compose the FMoD vector into a histogram, using a Bag-of-Features representation [9]. That is, all $(4^d-1)/3$ block vectors of the frame are clustered into c representative block summaries, where c is the codebook size parameter. Each vector is subsequently assigned to the nearest representative block summary, in terms of Euclidean distance. The number of block vectors assigned to each of the c clusters is an entry in a c-dimensional vector. This vector is followingly transformed into a histogram by L_1-normalization, in order to produce the final c-dimensional frame feature vector.

By employing subjective visual inspection, PI-FMoD frame description was empirically found to perform better than FMoD in the context of key-frame extraction, since spatial information is not necessarily an important factor for the determination of representative shot frames. For instance, a static-camera shot showing an actor walking from the left frame border towards the right one, might be represented by a single key-frame in a satisfactory way. However, in this case, FMoD description would produce significantly different feature vectors for the first and the last shot frame, leading to the unnecessary extraction of multiple key-frames.

Whatever the employed descriptor, FMoD or PI-FMoD, two feature vectors of identical dimensionality are computed for the i-th video frame, $i = 1, \cdots, N_f$, i.e., one for the luminance frame $\mathbf{V}_i^L$ and one for the corresponding disparity map $\mathbf{V}_i^D$. The method used for fusion of luminance and disparity information is simple vector concatenation, before clustering. Thus, scene texture, illumination and geometry are all taken into account as factors in order to construct an informative video frame description. Given that the feature vector dimensionality needs to be as low as possible for reasons of computational cost, color may be disregarded (the input video is assumed to be grayscale), since it has not been conclusively proven as an important modality for successful summarization [8]. However, if it were deemed necessary, FMoD or PI-FMoD descriptors could be computed on the hue channel of the HSV frame representation, i.e., on a matrix $\mathbf{V}_i^C \in \mathbb{R}^{M \times N}$. The resulting vector could easily be integrated with the combined luminance and disparity feature vector through concatenation, in order to form a unified, complete frame description.

The exploitation of disparity information, and, therefore, scene geometry, potentially leads to the extraction of more representative key-frames, since employing luminance information alone leads to different results than exploiting both disparity and luminance. Figure 2 shows example frames from the "Wall" 3D shot, where the camera pans horizontally from right to left, showing first a wall close-up and subsequently a building in long view. Thus, the shot is heavily differentiated in disparity but is mostly homogeneous in luminance and color characteristics, since the wall and the building have similar texture and reflectance properties. Figure 3 shows two key-frames (K = 2) extracted from the "Wall"

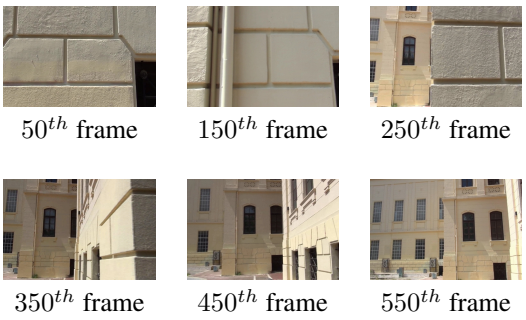

Fig. 2. Example frames from the left color channel of the "Wall" 3D shot.

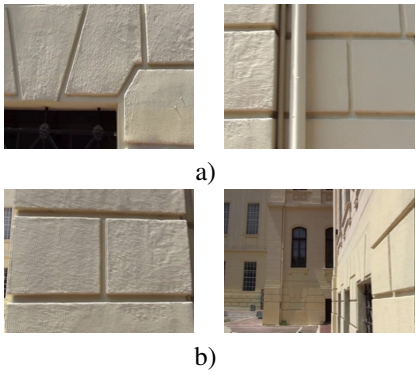

Fig. 3. a) Two left-channel key-frames computed using only luminance information, b) two left-channel key-frames computed by combining luminance and disparity information.

when disparity is ignored and when it is taken into account: a single key-frame would suffice if disparity was not exploited, but when employing disparity information two semantically meaningful key-frames may be found more easily.

The number of clusters K, i.e., the number of extracted key-frames per shot, is determined separately for each shot by evaluating different clusterings, one for each possible value of K, $K \in \mathbb{N}$, $K \in [2, \cdots, K_{max}]$. Thus, K-Means++ is performed $K_{max}-1$ times per shot. This adaptive approach does not induce significant computational overhead, since the number of frames per shot is typically less than 100 and clustering is performed very rapidly. The mean silhouette coefficient, one of the most simple, robust and well-performing cluster validity indices [14], is used as the metric for clustering evaluation. The selected value for K is the one corresponding to the clustering with the maximum silhouette.

In a subsequent post-processing filtering step, the extracted key-frames from all shots are partitioned into Sp clusters, by reapplying once the K-Means++ algorithm. S is the total number of extracted key-frames and the user-provided retention parameter p is a percentage that regulates the aggressiveness of frame elimination during this filtering process. The goal is to detect clusters of similar key-frames and remove all frames contained within the same cluster,

excluding the one closest to the respective centroid. Thus, a filtered key-frame set of smaller size is derived by considering the entire movie content. FMoD, which preserves spatial information, has been found to be a particularly effective descriptor for the detection of multiple similar shot / reverse shot instances, e.g., when two persons are shown alternatingly during a conversation between two characters, in order to reduce the informational redundancy inherent in this common film editing technique.

The filtered key-frames are temporally extended, using neighboring frames, to form key segments: assuming the i-th video frame is a key-frame, the video segment extending from the $(i-L)$-th frame up to the $(i+L)$-th frame is marked as a key segment. L is a user-provided parameter and the value $L = 20$ has been shown to perform well in our experiments. Thus, the initial duration of all key segments is $D = 2L + 1$. Subsequently, each key segment is contained within the boundaries of its shot and any temporally overlapping key segments are merged. The finally derived key segments are then concatenated in temporal order to form the video skim. Since the original video is a stereoscopic movie, meaning two color channels are available for each frame, the produced skim is also stereoscopic.

Given a set of stereoscopic key segments, annoying *depth jump cuts* may occur at key segment temporal concatenation points, due to disparity mismatches among consecutive segments [15], which cause visual fatigue. Such mismatches indicate severe differences in frame depth characteristics. Therefore, in the final stage of the proposed video summarization pipeline, a previously developed depth jump cut detection and characterization algorithm is applied on the produced video skim and a depth continuity characterization is derived per frame [16]. That is, a depth jump cut is either "absent" (A), "mildly uncomfortable" (MU), "uncomfortable" (U), or "highly uncomfortable" (HU).

In case no depth jump cut is present at a key segment concatenation point, no further processing is needed. Furthermore, if a U or a HU depth jump cut is detected, a luminance fade out / fade in process is applied to the shot cut, in order to eliminate the source of discomfort during stereoscopic viewing of the video skim. In case a MU depth jump cut is present, a less drastic heuristic technique is employed and described below, aimed at minimizing the presence of such defects in the final video skim.

Between two consecutive key segments $S_i, S_{(i+1)}$ that cause a MU depth jump cut, the last $\lceil D/4 \rceil$ frames of S_i and the first $\lceil D/4 \rceil$ frames of $S_{(i+1)}$ are exhaustively investigated in pairs in order to estimate the best possible concatenation point, i.e., the frame pair where the Euclidean distance between two of the disparity channel frames $\mathbf{V}_f^D$ and $\mathbf{V}_t^D$, where $\mathbf{V}_f^D, \mathbf{V}_t^D$ belong to the corresponding subsets of $S_i, S_{(i+1)}$ segments, is minimal.

3. EXPERIMENTS

In order to evaluate the proposed stereoscopic movie summarization pipeline, an objective evaluation scheme was employed. It was performed on 3 stereoscopic Hollywood movies released in 2011, hereby named "Movie1", "Movie2" and "Movie3". Disparity estimation had been applied prior to the evaluation, using a publicly available implementation of the SGBM algorithm [17].

Video skims derived with a combination of PI-FMoD / FMoD descriptors were compared against skims derived with image histogram descriptors, for various values of the retention parameter p. For each value of p, multiple FMoD-derived and histogram-derived skims were evaluated, by taking into account different combinations of luminance, color and stereoscopic disparity modalities. For each such combination, all employed frame histograms (one per modality) were computed with 256 bins and concatenated into a single vector. For PI-FMoD, codebook size c was set to $40N_m$, where $N_m \in \mathbb{N}, N_m \in 1, 2, 3$ is the number of employed modalities at each case. Moreover, d was set to 6, in the case of FMoD, and to 5, in the case of PI-FMoD. These parameter values were found to lead to good results without inducing unacceptably high computational cost.

The objective metric employed in our evaluation is the mean silhouette coefficient Sil of the clustering that is performed during the post-processing filtering stage. It holds that $Sil \in \mathbb{R}, Sil \in [0, 1]$ and that a higher value suggests a better clustering. Thus, the proposed video descriptor and the commonly employed histogram descriptors are compared with regard to their performance in clustering, instead of directly with regard to their performance in video summarization, in order to bypass the inherent ambiguity and the subjective nature of the summarization problem.

The 3 scores achieved by each video skim and the corresponding video description method (one for each of the 3 movies) were averaged to compute the aggregate results. In the following notation, L suggests the exploitation of the luminance modality during the description process, C the exploitation of the color / hue modality, D the exploitation of the stereoscopic disparity modality, while LD and LCD refer to the combination of multiple descriptors computed on the corresponding modalities. The results of the objective evaluation are shown in Table 1.

As it can be seen, the proposed video descriptor outperforms the typically employed histogram-based description method and the best results are achieved when all available image modalities (luminance, stereoscopic disparity, color) are exploited. This implies that the richer informational content of FMoD descriptors, in comparison to histograms, facilitates the determination of more compact and well-separated clusters in the higher-dimensionality feature space that is formed by the concatenation of multiple modalities. Additionally, the mean silhouette coefficients suggest a better clustering when less movie-wide clusters are being

Table 1. A comparison of the aggregate mean silhouette coefficients for different video description methods and different values of the retention parameter p.

Method	0.5	0.6	0.7	0.8
FMoD-L	0.22	0.22	0.20	0.16
FMoD-C	0.21	0.20	0.16	0.12
FMoD-LD	0.18	0.18	0.16	0.13
FMoD-LCD	**0.23**	**0.23**	0.21	0.16
Histogram-L	0.20	0.19	0.17	0.13
Histogram-C	0.13	0.13	0.13	0.10
Histogram-LD	0.15	0.15	0.14	0.12
Histogram-LCD	0.16	0.17	0.15	0.13

used (regulated by the value of the retention parameter p), thus resulting in a shorter, and thus arguably more enjoyable, video skim.

4. CONCLUSIONS

We have proposed a stereoscopic movie multimodal summarization pipeline, based on a novel video description method. Its necessary inputs are one (out of two) stereoscopic video channel and the corresponding disparity video, while its output is a short stereoscopic video skim. The investigated descriptor outperforms the prevalent histogram-based description approach, according to an objective performance measure, while the proposed pipeline also takes care of stereoscopic video quality issues (depth jump cuts) that may arise due to the skim construction process.

REFERENCES

[1] A. G. Money and H. Agius, "Video summarization: A conceptual framework and survey of the state of the art," *Journal of Visual Communication and Representation*, vol. 19, no. 2, pp. 121–143, 2008.

[2] W. Hu, N. Xie, L. Li, X. Zeng, and S. Maybank, "A survey on visual content-based video indexing and retrieval," *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 41, no. 6, pp. 797–819, 2011.

[3] P. Kelm, S. Schmiedeke, and T. Sikora, "Feature-based video key frame extraction for low quality video sequences," *10th Workshop on Image Analysis for Multimedia Interactive Services*, vol. 1, no. 1, pp. 25–28, 2009.

[4] Y. Zhuang, Y. Rui, T.S. Huang, and S. Mehrotra, "Adaptive key frame extraction using unsupervised clustering," *Proceedings of International Conference on Image Processing (ICIP)*, vol. 1, pp. 866–870, 1998.

[5] S. E. F. De Avilla, A. P. B. Lopes, A. L. Jr. Luz, and A. A. Araujo, "Vsumm: A mechanism designed to produce static video summaries and a novel evaluation method," *Pattern Recognition Letters*, vol. 32, no. 1, pp. 56–68, 2011.

[6] Marco Furini, Filippo Geraci, Manuela Montangero, and Marco Pellegrini, "Stimo: Still and moving video storyboard for the web scenario.," *Multimedia Tools Appl.*, vol. 46, no. 1, pp. 47–69, 2010.

[7] T. Wan and Z. Qin, "A new technique for summarizing video sequences through histogram evolution," *Signal Processing and Communications (SPCOM), 2010 International Conference on*, pp. 1–5, 2010.

[8] E J.Y. Cahuina and G. C. Chavez, "A new method for static video summarization using local descriptors and video temporal segmentation," *Graphics, Patterns and Images (SIBGRAPI), 2013 26th SIBGRAPI-Conference on*, pp. 226–233, 2013.

[9] G. Csurka, C. Bray, C. Dance, and L. Fan, "Visual categorization with bags of keypoints," *ECCV, Workshop on Statistical Learning in Computer Vision*, 2004.

[10] J. Almeida, N. J. Leite, and R. dS. Torres, "Vison: Video summarization for online applications," *Pattern Recognition. Letters*, vol. 33, no. 4, pp. 397–409, 2012.

[11] Y.S. Avrithis K.S. Ntalianis N. Doulamis, A. Doulamis 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.

[12] 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.

[13] D. Arthur and S. Vassilvitskii, "K-means++: the advantages of careful seeding," *SODA '07: Proceedings of the 18th annual ACM-SIAM symposium on discrete algorithms*, pp. 1027–1035, 2007.

[14] O. Arbelaitz, I. Gurrutxaga, J. Muguerza, J. M Pérez, and I. Perona, "An extensive comparative study of cluster validity indices," *Pattern Recognition*, vol. 46, no. 1, pp. 243–256, 2013.

[15] Bernard Mendiburu, *3D movie making. Stereoscopic digital cinema from script to screen*, Focal Press, 2009.

[16] S. Delis, N. Nikolaidis, and I. Pitas, "Automatic detection of depth jump cuts and bent window effects in stereoscopic videos," *IVMSP Workshop, 2013 IEEE 11th*, pp. 1–4, 2013.

[17] H. Hirschmuller, "Stereo processing by semiglobal matching and mutual information," *Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 2, pp. 328–341, 2008.